Dynamic Bayesian clustering of running activities

Italian Statistical Society 2019 conference
"Smart Statistics for Smart Applications"
Universita' Cattolica del Sacro Cuore, Milan

Stival Mattia and Bernardi Mauro
mattia.stival@phd.unipd.it

Department of Statistical Sciences
University of Padova

June 19, 2019

Introduction

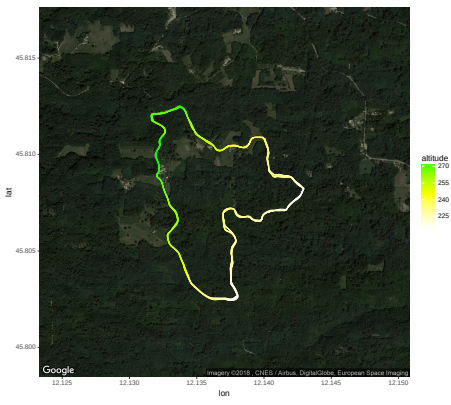
Introduction

- 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.

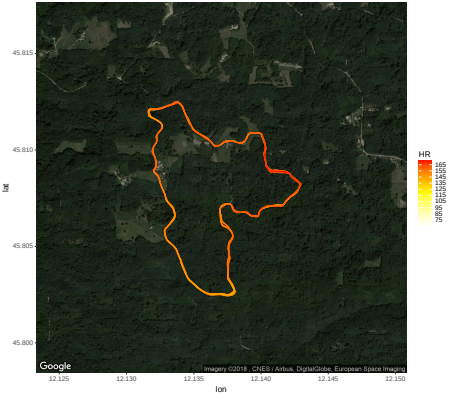
The dataset

- 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

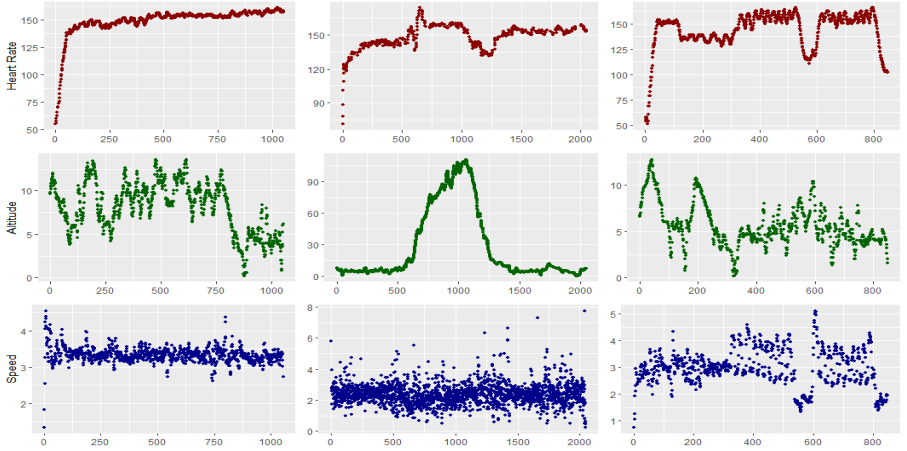


(A) `Altitude`



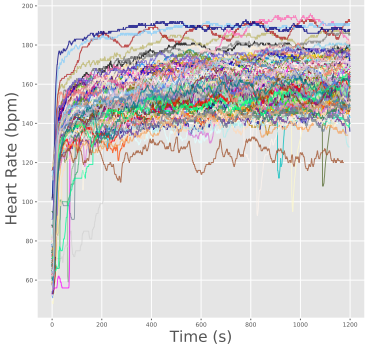
(B) `Heart Rate`

Exploratory analysis II

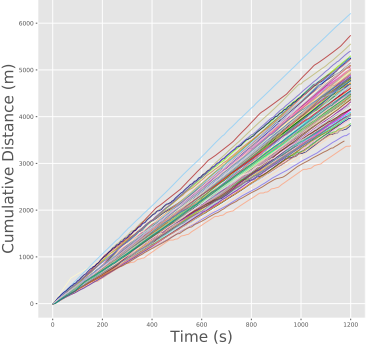


(A) `Activity 31` (B) `Activity 66` (C) `Activity 162`

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 trajectory 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.

Matrix variate form

The model can be represented in matrix-variate form

$$\mathbf{Y}_t = \sum_{g=1}^{G} \left(\boldsymbol{\Lambda}^{(g)} \otimes \boldsymbol{\Theta}_t^{(g)} \right) + \boldsymbol{\Upsilon}_t,$$

$$\boldsymbol{\alpha}_{t+1}^{(g)} = \mathbf{T}\boldsymbol{\alpha}_t^{(g)} + \boldsymbol{\psi}_t^{(g)},$$

where $\boldsymbol{\Theta}_t^{(g)} = \mathbf{Z}\boldsymbol{\alpha}_t^{(g)}$ is the signal and

$$\boldsymbol{\Lambda}^{(g)} = \left(\mathbb{1}\left(S_1 = g\right), \mathbb{1}\left(S_2 = g\right), \ldots, \mathbb{1}\left(S_N = g\right) \right),$$

where S_n is the mixture indicator of the group g for the n–th activity, $n = 1, \ldots, N$, and $g = 1, \ldots, G$.

