ESMA

Event Sequence Analysis

Version 3M4

MANUAL FOR IMPLEMENTATION AND USE

Martin Schulz

Sauder School of Business
University of British Columbia
2053 Main Mall
Vancouver, BC V6T 1Z2
Canada
Martin.Schulz@sauder.ubc.ca

2004

INTRODUCTION

ESMA means "EreignisSequenzMusterAnalyse". This is a German expression which might be translated as "event sequence pattern analysis". The program ESMA analyzes sequences of events over time. The kind of data usually analyzed by ESMA are event data of a set of individuals (or other kinds of units). Each individual has a data record. The data record for each individual contains information about events experienced (and events not experienced) by each individual. ESMA reads the data records of a set of individuals from an input file, identifies the sequences of events present in the data, and computes various statistics of these sequences. ESMA also identifies patterns of event sequences. These patterns are sequences of events which occur most frequently in the analyzed data. Finally, ESMA produces a data set of sequences of events which allows users to estimate logit models of specific target events. I.e. it allows the user to model the effect of sequence patterns on the likelihood of occurrence of a specific target event.

ESMA builds upon (and is an extension of) previous work on event history analysis (e.g. Tuma and Hannan, 1984). Classical event history analysis focusses on estimating transition rates in state spaces. Classical event history models do not consider sequences of events. This means that in some cases classical event history models are less efficient because they span the whole state space (i.e. all possible transitions between states). In contrast, ESMA tries to identify which transitions are empirically present and identifies the main sequences in which they occur.

ESMA considers three characteristics of events: Status, time, and name. Status is 1 if the event occurred, and 0 otherwise. Time gives the point in time when the event occurred (if it occurred; if it did not occur it should carry the censoring time). Name is the kind of the event occurred. For example, an analysis of family formation processes might be interested in identifying events such as marriage, moving into joint household, birth of first child, entering the labor market, etc. The user has to assign numbers (between 1 and 98) to these kinds of events (I.e. event names are represented as numbers in ESMA; but note that ESMA allows the user to provide labels for the names -- see below).

Thus, each event can be represented as a triple (S,T,E), where S is status, T is time point, and E is event kind. The event data analyzed by ESMA have the following form:

```
W S1 T1 E1 S2 T2 E2 S3 T3 E3 ......
```

I.e.: W is a weight attached to a specific observation (this is useful for weighted data -- you have to set it to 1.0 if you have unweighted data), S1 is the status of the first event, T1 is the time of occurrence of the first event, E1 is the kind of the first event, etc. These data are read in by ESMA and are then analyzed.

ESMA reads these data in unformatted forma from the file "EVENTD". It expects that the event data in that file are separated by comma or blank (the technical term for that is "list-directed format"). For example, the following data line would be a valid data line in the EVENTD file:

0.8 1 20.3 1 1 23.4 2 0 28.3 3

This could be data from the life history of an idividual which left school at age 20.3 (event 1), married at age 23.4 (event 2), and has not yet divorced (event 3). The time the individual was last observed was at its age 28.3. To compensate for a bias in the sample this individual has been assigned a weight of 0.8 in the data.

An interesting feature of ESMA is that it considers "quasi-simultaneity". Events can occur simultaneously with other events. They can also occur nearly simultaneous with other events. This is called "quasi-simultaneity". Quasi-simultaneity occurs, for example if respondents do not recall the exact dates of occurrence of past events, or other cases in which the event times have measurement errors. ESMA takes explicitly care of this. You can tell ESMA that you think that your event dates are a little imprecise by telling it (see the FUZZY parameter below) that your time data are fuzzy and that you think that it should consider two events as simultaneous if they occur less than FUZZY time units apart (I think Courgeau and Lelièvre termed the expression "fuzzy time" first). In that case ESMA regards events which occur quasi-simultaneous as special cases of events. They are treated as specific events which are parts of event sequences. In the printout they are labeled as S1, S2, S3, etc (see the "Analyse der Simultanklassen in the LIST file").

