

A BIVARIATE DENSITY ESTIMATION APPROACH USING DATA FROM THE SURVEY OF LABOUR AND INCOME DYNAMICS: AN EXAMPLE OF TESTING FOR TEMPORAL ORDER OF JOB LOSS AND DIVORCE

Norberto Pantoja-Galicia¹, Mary E. Thompson² and Milorad Kovacevic³

ABSTRACT

The relationship between job loss and divorce is a topic of interest among the social scientists. The longitudinal data from the Survey of Labour and Income Dynamics (SLID) gives the possibility to approach this issue in a variety of ways. Dates of changes of the marital status (with associated type of change) as well as dates of employment can be obtained for a longitudinal person in a six years panel. For a person who is employed and married let T_1 and T_2 be the times of losing a job and becoming divorced respectively. A formal nonparametric test for a partial order relationship proposed by Thompson and Pantoja Galicia (2002) requires estimation of the joint density for T_1 and T_2 . This density estimation is modified to account for complexities of the sample design, and to allow the observed times to events to be interval censored.

KEY WORDS: Interval censored data, Longitudinal surveys, Nonparametric bivariate density estimation, Temporal order.

RÉSUMÉ

La relation entre la perte d'emploi et le divorce est un sujet qui suscite beaucoup d'intérêt chez sociologues. Les données longitudinales provenant de l'Enquête sur la dynamique du travail et du revenu (EDTR) donnent la possibilité d'aborder cette question de plusieurs manières différentes. Les dates de changement de statut social (avec les types associée) ainsi que les dates concernant l'emploi peuvent être obtenues pour une personne mariée avec un emploi établissons T_1 and T_2 les temps respectifs de la perte d'emploi et du divorce. Un test non paramétrique formel pour un relation d'ordre partielle a été proposé par Thompson et Pantoja Galicia (2002) mais requiert l'estimation de la densité conjointe pour T_1 and T_2 . Le test que nous proposons est modifié pour tenir compte de la complexité du plan d'échantillonnage et pour permettre aux temps jusqu'aux événements d'être censurés par intervalle..

MOTS CLÉS : Densité non-paramétrique bivariée; données censurés par intervalle; enquêtes longitudinales; ordre temporel.

1. INTRODUCTION

The interest in exploring a relationship between two events has led scientists to approach several methods of achieving this purpose. Thompson and Pantoja Galicia (2002), based on the notions of temporal association and time order, propose procedures to investigate these relationships. One of their methods examines whether one of the two events tends to precede the other closely in time. If this is the case, a causal interpretation of an association between these events might be more plausible.

Since longitudinal surveys allow the sequence of events for individuals to be observed, their role is crucial for this purpose. In this context, Thompson and Pantoja Galicia (2002), propose a formal nonparametric test for a partial order relationship using data from complex surveys. This test requires estimation of the joint density of the times to occurrence of the events in question. In the process of estimating this density function, consideration needs to be given to the fact that the observed times to occurrence of events are sometimes not completely known or are known to be within a trusted

¹ Norberto Pantoja-Galicia, University of Waterloo, Waterloo ON, Canada, N2L 3G1, (npantoja@uwaterloo.ca)

² Mary E. Thompson, University of Waterloo, Waterloo ON, Canada, N2L 3G1, (methomps@uwaterloo.ca)

³ Milorad Kovacevic, Statistics Canada, 120 Parkdale Ave. Ottawa ON, Canada, K1A0T6, (Milorad.Kovacevic@statcan.ca)

interval of time. This produces data that is interval censored in nature. In addition, the complexities of the sample design need to be taken into account.

The purpose of the paper involves two parts. In the first one, we will give a nonparametric approach to the estimation of the required density function. In the second, we will outline an implementation of the nonparametric test using data from the Survey of Labour and Income Dynamics as a possible illustration.