Dimension of the model

Matrix/Vector	Dimension	Matrix /Vector	Dimension
$\mathbf{Y}_t$	$P \times N$	$\boldsymbol{\alpha}_t^{(g)}$	$2P \times 1$
$\boldsymbol{\Upsilon}_t$	$P \times N$	$\boldsymbol{\psi}_t^{(g)}$	$2P \times 1$
$\boldsymbol{\Lambda}^{(g)}$	$1 \times N$	$\mathbf{Z}$	$P \times 2P$
$\boldsymbol{\Theta}_t^{(g)}$	$P \times 1$	$\mathbf{T}$	$2P \times 2P$

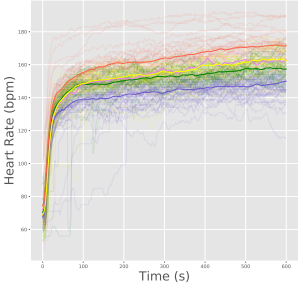
Estimation

- A fully conjugate Bayesian approach is adopted to estimate the parameters (Durbin & Koopman (2012))
- The Gibbs sampling algorithm consists, essentially, of a three-steps procedure:
  1. SSM recursion;
  2. parameters update;
  3. mixture allocation.

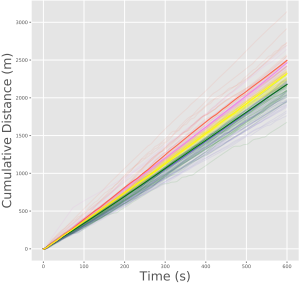
An application as example

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.

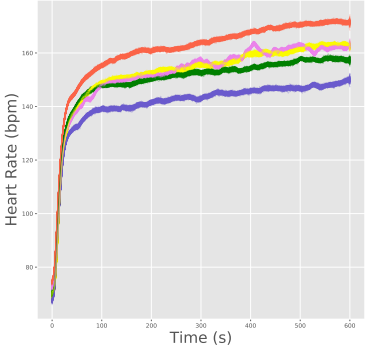


(A) `Heart Rate`

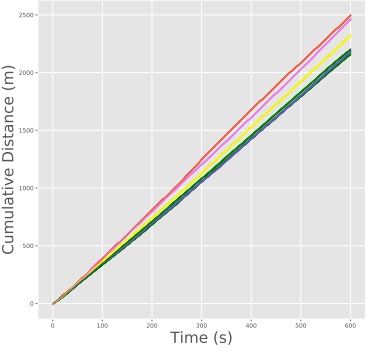


(B) `Cumulative Distance`

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 propose some further developments:

1. including control variables in the model;
2. face the problem of sparsity and curse of dimensionality;
3. allowing on-line update of the model's parameters;
4. adoption of a Bayesian non-parametric approach for dynamic state space models;
5. monitor more individuals over time.

References

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.

Cassese, A., Zhu, W., Guindani, M, Vannucci, M., et al.:
A bayesian nonparametric spiked process prior for dynamic model selection.
Bayesian Analysis, 2018.

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

Egidi, L., Gabry, J.:
Bayesian hierarchical models for predicting individual performance in soccer.
Journal of Quantitative Analysis in Sports, 14(3):143–157, 2018.

Hjort, N.L., Holmes, C., Müller, M., Walker, S.G.:
Bayesian nonparametrics, volume 28.
Cambridge University Press, 2010.

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.

References III


Li, F., Zhang, X.:
Bayesian Lasso with neighborhood regression method for Gaussian graphical model.
Acta Math. Appl. Sin. Engl. Ser., 33(2):485–496, 2017.


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.


Müller, P., Quintana, F.A., Jara, A., Hanson, T.:
Bayesian nonparametric data analysis.
Springer, 2015.

References IV

- Nieto-Barajas, L., Contreras-Cristán, A.:
  A Bayesian nonparametric approach for time series clustering.
  Bayesian Anal., 9(1):147–169, 2014.
- Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A.:
  The deviance information criterion: 12 years on.
  J. R. Stat. Soc. Ser. B. Stat. Methodol., 76(3):485–493, 2014.
- Tibshirani, R.:
  Regression shrinkage and selection via the lasso: a retrospective.
  J. R. Stat. Soc. Ser. B Stat. Methodol., 73(3):273–282, 2011.

Thanks for your attention!