IMPLEMENTATION AND USE

This version of ESMA is able to run on IBM-compatibles.The source code "ESMA3M4.FOR" is provided and can be modified to expand arrays and to run it on a mainframe (you should know a bit about Fortran and your compiler if you do that). The file "ESMA3M4.EXE" is the compiled program.

For installation on DOS/Windows systems I recommend that you:

1. create a specific subdirectory on your hard disk, called "esma"

2. copy the contents of the distribution disk to that subdirectory

You run ESMA by typing **esma3m4** to the DOS prompt. ESMA does not write to the screen. It puts all its output in the LIST file. The operating system informs you when ESMA is done (e.g. with the message "Execution terminated : 0"). After execution edit or print out the LIST file.

ESMA3M4 demands two input files "EVENTD" and "CESMA". ESMA will not run if it does not find these two files in the current directory[1]. ESMA creates 3 files: "LIST", "LOGITD", and "HELPFL". The function of these files will be explained now.

LIST contains the output from ESMA to be printed on a printer. **HELPFL** can be deleted after program execution. **LOGITD** contains the test data set for logit models etc. These data are written with format (`4I6,30I1`). Input this data set into a package which can estimate logit or probit models (e.g. GLIM, or SAS proc logistic).

EVENTD contains the event data in the usual form, that is sequences of triples of S T E (status, time, and event number). However, the first variable in each line must be the weight of each case. Thus, a line of data input might be:

[1] Please note that you can change filenames, as well as their path by changing the corresponding FORTRAN code in the subroutine XOPEN.

```
0.4  1 18.2 2  1 20.2 4  0 24.4 1  0 24.4 3  0 24.4 5
```

Where 0.4 is the weight of this case. The first triple (`1 18.2 2`) is the first event (status=1 => event occurred; time=18.2 => event occurred at age 18.2 years; eventnr=2 => the event was coded as event number 2 (only 2-digit numbers are allowed, and each event can occur only once during a lifetime of a case, i.e. there should not be 2 triples which have the same event numbers)). The next event occurred at age 20.2 and had event number 4. All other events were not experienced by this individual. They have the censoring date (the interview was taken at age 24.4) as time.

All variables in EVENTD must be separated by blanks or commas. Here are three valid data lines:

```
1.0 1 24.00 1 0 36.83 2 1 24.25 3 0 36.83 4
1.0 1 20.50 1 1 27.58 2 1 20.92 3 0 39.42 4
1.0 1 22.25 1 0 35.67 2 1 22.67 3 0 35.67 4
```

You also could have commas instead of blanks between the values as in the following case:

```
1.0,1,24.00,1,0,36.83,2,1,24.25,3,0,36.83,4
1.0,1,20.50,1,1,27.58,2,1,20.92,3,0,39.42,4
1.0,1,22.25,1,0,35.67,2,1,22.67,3,0,35.67,4
```

CESMA contains the control parameters. The first 11 parameters are read in list-directed format, that is, they can be separated by commas, carriage return etc. The following parameters are read list directed (I indicates Integer and R indicates Real variables): NEV(I), MIND(I), FACT(R), ZIEL(I), QS(I), DELT(R), TMIN(R), TMAX(R), FUZZY(R), NRES(I), and the list of selected event numbers (I). After this set of parameters the labels of the events are read in a fixed format (I3,A3). Note that ESMA wants to have all these parameters set, and it reads them in this sequence. The program disk contains an example file of CESMA.

The meaning of the parameters is as follows:

NEV: Number of events per case (i.e. input line). Remember that each event consists of 3 pieces of related information: Status, Time, and Event Name. For example, if NEV=4 then ESMA expects to find in the file EVENTD 13 numbers: One for the weight, and then 4 triples of STE data. NEV has to be smaller than 30. Note that ESMA expects a rectangular event data structure in the file EVENTD. That is, every unit needs to have the same number of events on its record. Events not experienced by a unit have to be coded as censored (99).

