Advances in Survival Analysis and Optimal Scaling Methods

Willems, S.J.W.

Combining optimal scaling and survival analysis techniques to identify possible predictors for unemployment duration

2

In this paper we propose a new approach in survival analysis in a non-traditional field of application; unemployment data. A common practice is to use factor analysis to first summarize survey data, and then fit a binomial logistic regression model to estimate the regression weights, which are used to identify factors associated with work status at a prespecified time point. In this paper, a combination of optimal scaling and survival analysis methods is proposed as an alternative to find possible predictors for unemployment duration. This combination of techniques is illustrated and compared to the traditional approach. Data from the Dutch Employee Insurance Agency are used to illustrate the method.

2.1 Introduction

Identifying possible predictors for unemployment duration can be useful to, for example, provide appropriate counseling or predict the costs for unemployment benefits. To identify these predictors, data on job seekers are, among others, collected via questionnaires and registries, and these are used to find associations between characteristics of the unemployed and their probability of finding a new job. A data preparation step is usually performed to summarize many survey items into fewer composite scores. These resulting scores can be used to model unemployment duration or employment status. An abundance of methods has been developed to summarize survey data and to model employment status. In this paper, the popular choice of combining factor analysis with logistic

This chapter is published as Willems, S. J. W., Fiocco, M., and Meulman, J. J. (2019) *Combining optimal scaling and survival techniques to identify possible predictors for unemployment duration.* International Journal of Statistics & Economics, 20(3), 1–22.

regression will be discussed and an alternative combination, optimal scaling principal component analysis with survival analysis, will be proposed.

The aim of a data preparation step is to summarize many items into a few composite scores. Factor analysis seems to be a popular choice for this data preparation step. This method is used to obtain a composite score for items with a common underlying factor. Each item within a factor is given a factor loading, which indicates the contribution of this item to the factor. These factor loadings are then used as weights to compute the weighted average of the factor items, the so-called factor score. Often, it is predefined which items have the same underlying factor, and that each item has just one underlying factor. Hence, each factor is a summary of a group of items, and each item belongs to one factor. This makes interpretation of the factor model straightforward, and therefore a popular choice. When factor analysis is used as a data preparation step, the resulting factor scores are used in a statistical model to identify possible predictors. Including only the factor scores instead of all individual items reduces the correlation between the predictors considerably. As a consequence, the problem of collinearity can be controlled. However, correlation between the variables is only reduced by factor analysis, but not eliminated completely since there might still be high correlation among the resulting factor scores. Furthermore, this method assumes the items are numeric, so it works optimally for numeric items. Ordinal category levels are coded by integers, which are often treated as numeric data by the model. In this way, a linear relation between the categorical item scores and the factor score is assured. However, since the category levels cannot be assumed to be equally spaced, a nonlinear relation possibly gives a better fit for this type of data.

As an alternative to factor analysis, optimal scaling principal component analysis (OS-PCA) can be applied to summarize items into fewer summary scores. This techniques can maintain the properties of ordinal categorical data, while finding composite scores that are completely uncorrelated. This method is also often referred to as nonlinear principal component analysis (NLPCA) (Linting et al., 2007; Meulman et al., 2004). The purpose of OS-PCA is to reduce the dimensionality of a dataset by summarizing the original variables into a smaller set of uncorrelated variables, called principal components. In this way, all collinearity between the composite scores is removed.

For each component, items get component loadings, which are used to calculate object scores for each component; the weighted average of the items. Only the most important components (the principal components) are included in the final model. The chosen number of components indicates the new dimensions of the data. Since all components are a weighted average of all items, interpretation of OS-PCA results is more challenging than the interpretation of a factor model in which each factor underlies a subset of the items.

Additionally to finding principal components, OS-PCA provides the option to optimally transform the category levels by giving them new values, called quantifications. The quantifications are chosen in an optimal way such that the nonlinear relationship is linearized. This option can be used to maintain the properties of ordinal data. While constructing principal components, the OS-PCA method aims to account for as much of the variance in the original dataset as possible. As a result, the principal components will also reveal the correlation structure between items, and thus provide a better understanding of how items in the survey are related. Since OS-PCA will remove the collinearity and will maintain the ordinal properties of the survey data, it might be a valuable alternative to factor analysis. The linear version of OS-PCA, PCA, has already been applied in the context of reemployment data (Wanberg et al., 2002).

Once the item data are summarized into composite scores, the next step is to identify variables associated with the probability of reemployment. For this step of the analysis, a binomial logistic regression analysis is a popular approach. This technique can be used to estimate the probability of reemployment within a prespecified time period, for example, to predict whether a person is reemployed within one year. This model is very useful if researchers are interested in the probability of an event within a specific time period. However, the binary outcome and prespecified time point might be too restricting if interest lies in the estimation of the reemployment probability over a time interval. In this case, techniques from the survival field can be used instead of a binomial logistic regression model. Survival analysis techniques estimate the time to an event of interest based on a set of variables. Hence, for reemployment data, survey results can be used to assess the probability of being reemployed over a range of time, instead of at one specific time point. Survival analysis techniques are often used in a medical setting to, for example, compare the effect of different treatments on the the survival time of patients. There are several instances where these techniques have been used in the field of reemployment prediction (Boršič and Kavkler, 2009; Kavkler et al., 2009; Tutkun and Karasoy, 2016; Wanberg et al., 2002), but it is not widely used. Logistic regression appears to be the default method.

In this paper, the two combinations of methods will be introduced in more detail, and they will be illustrated with an application on reemployment data from the Dutch Employee Insurance Agency (from hereon referred to by the Dutch abbreviation UWV). The paper is organized as follows. Details on the UWV dataset are provided in section 2.2, together with some details on the factor and logistic regression analysis results. The OS-PCA and survival analysis techniques are discussed in section 2.3. Results of the application of these techniques on the UWV data are shown in section 2.4. Advantages and disadvantages of the proposed methodology are discussed in section 2.5 where these methods are

compared to factor analysis combined with logistic regression.

