Capturing similarity in multivariate dyadic time series: a statistical framework

A Data Management Plan created using DMPonline.be

Creators: Chiara Carlier, Eva Ceulemans

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: Chiara Carlier, Eva Ceulemans

Data Manager: Chiara Carlier, Eva Ceulemans

Project Administrator: Eva Ceulemans

Grant number / **URL:** 1153723N

ID: 195091

Start date: 01-11-2022

End date: 31-10-2026

Project abstract:

Similarity is a key concept in the understanding of relationships between people. In the past, this has been investigated by isolating single variables and computing a specific similarity measures across a sample of dyads. Currently, behavioral research practices are expanding in two directions: from an
isolated variable-centered to a multivariate dyad-centered approach and from one single measurement per dyad to investigating longitudinal time series. However, the methodological field is lagging behind and the appropriate tools to capture similarity in these more complex data sets are lacking or ill-developed. We identified two clusters of problems: 1) there are a myriad of similarity measures and pre-processing methods to choose from that all have their pros and cons, 2) there are no broadly applicable and statistically sound ways to infer meaning and significance. The aim of this project is to solve these problems by constructing a statistical similarity framework for multivariate dyadic time series. First, we will investigate a wide variety of similarity measures and pre-processing methods and draft guidelines to help researchers make informed choices. Second, we will dive into non-parametric testing and cluster analysis to provide researchers with a widely applicable inference and interpretation framework. Last, we will pave the way for future researchers by implementing the results of this project in open-source software packages and applying them to empirical data sets.

Last modified: 02-03-2023

Capturing similarity in multivariate dyadic time series: a statistical framework Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

The current project will reuse existing data sets consisting of background surveys, lab interactions and experience sampling method surveys completed by romantic couples.
During the project we will generate R code for certain analyses and simulations and software in the form of R packages and apps.

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

The responsible persons for the data will be Chiara Carlier en Eva Ceulemans.
During the research, the data are being stored on the KU Leuven network drives and OneDrive and shared if needed through Sharepoint. After the research we will delete the used data sets on our drives, since they are further stored by their own administrators. R scripts that were shared (e.g. through OSF) will stay on these servers. Software and apps that were made available during the project, will also stay available online.

What's the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

We do not wish to deviate.

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

We will only perform secondary analyses on data that has been collected under ethical approval of the KU Leuven. The data sets that will be used in this project are pseudonymized by others with the help of participant numbers and do not contain any highly sensitive information such as names or addresses. Therefore, no additional security measures are in place than the standard ones in our organization (e.g. Bitlocker encrypted drives).

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

No other issues.

Capturing similarity in multivariate dyadic time series: a statistical framework
DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Yes

Capturing similarity in multivariate dyadic time series: a statistical framework
GDPR

GDPR

Have you registered personal data processing activities for this project?

- Yes

Capturing similarity in multivariate dyadic time series: a statistical framework
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .cvs,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA
ESM study	Experience sampling method (ESM) surveys where particpants rated their own and partner's emotions and emotion regulatory behavior, and context information	reuse	digital	observational	.sav, .xlsx, .csv	<100MB
VMR Study	Lab interactions with self-reported experience (video-mediated recall of 12 discrete emotions) and continuous behavior observation coding (16 SPAFFcodes)	reuse	digital	observational	.sav, .csv	<100MB
Lab study	Lab interactions with self-reported experience (video-mediated continuous recall of valence and repeated recall of discrete emotions)	reuse	digital	observational	.sav, .csv	<100MB
Cult 14 Study	Responses to the International College Survey	reuse	digital	observational	.xlsx, .txt	<100MB
Cult20 Study	Emotion ratings in response to vignettes	reuse	digital	observational	.xlsx	<100MB
Subject information	Background questionnaires on demographics, relationship quality, coping and depressive symptoms	new+reuse	digital	observational	.csv, .xlsx	<100MB
New Study	Emotion ratings in response to photos and vignettes and daily diary surveys on emotional experience in negative events	new	digital	observational	.csv, .xlsx	<100MB
Analytical syntaxes	R syntaxes for computing indices, analyzing, and pre-processing data	new	digital	software	.R, .Rmd	<100MB
Generating syntaxes	R syntaxes for generating data and doing simulations	new	digital	software	.R, .Rmd	<100MB
Simulated data	Outcomes of previous data source	new	digital	simulation data	.csv, .xlsx	<100MB
R packages	Newly compiled R packages for applying methods	new	digital	software	.R, .Rmd	<100MB

