A dynamic matrix variate clustering approach for monitoring running activities

Aueb Sports Analytics Workshop 2019
Athens

Stival Mattia
mattia.stival@phd.unipd.it

Department of Statistical Sciences
University of Padova

25-26 November, 2019

Introduction

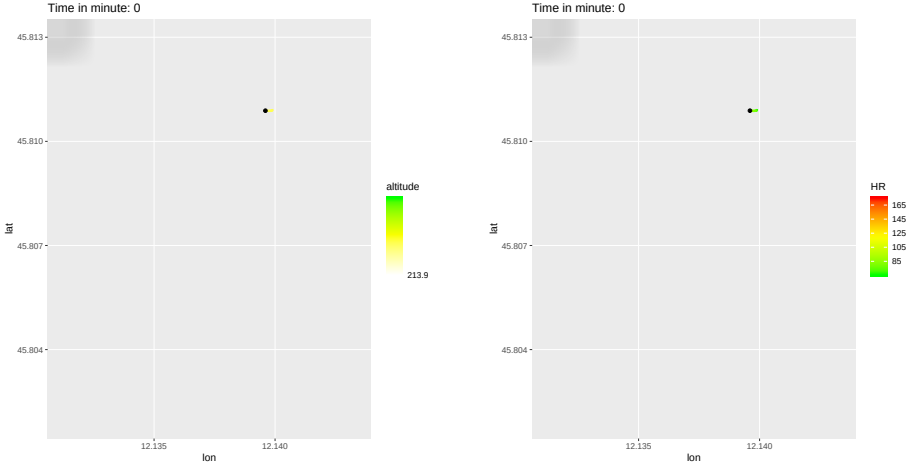
Introduction

- The evolution of new technologies provides an ever growing amount of data in all the aspects of everyday life and it is rapidly changing the way people make use of information.
- Athletes of several disciplines, such as running, swimming and cycling, use sport devices that collect geo-localized biometrical and physical data over time.
- These data are useful for analyzing the performances, in order to check personal physical conditions and to plan future training activities.
- Some recent R-packages develop for analyze these kind of data are
  1. `trackeRapp: Interface for the Analysis of Running, Cycling and Swimming Data from GPS-Enabled Tracking Devices` [Kosmidis and Hornak, 2019]
  2. `trackeR: Infrastructure for Running, Cycling and Swimming Data from GPS-Enabled Tracking Devices` [Frick and Kosmidis, 2017]

The dataset

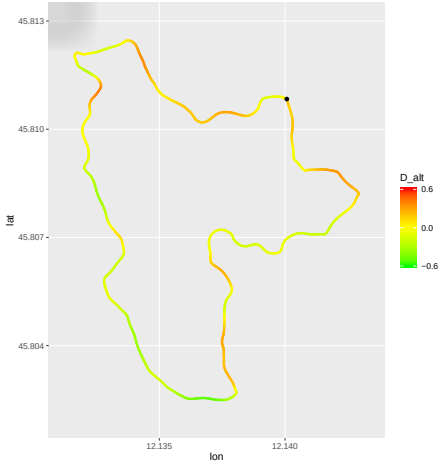
- The dataset is a collection of more than 2800 running activities recorded by the *smart-watches* of 17 subjects.
- Each subject collects their activities during time, and each activity is an high frequency geo-localized multivariate time series characterized by complex behaviors and the presence of missing data.
- The variables collected over time are `Heart Rate` (`bpm`), `Speed` (`m/s`), `Cumulative Distance` (`m`), `Altitude` (`m`), etc.
- Data were made available by the users of the on-line platform Strava (www.strava.com), which is extensively used for storing, sharing and analyzing sport data.