2.2 Development of UWV's Work Profiler 1.0

In the Netherlands, job seekers who recently became unemployed can apply for unemployment benefits and counseling at the UWV agency. Since the budget of this institute is limited, it aims to reduce counseling service expenses. One of the strategies is to reduce face-to-face counseling by replacing it with computerized services. For this replacement, UWV has developed an online instrument, the Work Profiler (Wijnhoven and Havinga, 2014). The purpose of this instrument is to select those individuals who experience difficulties in finding a job and could therefore benefit from unemployment counseling. Once these individuals are selected, the Work Profiler should additionally provide a quick diagnosis of the main obstacles faced by these persons, such that appropriate services can be provided for each individual. To make the selection and to give the diagnosis, the Work Profiler makes use of an online questionnaire given to the unemployed persons. The replacement of the selection and diagnosis procedures by an online tool will greatly reduce the costs for face-to-face counseling.

The UWV Centre for Knowledge (Kenniscentrum UWV) and the School of Medical Sciences of the University Medical Centre Groningen (UMCG) in the Netherlands collaborated to develop the first version of the Work Profiler (Brouwer et al., 2015). Aim of this research was to identify possible baseline predictors for resuming work within 12 months after becoming unemployed. It consisted of three steps: a literature study, a cross-sectional study and a longitudinal study. First, many factors that possibly influence the probability of finding a new job were listed from literature. A 500-item questionnaire was created with items corresponding to these factors. Then, during the cross-sectional study the questionnaire's length was reduced to 155 items, each corresponding to one of the 70 remaining factors found to be most relevant. These 155 items were included in the survey used in the third step, the longitudinal study. Newly unemployed were asked to fill in this survey, and after 12 months their work status was registered. The mean scores of the items corresponding to each factor were included as variables in a logistic regression analysis. This analysis identified 10 predictive factors associated with work resumption within one year. An additional 11th factor was retained in the model at UWV's request. All items corresponding to these 11 predictors were included in the first version of the Work Profiler, resulting in a survey consisting of 20 items. This version has been used since 2013 as part of UWV's services.

Since the realization of Work Profiler 1.0, a new study has started to further develop this instrument. Work Profiler 1.0 was extended to include more factors

and used to collect data for the second longitudinal study. In the remaining of this section, the data collection procedure for this new study will be described in more detail and the first analyses on this data will be discussed.

2.2.1 Data collection

Participants

The data used for this research is from job seekers associated with one of the 11 participating UWV offices. The cohort of this study includes persons younger than 64 years who claimed unemployment benefits between March 1st 2014 and October 31st 2014. Individuals should be eligible for at least three months of benefits. Six to ten weeks after the start of their unemployment benefits, the extended version of Work Profiler 1.0 (in Dutch) was provided via UWV's online system. Hence, only persons who had access to the online system could participate (almost everyone eligible for unemployment benefits). Unemployed whose unemployment benefits ended within 10 weeks, e.g. due to being reemployed or due to other reasons, were excluded from the analysis. The data consists of 32,623 observations.

Factors and items

Since the aim was to further develop the Work Profiler 1.0, all 20 items of the 11 factors from this version were included in the new one. To find more predictors for unemployment duration, items corresponding to other factors were added as well. This resulted in a total of 32 factors (15 hard factors and 17 soft factors) measured by 55 items. Most hard factors, like age, gender, and education were already available from UWV's registries, while the majority of the soft factors, which indicate a person's psycho-social situation in relation to reemployment, was measured by the questionnaire. All hard factors are displayed in Table 2.1, and all soft factors and the number of corresponding items are given in Table 2.2. The classification of the items into factors is based on the study by Brouwer et al. (2015).

All soft factors are ordinal categorical variables, i.e. there is an ordering among the possible answer options. Most of the soft factors were measured on 5-point Likert scales. For example, to indicate to what extend newly unemployed agreed with a statement, the provided options were as follows: *strongly disagree*, *disagree*, *don't disagree/agree*, *agree*, or *strongly agree*. Most of the remaining items had similar or comparable scales, like a 5-point scale to indicate the importance of an aspect, or a 1–10 scale to grade certain aspects of life. Also, many hard factors were measured on an ordinal scale. Exceptions were factors like age, nationality, and industry.

Hard factor	Measurement level	# categories
Additional income - work	nominal	2
Additional income - benefits	nominal	2
Age group	ordinal	5
Duration last job (years)	numeric	-
Education level	nominal	11
Extend of unemployment (hours)	numeric	-
Former working hours	numeric	-
Resigned themselves (fraction)	numeric	-
Gender	nominal	2
Household position	nominal	6
Industry	nominal	12
Maximum duration of benefits (weeks)	numeric	-
Nationality	nominal	4
Profession level	ordinal	5
UWV office	nominal	11

Table 2.1: *Hard factor candidates for Work Profiler 2.0, with their measurement levels and number of categories (if applicable).*

The category frequencies for some of the measured covariates are given in Table 2.3.

Unemployment duration

The aim of further developing Work Profiler 1.0 is to identify possible predictors for unemployment duration and to make good predictions for the probability of being reemployed. Therefore, for each person in the dataset the starting dates of unemployment and reemployment (if applicable) were registered. Starting dates of unemployment were specified by the unemployed when they claimed their benefits. If applicable, the dates of reemployment were either provided by the persons claiming benefits, or retrieved by UWV. If a person was reemployed while being eligible for the unemployment benefits, these benefits were stopped, and hence the reemployment date was registered in UWV's administration on unemployment benefits. If a person did not find a job while being eligible for unemployment benefits, the date of reemployment was checked in the POLIS registry in which all gainful employment in the Netherlands is registered. In this way, for each person in the dataset it was determined whether he/she had found a job within one year. The duration of unemployment was determined from the starting dates of unemployment and reemployment.

