Chemometrics and Intelligent Laboratory Systems 154 (2016) 194–206

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Two-mode K-spectral centroid analysis for studying multivariate longitudinal profiles☆

Joke Heylen, Iven Van Mechelen, Eiko I. Fried, Eva Ceulemans *

KU Leuven, Research Group of Quantitative Psychology and Individual Differences, Tiensestraat 102-bus 3713, 3000 Leuven, Belgium

ARTICLE INFO

Article history:
Received 30 October 2015
Received in revised form 3 March 2016
Accepted 8 March 2016
Available online 17 March 2016

Keywords:
K-spectral centroid analysis
Biclustering
Time profiles
Multiway data

ABSTRACT

In many scientific areas, researchers collect multivariate time profile data on the evolution of a set of variables across time for multiple persons. For instance, clinical studies often focus on the effects of an intervention on different symptoms for multiple persons, by repeatedly measuring symptom severity for each symptom and each person. To pursue an insightful overview on how these time profiles vary as a function of both symptoms and persons, we propose two-mode K-Spectral Centroid (2M-KSC) analysis, which is a multivariate extension of K-Spectral Centroid analysis. Specifically, 2M-KSC assigns the persons to a few person clusters and the symptoms to a few symptom clusters and imposes that the time profiles that correspond to a specific combination of a person cluster and a symptom cluster have the same shape, but may vary in amplitude scaling. An algorithm for fitting 2M-KSC is proposed and evaluated in a simulation study. Finally, the new method is applied to time profiles regarding the severity of depression symptoms during a citalopram treatment.

1. Introduction

In many research areas interest in understanding how multiple variables change over time increases. Good examples, on which we focus in this paper, are intervention studies, targeting specific medical or psychological problems (e.g., [1,2]). In such studies, one often measures the evolution of multiple symptoms for multiple persons at consecutive time points, where a score of zero reflects the absence of a symptom. For instance, the NIH-supported STAR*D study (data version 3.0) [3,4], which we will revisit in this paper, mapped the severity of fourteen depressive symptoms for clients with major depressive disorder (MDD), and receiving a citalopram treatment, across several weeks.

The evaluation of such time profiles allows the researcher to address several important questions, including: How fast does the effect of the intervention kick in for different symptoms? Do relapses occur for some symptoms? When is full effect of the treatment reached? Are these effects across time the same for the different symptoms or can a few symptom groups be discerned each reacting differently? Similarly, what about individual differences: Is there evidence that the shape of the time profiles depends on the persons involved, and, if so, which types of persons react similarly? Obviously, these questions are important, since they allow predicting how a specific individual with a particular symptom profile would react to the intervention under study.

To further clarify these research questions and the associated modeling challenges, it is instructive to briefly review the different modeling approaches for analyzing time profiles. To this end, it is useful to distinguish between three modeling levels (for similar distinctions, see [5,6]): the *phenotype level*, the *constituent level*, and the *generating level*. Approaches at the *phenotype level* model examine which time profiles have the same manifest appearance, for instance, by clustering them into a few types. Approaches at this level differ in which profile characteristics are taken into account or sidelined when deciding whether profiles have the same shape or not. Specifically, one may take timing differences (i.e., phase variability, [7,8]) between the time profiles into account, by deciding that profiles that are time shifted (complete profile is shifted by a few time points) or warped (compressing some parts of the profile while stretching out others) versions of another differ in shape. If such differences are sidelined, however, they are removed before conducting the shape comparison. This implies that within each type, room is left for heterogeneity with respect to differences in these profile characteristics. The same holds for intensity differences (i.e., amplitude variability) between the time profiles, such as intensity shifting (complete profile is shifted in intensity by adding a scalar) or amplitude scaling (complete profile is deflated or inflated by multiplying it with a scalar).

At the *constituent level* approaches focus on the underlying constituents or components of the time profiles. For example, growth curve and trajectory models (e.g., [9,10,11]) can be situated at this level, as they model time profiles as a weighted sum of linear, quadratic, etc. basis functions and thus summarize the profiles in terms of intercepts and

☆ Selected papers from ThRee-way methods In Chemistry And Psychology conference, 7th edition, 2015.

* Corresponding author. Tel.: +32 16 325881.
E-mail addresses: joke.heylen@kuleuven.be (J. Heylen), iven.vanmechelen@kuleuven.be (I. Van Mechelen), eiko.fried@kuleuven.be (E.I. Fried), eva.ceulemans@kuleuven.be (E. Ceulemans).

http://dx.doi.org/10.1016/j.chemolab.2016.03.009
0169-7439/

slopes. Another example of a *constituent level* method is the method of Heard, Holmes, Stephens, Hand and Dimopoulos [12], as it models time profiles as weighted combinations of prespecified nonlinear basis functions and clusters profiles based on these weights.

At the *generating level*, one is interested in the mechanism that generates the time profiles and aims to discover the underlying laws. For instance, approaches using differential equations, that relate observed scores on a variable (e.g., symptom severity) to its rate of change (e.g. [13]), or Markov approaches, where the present state of a variable (e.g., symptom severity) is dependent on the immediately preceding state only (e.g. [14]), fall within this level.

If we return to our research questions – does the shape of the time profiles vary as a function of the persons and symptoms under study, and can person types and symptom groups be induced that have similar time profiles –, it is clear that the resulting modeling challenges pertain to the manifest appearance of the profiles and thus are located at the *phenotype level*. Moreover, in the case of symptom profiles, timing differences should be taken into account when categorizing the profiles as similar or not, since they can be meaningfully interpreted as delayed or accelerated reactivity to the intervention. In addition, vertical profile or severity shifts should be taken into account as well, not in the least because zero severity values have a clear meaning (viz., symptom absence), which disappears after an upward profile shift. Differences in amplitude scaling can be sidelined, however, because they might be due to differences in the overall severity of the symptoms or their wording (e.g., suicidal thoughts vs. waking up too early) or to inter-individual differences in response style; note that such scaling differences do not affect zero scores.