Exploratory analysis I

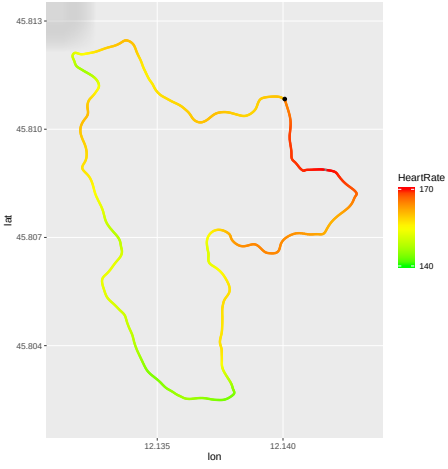


[Kahle and Wickham, 2013]

Exploratory analysis I



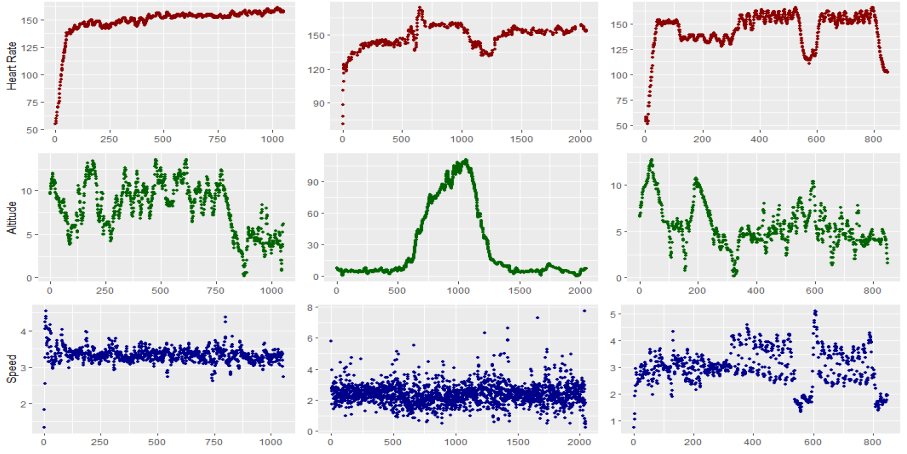
(a) `Altitude`



(b) `Heart Rate`

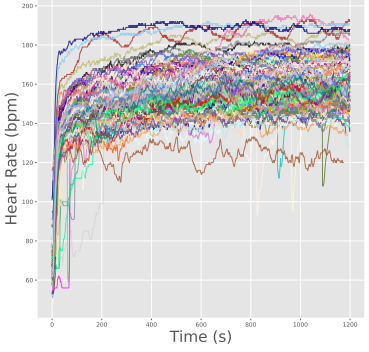
[Kahle and Wickham, 2013]

Exploratory analysis II

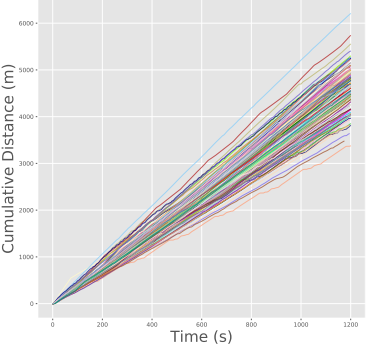


(A) `Activity 31` (B) `Activity 66` (C) `Activity 162`

Exploratory analysis III



(A) `Heart Rate`



(B) `Cumulative Distance`

Model

Motivation & Scope

We propose a new Bayesian clustering approach that relies on matrix-variate state-space models. The reasons are listed below:

1. the identification of activities which require similar efforts is crucial for monitoring individual enhancement over time;
2. the matrix-variate state-space form of the model allows us to consider the dependence within each activity (a multivariate time series) and between activities (a sequence of time series);
3. the advantage of performing time series clustering within a state-space framework is threefold:
   (a) consider complex dependencies (trend and periodic patterns);
   (b) automatic treatment of missing values;
   (c) analyzing data on-line, while they are collected.

Scope: monitoring athletes performances in training activities over time (athletes performance is the latent variable of interest).

