The Pennsylvania State University
The Graduate School
College of Health and Human Development

WINDOWED EMBEDDING DIMENSION
SELECTION OF NONSTATIONARY DYNAMICAL SYSTEMS

A Thesis in
Human Development and Family Studies

by
Allison L. Gray

© 2017 Allison L. Gray

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2017
The thesis of Allison L. Gray was reviewed and approved* by the following:

Timothy R. Brick
Assistant Professor of Human Development and Family Studies
Thesis Advisor

Zita Oravecz
Assistant Professor of Human Development and Family Studies

Lisa Kopp
Associate Professor of Human Development and Family Studies
Graduate Professor in Charge

*Signatures are on file in the Graduate School
Abstract

**Objectives:** A key objective of modeling longitudinal change is to identify the characteristics of within person change before identifying mechanisms of within person change (Baltes & Nesselroade, 1979). Exploratory methods can provide insights to underlying patterns of change within any person or process. As more emphasis is placed on intensive longitudinal data collection, the need for flexible methods to explore such data also grows. **Method:** Recurrence Quantification Analysis (RQA) reveals patterns of repeating sequences within a process, and it assumes a set embedding dimension that applies to the entire sequence. However, dimensionality can change over the course of a process (i.e. emotion regulation over the course of the lifespan). I present a novel method called Windowed embedding dimension selection (WEDS), which characterizes the dimensionality and complexity of a singular process at localized points in time. **Results:** Introducing a sliding window to the selection of embedding dimension creates a dynamic view of underlying dimensions in a longitudinal process. A simulated example will illustrate the method for didactic purposes. An empirical example will utilize data from a Speed Dating Study where participants repeatedly held conversations all while wearing physiological monitoring watches. Twenty minutes of heart rate data will be used to showcase the ability and need to track change in optimized embedding dimension over time. **Discussion:** In itself, windowed embedding dimension selection is a great assumption check. It offers an exploratory tour of any process before analyses. As a diagnostic method, researchers are better able to pick or check the assumptions of any analytical method once they know how the process of interest is changing over time.

**Keywords:** time series analysis, exploratory methods, embedding dimension, time delay embedding, dynamical systems modeling
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>v</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>vi</td>
</tr>
<tr>
<td>Chapter 1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Background</td>
<td>2</td>
</tr>
<tr>
<td>The Present Study</td>
<td>7</td>
</tr>
<tr>
<td>Chapter 2 Two Examples</td>
<td>8</td>
</tr>
<tr>
<td>Simulated Example</td>
<td>8</td>
</tr>
<tr>
<td>Empirical Example</td>
<td>10</td>
</tr>
<tr>
<td>Chapter 3 Discussion</td>
<td>16</td>
</tr>
<tr>
<td>Conclusions</td>
<td>17</td>
</tr>
<tr>
<td>References</td>
<td>18</td>
</tr>
</tbody>
</table>
List of Figures

Figure 1. False Nearest Neighbors gains information from nearness criteria. In the top panel, three data points are plotted in a unidimensional space. Once another dimension is added, information is gained as those points move apart. The two red points are considered false neighbors.

Figure 2. The entire simulated time series ($N = 1200$) is simulated from a cosine wave ($N = 400$), a noisy damped oscillator ($N = 400$), and a Gaussian random walk ($N = 400$).

Figure 3. Windowed embedding dimension selection ($w = 400$) yields a variety of optimized embedding dimensions ranging from $m = 2$ to $m = 9$. The panels of the plot are aligned such that each point in the top panel corresponds to the midpoint a window, width of 400, of the original series.

Figure 4. The optimized embedding dimension over a variety of window sizes for the simulated time series. Each point on the plot corresponds to the midpoint of the window.

Figure 5. The time series of 20 minutes ($N = 1200$) of heart rate data collected from one male participant starting one minute before his first date, a four minute first date, and fifteen minutes afterwards. The shaded rectangle corresponds to the time in the date.