2. A TEST FOR TEMPORAL ORDER

2.1 Close precursor

The following approach is proposed by Thompson and Pantoja Galicia (2002):

Let F denote a survivor function. If T_1 and T_2 are duration times for events E_1 and E_2 respectively, we say T_1 is a close precursor of T_2 if for some positive numbers δ and $\kappa(t_1)$ we have

$$\frac{F_2(t_1 + \kappa(t_1)|T_1 = t_1)}{F_2(t_1|T_1 = t_1)} < \frac{F_2(t_1 + \kappa(t_1))}{F_2(t_1)} - \delta. \quad (1)$$

In other words, we say T_1 is a close precursor of T_2 if the occurrence of the first event E_1 at time T_1 decreases the probability that we have to wait longer than $\kappa(t_1)$ to observe the occurrence of the second event E_2 .

The idea is that the more closely the time to the second event (T_2) tends to follow the time to the first one (T_1), the greater the plausibility for a causal connection might be.

2.1 Nonparametric test for close precursor

For each t_1 and suitable $\kappa(t_1)$; a formal test for close precursor is given by

$$S = \int \left(\frac{\hat{F}_2(t_1 + \kappa(t_1)|T_1 = t_1)}{\hat{F}_2(t_1|T_1 = t_1)} - \frac{\hat{F}_2(t_1 + \kappa(t_1))}{\hat{F}_2(t_1)} \right) d\hat{F}_1(t_1). \quad (2)$$

In order to test the null hypothesis $H_0: S = 0$ (its mean is 0), the value of S can be compared with its estimated standard error multiplied by two.

3. BIVARIATE DENSITY ESTIMATION

3.1 A Kernel density estimation approach

It has been previously pointed out that there are situations in which the exact time of occurrence of an event is not completely observed, but observed only within a time interval. In this case, the time of occurrence is interval censored. In this situation, estimation of the joint distribution of T_1 and T_2 to apply the test described in the previous section requires a special treatment.

Estimation of univariate and multivariate density functions, in the case of independent and identically distributed random variables, is presented for example by Silverman (1986), Scott (1992), Simonoff (1996), and Venables and Ripley (2002) with material on kernel density estimation. Turnbull (1976), Gentleman and Geyer (1994) and Li, Watkins and Yu (1997) have proposed nonparametric estimators for the distribution function with univariate interval censored data. Density estimation for univariate interval censored data has been covered by Duchesne and Stafford (2001) and Braun, Duchesne and Stafford (2005). In the context of complex surveys research, density estimation is examined by Bellhouse and Stafford (1999), Bellhouse, Goia and Stafford (2003), and Buskirk and Lohr (2005).

In this section, using the ideas proposed by Duchesne and Stafford (2001) as a starting point, we present an extension of their method to the bivariate case. We take into account the interval censoring nature of the data as well as the complexities of the survey design.

For non-censored data Y_1, Y_2, \dots, Y_n , the standard kernel density estimate is given by the expression

$$\hat{f}_{nc}(y) = n^{-1} \sum_{i=1}^n K_h(Y_i - y) = E_{F_n} [K_h(Y - y)], \quad (3)$$

where nc stands for noncensored, $K_h(u) = h^{-1}K(h^{-1}u)$ is a kernel function with bandwidth h and F_n is the empirical distribution of the sample.

Let $X_i \in I_i = (A_i, B_i)$ be randomly interval censored data with real numbers A_i and B_i , and $i = 1, \dots, n$. Duchesne and Stafford (2001) approach the corresponding kernel density estimation in the following manner:

$$\hat{f}(x) = n^{-1} \sum_{i=1}^n E [K_h(X_i - x) | X_i \in I_i]. \quad (4)$$

In order to estimate the required density they propose to compute the following expression

$$\hat{f}(x) = n^{-1} \sum_{i=1}^n E_{\hat{f}} [K_h(X_i - x) | X_i \in I_i], \quad (5)$$

which may be solved using this iterative approach:

$$\hat{f}_j(x) = n^{-1} \sum_{i=1}^n E_{\hat{f}_{j-1}} [K_h(X_i - x) | X_i \in I_i], \quad (6)$$

where expectation is with respect to the conditional density $\hat{f}_{j-1;i}(u) = \delta_i(u) \hat{f}_{j-1}(u) / c_{j-1;i}$ over I_i ; and where $\delta_i(u)$ is an indicator function for I_i ; and $c_{j-1;i}$ is the unconditional expectation under \hat{f}_{j-1} .