MIND: Instructs ESMA to print out event sequences which were experienced by at least MIND cases. In other words: print sequences which have at least MIND number of observed cases in the last event of the sequence. If MIND is set to 1 the whole event tree is printed out (which can lead to lots of printout because all rare sequences are also printed).

FACT: ESMA identifies patterns of event sequences. Event sequences are those sequences of events which occur with a relatively high frequency in the data. Of course, the observed frequency of an event sequence depends on its length. For example, it is more likely to observe a sequence consisting of two events than one of, say 10 events. ESMA compares event sequences of equal length and assigns

the most frequent ones pattern status. For that it uses a mathematical function which considers the length of the sequence. An event sequence is an event sequence pattern if its frequency in the data exceeds the following number:

$$B = 10 + Ne^{-2.3\,FACT + \sqrt{(L+1)}}$$

where N is the number of cases in the input data, and L is the length of the sequence analyzed (e.g. the sequence {wedding, first child born, divorced} has the length of 3). By modifying FACT you can increase or decrease the pattern criterion.

ZIEL: This parameter tells ESMA to consider a specific event in the data as the "target" event. This is important if one wants to find out how sequences of events affect the likelihood a particular target events (e.g birth of 3rd child, or divorce, etc). You can assign one target event only. ZIEL is the event name (i.e. its event number, see below) of the target event.

QS: This parameter tells ESMA how you want to treat cases in which the target event occurs quasi-simultaneous with other events. For example you might be interested to find out if marriage affects leaving the labor market and you want to exclude those cases in which both events occur at the same time. In that case set QS=0. In most cases you want to set QS=1.

DELT: This parameter sets the survival time interval for the computation of survival probabilities (G(DELT)). If you set DELT=2 ESMA computes the probability of an individual (which has experienced a given sequence of events) to survive two time units (e.g. years) without experiencing another event. See the legend in the printout for more information.

TMIN: This is the start time of the process clock. For example, if you are analyzing labor force participation of individuals you could set it to 10, assuming that none of your individuals has ever worked before age 10. If in doubt set TMIN to 0.0.

TMAX: This is the corresponding end time of the process clock. You can use this parameter to restrict the analysis to a specific age range (e.g. exclude retirement events).

FUZZY: This parameter sets the fuzzy time. Fuzzy time controls if two events occurring close together are considered as quasi-simultaneous. If you have precise time data and no measurement error then set FUZZY to 0.0.

NRES: You can select specific events from your data to be included in the analysis (thereby exclude the other events). If you want to include all events in the analysis set NRES equal to the value of NEV. The event names to be included in the analysis have to be listed on the next line in the CESMA file. For example, if you want to include only the events 1, 3, and 5, you would put the following two lines into CESMA:

```
3
1, 3, 5
```

LABELS: The last information in the CESMA file are the labels for the events. These have to be specified in a fixed format as follows: columns 1-3: event name (i.e. its number); 4-6: event label. For example, if you want to assign labels for events 1, 3, and 5, you might use the following lines:

```
1MAR
3COH
5BIR
```

Note that you can provide as many labels as you want -- not only for the events selected by the NRES parameter.

A note on event names: Valid event names for ESMA are integers between 1 and 98. Number 99 is reserved for the internally produced event "censoring". You might provide a label for that too. Furthermore, note that ESMA generates event-names for quasi-simultaneous events too. They are assigned event numbers 501, 502, 503, ... Their labels are S1, S2, S3, ...

A FINAL NOTE

You are granted to right to use the program ESMA. You do not have the right to sell the program code to other people. All publications of results produced with ESMA should indicate that computations were made with the program ESMA Version 3M4, and that the author is Martin Schulz. I would appreciate to receive copies of publications which use ESMA. If you have questions, write me, or send me e-mail. My address is:

Martin Schulz
Sauder School of Business
University of British Columbia
2053 Main Mall
Vancouver, BC V6T 1Z2
Canada
Martin.Schulz@sauder.ubc.ca