Model specification I

- Let $y_{p,n,t}$ denotes the observation at time t, for $t = 1, 2, \ldots, T_n$, of the p–th scalar random variable for activity n, for $p = 1, 2, \ldots, P$ and $n = 1, 2, \ldots, N$.
- We assume that the activities can be clustered in one of G different groups depending on the observed variables.
- The activities which belong to the group g share the same trajectories for all the observed variables.
- As an example, we assume a random walk with stochastic drift to describe the dynamic evolution of the p–th observed variable of activities belonging to the group g, for $g = 1, \ldots, G$, $p = 1, \ldots, P$.

Model specification II

We specify the following state-space model for the dynamic evolution of $y_{p,n,t}$, the p–th scalar random variable at time t for activity n, which belong to the (unknown) group g:

$$
\begin{aligned}
y_{p,n,t} &= \mu_{p,t}^{(g)} + \varepsilon_{p,n,t} \\
\mu_{p,t+1}^{(g)} &= \mu_{p,t}^{(g)} + \beta_{p,t}^{(g)} + \eta_{p,t}^{(g)} \\
\beta_{p,t+1}^{(g)} &= \beta_{p,t}^{(g)} + \zeta_{p,t}^{(g)},
\end{aligned}
$$

with independent Gaussian innovations, $\mu_{p,0}^{(g)} \sim N(\hat{\mu}_{p,0|0}^{(g)}, P_{\mu,p,0|0}^{(g)})$ and $\beta_{p,0}^{(g)} \sim N(\hat{\beta}_{p,0|0}^{(g)}, P_{\beta,p,0|0}^{(g)})$.

Matrix variate form

The model can be represented in matrix-variate form

$$\mathbf{Y}_t = \mathbf{Z}\mathbf{A}_t\mathbf{S}^\intercal + \boldsymbol{\Upsilon}_t, \quad \boldsymbol{\Upsilon}_t \sim \mathbf{MN}_{P,N}(\mathbf{0}, \boldsymbol{\Sigma}^R \otimes \boldsymbol{\Sigma}^C)$$
$$\mathbf{A}_{t+1} = \mathbf{T}\mathbf{A}_t + \boldsymbol{\Xi}_t, \qquad \boldsymbol{\Xi}_t \sim \mathbf{MN}_{Q,G}(\mathbf{0}, \boldsymbol{\Psi}^R \otimes \boldsymbol{\Psi}^C)$$

with $\mathbf{A}_0 \sim \mathbf{MN}_{Q,G}(\hat{\mathbf{A}}_{0|0}, \mathbf{P}^R_{0|0} \otimes \mathbf{P}^C_{0|0})$. .

Matrix variate form

The model can be represented in matrix-variate form

$$\boxed{\mathbf{Y}_t} = \mathbf{Z}\mathbf{A}_t\mathbf{S}^\intercal + \boldsymbol{\Upsilon}_t, \quad \boldsymbol{\Upsilon}_t \sim \mathbf{MN}_{P,N}(\mathbf{0}, \boldsymbol{\Sigma}^R \otimes \boldsymbol{\Sigma}^C)$$
$$\mathbf{A}_{t+1} = \mathbf{T}\mathbf{A}_t + \boldsymbol{\Xi}_t, \qquad \boldsymbol{\Xi}_t \sim \mathbf{MN}_{Q,G}(\mathbf{0}, \boldsymbol{\Psi}^R \otimes \boldsymbol{\Psi}^C)$$

with $\mathbf{A}_0 \sim \mathbf{MN}_{Q,G}(\hat{\mathbf{A}}_{0|0}, \mathbf{P}^R_{0|0} \otimes \mathbf{P}^C_{0|0})$.

$\boxed{\mathbf{Y}_t}$ is a matrix of dimension $P \times N$ that stores the observations of the P variables for the N activities, for all $t = 1, \ldots, T$. .