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

In this project, five different empirical data sets will be reused: one experience sampling study (ESM), one negative lab interaction with video-mediated recall (VMR), one cross-sectional negative lab interaction (lab) and two cross-cultural data sets with 14 and 20 measured emotions (Cult14 and Cult20).
For the **ESM data** see : Sels, L., Ruan, Y., Kuppens, P., Ceulemans, E., & Reis, H. (2020). Actual and Perceived Emotional Similarity in Couples' Daily Lives. *Social Psychological and Personality Science*, *11*(2), 266–275. https://doi.org/10.1177/1948550619845927
For the **VMR data** see: Boiger, M., Kirchner-Häusler, A., Schouten, A., Uchida, Y., & Mesquita, B. (2020). Different bumps in the road: The emotional dynamics of couple disagreements in Belgium and Japan. *Emotion*. https://doi.org/10.1037/emo0000910
For the **lab data** see: Sels, L., Ceulemans, E., & Kuppens, P. (2019). All's well that ends well? A test of the peak-end rule in couples' conflict discussions. *European Journal of Social Psychology*, *49*(4), 794–806. https://doi.org/10.1002/ejsp.2547
For the **Cult14 data** see: Kuppens, P., Ceulemans, E., Timmerman, M. E., Diener, E., & Kim-Prieto, C. (2006). Universal Intracultural and Intercultural Dimensions of the Recalled Frequency of Emotional Experience. *Journal of Cross-Cultural Psychology*, *37*(5), 491–515. https://doi.org/10.1177/0022022106290474
For the **Cult20 data** see: De Leersnyder, J., Mesquita, B., & Kim, H. S. (2011). Where Do My Emotions Belong? A Study of Immigrants' Emotional Acculturation. *Personality and Social Psychology Bulletin*, *37*(4), 451–463. https://doi.org/10.1177/0146167211399103

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes, human subject data

We will perform secondary analyses on data that has been collected under ethical approval of the KU Leuven. The data sets that will be used in this project are mentioned above and are pseudonymized by others with the help of participant numbers. We will also collect new data in romantic couples (see table). This study was ethically approved by SMEC: G-2022-5960-R2(MIN).

Questionnaires	Self-construed background questionnaire (gender, age, nationality, education, occupation, relationship duration, marital status, number of children, time spent with partner) CES-D (Radloff, 1977) Perceived Relationship Quality Component Inventory (Fletcher et al., 2000) Dyadic coping inventory (Bodenmann, 2008) Inclusion of others in the self-scale (Aron et al., 1992) Satisfaction with Life Scale (Diener et al., 1985) Ten Item Personality Inventory (Gosling et al., 2003) EROS extrinsic subscale (Niven et al., 2011) Ruminative response scale (Nolen-Hoeksema & Morrow, 1991)
Surveys	Participants' ratings of their own emotions following the presentation of emotion pictures taken from the InternationalAffective Picture System Participants' ratings of their own and expected partner's emotions following the presentation of emotional vignettes
Daily diary	Daily description of a negative event + Participants' ratings of their own and expected partner's emotions in response tothis event + some context

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes

All ratings of emotions and questionnaires are coded using a participant number. Information via which participants are directly identifiable (name and email address) is stored separately from other data collected: a separate password-protected document links participant numbers to this participant-identifying information. This document is only accessible by the researchers who need to know this information for the purpose of conducting the study.

