GaitTrends Software User's Manual

Klaudia Kozlowska, Miroslaw Latka

August 2, 2020

1   Introduction

GaitTrends repository contains the MATLAB scripts that we developed while working on the paper "Significance of trends in gait dynamics" which was submitted to PLOS Computational Biology. Please read the paper before running the code. If you use our software please cite: https://doi.org/10.1101/677948. The code is distributed under the GNU GPL v3 (or later) license and it was tested using MATLAB R2016b and R2018a on Windows 10.

There are several subfolders in the main directory:

- `data` – It contains the MAT-files with either experimental data or the output generated by the scripts from `data_preparation` folder. If you run these scripts the output files will be overwritten.

- `data_analysis` – The m files perform the analysis of both the experimental and surrogate data stored in the MAT-files in the `data` subfolder.

- `data_preparation` – The scripts generate MAT-files used for data analysis.

- `libs` – For convenience, we put here the third party open source libraries ARESLab and WFDB toolbox. We have not modified these libraries. Here you will also find the program surrogates from TISEAN Nonlinear Time Series Analysis library which uses iterative Fourier based scheme to generate surrogate data.

- `utils` – It contains the functions called by the scripts from both `data_preparation` and `data_analysis` folders.

The content of each subfolder will be described in detail in Description of folders section.

2   How to start

Clone the repository. Do not put GaitTrends folder deep in the Windows folder tree! Some of the third party functions such as TISEAN's surrogates cannot handle long file names. Each script contains the relative paths to functions and data. Do not rename `data` folder or any of its subfolders. The scripts that write output to files do not check whether the required destination folders exist. Before running any code, please permanently add to the MATLAB search path the folders located in libs subdirectory. Use option Add with Subfolders (MATLAB documentation: Change Folders on the Search Path).

3 Description of folders

3.1 data

This directory contains the following subfolders:

- `mat_data`

- `original_Dingwell_data`

- `surrogates`

- `trend_stats`

- `window_length_dependence`

3.1.1 mat_data

15 mat files in this subfolder were generated by `prepare_data.m` script. For three gait parameters: stride length (SL), stride time (ST), and stride speed (SS) this script was ran 5 times for each of five treadmill speeds (80, 90, 100, 110, 120% of preferred walking speed). Each mat file contains the cell with the description of the experimental data (subject and trial number) – *infosAll*, the corresponding experimental time series (*seriesAll*) and the MARS models (*modelsAll*). For convenience, the indices of the MARS knots and the values at these knots were stored in *knotIndicesAll* and *valuesAtKnotAll* cells, respectively. *residualsAll* keeps the MARS residuals obtained by subtracting the piecewise linear trends (*trendsAll*) from the experimental time series (*seriesAll*).

3.1.2 original_Dingwell_data

This folder contains the young subjects' gait data from Dingwell et al. study.

3.1.3 surrogates

We used three types of surrogate data for SL and ST (see our paper for more details):

- cross-correlated phase-randomized

- independently phase-randomized

- independently shuffled.

The first two types of surrogates were generated by `../data_preparation/prepare_surrogates.m` script. The random shuffling was done with `../data_preparation/prepare_shuffled_surrogates.m`. The surrogate data were generated for all five speeds (SPD1-5) and stored in the corresponding folders (`cross_correlated`, `independent`, `shuffled`). Each MAT-files contains two MATLAB structures: `data_surrogatesSL` and `data_surrogatesST`. These structures are organized in the same way as the mat files in the `mat_data` folder.

`../data/surrogates/surrogates_trend_distributions/` folder contains the MAT-files written by `multiple_surrogates.m`. This script calculates the normalized MARS trend durations and trend's slopes in cross-correlated (SL-ST) or independent phase-randomized surrogates.