Matrix variate form

The model can be represented in matrix-variate form

$$\mathbf{Y}_t = \boxed{\mathbf{Z}}\mathbf{A}_t\mathbf{S}^\intercal + \boldsymbol{\Upsilon}_t, \quad \boldsymbol{\Upsilon}_t \sim \mathbf{MN}_{P,N}(\mathbf{0}, \boldsymbol{\Sigma}^R \otimes \boldsymbol{\Sigma}^C)$$
$$\mathbf{A}_{t+1} = \boxed{\mathbf{T}}\mathbf{A}_t + \boldsymbol{\Xi}_t, \quad \boldsymbol{\Xi}_t \sim \mathbf{MN}_{Q,G}(\mathbf{0}, \boldsymbol{\Psi}^R \otimes \boldsymbol{\Psi}^C)$$

with $\mathbf{A}_0 \sim \mathbf{MN}_{Q,G}(\hat{\mathbf{A}}_{0|0}, \mathbf{P}^R_{0|0} \otimes \mathbf{P}^C_{0|0})$.

$\boxed{\mathbf{Z}}$ is a $P \times Q$ matrix of loadings, where Q denotes the number of latent states for each group, while $\boxed{\mathbf{T}}$ is a $Q \times Q$ transition matrix.

Working on the specification of $\boxed{\mathbf{Z}}$ and $\boxed{\mathbf{T}}$ allows us to consider different dynamics for different variables. .

Matrix variate form

The model can be represented in matrix-variate form

$$\mathbf{Y}_t = \mathbf{Z}\,\boxed{\mathbf{A}_t}\,\mathbf{S}^\top + \boldsymbol{\Upsilon}_t, \quad \boldsymbol{\Upsilon}_t \sim \mathbf{MN}_{P,N}(\mathbf{0}, \boldsymbol{\Sigma}^R \otimes \boldsymbol{\Sigma}^C)$$
$$\mathbf{A}_{t+1} = \mathbf{T}\mathbf{A}_t + \boldsymbol{\Xi}_t, \qquad \boldsymbol{\Xi}_t \sim \mathbf{MN}_{Q,G}(\mathbf{0}, \boldsymbol{\Psi}^R \otimes \boldsymbol{\Psi}^C)$$

with $\mathbf{A}_0 \sim \mathbf{MN}_{Q,G}(\hat{\mathbf{A}}_{0|0}, \mathbf{P}^R_{0|0} \otimes \mathbf{P}^C_{0|0})$.

$\boxed{\mathbf{A}_t}$ is a $Q \times G$ matrix of latent states, where the g–th column stores the Q latent states of the g-th group. .

Matrix variate form

The model can be represented in matrix-variate form

$$\mathbf{Y}_t = \mathbf{Z}\mathbf{A}_t \mathbf{S}^\intercal + \boldsymbol{\Upsilon}_t, \quad \boldsymbol{\Upsilon}_t \sim \mathbf{MN}_{P,N}(\mathbf{0}, \boldsymbol{\Sigma}^R \otimes \boldsymbol{\Sigma}^C)$$
$$\mathbf{A}_{t+1} = \mathbf{T}\mathbf{A}_t + \boldsymbol{\Xi}_t, \qquad \boldsymbol{\Xi}_t \sim \mathbf{MN}_{Q,G}(\mathbf{0}, \boldsymbol{\Psi}^R \otimes \boldsymbol{\Psi}^C)$$

with $\mathbf{A}_0 \sim \mathbf{MN}_{Q,G}(\hat{\mathbf{A}}_{0|0}, \mathbf{P}^R_{0|0} \otimes \mathbf{P}^C_{0|0})$.

$\mathbf{S}$ is a matrix of dimension $N \times G$ is a selection matrix (a latent random matrix), where the n-th row has one 1 in the g-th column if the n-th activity belong to the g-th group, and zeros elsewhere.

[El Assad et al., 2016] propose a similar model to dectect trajectories but, differently from their proposal, in our model $\mathbf{S}$ does not depend on time. .