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

Reused data: information on
Collected data: We will create a separate folder that contains the following information: **1. Readme**.docx: In this word-document, we will discuss which researchers were involved in the collection of the data (e.g., master students), the ethical approval (reference number & institution), a short overview of the study course and protocol, which questionnaires we administered and their variable labels, and short written information on the data cleaning process and steps. **2. Codebook**.xlsx: In this excel-document, we will provide pseudonymized baseline information about all participants that were enrolled in the study (e.g., age, gender, which phases of the study they completed and when, whether they completed the entire study or dropped-out, their compliance, other remarks, etc.). We also provide specific information about

the questionnaires we administered (i.e., number of items, reference, their variable labels, ranges and description). Finally, we provide some basic summary statistics (e.g., gender and age distribution, overall study compliance, etc.). **3. Folder** with all the study documents: Ethical application and approval, informed consent example, the instructions we gave participants. The PDF of all questionnaires (ESM, interview) will be included. **4. Pre-processing documents**: The very raw data (pseudonymized), a reproducible syntax / code to clean the data in to .xlsx files that are ready for data-analysis. The folder will never contain sensitive (identifiable) participants information such as names, contact details, audio-files of interviews, etc.
Simulated data: The simulated data will be generated and analyzed by means of codes that will be stored together with the simulation results (see data description), using a folder structure and files with self-explaining names. We will also add a readme file, to further explain the structure and files of the simulations.
Software: The R-package will contain open source functions and will be posted on CRAN and Github, with an associated documentation .pdf.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- No

Collected data: The meta-data will consist of a Read me.docx, Codebook.xlsx, syntax to process the raw data into files that are ready for analysis, and a folder with all the study documents (e.g., Informed consent, ethical approval, etc.) to further contextualize the collection of the data underlying our publications. See previous question for more detailed information.
The **simulated data** will be generated and analyzed by means of codes that will be stored together with the simulation results (see data description), using a folder structure and files with self-explaining names. We will also add a readme file, to further explain the structure and files of the simulations.
The **R-package** will contain open source functions and will be posted on CRAN and Github, with an associated documentation .pdf.

3. Data storage & back-up during the research project

Where will the data be stored?

Digital data will be stored in a restricted network share on the J: drive, which can only be accessed by the involved researchers. All drives are managed by KU Leuven personnel, bound by the KU Leuven general and ICT codes of conduct.
Offline copies of digital data will be stored on encrypted media. File transfers will be performed with Belnet Filesender using the encryption option.
To support pseudonymisation, technical and organisational controls are in place to protect files that contain mappings between pseudonyms (codes) and the original personally identifiable information. Access to these code files is controlled by a data manager (with the PI as a back-up). All other researchers who participate in the project have access to the pseudonymised data only.

How will the data be backed up?

Snapshots
Backups of project data on faculty network shares are made using “snapshot” technology, which is the online storage of incremental data changes. The standard backup regime is as follows:

- An hourly backup (at 8 AM, 12 PM, 4 PM and 8 PM) the last 6 of which are stored on our servers
- A daily backup, at midnight, the last 6 of which are stored on our servers
- A weekly backup, Saturday night at midnight, the last 12 of which are stored on our servers

The end user can use his own Windows PC to restore files to an older version using the “previous versions” function. According to the above backup scheme, it is possible to go back in time up to 12 weeks (~3 months).
Mirror
For the purpose of “business continuity” or “disaster recovery”, a mirror (exact copy) of all data is created in a second datacenter. The data are copied every hour to the second datacenter. In the event that the primary storage unit is corrupted, the ICTS team can get this copy online within the hour.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

Yes. The necessary funding for storage and backup for the contracted service has been foreseen.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

All ICT solutions at KU Leuven are subject to the university-wide ICT information security standards. The faculty’s ICT service organizes the raw network storage it procures from central ICT services in such a way that access permissions are limited, fixed, delegated to and audited by data managers who do not need to have an IT background.
Digital data will be stored in a restricted network share on the J: drive, which can only be accessed by the involved researchers.
Access to code files with pseudonyms is controlled by a data manager (with the PI as a back-up). All other researchers who participate in the project have access to the pseudonymised data only.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

The yearly cost for network share storage on the J: drive is 800 euro per TB, including backup and unused space.
The yearly cost for SharePoint storage is approx. 300 euro for 5 GB plus 40 euro per additional 5 GB.
The necessary funding for storage and backup for the contracted service has been foreseen.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be

preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All data (including syntaxes and analysis output) will be retained for the expected 5 year period after the end of the project.