3.1.4 `trend_stats`

There are two MAT-files in this folder: `SL_trends.mat` and `ST_trend.mat` which were generated by `trend_stats.m` script. For each treadmill speed, the mat files contain 4 cells. `trend_durations` and `trend_slopes` cells store the values of normalized trend durations and normalized slopes, respectively. The same properties for the long trends (whose normalized length was greater than 40) are saved in `long_trend_durations` and `long_trend_slopes` cells.

3.1.5 `window_length_dependence`

MAT-files `dingwell_alphas_windowed_*n_SPD*.mat` are the output files of `window_length_dependence.m` script from `data_preparation` folder. The analogous calculations were performed using `window_length_dependence_fbm.m` for two ensembles of 500 fractional Brownian motion random walks with the Hurst exponent $H = 0.40$ and $H = 0.75$. In the latter case, the output files were named `fbm_0.4_alphas_windowed.mat` and `fbm_0.75_alphas_windowed.mat`, respectively.

3.2 `data_analysis`

The following scripts may be found in this folder:

- `alphas_calculations.m`

- `alphas_calculations_surrogates.m`

- `alphas_windowed.m`

- `compare_distributions.m`

- `correlations.m`

- `correlations_surrogates.m`

- `trend_speed_control.m`

- `trend_speed_control_surrogates.m`

3.2.1 `alphas_calculations.m`

This script uses DFA1-3 and madogram algorithms to calculate the scaling indices of Dingwell's experimental time series and those of the corresponding MARS residuals. The input files must be generated first using `prepare_data.m` script. For a given treadmill speed, the script also determines whether the scaling is anti-persistent (i.e. the median or mean of scaling exponents is statistically smaller than 0.5). Set `attributeNumber` to 1, 2, and 3 to select SL, ST, and SS, respectively.

3.2.2 `alphas_calculations_surrogates.m`

This script uses DFA1-3 and madogram algorithms to calculate the scaling indices of surrogates of Dingwell's experimental time series and those of the corresponding MARS residuals. The input files must be generated first using `prepare_surrogates.m` script. For a given treadmill speed the scripts also determines whether the scaling is anti-persistent (i.e. the median or mean of scaling exponents is statistically smaller than 0.5) Set `surrogates_type` variable to select the type of surrogates (1 – independent, 2 cross-correlated, 3 – shuffled). The boxplots of madogram scaling exponents of gait surrogates are displayed.

3.2.3 `alphas_windowed.m`

This script visualizes the dependence of scaling exponents of Dingwell's gait time series and fractional Brownian motion on data window length. The boxplots are displayed for the DFA and madogram scaling exponents. Change `fileName` variable to analyze either experimental data or fractional Brownian motion time series. Both types of time series are stored in `../data/window_length_dependence` folder. For example with `fileName = dingwell_alphas_windowed_SL_SPD1.mat` the boxplots for stride length at preferred walking speed will be shown. For `fileName = fbm_0.4_alphas_windowed.mat` the analysis will be performed for the fractional Brownian motion with the Hurst exponent equal to 0.4.

3.2.4 `compare_distributions.m`

Two-sample Kolmogorov-Smirnov test is used to compare the distributions of trend durations and slopes in experimental data and their phase-randomized surrogates. The script uses MAT-files located in `../data/surrogates/surrogates_trend_distributions/` folder. These input files must be created first by running `../data_preparation/multiple_surrogates.m`. The number of cases in which distributions are not statistically different is printed in the MATLAB's command window.

3.2.5 `correlations.m`

Calculations of Pearson's correlation coefficient between:

- ST and SL

- their piecewise linear MARS trends

- MARS residuals.

For trends, the boxplots of correlation coefficients are displayed for all five treadmill speeds. The input files must be generated first using `prepare_data.m` script.

3.2.6 `correlations_surrogates.m`

The correlations for surrogate data are calculated in the same way as in `correlations.m`. The input files must be generated first using `prepare_surrogates.m` script. When the variable `cross_correlated` is set to true, cross-correlated surrogates are used. Otherwise, the independent surrogates are selected.