Figure 6. Windowed embedding dimension selection ($w = 400$) yields optimized embedding dimensions ranging of either $m = 2$ or $m = 3$ in the heart rate series. The panels of the plot are aligned such that each point in the top panel corresponds to the midpoint of each window, width of 400, of the original series.

Figure 7. The optimized embedding dimension over a variety of window sizes for the time series of heart rate. Each point on the plot corresponds to the midpoint of the window.
Acknowledgements

Thanks very much to the study participants for taking the time to participate in the experiments included in this study. Research support was provided from College of Health and Human Development’s Discovery to Innovation Project (HHD-D2I). Thanks also to Dr. Steven Boker for code to simulate a damped oscillator.

Thanks to my supportive second reader, Dr. Zita Oravecz, who has cheered me on throughout this process. Thanks to my always helpful adviser, Dr. Timothy Brick, who has supported me since day one of graduate school.
Chapter 1

Introduction

Exploratory within-person analyses must be prioritized when attempting to characterize the change of a process over time. Technology is rapidly evolving to aid researchers in more intensive yet less burdensome methods of data collection creating the opportunity to measure changes that occur at a much faster time scale than before. In cases where developmental theories are vague in regards to time and pace of change, researchers should not shy away from exploratory methods in order to inform and revise theory. To push developmental research forward, theories must account for timescale of observable change. In order to improve developmental theories in this new, data-rich context, statistical methods must be able to (1) account for rich, intensive longitudinal data, (2) provide the researcher with some sense of a pattern of change over time, and (3) provide the researcher with a time scale or range of timescale at which change is occurring.

Many commonly used methods assume stationarity and linearity in the system. However, in exploratory analyses, we should relax as many assumptions as possible. In line with the first objective of longitudinal research, we must first identify intraindividual change before proceeding to tests of causality and determinants (Baltes & Nesselroade, 1979). Identification includes understanding if and how the process evolves over time. Capturing the underlying complexity of the system will lead to better predictions and intraindividual conclusions. This paper introduces Windowed Embedding Dimension Selection (WEDS), a flexible, robust, and easy-to-apply exploratory method which captures patterns of change in complexity of time series data.
We first introduce the **Background** necessary to use windowed embedding dimension selection and how it fits into the growing field of dynamical systems analysis. Next, a **Simulated Example** highlights how this method is implemented and interpreted. In the **Empirical Example** section, we apply the method to heart rate data of one participant from a Speed Dating Study. In **Discussion**, we explain the benefits of this method as an exploratory approach. Accompanying code and a tutorial of windowed time delay embedding can be found on [quantdev.ssri.psu.edu/tutorials](http://quantdev.ssri.psu.edu/tutorials).

**Background**

Intensive longitudinal data are increasingly easier to collect. Technology has taken much of the burden away from participants (depending on types of data collected). More researchers are collecting longitudinal data, and there are more opportunities to push existing theories of change and development. In addition, computational speeds are increasing, allowing researchers to run multiple complex dynamic models on much larger data sets. The combination of richer longitudinal data and higher computational speeds creates the ideal environment for advanced analyses, but researchers must not ignore the need to explore their observed processes before or after applying statistical tests and models.

**Dynamical Systems**

Processes within a person can take a variety of shapes—oscillation, growth, or stability—that often depend on the time scale at which data were collected (Luke, 1999). In the study of development, it is common to approach longitudinal research from a dynamical systems approach (Boker, 2013; van Geert, 2011). The methods used to analyze longitudinal behavioral
data must reconcile the structure and bias of the measurement occasions and the true shape of change over time.

The field of dynamical systems has been adapted from non-person applications to the behavioral sciences with applications to within person processes (Thelen & Smith, 1994; Boker, 2013). Focusing on variation over time, dynamical systems models are able to capture the underlying dynamics of the process better than traditional methods focusing on group differences, and as a result have been becoming more popular in development research (Deboeck, Boker, & Bergeman, 2008; van Geert, 2011).