Soft factor	# of items	# categories
Acceptance readiness - full time	1	5
Acceptance readiness - time	2	5, 5
Acceptance readiness - work	2	5, 5
External variable attribution	3	5, 5, 5
Childcare problems	1	5
Financial need/problems	2	3, 5
Hours per week capable of work	1	5
Job search attitude - advantageous / pleasant	2	5, 5
Job search attitude - utility / necessity	2	5, 5
Job search behavior - applications	1	4
Job search behavior - direct contact employers	3	4, 4, 4
Job search intention	3	5, 5, 5
Perceived health	5	5, 5, 5, 5, 10
Self-efficacy - preparation applications	3	5, 5, 5
Subjective norm	1	14
Balance pros and cons for not working	1	4
View return to work	3	5, 5, 5

Table 2.2: *Ordinal soft factors candidates for Work Profiler 2.0, with the number of items used to assess them and the number of answer categories for each of these items.*

Covariate and category levels	# observations
Age	
- ≤ 27	4,213
- > 27 and ≤ 39	9,669
- > 39 and ≤ 49	8,777
- > 49 and ≤ 54	4,084
- > 54	5,880
Gender	
- male	15,641
- female	16,982
Household position	
- Living alone	6,024
- Married / cohabiting, no children	8,440
- Married / cohabiting, youngest child ≥ 7 years	7,330
- Married / cohabiting, youngest child < 6 years	5,832
- Single parent	2,308
- Other	2,689
Profession level	
- Elementary	2,607
- Lower	9,290
- Middle	12,443
- Higher	7,005
- Academic	1,227
- *Missing*	51
Nationality	
- 1st Dutch, 2nd no or western	31,220
- 1st or 2nd Polish	406
- 1st or 2nd non-western	349
- 1st western (other countries)	619
- *Missing*	29

Table 2.3: *Number of observations in each category of some of the covariates in the dataset, $n = 32{,}623$.*

2.2.2 Statistical analysis to develop Work Profiler 2.0

The Netherlands Organisation for applied scientific research (from hereon referred to by the Dutch abbreviation TNO) has conducted an extensive analysis of the data collected by UWV. The results of this analysis were described in an internal report which is not yet published. In this section a short description of the analysis performed by TNO will be given.

Recall that Brouwer et al. (2015) used the mean of all items corresponding to a factor as the factor score. TNO extended the data preparation step by performing a factor analysis on the items corresponding to the soft factors. Restrictions were put on the factor model to enforce some assumptions about the model. For example, the factor classification of each item was prespecified according to the research by Brouwer et al. (2015) and correlations between the factors were allowed for in the factor model, since it was expected that some of the factors are related. The factor loadings resulting from this analysis were used as weights to compute the factor scores; the weighted means of the items corresponding to each factor.

A univariate logistic regression model was fit for each factor to study the association with work status after one year. All relevant factors were included in the multivariate logistic regression model, either as a numeric variable or as a categorical variable. Then, by using a combination of forward and backward selection, the variables were removed from the model to derive a parsimonious model with an easy interpretation. TNO's final model consists of 18 factors.

In order to get more insight in the probability of finding a new job during the first year, logistic regression models were fitted for work status at six and nine months in a similar way as for 12 months.

2.3 Alternative method to analyze reemployment data

As discussed in the introduction, a combination of OS-PCA and survival analysis to identify possible predictors for unemployment duration are introduced as an alternative to the combination of factor analysis and logistic regression.

2.3.1 Optimal scaling principal component analysis

The extended version of Work Profiler 1.0 consists of 55 items intended to measure 32 factors. Many of these factors are closely related, which implies their factor scores to be correlated. The correlation between scores is a common phenomenon in survey data and may lead to problems in the estimation of a statistical model due to the presence of collinearity. For example, TNO found two pairs of hard factors to be collinear. To prevent the problem of collinearity,

it is preferable to include only weakly correlated or uncorrelated variables in such models. Some data preparation procedures can help to summarize items in less correlated scores.

Principal component analysis (PCA) was developed to reduce the dimensionality of a dataset by summarizing numeric variables into a smaller set of uncorrelated summary variables, the principal components (Jolliffe, 2002). While constructing the principal components, the PCA method aims to account for as much variance of the original variables in the dataset as possible. So, the resulting components are uncorrelated, but still retain much of the correlation between the original item scores. In this way, the method reveals the correlation structure of the original variables.

Since the PCA algorithm is based on calculating correlations, it can only be applied to numeric data. To deal with categorical variables, optimal scaling principal component analysis (OS-PCA) was developed. This technique transforms categorical variables into numeric variables (quantifications) while simultaneously calculating the principal components. OS-PCA uses the quantifications to calculate correlations. Details of this method were described by Meulman et al. (2004), Linting et al. (2007), and Linting and van der Kooij (2012). OS-PCA is currently available in the Categories package (Van der Kooij and Meulman, 1999) of the statistical software SPSS (SPSS Inc., 2008).

Several restrictions can be put on the OS transformation. For example, to preserve the ordering of ordinal category levels, one would choose a monotone transformation. As output, OS-PCA provides the quantifications of the categorical variables. Furthermore, correlation between each item and each component are given by component loadings. These loadings indicate how well items are explained by the component, i.e. how much information about each item is included in the component. Items that correlate strongly with a component (high component loading), are represented well by this component. For each of the components, OS-PCA will also provide the variance accounted for (VAF), which indicates the total variance in the data explained by the component. Components are ordered according to their VAF. Hence, the first few components explain most of the variance in the data and are therefore the most important components, i.e. the *principal* components. The score of each object on each of the components (object scores) can be calculated from these results.

The items corresponding to the soft factors in the Work Profiler survey measured a person's psycho-social situation. These items were highly correlated. Therefore, OS-PCA was applied as a data preparation step to remove the correlation among these factors and to reveal their correlation structure. It was expected that for many of these covariates, the effect of the category levels was monotonically associated with the outcome. For example, it seems reasonable to expect that the more a person agrees with the statement *I am highly motivated*

to find work the coming month, the sooner this person will be reemployed. Therefore, it was preferred to keep the ordering of the category levels in the model. Furthermore, these effects were not expected to grow linearly with the category levels, so no linear restrictions were enforced. The components resulting from the OS-PCA preparation step can be used as input for a statistical model.

2.3.2 Survival analysis

Most of the previous research done to develop Work Profiler 1.0, and TNO's recent research to develop the second version was focused on predicting work status at 12 months. However, TNO extended their research and performed logistic regression for work status at six and nine months as well to get a better understanding of the change of reemployment probability over time. As an alternative, survival analysis can be used to analyze reemployment data if the time aspect is of interest.

