**Inferring Order of Events from Paired Comparisons**

Farrokh Alemi, Ph.D.

**Procedures for Temporal Analysis in Network Construction.** Network based mediation requires knowledge of the temporal order of independent variables. In EHR data, it is relatively easy to establish the temporal order for a pair of variables, as the entrees in EHR are timestamped. The difficulty arises as there is a great deal of variability among individuals. For some one variable occurs before another and for others the reverse happens. A procedure is needed that can address this variability. If $w\_{ij} $is the number of times that variable j occurs after variable i, and *Wi* is the number of times it occurs later than all other variables, then starting from an arbitrary vector **p**, the algorithm iteratively performs the update for all variables so that:

$$p\_{i}^{'}=W\_{i}\left(\sum\_{j\ne i}^{}\frac{w\_{ij}+w\_{ji}}{p\_{i}+p\_{j}}\right)^{-1} Normalized so p\_{i}={p\_{i}^{'}}/{\sum\_{j=1}^{n}p\_{i}^{'}}$$

the estimates are normalized so that This procedure is based on a well-known technique in psychology for inferring the value of an option from comparisons of pairs of alternatives. It has also been successfully used to rank sport teams, even if some pairs of teams have not played each other. The procedure is widely available, including in an R package [[[1]](#endnote-1)].

The problem with paired comparisons is that these statistics can be contradictory. For some individuals, the variable “i” occurs before, and for other individuals, it occurs after variable “j”. To be consistent, one could think of the age at which a variable occurs as normally distributed and rely on the expected values of the variables to order them. On average, diagnoses and related treatments tend to occur at certain ages. For example, dementia and Alzheimer’s tend to occur later in life [[[2]](#endnote-2)]. Cardiovascular illnesses tend to have their onset when individuals are in their sixties and seventies [[[3]](#endnote-3)]. Substance abuse tends to start when individuals are 20 years old [[[4]](#endnote-4)]. While there are some tendencies for diagnosis to occur at certain ages, there is also a great deal of individual variability. Some individuals never have the illness; some have an early onset of late-life disease. While drug abuse tends to occur in young people, it also occurs to people at end of their lives [[[5]](#endnote-5)]. These variations in occurrences of variables create intransitive circular triads. A circular triad may produce a temporal order where k occurs before j, which occurs before i, which illogically also occurs before k.

|  |
| --- |
| **Figure 5: Variables Occurring in Reverse Temporal Order of Expected Values** |

Obviously, time cannot be circular. A theory is needed for why these circular temporal orders occur and how should they be resolved. One way to explain away the concern with circular triads is as the consequence of random variation. These random variations could be the result of independent, unobserved causes that may, for example, cause diabetes among patients who are not obese. Assume that every variable occurs at an expected time, plus or minus a random element. Imagine two variables whose time of occurrence have overlapping normal distributions (see **Figure 5** for variables *U* in blue and *V* in red). For some individuals, the two variables may occur in reverse order of their expected values. In **Figure 5**, for individual 2, *V2* occurs after *U2*, in contradiction to the order of expected values of ***U*** and *V*. These contradictions are more likely to occur if the two variables occur close in time (leading to smaller differences in their expected values), or if one of the variables occurs at wide range of ages (leading to larger standard deviation for at least one of the distributions). A method is needed to decide when two variables have a sufficient number of circular triads to make distinguishing their time of occurrence unreliable. The Coefficient of Internal Reliability, $ϑ, $is a statistic that can help investigators make these decisions [[[6]](#endnote-6)]. It is calculated from the frequency of circular triads as:

$$ϑ=1-\frac{C}{C\_{max}} C\_{max}=\frac{r\left(r^{2}-1\right)}{24} for odd r, C\_{max}=\frac{r\left(r^{2}-4\right)}{24} for even r$$

Where C is the count of circular triads in the matrix of binary temporal comparison of pairs of variables and r is the number of variables.

Circular triads can be reduced by re-defining variables so that they are, on average, several months apart. A technique called “double sorted” matrix can identify the pair of variables that are most responsible for circular triads [[[7]](#endnote-7)]. Once the pair of variables responsible for most circular triads has been identified, then one can revise the definition of one of the two variables. If variables *V* (in red) and *U* (in blue) are responsible for significant circular triads, then variable *V* is split into two variables and variable *U* is used as an anchor to do so. The two new variables will be “*V* prior to *U*”, and “*V* after *U*.” For example, if obesity and diabetes are responsible for significant circular triads, then obesity is defined as two variables: “obesity prior to diabetes” and “obesity post diabetes.” The circular variables are redefined in this fashion until the overall Coefficient of Internal Reliability, $ϑ,$ is no longer statistically significant and the order of mediators can be established reliably.

1. Bradley , Terry ME. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. Biometrika 1952. 39 (3/4): 324–345. [↑](#endnote-ref-1)
2. von Strauss E, Viitanen M, De Ronchi D, et al. Aging and the occurrence of dementia: findings from a population-based cohort with a large sample of nonagenarians. Arch Neurol. 1999 May;56(5):587-92. [↑](#endnote-ref-2)
3. Rodgers JL, Jones J, Bolleddu SI, Vanthenapalli S, Rodgers LE, Shah K, Karia K, Panguluri SK. Cardiovascular Risks Associated with Gender and Aging. J Cardiovasc Dev Dis. 2019 Apr 27;6(2):19. [↑](#endnote-ref-3)
4. Guerin AA, Kim JH. Age of Onset and Its Related Factors in Cocaine or Methamphetamine Use in Adults from the United States. Int J Environ Res Public Health. 2021 Nov 22;18(22): 12259. [↑](#endnote-ref-4)
5. Koechl B, Unger A, Fischer G. Age-related aspects of addiction. Gerontology. 2012;58(6):540-4. [↑](#endnote-ref-5)
6. Kendall MG, Smith BB. On the method of paired comparisons. Biometrika 31: 324-345, 1940. [↑](#endnote-ref-6)
7. Brown TC, Peterson GL. An inquiry into the method of paired comparison: reliability, scaling, and Thurstone’s law of comparative judgment. United States Department of Agriculture, Forest Service, Rocky Mountain Research Station, RMRS-GTR-216WWW, January 2009. [↑](#endnote-ref-7)