**Assumptions of Dynamical Systems**

As with all models, dynamical systems models have certain assumptions. A frequent assumption of dynamical systems models is ergodicity (Molenaar, Sinclair, Rovine, Ram, & Corneal, 2009). There are a number of ways to meet the assumption of ergodicity (for an in-depth discussion of ergodicity and its requirements see Molenaar, 2004), one of which requires the process to be stationary, or variation must not change over time (Molenaar, 2008; Molenaar et al., 2009). One must meet strong assumptions of ergodicity before making between person conclusions given within person analyses (Molenaar, 2008). In practice, the assumption of stationarity may not hold as often as we claim, which leads to attempting to fit a model not suited for the structure of the dataset. Before we impose assumptions that limit the modelable shape and behavior of developmental processes, it is imperative to explore how the underlying process behaves without bias or assumptions.

Understanding characteristics of a singular process before making between person comparisons is often overlooked in developmental research. In order to strengthen our ability to make conclusions on between person comparisons or offer insight about potential interventions,
we must understand the process of interest in a singular case (Molenaar, 2004). From a developmental or dynamics centered approach, variation over time within person, otherwise considered measurement noise, is the focus of study. Intensively measuring one process or individual best characterizes change over time.

**Time Delay Embedding and False Nearest Neighbors**

Time delay embedding is a common practice used in the dynamical systems literature to handle temporal dependencies in longitudinal data (i.e. time series data) (Pettersson, Boker, Watson, Clark, & Tellegen, 2013). It has a strong history in physics and mathematics (i.e. chaotic theory) because of its ability to recover the characteristics of a dynamical attractor space (Abarbanel, Carroll, Pecora, Sidorowich, & Tsimring, 1994; Takens, 1981). Some of the earliest work on dynamical systems analyses characterized multidimensional systems as governed by ‘manifolds,’ or attractor spaces, and they can be recovered by reconstructing the underlying phase space of the system (Whitney, 1936). In a practical sense, this is done by creating a matrix where each row is a small set of observations, rather than just one timepoint, and modeling each set as one observation.

The driving motivation behind time delay embedding is to recreate a multivariate landscape from a univariate time series. The measured data are only a snapshot of the complex system that actually occurred during the observation period. The process of recreating the multivariate landscape is called *phase space reconstruction*, which attempts to recover the underlying attractors driving the system’s behavior over time (Kennel, Brown, & Abarbanel, 1992). This notion of underlying attractors is similar to the notion of latent factors in the structural equation modeling literature. These attractors were not measured or observed, but we are able to recover them given the observations we did obtain.
There are multiple methods to reconstruct the phase space, such as singular value decomposition (Golub & Reinsch, 1970; De Lathauwer, De Moor, & Vandewalle, 2000) and topological mapping (Liebert, Pawelzik, & Schuster, 1991a), and there are a wide variety of distance metrics to use, such as Euclidean or Manhattan distance. For a walkthrough of various methods, refer to (Cao, 1997) and (Kennel et al., 1992). This paper will proceed using the false nearest neighbors method.

False nearest neighbors determines distances between all pairs of points when taking another dimension into account. In Figure 1, the two red points are relatively close to each other in a unidimensional space. These points share information because they are close to either other. However, once a second dimension is added, in the second panel of Figure 1, the points are far apart from one another. Because the points have moved further apart, we gain more information about them than we knew before in a unidimensional space. Each new dimension exposes some degree of new information about the data points. Once the added information dips below a certain percentage, the algorithm stops adding embedding dimensions. The process is similar to determining the optimal number of principal components by looking at a scree plot. A more in depth explanation of false nearest neighbors can be found in Liebert, Pawelzik, & Schuster (1991) and (Kennel et al., 1992).
Figure 1. False Nearest Neighbors gains information from nearness criteria. In the top panel, three data points are plotted in a unidimensional space. Once another dimension is added, information is gained as those points move apart. The two red points are considered false neighbors.