Survival analysis techniques study the distribution of time to a certain event of interest (see for example Klein and Moeschberger (2003)). This could be any type of event, for example death (hence *survival* analysis), recovery, or reemployment. Typical situations in which survival analysis methods are used, are those where the time to the event of interest is not observed for some individuals. There could be many reasons why the event was not observed, for example subjects are lost to follow-up, or another event occurs which prevents the event of interest (competing risks (Putter et al., 2007)). If the event of an individual is not observed, this is a censored observation. For these observations, it is only known at which time (the *censoring time*) the event had not occurred yet. For these observations, the corresponding event time was longer than the censoring time. The censoring times provides some information about the distribution of the event times and are therefore included in the survival analysis.

There are several survival models which provide the possibility to include covariates to estimate the effect of covariates associated with event times. A popular model is the Cox proportional hazards model proposed (Cox, 1972), in which the hazard function is estimated and the effect of a covariate on the hazard is quantified by its hazard ratio (HR). The HR is the ratio of survival probabilities between subjects with different values for a particular covariate. A HR close to 1 indicates no effect from this specific covariate on the hazard. A subject with HR larger than 1 will experience the event faster than someone in the reference group, while an individual with a HR smaller than 1 will need more time than those in the reference group. Model based estimated survival probabilities can be plotted over time for different individuals.

To develop the Work Profiler, both Brouwer et al. (2015) and TNO applied

logistic regression to find possible predictors for work status at one year, i.e. a binary outcome. Additionally, TNO fitted logistic regression models at six and nine months. Survival analysis techniques can be applied to the UWV dataset to investigate reemployment probabilities during the whole first year. In this analysis, the exact unemployment duration is used as the outcome variable instead of the status at a predetermined time point. To illustrate the use of survival analysis for the duration of unemployment, a Cox proportional hazards model was fitted on the UWV data. In the model, the hard factors, the principal components that summarize the soft factors, and some interactions to find covariates associated with unemployment duration were included. In this context, a HR close to 1 indicates no effect for the corresponding risk factor on the probability of reemployment. Compared to the reference group, a HR smaller than 1 predicts a longer unemployment duration, while a HR larger than 1 indicates a shorter unemployment duration. Estimated probabilities of remaining unemployed can be plotted over time for different types of individuals.

2.4 Statistical analysis

2.4.1 Optimal scaling principal component analysis

OS-PCA was applied to the UWV dataset to investigate the correlation structure among the items of the soft factors and to reduce the observed variables to a smaller number of uncorrelated principal components. All soft factors were included in the analysis except for the items of the factors *Hours capable of work*, *Acceptance readiness - Full Time* and *Child care hindrance*, since they are strongly correlated with the hard factor *Household position*.

The choice of the number of principal components for the OS-PCA analysis was based on the combination of the VAF by each of the components and their interpretability. The items *Financial Hindrance* and *Job Search Behavior - Applications* were removed from the model, because the total VAF by these items in the final model was smaller than 0.25 (the suggested minimum by Linting and van der Kooij (2012)), which means that these items were poorly explained by the OS-PCA result. In the final OS-PCA model, 31 items were analyzed on an ordinal analysis level and summarized into seven components. In total, this model accounted for 60.9% of the variance in the original dataset. The contribution of each component to the total VAF is shown in Figure 2.1. This plot shows that the first component explained a large part of the total VAF. It accounted for around 22% of the total variance, while the other components explained 8.3% or less.

The component loadings for all items on all components are given in Table 2.4.

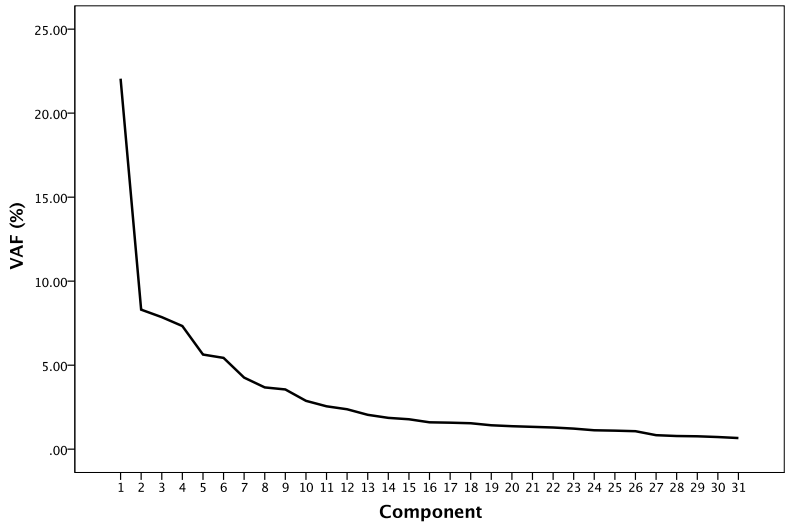

Figure 2.1: *Variance accounted for (VAF) by each component in the OS-PCA model of 7 dimensions on the 31 items corresponding to soft factors.*

The factors whose items were highly correlated with the components are the ones with a high absolute component loading. Linting and van der Kooij (2012) used 16% VAF (absolute value component loading > 0.4, printed in bold gray in Table 2.4) as a cut-off value. However, 20% VAF (absolute value component loading > 0.447, printed in bold black in Table 2.4) may be a more reasonable choice. The first principal component is associated to persons who feel healthy, have high expectations, are highly motivated, and know how to find a job. These factors appeared to be highly correlated, and most of the variance in the dataset was explained by differences in these factors. There is a striking resemblance between the factors that correlated strongly with the first components, and the Theory of Planned Behavior (Ajzen, 1991) and the Motivation Model (Vroom, 1964). According to the Theory of Planned Behavior, high scores on factors like subjective norm, job search attitude, and job search self-efficacy indicate a strong job search intention, which will usually result in job search behavior. As shown in Table 2.4, items on these factors were strongly correlated with the first component, and with each other. This result supports the Theory of Planned Behavior. Furthermore, the high component loadings on the items on *view return to work* showed that unemployed who have high expectations are more motivated to find a job, also showed more job search behavior. This is in accordance with the Motivation Model.