Matrix variate form

The model can be represented in matrix-variate form

$$\mathbf{Y}_t = \mathbf{Z}\mathbf{A}_t\mathbf{S}^\intercal + \boldsymbol{\Upsilon}_t, \qquad \boldsymbol{\Upsilon}_t \sim \mathbf{MN}_{P,N}(\mathbf{0}, \boldsymbol{\Sigma}^R \otimes \boldsymbol{\Sigma}^C)$$
$$\mathbf{A}_{t+1} = \mathbf{T}\mathbf{A}_t + \boldsymbol{\Xi}_t, \qquad \boldsymbol{\Xi}_t \sim \mathbf{MN}_{Q,G}(\mathbf{0}, \boldsymbol{\Psi}^R \otimes \boldsymbol{\Psi}^C)$$

with $\mathbf{A}_0 \sim \mathbf{MN}_{Q,G}(\hat{\mathbf{A}}_{0|0}, \mathbf{P}^R_{0|0} \otimes \mathbf{P}^C_{0|0})$.

The elements $\boldsymbol{\Upsilon}_t$, $\boldsymbol{\Xi}_t$, and $\mathbf{A}_0$ are matrix-variate random normal, with separable covariance matrix [Gupta and Nagar, 1999].

Stacked form

The model has a stacked representation, which is

$$Y_t = \mathbf{Z}_{vec}\alpha_t + \upsilon_t, \quad \upsilon_t \sim MVN_{PN}(\mathbf{0}, \mathbf{\Sigma}^C \otimes \mathbf{\Sigma}^R)$$
$$\alpha_{t+1} = \mathbf{T}_{vec}\alpha_t + \xi_t, \quad \xi_t \sim MVN_{QG}(\mathbf{0}, \mathbf{\Psi}^C \otimes \mathbf{\Psi}^R)$$

with $\alpha_0 \sim MVN_{QG}(\hat{\alpha}_{0|0}, \mathbf{P}^C_{0|0} \otimes \mathbf{P}^R_{0|0})$.

The involved elements are

$$Y_t = \text{vec}(\mathbf{Y}_t), \quad \mathbf{Z}_{vec} = \mathbf{S} \otimes \mathbf{Z}, \quad \alpha_t = \text{vec}(\mathbf{A}_t),$$
$$\mathbf{T}_{vec} = \mathbf{I}_G \otimes \mathbf{T}, \quad \upsilon_t = \text{vec}(\Upsilon_t), \quad \xi_t = \text{vec}(\Xi_t).$$

With this specification of the model, standard routines for state space models with multivariate observations can be used [Durbin and Koopman, 2012].

Estimation (non-technical)

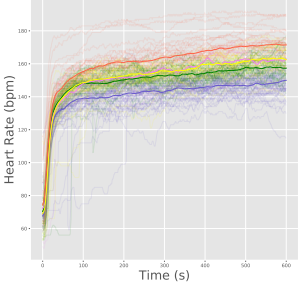
- A data augmentation (both $\mathbf{A}_t$ and $\mathbf{S}$ are latent random variables independent of each other) and fully conjugate Bayesian approach is adopted to estimate the parameters.
- The Gibbs sampling algorithm consists, essentially, of a three-steps procedure. In these steps, we move iteratively from a state space representation to a representation where each activity (a multivariate time series) is treated as a multivariate Gaussian. Specifically:

  1. SSM recursion: Kalman filter, Kalman smoother, simulation smoothing;
  2. model's parameters updating (variances);
  3. mixture allocation and probabilities updating.

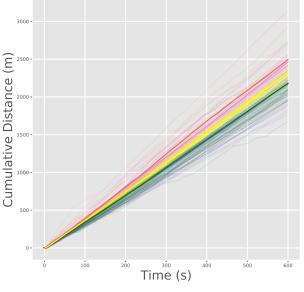
An application as example

Results I

As an example, we present the results of the model specifying $G = 5$, for the first $T = 600$ seconds of $N = 148$ running activities of one individual and $P = 2$ observed variables, with diagonal covariance matrices.