Windowing Embedding Dimension Selection

Time delay embedding is applied to the entirety of the time series, and it produces a global level of dimensionality for the process. It is assumed that the number and location of attractors in the reconstructed phase space remains constant over time. As developmental researchers, we are interested in how individuals change over time both at the observed and unobserved levels (Boker, 2013). Further, we want to be able to track when and how the process is changing in order to characterize how something like emotional regulation, for example, behaves across the life span. The timescale at which change is happening could also vary such that the latent space
changes over time, e.g. over decades of life or seconds during a tightly controlled experiment, e.g. a cognitive memory task. The ability to track when and how these changes occur will allow us (1) to tighten our developmental theories and (2) provide more refined timing for potential interventions.

In order to modify time delay embedding to track changes in the underlying phase space over time, we introduce a windowing method that calculates the optimal embedding dimension, via false nearest neighbors, on a subset, or window \( w \), of the time series. The window slides across the series, stepping by one in our case, and re-optimizes embedding dimension for a slice of the time series of size \( w \). As a result, WEDS returns a series of optimized embedding dimensions of length \( N - w \).

The size of the window is related to the total number of observations. Similar to other windowing methods, the size of the window must handle the tradeoff between sensitivity and reliability (Boker, Xu, Rotondo, & King, 2002). A window too small will not converge due to a phase space larger than the length of the size of the window while a window too large will not be sensitive enough to detect faster changes (and may be no different from performing false nearest neighbors on the entire series). As an exploratory measure, it is important to consider a range of potential window size values to best capture the scale and change in complexity apparent in the data.

**The Present Study**

This paper demonstrates the utility and flexibility of WEDS with two examples. The first example will use simulated data for didactic purposes. The second example will use one participant’s heart rate data from a larger study.
Chapter 2

Two Examples

All data cleaning, wrangling, analyses, and visualizations were performed in R (R Core Team, 2017). Data were transformed into long format using dplyr (Wickham, Francois, Henry, & Muller, 2017) and tidyr (Wickham & Henry, 2017). Custom functions were made by modifying the crqa package (Coco & Dale, 2013) and the tseriesChaos package (Antionio, Hegger, Schreiber, & Di Narzo, 2013). The simulated example was made in part with the deSolve package (Soetaert, Petzoldt, & Setzer, 2010). All figures/visualizations were made using ggplot2 (Wickham, 2009). A tutorial of windowed embedding dimension selection can be found on the Penn State Quantitative Developmental Systems Methodology Core’s website (quantdev.ssri.psu.edu/tutorials).

Simulated Example

Methods

A simulated time series consists of 1200 time points assumed to be equally spaced as shown in Figure 2. To mimic variability produced from a change process with three moments/phases/regimes, the simulated series consists of three underlying shapes of processes, each with its own optimal embedding dimension. The first third of the simulated series is a cosine wave. This portion alone \((n = 400)\) optimizes to an embedding dimension of two \((m = 2)\). The second portion \((n = 400)\)–a noisy damped oscillator–optimizes to have a higher embedding dimension \((m = 5)\). The third portion \((n = 400)\) of the series consists of Brownian noise, and it optimizes to a much higher embedding dimension \((m = 9)\). All three segments were concatenated together and analyzed using custom made functions.
Figure 2. The entire simulated time series \((N = 1200)\) is simulated from a cosine wave \((N = 400)\), a noisy damped oscillator \((N = 400)\), and a Gaussian random walk \((N = 400)\).

Results

Optimizing embedding dimension of the entire time series \((N = 1200)\) yields an \(m\) of 4. In order to capture dynamics of the system via recurrence quantification analysis, the phase space reconstruction would ideally need four dimensions. It is not always obvious where regime changes occur in any process, so it is important to explore via windowing.