Table 2.4 suggests that the second component is related to health perception; individuals who score high on the five perceived health items, will also get a

high score on the second component.

Short interpretations of the other components are as follows. Component 3 and 6 contain mainly variables about a person's acceptance readiness for different type of work or different working hours compared to a previous job. Component 4 contrasts persons who score high on external variable attribution and those who are not ready to accept different working hours. Component 5 is about job search behavior, while component 7 indicates a person's job search attitude. The objects scores for each person for each principal component was calculated using this OS-PCA model.

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7
Acceptance - time 1	0.330	-0.166	**0.434**	**-0.414**	0.166	0.339	0.151
Acceptance - time 2	0.322	-0.170	**0.410**	**-0.429**	0.209	0.315	0.147
Acceptance - work 1	0.122	-0.261	**0.436**	-0.253	-0.030	**0.434**	0.121
Acceptance - work 2	0.228	-0.183	0.303	-0.184	-0.026	**0.529**	0.112
External variable 1	-0.254	-0.319	**0.435**	**0.412**	0.195	-0.244	0.101
External variable 2	-0.039	-0.310	0.336	**0.473**	0.258	-0.183	0.108
External variable 3	-0.103	-0.225	0.333	**0.432**	0.242	-0.123	0.083
Financial need	0.363	-0.360	0.224	-0.232	-0.027	-0.264	-0.276
Search attitude - a/p 1	0.286	-0.266	-0.127	-0.154	-0.333	-0.309	**0.635**
Search attitude - a/p 2	0.310	-0.298	-0.111	-0.161	-0.397	-0.312	**0.570**
Search attitude - u/n 1	**0.569**	-0.353	0.131	-0.080	-0.284	-0.231	-0.246
Search attitude - u/n 2	**0.598**	-0.315	0.060	-0.093	-0.316	-0.234	-0.168
Search beh - contact 1	0.387	-0.040	-0.241	-0.205	**0.553**	-0.174	0.045
Search beh - contact 2	0.322	-0.134	-0.161	-0.189	**0.580**	-0.243	0.062
Search beh - contact 3	0.280	-0.103	-0.153	-0.246	**0.620**	-0.254	0.107
Search intention 1	**0.640**	-0.260	0.011	0.265	-0.032	0.071	-0.181
Search intention 2	**0.592**	-0.177	-0.279	0.279	0.016	0.212	-0.047
Search intention 3	**0.689**	-0.241	-0.046	0.267	-0.036	0.089	-0.155
Perceived health 1	**0.615**	**0.423**	0.253	0.134	-0.022	-0.038	0.082
Perceived health 2	**0.652**	**0.490**	0.277	0.125	0.009	-0.082	0.072
Perceived health 3	**0.639**	**0.521**	0.287	0.146	0.004	-0.087	0.097
Perceived health 4	**0.578**	**0.458**	0.269	0.184	-0.034	-0.055	0.048
Perceived health 5	**0.663**	**0.482**	0.304	0.118	0.004	-0.112	0.075
Self-effic - prep appl 1	**0.484**	-0.221	-0.307	0.365	0.005	0.288	0.107
Self-effic - prep appl 2	**0.484**	-0.214	-0.329	0.381	0.087	0.259	0.123
Self-effic - prep appl 3	**0.523**	-0.172	-0.335	0.377	0.050	0.269	0.084
Subjective norm	**0.448**	-0.287	0.189	-0.119	-0.076	-0.209	-0.276
Balance pros cons	-0.397	0.157	-0.164	0.156	0.046	0.091	0.320
View return to work 1	**0.524**	0.212	-0.346	-0.325	0.022	0.017	-0.042
View return to work 2	**0.521**	0.220	-0.327	-0.223	-0.071	0.063	-0.089
View return to work 3	**0.537**	0.091	-0.326	-0.213	0.079	-0.020	-0.013

Table 2.4: *Component loadings for the final OS-PCA model. All loadings with absolute value more than* $\sqrt{0.16} = 0.4$*, i.e.* $> 16\%$ *VAF, are printed bold and in gray. The ones accounting for more than 20% variance are printed bold and in black. See Table 2.2 for the variables' full names.*

2.4.2 Survival analysis

A Cox proportional hazards model was initially fitted on all hard factors (see Table 2.1) and the seven principal components found by the OS-PCA as described in subsection 2.4.1. Since one of the aims of UWV's research is to remove items that do not predict unemployment duration, a parsimonious model was fitted by removing some of the variables from the model containing all factors. The hard factor on education level was removed because this is closely related to someone's profession level, and it probably looses its importance when work experience increases. Items on former working hours and on the extend of unemployment were also removed since they were closely related to the number of days per week a person is able to work. Variables that had no significant effect on unemployment duration, and significant variables whose hazard ratio was close to 1, i.e. which had only a very small effect on unemployment duration, were also removed. Interactions between gender and household position, and between gender and the number of days per week able to work were included in the model.

Variables associated with unemployment duration, according to the final model, along with their corresponding HRs are given in Figure 2.2. Age is one of the most important indicators for unemployment duration. The older an individual is, the more time it will take to find a new job. Figure 2.3 shows survival curves which describe the model based probability of remaining unemployed as a function of time for different individuals. The covariate and principal component values used to make the curves shown in Figure 2.3 are displayed in Table 2.5. Note that all survival curves are equal to 1 from the beginning of the unemployment status, until around 70 days. This is due to how the data are collected. Only individuals who received unemployment benefits for at least ten weeks were included in the analysis (see subsubsection 2.2.1). Therefore, for this group, nobody was reemployed within 70 days, leading of a 100% survival (all people are unemployed) for this time period.

In Figure 2.3 survival curve estimates of several types of unemployed persons are shown. In Table 2.5 all values of the covariates used to plot the survival curves are provided.