Now, let us generalize to the bivariate case. Considering noncensored data, a version of the traditional kernel density estimator comes next as

$$\hat{f}_{nc}(x_1, x_2) = n^{-1} \sum_{i=1}^n K_{h_1, h_2}(X_{i,1} - x_1, X_{i,2} - x_2) = n^{-1} \sum_{i=1}^n K_{\underline{h}}(\underline{X}_i - \underline{x}), \quad (7)$$

where $K_{\underline{h}}(\underline{u})$ is a bivariate kernel with bandwidth $\underline{h} = (h_1, h_2)$ and $\underline{X}_i = (X_{i,1}, X_{i,2})$, $\underline{x} = (x_1, x_2)$.

In the context of interval censored data, \underline{X}_i lies within the 2-dimensional interval $\underline{I}_i = (A_{i,1}, B_{i,1}) \times (A_{i,2}, B_{i,2})$. Therefore, for the bivariate case, an extension to the idea of Duchesne and Stafford (2001) gives the corresponding kernel density estimate in terms of iterated conditional expectation as follows

$$\hat{f}_j(\underline{x}) = n^{-1} \sum_{i=1}^n E_{\hat{f}_{j-1}} [K_{\underline{h}}(\underline{X}_i - \underline{x}) | \underline{X}_i \in \underline{I}_i]. \quad (8)$$

As it is in the univariate case, expectation is also with respect to the conditional density $\hat{f}_{j-1;i}(\underline{u}) = \delta_i(\underline{u}) \hat{f}_{j-1}(\underline{u}) / c_{j-1;i}$ over \underline{I}_i ; and where $\delta_i(\underline{u})$ is an indicator function for \underline{I}_i ; and $c_{j-1;i}$ is the unconditional expectation under \hat{f}_{j-1} .

Duchesne and Stafford (2001) estimate the conditional expectation using an importance sampling scheme. In the same way, we apply this approach to our bivariate scenario. Let us define

$$\mu_{j-1|\underline{I}}(\underline{x}) = E_{\hat{f}_{j-1}} [K_{\underline{h}}(\underline{X} - \underline{x}) | \underline{X} \in \underline{I}]. \quad (9)$$

So, we may use

$$E_{\hat{f}} [K_{\underline{h}}(\underline{X} - \underline{x}) | \underline{X} \in \underline{I}] = E_g [K_{\underline{h}}(\underline{X} - \underline{x}) w(\underline{X})], \quad (10)$$

where g is some distribution, easy to sample from, over the interval \underline{I} , and $w(\underline{X}) = \hat{f}_{j-1}(\underline{X})/g(\underline{X})$ is the importance sampling weight.

The desired conditional expectation, $\hat{\mu}_{j-1|\underline{I}}(\underline{x})$, may be approximated by the following expression:

$$\hat{\mu}_{j-1|\underline{I}}(\underline{x}) = \sum_{b=1}^B [K_{\underline{h}}(\underline{X}_b - \underline{x}) w_b^u], \quad (11)$$

where $w_b^u = w(\underline{X}_b) / \sum_{k=1}^B w(\underline{X}_k)$ and the \underline{X}_b are generated using a uniform sampling scheme derived from the *orthogonal array-based Latin hypercubes* described by Tang (1993).

Once an estimate of the conditional expectation can be obtained using the previously mentioned importance sampling approach, the density estimator may be obtained in the following manner:

$$\hat{f}_j(\underline{x}) = n^{-1} \sum_{i=1}^n \hat{\mu}_{j-1|\underline{I}}(\underline{x}). \quad (12)$$

Note that this estimator does not take into account the complexities of the survey design.