Among the existing approaches at the phenotype level, the method that most closely meets our modeling needs is K-Spectral Centroid (KSC) analysis ([15]; for an application in emotion psychology, see [16]). KSC clusters time profiles based on their shape, while allowing for amplitude scaling differences among the profiles that belong to the same cluster. However, KSC is a univariate method, in that it models differences in the time profiles of one symptom, or one variable in general. Therefore, the aim of this paper is to develop a multivariate extension of KSC, called two-mode KSC (2M-KSC), that allows modeling how time profiles vary as a function of both the persons and the symptoms under study. Specifically, 2M-KSC assigns the persons to a few persons clusters and the symptoms to a few symptom clusters and imposes that the time profiles that correspond to a specific combination of a person cluster and a symptom cluster have the same shape, but may vary in amplitude scaling.

The remainder of this paper is organized as follows: In the next section, the new 2M-KSC model is introduced. In Section 3, we discuss the 2M-KSC loss function and an algorithm for estimating the model parameters. Next, we elaborate on model selection. Section 4 reports a simulation study to evaluate the performance of this algorithm. In Section 5 we apply 2M-KSC to dataset version 3.0 of the STAR*D study. Finally, in Section 6, we demonstrate the usefulness of 2M-KSC in other domains of application and compare our method with existing, related phenotype methods.

2. Model

As stated above, 2M-KSC is a model for multivariate time profiles. More specifically, 2M-KSC assumes that J symptoms are measured at T time points for I persons. The T time points are comparable across the persons and the symptoms, implying that the data can be meaningfully arranged in a three-way three-mode data array $\underline{\mathbf{X}}$. Throughout this subsection we will make use of the hypothetical data set in Fig. 1, which consists of time profiles of the day-to-day severity of 4 depression symptoms collected for five MDD persons, across 10 treatment days. This data set can be perfectly reconstructed by a 2M-KSC model.

2M-KSC simultaneously clusters the I persons into K person clusters and the J symptoms into C symptom clusters. This clustering is exclusively based on the shape of the time profiles under study, discarding any amplitude scaling differences (while taking into account time shifts, warps, and severity shifts, see Introduction). All the time profiles that correspond to a specific combination of a person cluster and a symptom cluster are modeled with one particular reference profile, which reflects their typical evolution over time. Furthermore, each observed time profile receives an amplitude score, indicating its overall intensity relative to its corresponding reference profile. Specifically, this amplitude score indicates how much the reference profile has to be inflated or deflated to obtain the observed profile.

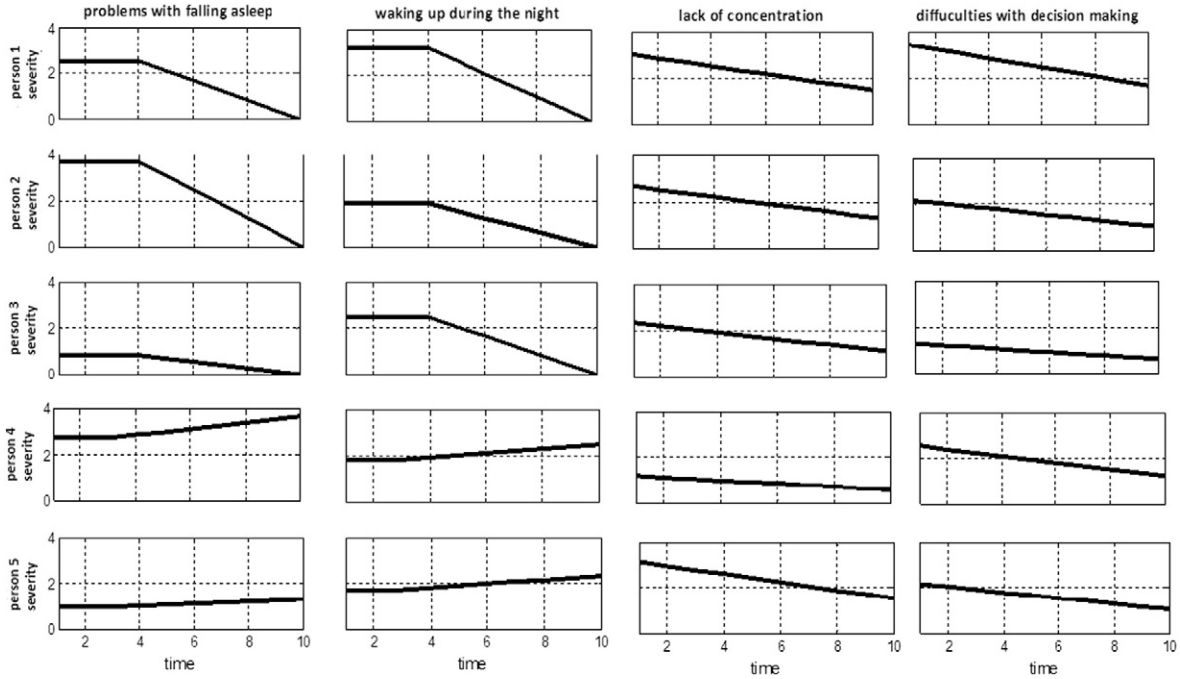

Fig. 1. Hypothetical time profiles of four depression symptoms for five MDD persons across ten treatment days.