First we compare the probabilities of remaining unemployed for young Dutch men and women who have little work experience and who live alone were studied. In Figure 2.3a the probability of remaining unemployed for this type of person with a high job search intention (value 1 for the first principal component) is shown. Figure 2.3b shows a person with the same characteristics, but with a low job search intention (value -1 on the first principal component). These plots show that there is almost no gender difference. However, the effect of the

first principal component is large; the probability of finding a job within one year is around 80% for individuals with a high intention to find a job, while this probability is much lower, around 55%, for those with a low score on the first principal component. This means that this component is strongly associated with unemployment duration, which was already indicated by the HR corresponding to this component.

The effect of household position on Dutch men and women in their 40s who are highly motivated to find a job, but can only work for four days per week (see Table 2.5) was studied. Figure 2.3c and Figure 2.3d show the probabilities of remaining unemployed for these men and women respectively. These plots show, for example, that finding a job takes more time for married or cohabiting mothers with young children. This phenomena is quantified by the HR of the interaction of gender and household postion (HR < 1).

Figure 2.3e and Figure 2.3f show the probability of reemployment for men and women in their 40s at different profession levels, who are married or cohabiting and have a young child of at most six years, with an average level of intention to find a new job (see Table 2.5 for exact covariate values). The plots suggest that the probability of finding a new job is approximately similar for mothers at all profession levels, while for fathers, profession seems to be associated to the reemployment probability. The plots show that overall the probability of begin reemployed is smaller for mothers than for fathers.

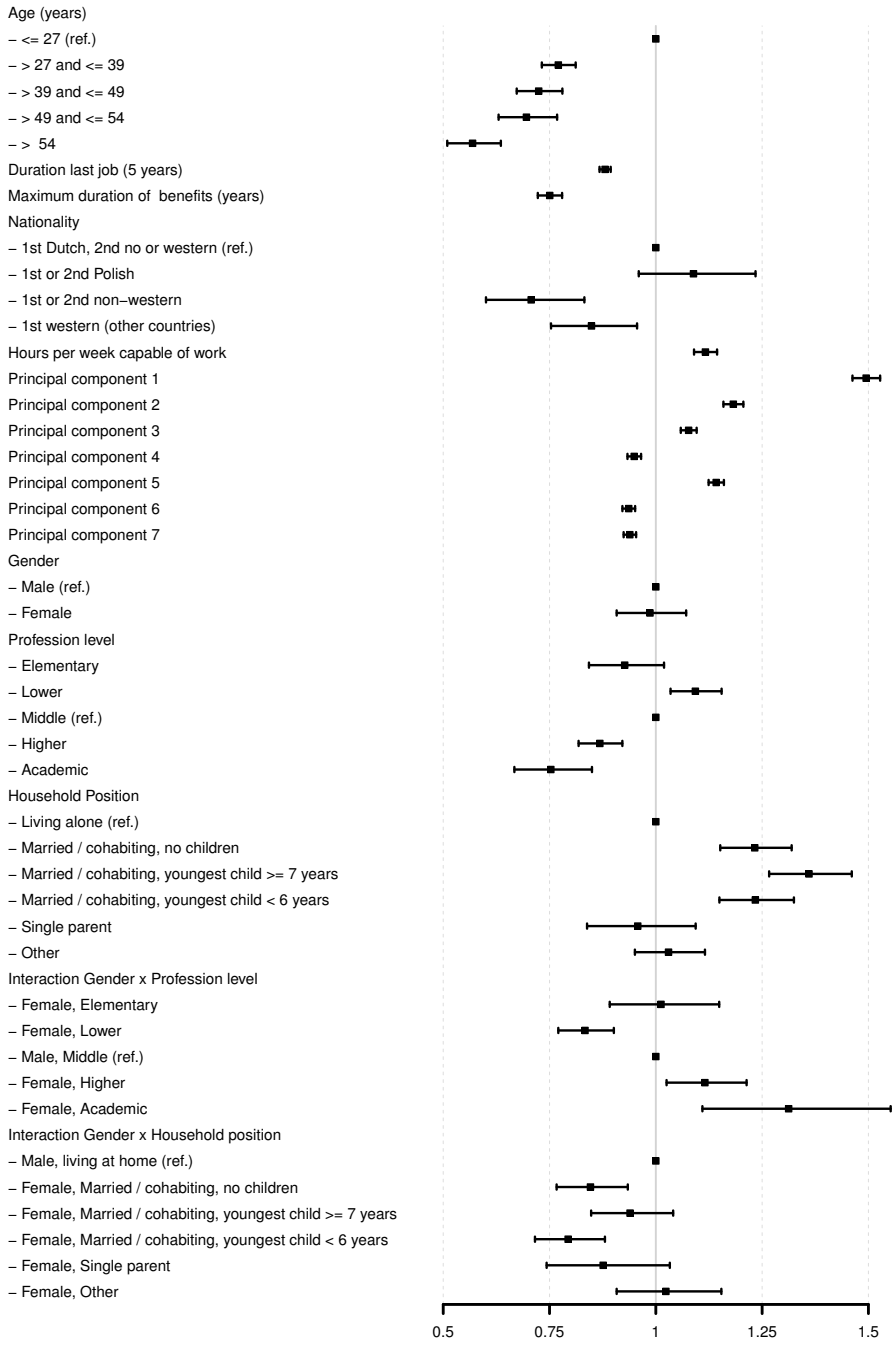

Figure 2.2: *Estimated hazard ratios and their 95% confidence intervals for the covariates in the final Cox model fitted on the hard factors and principal components. The reference category is indicated for each categorical variable.*

	Age (years)	Duration last job (years)	Gender	Household position	Maximum duration of benefits (years)	Nationality	Profession level	Days per week capable of work	PC 1	PC 2-7
Figure 2.3a	≤ 27	2.5	*All categories are shown*	Living alone	0.75	1st: Dutch 2nd: no/western	Middle	5	**1**	0
Figure 2.3b	≤ 27	2.5	*All categories are shown*	Living alone	0.75	1st: Dutch 2nd: no/western	Middle	5	**-1**	0
Figure 2.3c	$40-49$	10	**Male**	*All categories are shown*	2	1st: Dutch 2nd: no/western	Middle	4	1	0
Figure 2.3d	$40-49$	10	**Female**	*All categories are shown*	2	1st: Dutch 2nd: no/western	Middle	4	0	0
Figure 2.3e	$40-49$	10	**Male**	Married/ cohabiting, youngest child < 6 years	2	1st: Dutch 2nd: no/western	*All categories are shown*	4	0	0
Figure 2.3f	$40-49$	10	**Female**	Married/ cohabiting, youngest child < 6 years	2	1st: Dutch 2nd: no/western	*All categories are shown*	4	0	0