3.2 Incorporation of the survey weights

In the case of complex surveys administered by Statistics Canada such as the Survey of Labour and Income Dynamics (SLID), the national statistical agency prepares longitudinal weights that take into account some of the complexities of the survey design. Let w_i^l be the longitudinal weight derived by Statistics Canada for the i^{th} individual on the survey sample. These survey weights compensate for some of the complexities of the survey design: non response, selection bias, stratification and postratification.

Let w_i^n be the standardized or normalized weight for the i^{th} individual on the survey sample such that $\sum_{i \in S} w_i^n = 1$, where S corresponds to the final subsample. According to the idea of replacing population totals by weighted totals and accounting for some complexities of the survey design, we estimate the required density function using the expression:

$$\hat{f}_j^w(\underline{x}) = \sum_{i \in S} \hat{\mu}_{j-1|\underline{I}}(\underline{x}) w_i^n. \quad (13)$$

4. EXAMPLE

4.1 The Survey of Labour and Income Dynamics

The Survey of Labour and Income Dynamics (SLID) is a longitudinal survey composed of panels of six years in length. The purpose of this survey is to track the experiences of individuals in the labour market, their income and changes in family life. The first panel started in 1993 and consisted of about 15,000 households, which account for approximately

40,000 people (31,000 persons who are over 16 years of age). Subsequent panels increased the number of households by about 2,000 units. The sample is taken from the Labour Force Survey (LFS), and the dwelling place is the last-stage sampling unit. All household members of the selected dwellings are included in the LFS sample. Further details regarding design as well as other important issues of the Survey of Labour and Income Dynamics can be found at different documents published by Statistics Canada. We include a list in our references.

The Survey of Labour and Income Dynamics provides important elements to explore a relationship between job loss and separation or divorce. This topic is of interest in the social sciences, as discussed by Charles and Stephens (2004), Huang, J. (2003) and Yeung and Hofferth (1998).

4.2 Job loss and divorce

Let T_1 denote the time to occurrence of event E_1 . In this case E_1 is the termination of the job of the subject. In a similar way, let T_2 be the time to occurrence of event E_2 . This event is either separation or divorce, whichever applies or comes first (after the termination of the marriage or common-law relationship of the individual).

In a certain panel of SLID we can obtain a vector of all the dates of changes of marital status for each individual with associated type of change. Similarly, a vector of job history can be obtained for dates of changes in employment status. Once this information is retrieved we are in a position to generate an appropriate data set for our own subsample. We take into account individuals from the first panel of the SLID and consider the time origin to be the date of the first interview, i.e. day zero. We also consider respondents who are both married or in a common-law relationship, and employed at the time origin. Our sample involves people with only one marriage (or common-law union) and one job during the life of the panel. This job is reported to have ended during the life of the panel due to "involuntary" reasons and we condition on observing both events, during the time window from January 1993 to April 1999.

If we define the following notation:

- D_U : Date at which the union started. Either marriage or common-law (whichever came first).
- D_J : Date at which the job started.
- D_I : Date of the first interview.
- D_T : Date of the termination of the panel.
- D_{E1} : Date of the occurrence of event one (end of job).
- D_{E2} : Date of the occurrence of event two (end of the marriage or common-law relationship).

Then, according to the description of our sample, these dates are restricted to $D_U \leq D_I$ and $D_J \leq D_I$. Also $D_I < D_{E1} \leq D_{E2} < D_T$ or $D_I < D_{E2} \leq D_{E1} < D_T$.

4.3 Determination of the Event Times

Memory plays an important role in survey responding. Whenever a date of an event is reported, there exists the potential for dating errors. *Forward telescoping* is a type of memory error which involves reporting the occurrence of events more recently than they actually happened. The events are seen as closer in time than they really are, according to the interview's vantage point. As stated by Tourangeau et al. (2000), this phenomenon has been studied by survey methodologists and cognitive psychologists since Neter and Waksberg (1964) first documented it. In the opposite direction, *backward telescoping* is another possibility for a dating error. Tourangeau et. al (2000, chapter 4) present a vast review of the literature documenting these sort of memory errors. From the same source (page 11), we quote:

"Reporting errors due to incorrect dating seem to arise through several distinct mechanisms. People may make incorrect inferences about timing based on the accessibility (or other properties) of the memory, incorrectly guess a date within an uncertain range, and round vague temporal information to prototypical values (such as 30 days)."