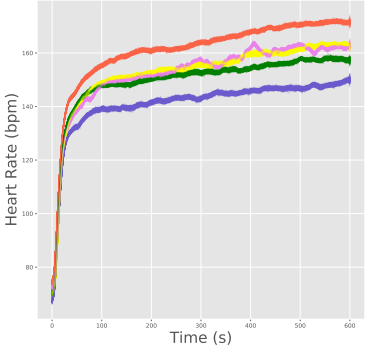


(A) `Heart Rate`

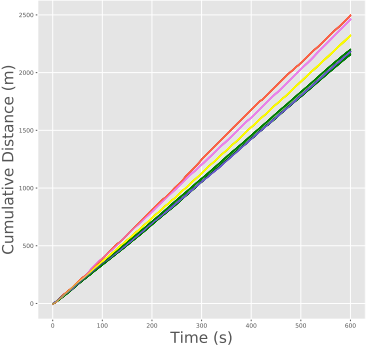


(B) `Cumulative Distance`

Results II



(A) Heart Rate



(B) Cumulative Distance

Conclusion and further developments

Conclusion and further developments

The exemplary nature of this application should be highlighted.

We are working on:

a. algorithm for filtering and smoothing for the model in matrix form [Carvalho and West, 2007]
b. specification of covariance matrix to capture dependence between groups.
c. Online learning and parameter updating while new activities are recorded [Lopes and Carvalho, 2013]

Other possible developments:

1. including an intercept term and control variables in the model;
2. adoption of a Bayesian non-parametric approach for dynamic state space models to capture enhancements of physical conditions.

References & future readings

Any questions or suggestions?
Thanks for your attention!

References I

Bartolucci, F., Murphy, T.B.:
A finite mixture latent trajectory model for modeling ultrarunners behavior in a 24-hour race.
Journal of Quantitative Analysis in Sports, 11(4):193–203, 2015.

Carvalho, C. M. and West, M. (2007).
Dynamic matrix-variate graphical models.
Bayesian Anal., 2(1):69–97.

Durbin, J., Koopman, S.J.:
Time series analysis by state space methods, volume 38 of Oxford Statistical Science Series.
Oxford University Press, Oxford, second edition, 2012.

References II

El Assaad, H., Samé, A., Govaert, G., and Aknin, P. (2016).
A variational expectation-maximization algorithm for temporal data clustering.
Comput. Statist. Data Anal., 103:206–228.

Frick, H. and Kosmidis, I. (2017).
trackeR: Infrastructure for running and cycling data from gps-enabled tracking devices in R.
Journal of Statistical Software, 82(7):1–29.

Gupta, A. K. and Nagar, D. K. (1999).
Matrix variate distributions.
Chapman and Hall/CRC.

References III

Huang, L., Bai, J., Ivanescu, A., Harris,T., Maurer, M., Green, P., Zipunnikov, V.:
Multilevel Matrix-Variate Analysis and its Application to Accelerometry-Measured Physical Activity in Clinical Populations.
Journal of the American Statistical Association, pages 1–12, jun 2018.

Kahle, D. and Wickham, H. (2013).
ggmap: Spatial visualization with ggplot2.
The R Journal, 5(1):144–161.

Kosmidis, I. and Hornak, R. (2019).
trackeRapp: Interface for the Analysis of Running, Cycling and Swimming Data from GPS-Enabled Tracking Devices.
R package version 1.0.

References IV

Lopes, H.F., Carvalho, C.M.:
Online Bayesian learning in dynamic models: an illustrative introduction to particle methods.
In Bayesian theory and applications, pages 203–228. Oxford Univ. Press, Oxford, 2013.

Viroli, C. (2012).
On matrix-variate regression analysis.
J. Multivariate Anal., 111:296–309.

Wang, H. and West, M. (2009).
Bayesian analysis of matrix normal graphical models.
Biometrika, 96(4):821–834.