Table 2.5: *Covariate values used to estimate the survival curves shown in Figure 2.3. The only difference between the individuals in Figure 2.3a and Figure 2.3b is the value of Principal Component 1 (printed in bold). The differences between the individuals in Figure 2.3c and Figure 2.3d, and the individuals in Figure 2.3e and Figure 2.3f is gender (printed in bold).*

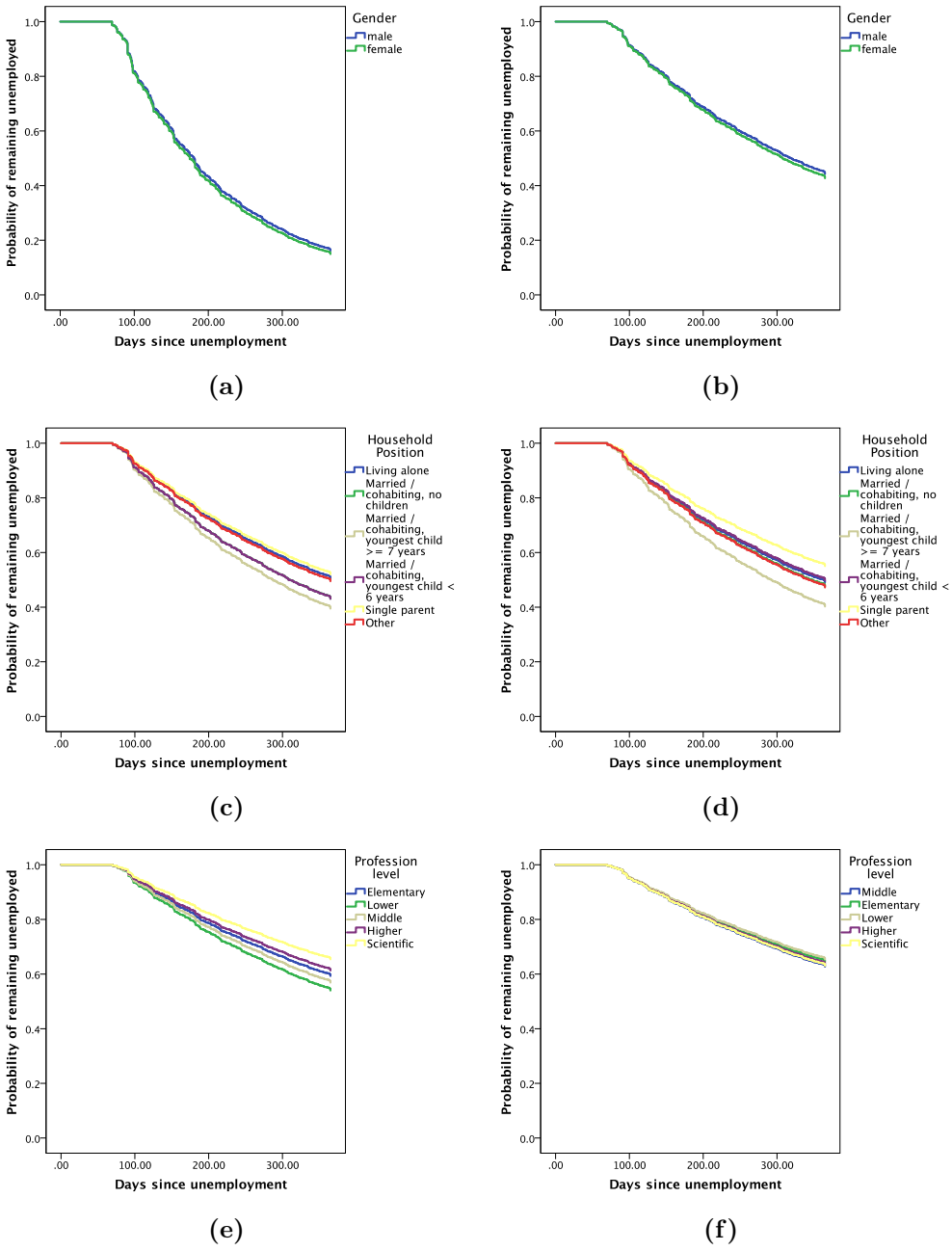

Figure 2.3: *Estimated probabilities of remaining unemployed for the six individuals with covariate values as given in Table 2.5. The upper panel shows the effect of the first principal component for young persons with little work experience and who live alone. The middle and lower panel respectively show the survival curves corresponding to different household positions and profession levels for both genders.*

2.4.3 Conclusions

The two-step analysis performed on the UWV data provided an interesting insight of the characteristics of the unemployed. The OS-PCA analysis revealed the correlation structure of the survey items from the extended version of the Work Profiler. The survival analysis provided possible predictors for unemployment duration.

The OS-PCA resulted in a model with seven components. The first component seemed to have the strongest association with unemployment duration. This component shows strong similarities with two theories on job search behavior: the Theory of Planned Behavior and the Motivation Model. Soft factors related to unemployment duration provided by the methodology discussed in this paper are the same factors suggested by the two theories. Additionally, health perception seemed to be important indicator for job search behavior.

Survival analysis results indicate age as one of the best indicators for unemployment duration. The model shows that older unemployed have the smallest probabilities of finding a new job. Furthermore, gender seems to have no significant effect on unemployment duration. However, some household positions are more disadvantageous for women when it comes to finding a new job, and the association of profession level with unemployment duration is different among the two genders.

This analysis was based on a dataset that only contained unemployed who claimed unemployment benefits and remained unemployed for at least ten weeks. Hence, it did not take into account that some characteristics may cause a very short (< 10 weeks) unemployment duration.