When using a window size of 400 \((w = 400)\), the optimized embedding dimension ranges from \(m = 2\) to \(m = 9\), as shown in Figure 3, with fluctuations throughout. In the plot, we can see that windows that only contain the sine wave optimized consistently at \(m = 2\). The \(m\) is highest in windows that contain the transition to the damped oscillator and the majority of the damping process. Given this window size, these are the most complex portions of the whole process because a higher optimized embedding dimension signals more complexity. The range of optimized embedding dimensions \((m = [2, 9])\) is an indicator of lack of stationarity and complexity, so any tests on this series should not assume stationarity.
Figure 3. Windowed embedding dimension selection ($w = 400$) yields a variety of optimized embedding dimensions ranging from $m = 2$ to $m = 9$. The panels of the plot are aligned such that each point in the top panel corresponds to the midpoint a window, width of 400, of the original series.

In addition, more possibilities in changes of embedding dimensions appear with a wide variety of window sizes, as shown in Figure 4. WEDS is an exploratory method, and there is no set recommendation for the best size window to use. Using a range of window sizes, $w = 300$ to $w = 1050$, Figure 4 illustrates that, given the variety of window sizes, there is no consensus on one embedding dimension. This pattern signals that this process has a high degree of complexity. The range of embedding dimensions is an indicator of lack of stationarity and complexity, or intraindividual variability in process, within one process.

**Empirical Example**

**Methods**

These data were collected during a Speed Dating Study. Participants first rated many characteristics about themselves—demographic information, personality (Gosling, Rentfrow, &
Swann, 2003), interpersonal characteristics, and information regarding last romantic relationship. Participants wore a research grade physiology watch to measure heart rate, body temperature, electrodermal activity, etc. over time (Garbarino, Lai, Bender, Picard, & Tognetti, 2015).

**Figure 4.** The optimized embedding dimension over a variety of window sizes for the simulated time series. Each point on the plot corresponds to the midpoint of the window.
This example will use twenty minutes of heart rate data from one male participant. Heart rate was down-sampled so that one measurement occurs per second. The first minute (N = 60) occurred before his first date of the Speed Dating session. The first date lasted four minutes (N = 240), and during the remaining fifteen minutes (N = 900) he completed surveys and waited for the next date.

**Participants**

Data were collected during a session of a Speed Dating Study (N = 10). In order to be considered for the study, participants were students of Penn State, identified as either male or female, heterosexual, and reported that they were not currently in a romantic relationship. Men and women were kept in separate waiting rooms. This example will use twenty minutes of heart rate data from one male participant, who followed experimental protocol correctly.

**Procedure**

Each participant was fitted with an Empatica E4 watch immediately following the consent process. The watch remained on during the entirety of the study session lasting just over two hours. Before the session began, participants filled out a survey assessing personality type, current emotional state, information about their previous relationship, and demographic information. Participants then held a series of four-minute conversational "dates." Each man talked to each woman and vice versa such that each participant had five conversational “dates.” After each date, participants completed a post-conversation survey assessing the quality of the conversation and various personality and emotion characteristics of their partner. Participants were instructed to push the event button located on the E4 at the beginning and end of each four-minute date.
Measures

Heart Rate. Heart rate is collected passively as each participant wears the Empatica E4 watch (Garbarino et al., 2015). The E4 measures heart rate via a photoplethysmography sensor located on the underside of the watch which sits on the top of the wrist. The PPG sensor samples at a rate of 64 Hz. The E4 is moderately affected by movement, so all participants were seated. However, there were no explicit instructions about limiting movement in order to retain some level of ecological validity. This example uses twenty minutes of heart rate from a pair shortly before their first date, during the date, and fifteen minutes after the date. Figure 5 shows the time series for this participant, and the shaded rectangle indicates the time they spent on the speed date.

![20 Minutes of Heart Rate](image)

**Figure 5.** The time series of 20 minutes ($N = 1200$) of heart rate data collected from one male participant starting one minute before his first date, a four minute first date, and fifteen minutes afterwards. The shaded rectangle corresponds to the time in the date.
Results

Optimizing the series of heart rate yields an embedding dimension of two \((m = 2)\). Given the experimental design, it is expected that the change in heart rate will not be constant over time. Once applying a window of size 400 \((w = 400)\), iterations of the embedding dimension selection are either \(m = 2\) or \(m = 3\), as shown in Figure 6.