Where will these data be archived (stored and curated for the long-term)?

Digital data will be stored at a restricted area of the K: archive drive, which can only be accessed by the involved researchers and the unit's data manager. All drives are managed by KU Leuven personnel, bound by the KU Leuven general and ICT codes of conduct.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

The yearly cost for network share storage on the K: archive drive is 200 euro per TB, including backup and unused space.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in a restricted access repository (after approval, institutional access only, …)
- Yes, in an Open Access repository

Collected data: The full pseudonymized dataset will be available for colleague researchers for meta-analytic projects upon request (and signing a data-sharing agreement). For reviewers and readers of our published articles a trimmed and pseudonymized dataset (i.e., only containing the variables required to reproduce our results) will be uploaded in a .csv or .xlsx format to the Open Science Framework under a CC-BY license alongside full dataset documentation. Participants' personal information (e.g., contact information, names, etc.) will never be shared.
Simulated data: Simulated data will be made available upon request by other researchers. Codes will be uploaded to the Open Science Framework and/or Github.
Software: the R-package and associated functions will be uploaded on CRAN and Github under a GPL license.

If access is restricted, please specify who will be able to access the data and under what conditions.

Collected data: For meta-analytic projects researchers can request the full pseudonymized dataset via email, and will receive these after signing a data-sharing agreement. For published papers, a trimmed and pseudonymized dataset (i.e., only containing the variables required to reproduce our results) will be uploaded in a .csv or .xlsx format to the Open Science Framework under a CC-BY license alongside full dataset documentation. Participants' personal information (e.g., contact information, names, etc.) will never be shared.
Simulated data: Simulated data will be made available upon request by other researchers. Codes will be uploaded to the Open Science Framework and/or Github.
Software: the R-package and associated functions will be uploaded on CRAN and Github under a GPL license.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- No

Where will the data be made available? If already known, please provide a repository per dataset or data type.

Collected data: For meta-analytic projects researchers can request the full pseudonymized dataset via email, and will receive these after signing a data-sharing agreement. For published papers, a trimmed and pseudonymized dataset (i.e., only containing the variables required to reproduce our results) will be uploaded in a .csv or .xlsx format to the Open Science Framework under a CC-BY license alongside full dataset documentation. Participants' personal information (e.g., contact information, names, etc.) will never be shared.
Simulated data: Simulated data will be made available upon request by other researchers. Codes will be uploaded to the Open Science Framework and/or Github.
Software: the R-package and associated functions will be uploaded on CRAN and Github under a GPL license.

When will the data be made available?

Upon publication of the research results

Which data usage licenses are you going to provide? If none, please explain why.

Collected data: For meta-analytic projects researchers can request the full pseudonymized dataset via email, and will receive these after signing a data-sharing agreement. For published papers, a trimmed and pseudonymized dataset (i.e., only containing the variables required to reproduce our results) will be uploaded in a .csv or .xlsx format to the Open Science Framework under a CC-BY license alongside full dataset documentation. Participants' personal information (e.g., contact information, names, etc.) will never be shared.

Simulated data: Simulated data will be made available upon request by other researchers.
Codes will be uploaded to the Open Science Framework and/or Github.
Software: the R-package and associated functions will be uploaded on CRAN and Github under a GPL license.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

A permanent identifier is added to data upon deposit in a repository.

Collected data will be uploaded to the Open Science Framework under a CC-BY license alongside full dataset documentation. Participants' personal information (e.g., contact information, names, etc.) will never be shared.
Codes will be uploaded to the Open Science Framework and/or Github.
R-package and associated functions will be uploaded on CRAN and Github under a GPL license.

What are the expected costs for data sharing? How will these costs be covered?

Sharing data at the Open Science Framework, CRAN and Github is free.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Chiara Carlier en Eva Ceulemans

Who will manage data storage and backup during the research project?

Chiara Carlier, Eva Ceulemans and Kristof Meers

Who will manage data preservation and sharing?

Chiara Carlier, Eva Ceulemans

Who will update and implement this DMP?

Chiara Carlier, Eva Ceulemans