The analysis proposed in this paper aimed to illustrate how OS-PCA can be used to assess the correlation structure between survey items and how survival methodology can be applied to investigate covariates associated with reemployment duration. This is a preliminary analysis and is not meant for policy making. The analysis could be extended, for example by considering different settings in the OS-PCA, or by including some of the hard factors in the data preparation step as well. Depending on the research question, the analysis could be extended to also compare, for example, the probability of reemployment for different sectors or different regions.

2.5 Discussion

In this paper, two alternative methods for a two-step analysis of reemployment data were discussed. First, OS-PCA was proposed as an alternative to factor analysis to summarize item scores in survey data. Next, survival analysis was proposed as an alternative to logistic regression to analyze the probability of the

occurrence of an event when the time duration is of interest. In this section the advantages and disadvantages of these methods are compared

Usually many survey items are very closely related, which results in highly correlated data. Correlated data may cause problems when fitting regression models. Therefore, the item scores are often summarized in several composite scores which are not or only a little correlated. Using these composite scores as input in the model reduces the number of variables in the model. This usually simplifies the model interpretation.

Several methods have been developed to summarize items into summary scores. In this paper, two methods were illustrated: factor analysis and OS-PCA. Factor analysis summarizes related items into composite scores, resulting in factor scores. Each item corresponds to only one factor and the factors may be correlated. OS-PCA estimates several principal components. These components are a weighted average of all items and are uncorrelated. Component loadings indicate the importance of each item in each component.

Since factor scores are a weighted average of a subset of the items, they are usually easier to interpret than principal components in which all items are included in the composite score. As a result, the estimated factor model is easy to interpret. Furthermore, if researchers aim to reduce the number of items in a survey, factor analysis will allow them to remove factors and their corresponding items one by one. Since all items only belong to one factor, removing all items corresponding to one factor will not influence the scores of the other factors. On the other hand, in the OS-PCA setting, removing one item will influence all component scores, since these are weighted averages of all items. Hence, reducing the length of a survey can more easily be done using factor analysis. However, a disadvantage of factor analysis is that factor scores may still be correlated with each other. Therefore, the collinearity problem may still occur when factor scores are included in a statistical model. Principal components analysis overcomes this problem since the estimated components are uncorrelated.

A strong advantage of OS-PCA is its ability to transform categorical item scores nonlinearly, as opposed to the numerical interpretation in factor analysis. The nonlinearity transformation allows for the evaluation of categorical data without loosing data properties. In case unequal distances are expected between the categorical levels, one may consider using the OS-PCA method for dimension reduction.

Based on the differences between the two techniques, the choice between them mainly depends on the aims of a study. If the model should be easy to interpret and allow for removal of items, factor analysis is probably the best analysis method. If interpretation is less important, but keeping the properties of the ordinal categorical data is preferable, OS-PCA might be a solution for

dimension reduction.

Once survey data has been summarized, the summary scores can be used to predict the outcome of interest. In the context of reemployment data this could be the probability of being reemployed after a prespecified time point, or the actual duration of unemployment. The choice between logistic regression and survival analysis depends mostly on the outcome of interest. If interest lies in the reemployment probability at a specific time point, logistic regression is the appropriate model. However, if a binary outcome at a specific time point seems too strict, survival analysis is the proper methodology to be applied, since this method estimates the distribution of the unemployment duration. Although the research question should be leading in choosing the analysis method, other model properties might play a role as well.

Prediction error estimation may be one of these properties. If researchers want an easily interpretable prediction error, logistic regression may be the better choice. Since prediction is performed at a specific time point, one can check whether the prediction was correct at that time point. For example, if the probability of being reemployed within the set time was estimated to be $> 50\%$ for a person with particular characteristics, and this person had actually found a job before that time, one could say the prediction was correct. In this way, the ratio of correct predictions among all predictions gives an indication of the prediction error. A ratio of 0.5 indicates a bad performance, comparable to flipping a coin. A score of 1 indicates perfect prediction. Instead of 50%, other cut-off values could be chosen as well. The optimal cut-off value can be determined by minimizing the prediction error over all possible cut-off values.

In survival analysis, estimation of the prediction error is slightly more complicated. Since the actual time duration is included in the model, the prediction error is evaluated at a grid of time points. The Brier score (Graf et al., 1999), for example, computes at each time point the mean squared difference between the estimated survival probabilities and the actual outcomes (still alive/unemployed $= 0$, dead/reemployed $= 1$). Hence, a low Brier score is preferable. Although the Brier Score will give a good indication of the prediction error over the time grid, it is not possible to interpret it as the ratio of correctly predicted outcomes. The C-index (Harrell et al., 1996) might be used to quantify the ratio of concordant pairs of observations, i.e. the ratio of pairs of observations whose events were predicted to be in the same order as they actually occurred. This index is useful, but harder to interpret compared to the ratio of correct predictions used in logistic regression.

Another aspect that may play a role in the model choice is the possible presence of missing data. Since observing an event takes time, the events of some subjects may not be observed. In the context of reemployment data for

example, some reasons for not observing the time for reemployment might be emigration to another country, retirement, or death. In a logistic regression analysis, the outcome would be missing for these subjects, which makes the analysis more complicated. Survival analysis methods, however, were designed to include this type of subjects in the analysis. When time to event is not observed, the time until last contact would be included in the model. Depending on the type of incomplete information, the observation will be used as censored (as emigration) or as competing event (as retirement and death). It will include all the known information in the model as the censoring time, i.e. the date of emigration, retirement, or death, and will estimate the model while using this information.

In this paper, the advantages and disadvantages of the proposed methods to analyze reemployment data were discussed and the research questions that can be answered by each method were characterized. Aspects like interpretability, collinearity of variables, prediction error estimation, and missing data must be considered when presenting the method. Although this paper's focus is on the combinations of factor analysis with logistic regression and OS-PCA with survival analysis, researchers are not limited to only these two combinations.

Acknowledgements

The Dutch Employee Insurance Agency (UWV) is gratefully acknowledged for making the data available for this chapter. In particular, Harriët Havinga and Gijsbert van Lomwel (Kenniscentrum UWV) are acknowledged for their help on the interpretation of the covariates and model results.