![Figure 6](image)

**Figure 6.** Windowed embedding dimension selection \((w = 400)\) yields optimized embedding dimensions ranging of either \(m = 2\) or \(m = 3\) in the heart rate series. The panels of the plot are aligned such that each point in the top panel corresponds to the midpoint of each window, width of 400, of the original series.

When exploring a wide range of windows, as shown in Figure 7, the optimized embedding dimension does not stray from either \(m = 2\) or \(m = 3\). Therefore, using the windowed approach and exploring a wide range of windows, we can determine that this heart rate series is relatively stationary. The range of optimized embedding dimensions indicates that this series has low levels of complexity. The windows within which the optimized embedding dimension is
higher \((m = 3)\) than the global level \((m = 2)\) could signal a different process (e.g. recovery from a stressful first date) or a different state of one underlying process.

**Figure 7.** The optimized embedding dimension over a variety of window sizes for the time series of heart rate. Each point on the plot corresponds to the midpoint of the window.
Chapter 3
Discussion

Benefits of the Method

In this simulated example, windowing the embedding dimension selection process yields a much wider range of potential options. The windowing process was able to detect differences in the underlying system, i.e. differentiating the first third from the second. The optimized embedding dimension for the entire series \( (m = 4) \) would have underestimated portions of the series where the dynamics were more complex.

In the empirical example, the embedding dimension for the entire series \( (m = 2) \) did not vary much given a wide range of windows, so his heart rate series is relatively stationary. However, there were portions of the series where a higher embedding dimension \( (m = 3) \) would be more suitable when using smaller window sizes \( (w = 300, 350, 400, 450, 500, 550) \). The portions of the original series where a higher embedding dimension is preferred seem to align with the portion of time where the participant was transitioning from the date to the waiting room—a time of down regulation. The optimized embedding dimension series maps back onto the study design, and it signals a separate regime for the participant, i.e. increased arousal due to nervousness of the date versus recovery from the date.

Using a variety of window sizes, the variability of embedding dimension increases. As the window size approaches the length of the original series, the mode of the optimized embedding dimensions becomes the embedding dimension of the entire series. As window size decreases, it becomes more vulnerable to noise in the process, and there might be errantly low or high embedding dimensions given artifacts in the time series. The size of the window affects the
variability of embedding dimension choices and the range of values, i.e. more versus less smoothing, so there is a trade-off between sensitivity and reliability.

**Cautionary Notes**

Windowed Embedding Dimension Selection is an exploratory method. It is not a significance test in itself, and provides neither test statistic nor p-value. Rather this method should be used in tandem with other exploratory and summary methods to use exploratory methods to inspect the process before modeling.

As previously stated, this is not only way to select an optimal embedding dimension. Here, we used false nearest neighbors. However, there are other methods such as singular value decomposition and topological mapping. To be certain, it is recommended to optimize embedding dimension using at least two methods to validate the underlying complexity.

**Conclusions**

As it becomes easier to collect intensive longitudinal data on any given process of interest, researchers must use more exploratory analytical methods to understand what is happening within each individual before making conclusions at the between-persons level. With the rise in intensive longitudinal data collection, dynamical systems methods are increasingly suited to modeling the measured processes. However, it is important to understand the characteristics of any given process before making the assumptions needed to fit many of these dynamical models. Exploratory studies will help to improve current developmental theories with an emphasis on the timescale and rate of change.
References


http://doi.org/10.1080/00273170802490616


http://doi.org/10.1109/MOBIHEALTH.2014.7015904


http://doi.org/10.1016/S0092-6566(03)00046-1


http://doi.org/10.1209/0295-5075/14/6/004


http://doi.org/10.1209/0295-5075/14/6/004


http://doi.org/10.2307/1968482