These issues might be reflected in survey data as measurement error. In reports like those from Huttenlocher et al. (1990) it is also shown that respondents round estimates to times that are stand-ins for calendar units, that is seven days or thirty

days. On this basis, we might decide whether to trust the reported dates of the events to be within a week or a month instead of a specific day. This leads us to have times to occurrence that are interval censored for the events of our interest.

In a simple approach, of which the intention is to serve as an illustration, we considered the reported dates to be trusted within an interval of thirty days and obtained these intervals, that we called I_1 and I_2 , in a straightforward manner. Following the corresponding notation we have: $I_1 = (D_{E1} - D_1 - 14, D_{E1} - D_1 + 15)$ and $I_2 = (D_{E2} - D_1 - 14, D_{E2} - D_1 + 15)$.

4.4 Application

The size of the sample satisfying the conditions described in section 4.2 is of 70 individuals, who represent about 49,000 people of the total target population. The corresponding intervals I_1 and I_2 have been determined for every subject in our sample. According to section 3.1, estimation of the joint density for T_1 and T_2 was obtained using the product normal kernel.

Incorporation of the survey weights was implemented as described in section 3.2, so that the complexities of the survey design are taken into account. Figure 1 shows the contour plot of the estimated joint density of T_1 (time until the end of job) on the horizontal axis and T_2 (time until separation or divorce) on the vertical axis. This picture evidently shows the expected order of the times.

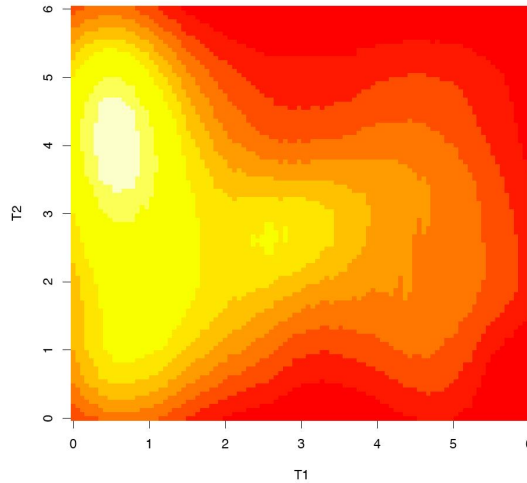


Figure 1: Contour Plot of the Estimated Joint Density of T_1 and T_2 .

Concerning the test for *close precursor* referred to in section 2.1, if we let $\kappa(t_1)$ to have a constant value of 6 months, for every t_1 , S will have a value of 0.034 with a standard error of 0.0025. The comparison of S with twice its estimated standard error, gives us evidence to reject the null hypothesis that its mean is 0. Therefore, we can argue that T_1 is a close precursor of T_2 , i.e. losing a job at time T_1 decreases the probability that you have to wait longer than 6 months to observe a separation or divorce.

The standard error of S was assessed according to the method proposed by Rao and Wu (1988) by using bootstrap replicates of the survey weights provided by Statistics Canada.

5. FINAL COMMENTS

Some important specifications are considered in the selection of our sample. First, conditioning on being at risk for both events at the day of the first interview. This sacrifices the possibility of including individuals who become married and/or employed after day 0, for whom different distributions will apply. Second, observing both events in the time window of the life of the panel. This conditioning is reasonable and gives us a sensible estimate since we are looking at close following or close order. Third, the fact of selecting those individuals whose job is reported to have ended during the life

of the panel due to “involuntary” reasons. This inclusion is intended to avoid as much as possible a potential effect of divorce as a trigger for job loss.

The choice of the bandwidth employed for the kernel density estimation presented in section 2 has been done in a subjective manner and therefore selection of an optimal bandwidth is left as a topic for further research in the context presented here.