3.2.7 `trend_speed_control.m`

This script calculates:

- coefficient of variation (COV) for gait parameters (SL/ST/SS)

- trend speed and its COV

- trend speed control.

The boxplots of these quantities are plotted for each treadmill speed. First, the input files must be generated using `prepare_data.m` script.

3.2.8 `trend_speed_control_surrogates.m`

The modification of `trend_speed_control.m` which handles surrogate data. Set `surrogates_type` variable to 1, 2, and 3 to select independent, cross-correlated, and shuffled data, respectively. The input files must be generated first using `prepare_surrogates.m` script.

3.3 `data_preparation`

There are the following scripts in this folder:

- `multiple_surrogates.m`

- `prepare_data.m`

- `prepare_shuffled_surrogates.m`

- `prepare_surrogates.m`

- `trend_stats.m`

- `trend_stats_surrogates.m`

- `window_length_dependence.m`

- `window_length_dependence_fbm.m`

3.3.1 `multiple_surrogates.m`

This script loads gait data from Dingwell's MAT-files and calculates multiple phase-randomized surrogates using the TISEAN library. Then it finds piecewise linear MARS trends in surrogate time series. For a given treadmill speed, the normalized trend durations and normalized trend slopes are saved to `../data/surrogates/surrogates_trend_distributions/` folder. Before running the script, please set: 1) treadmill's speed (`SPD`) 2) number of surrogates for each experimental time series (`repetitions`) 3) type of surrogates (`cross-correlated true, independent false`)

This script uses `surrogates.exe` file from TISEAN library. Two text auxiliary files: `in.txt` and `out.txt` are used as input and output to this function.

3.3.2 `prepare_data.m`

This script loads gait data from Dingwell's MAT-files and calculates piecewise linear MARS trends. You need to set `SPD` and `attributeNumber` variables in order to select speed (80, 90, 100, 110, 120 [%PWS]) and gait parameter (1 – SL, 2 – ST, 3 – SS). You can modify MARS parameters using `params` variable. The script saves MAT-files with: gait (SL/ST/SS) time series, time stamps, MARS piecewise linear trends, MARS residuals, MARS models, MARS knot indices, and values at MARS knots to `../data/mat_data` folder.

3.3.3 `prepare_shuffled_surrogates.m`

The modification of `prepare_data.m` which generates shuffled surrogate data. You need to specify `SPD` only since surrogates of SL and ST are both stored in the same MAT-file `../data/surrogates/shuffled/` folder.

3.3.4 `prepare_surrogates.m`

The modification of `prepare_data.m` which generates phase-randomized surrogate data using TISEAN library. Set variable `cross_correlated` to true to use cross-correlated SL and ST surrogates. For independent surrogates set this variable to false. You need to specify `SPD` only since surrogates of SL and ST are both stored in the same MAT-file either in `../data/surrogates/cross_correlated/` or `../data/surrogates/independent/` folders.

3.3.5 `trend_stats.m`

This script calculates normalized trend durations and normalized trend slopes for each treadmill speed (`SPD`). It also calculates these two quantities for trends lasting longer that the chosen threshold (`threshold variable`). The script uses MAT-files located in `../data/mat_data/` folder. These input files must be created first by running `prepare_data.m`. Change `attributeNumber` variable to select gait parameter (1 – SL, 2 – ST, 3 – SS). The script plots histograms of normalized trend duration and normalized slope of long trends. The output MAT files are saved in `../data/trend_stats/` folder.

3.3.6 `trend_stats_surrogates.m`

The modification of `trend_stats.m` which handles phase-randomized surrogate data. Set variable `cross_correlated` to true to use cross-correlated SL and ST surrogates. For independent surrogates set this variable to false. The output MAT-files are saved in `../data/trend_stats/` folder.

3.3.7 `window_length_dependence.m`

This script divides first 260 samples of ST and SL experimental time series (from `../data/mat_data/` into non-overlapping windows of length k (from $k = 40$ to $k = 260$ with step 20) and then for each such window calculates the scaling exponents using DFA1-3 and madogram algorithms. The corresponding exponents are stored in the first four columns of the matrix. The window length and the trial's id may be found in columns 5 and 6, respectively. Set `fileName` variable to choose the desired parameter (SL or ST) and the speed. E.g. `fileName = SL_SPD1.mat` corresponds to stride length at preferred walking speed while `fileName = ST_SPD1.mat` is the stride time series for the same speed. The output MAT files are saved in `../data/window_length_dependence/` folder. For a given scaling exponents, the script generates the boxplots of its values for all window lengths.

3.3.8 `window_length_dependence_fbm.m`

The calculations analogous to those implemented in `window_length_dependence.m` can be performed for an ensembles of fractional Brownian motion random walks with the chosen Hurst exponent. The output file is saved in `../data/window_length_dependence/` folder.

3.4 `libs`

This directory contains the following third party libraries:

- ARESLab (implementation of MARS algorithm)
- WFDB Toolbox (dfa – implementation of Detrended Fluctuation Analysis)
- TISEAN library (surrogate data generation).

3.5 `utils`

This directory consists of following functions:

- `boxplots_for_all_speeds.m`
- `calculate_trend_stats.m`
- `perform_MARS.m`
- `swtest.m`
- `variogram.m`.

3.5.1   `boxplots_for_all_speeds.m`

Makes boxplots of five vectors stored in a cell. Used to visualize the quantities calculated from Dingwell's experimental data.

Inputs:

- `cell_data` – cell with five vectors

- `title_str` – plot's title (string)

- `ylabel_str` – y-axis label (string).

Outputs:

- none.

3.5.2   `calculate_trend_stats.m`

Calculates the normalized duration and normalized slope of MARS trends in gait parameters (ST/SL/SS).

Inputs:

- `data` – the collection of cells returned by one of the following scripts:

  `../data_preparation/prepare_data.m`,

  `../data_preparation/prepare_surrogates.m`, and

  `../data_preparation/prepare_shuffled_surrogates.m`.

- `threshold` – if this optional argument has been passed the function also returns the normalized durations and slopes of all MARS trends longer than the threshold.

Outputs:

- `trend_durations` – normalized durations of MARS trends (vector)

- `trend_slopes` – normalized slopes of MARS trends (vector)

- `long_trend_durations` – normalized durations of MARS trends longer than threshold (vector)

- `long_trend_slopes` – normalized slopes of MARS trends longer than threshold (vector).

3.5.3   `perform_MARS.m`

Builds MARS model for Dingwell's gait time series (ST/SL/SS) using ARESLab.

Inputs:

- `X` – stride time stamps (vector)

- `Y` – gait time series (ST/SL/SS) (vector)

- `generateFigures` – if true make plot of time series and their MARS trends (logical)

- `plotTitle` – title of the plot (string)

- `MARS_params` – parameters of ARES model (struct).

Outputs:

- `outputData` – contains gait (SL/ST/SS) time series, time stamps, MARS piecewise linear trends, MARS residuals, MARS models, MARS knot indices, and values at MARS knots (struct).

3.5.4   `swtest.m`

Performs Shapiro-Wilk parametric hypothesis test of composite normality. It was developed by Ahmed Ben Sada.
   Inputs:

- `x` – data vector of length between 3 and 5000

- `alpha` – significance level (default set to 0.05).

Outputs:

- `H` – hypothesis (0 or 1)

- `pValue` – test p-value

- `W` – test statistics.

3.5.5   `variogram.m`

Computes madogram estimator of the scaling exponents. Implementation based on the algorithm described by Gneiting et al. – "Estimators of fractal dimension: Assessing the roughness of time series and spatial data." Statistical Science (2012): 247-277.
   Inputs:

- `x` – scalar time series (vector)

- `p` – variogram's order.

Outputs:

- `H` – scaling exponent estimate.