ACKNOWLEDGEMENTS

We would like to thank Statistics Canada, the National Program on Complex Data Structures (NPCDS), the Mathematics of Information Technology and Complex Systems (MITACS), and the National Council on Science and Technology (CONACYT). Discussions with James Stafford are gratefully acknowledged.

REFERENCES

- Bellhouse, D.R. and Stafford J.E. (1999), “Density Estimation from Complex Surveys”. *Statistica Sinica*, **9**, 407-424.
- Bellhouse, D.R., Goia, C.M. and Stafford J.E. (2003), “Graphical Displays of Complex Survey Data through Kernel Smoothing”. *Analysis of Survey Data*. P. 133-150. John Wiley & Sons Ltd.
- Braun, J., Duchesne, T. and Stafford, J. E. (2005), “Local Likelihood Density Estimation for Interval Censored Data”. *The Canadian Journal of Statistics*, **33**, 39-60.
- Buskirk, T.D. and Lohr S.L. (2005). “Asymptotic properties of kernel density estimation with complex survey data”. *Journal of Statistical Planning and Inference*. **128**, 165-190.
- Charles, K. K. and Stephens, M. Jr., (2004) “Job Displacement, Disability and Divorce”. *Journal of Labour Economics*, **22**, 489-522.
- Duchesne, T. and Stafford, J. E. (2001), “A Kernel Density Estimate for Interval Censored Data”. *Technical Report No. 0106*, University of Toronto.
- Gentleman, R. and Geyer, C.J. (1994). “Maximum Likelihood for Interval Censored Data: Consistency and Computation”. *Biometrika*. **81**, 618-623.
- Huang, J. (2003), “Unemployment and Family Behavior in Taiwan”. *Journal of Family and Economic Issues*, **24**, 27-48.
- Huttenlocher, J., Hedges, L.V. and Bradburn, N.M. (1990). “Reports of Elapsed Time: Bounding and Rounding Processes in Estimation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **16**, 196-213.
- Li, L., Watkins, T. and Yu, Q. (1997). “An EM Algorithm for Smoothing the Self-consistent Estimator of Survival Functions with Interval-censored Data”. *Scandinavian Journal of Statistics*, **24**, 531-542.
- Neter, J., and Waksberg, J. (1964). “A Study of Response Errors in Expenditures Data from Household Interviews”. *Journal of the American Statistical Association*, **59**, 17-55.
- Rao, J.N.K. and Wu C.F.J. (1988), “Resampling Inference with Complex Survey Data”. *Journal of the American Statistical Association*, **83**, 231-241.
- Scott D. W. (1992). *Multivariate density estimation: theory, practice, and visualization*. New York: Wiley.
- Silverman B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman-Hall.
- Simonoff J. S. (1996). *Smoothing Methods in Statistics*. Springer.

Statistics Canada. SLID – A Survey Overview. www.statcan.ca/english/freepub/75F0011XIE/free.htm

Statistics Canada. SLID User's Guide. www.statcan.ca/english/freepub/75M0001GIE/free.htm

Statistics Canada. SLID electronic data dictionary. <http://www.statcan.ca/english/IPS/Data/75F0026XIB.htm>

Tang, Boxin (1993). "Orthogonal Array-Based Latin Hypercubes", *Journal of the American Statistical Association*, **88**, 1392-1397.

Thompson M. and Pantoja Galicia N. (2002), "Interval Censoring of Smoking Cessation in the National Population Health Survey". *Proceedings of Statistics Canada Symposium*.

Tourangeau, R., Rips, L. J. and Rasinski, K. (2000). *The Psychology of Survey Response*. New York: Cambridge University Press.

Turnbull B.W. (1976). "The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data". *Journal of the Royal Statistical Society. Series B (Methodological)*, **38**, 290-295.

Yeung, W. Jean and Hofferth Sandra L. (1998), "Family Adaptations to Income and Job Loss in the U.S". *Journal of Family and Economic Issues*, **19**, 255-283.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York, Springer.