Francesco Bartolucci* and Thomas Brendan Murphy

A finite mixture latent trajectory model for modeling ultrarunners' behavior in a 24-hour race

DOI 10.1515/jqas-2014-0060

Abstract: A finite mixture latent trajectory model is developed to study the performance and strategy of runners in a 24-h long ultra running race. The model facilitates clustering of runners based on their speed and propensity to rest and thus reveals the strategies used in the race. Inference for the adopted latent trajectory model is achieved using an expectation-maximization algorithm. Fitting the model to data from the 2013 World Championships reveals three clearly separated clusters of runners who exhibit different strategies throughout the race. The strategies show that runners can be grouped in terms of their average moving speed and their propensity to rest during the race. The effect of age and gender on the probability of belonging to each cluster is also investigated.

Keywords: clustering; expectation-maximization algorithm; non-ignorable drop-out; ultra running.

1 Introduction

The International Association of Ultrarunners (IAU) 24-Hour World Championships were held in Steenbergen, The Netherlands, from May 11th to 12th, 2013. In this race, 299 competitors ran for 24 h on a course that had a 2.314 km lap. The runners had a chip on their shoe, so their lap count could be recorded automatically and so that time that they finished each lap was also recorded. At the end of the 24-h time period, an alarm is sounded and all runners must stop running immediately. Then, the fraction of the final lap is recorded for each runner. Throughout the race, during any of these laps a runner can continue running, rest for a while and, obviously, he/she can leave the competition before the end. The strategy and performance of the runners in the event are potentially very different due to the age and gender of the participants.

In this paper, we study these data to uncover the strategies used by the runners in the race. We anticipate that the runners will fall into clusters depending on how their strategy might evolve throughout the race. The study of the pacing strategies used allows for comparing the strategies used to those in races over shorter distances. Further, the influence of age and gender on performance can be assessed.

In order to analyze these data and, in particular, cluster runners according to the adopted strategy, we introduce a latent trajectory model in the spirit of Roeder, Lynch, and Nagin (1999); see also (e.g. Muthén and Shedden 1999; Muthén 2004; Bollen and Curran 2006). In practice, the approach is based on a finite mixture of linear and multinomial logit regressions models. The linear regression component is for the speed observed at every lap completed by the runner as response variable. The multimonial logit regression component is for a categorical response variable that, again for every lap, indicates if the subject is regularly running, resting, or leaves before completing the lap. In this way we account for a form of non-ignorable drop-out.

The proposed model accounts for a number of aspects of modeling race duration data that have not been accounted for in many previous studies. In fact, previous studies aggregate the runner pace data over fixed distance intervals and analyze the data using analysis of variance methods (e.g. Hanley 2015). This approach removes important information when the aggregation is used and the analysis of variance does not account for the temporal dependence in the pacing data throughout the race. Our model accounts for the temporal dependence in the runner speeds within the race. We explicitly model the runner resting or stopping, which is a feature of ultra running data that does not appear in races over shorter distances. Thus, the clustering of the runners into distinct strategies is determined by the runner speed and propensity to stop rather than their speed alone which can vary hugely if resting is not explicitly accounted for in the model.

The paper is structured as follows. In Section 2 we introduce the data from the IAU 24-Hour World

***Corresponding author: Francesco Bartolucci,** Department of Economics, University of Perugia, Via A. Pascoli, 20, 06123, Perugia, Italy, e-mail: bart@stat.unipg.it
Thomas Brendan Murphy: University College Dublin, Dublin, Ireland

Championships and discuss previous studies on race pacing. In Section 3 we develop the model for analyzing the runners' strategies and discuss issues to do with model fitting and model selection. The results of fitting the model are presented in Section 4 and the paper concludes by discussing the analysis and the methodology in Section 5.

We make the data and the R code that we used to fit the model available to the reader upon request.

2 Preliminaries

In the following we provide more details on the background and the data studied in this article.

2.1 Background

The IAU 24-Hour World Championships were held in Steenbergen, The Netherlands, in 2013. Over the 24-h period a total of 299 competitors ran on a course that had a 2.314 km lap. The athletes in the race were selected from their home country using qualification criteria set by their home ultrarunning association, thus most participants would be considered to be elite ultrarunners with experience in events that are challenging in distance and duration.

The weather over the two race days averaged 10°C which was 3°C lower than the mean temperature for that time of year in the region. Further, there was a considerable amount of precipitation, 8.8 mm of rain over the 2 days which included the race, and the winds were about 6 m/s with gusts as strong as 17 m/s. So, the race conditions were considered to be very difficult. These weather conditions increased the necessity to stop and change strategy during the race and contributed to the high drop-out rate due to injury, hypothermia, exhaustion and other factors during the 24 h.

A number of factors have been shown to influence the performance of a runner and their probability of completion in extreme ultrarunning events of the type being analyzed herein. Zingg et al. (2013) found that 40–44-year-old men and 35–39-year-old women had the fastest pace in 24-h races. Further, they found that female athletes have an average speed that is approximately 10% slower than male athletes. Lambert et al. (2004) studied the decrease in pace in 100 km ultra races and showed that the top runners were able to maintain their speed for 50 km but declined by 15% from their starting speed by the finish; however, in the IAU 24-h race data the runners tend to cover much greater distances than those studied therein. Further, Kao et al. (2008) found that runners in a 24-h race lose 5.05±2.28% of their body weight, so the nutritional and hydration demands of such events are crucial and thus the races can have high attrition throughout.

A significant amount of research has been completed on pacing in athletic events including running, cycling, swimming, speed skating and triathlon. Abbiss and Laursen (2008) review the strategies used and characterize them into: negative pacing (where speed increases throughout the event), all-out pacing (where there is an initial burst of speed followed by a slowly decreasing speed), positive pacing (where there is a gradually decreasing speed), even pacing (where the speed is constant), parabolic pacing (where the speed gradually decreases through an event but increases at a later stage in the event) and variable pacing (where the speed varies throughout an event, usually due to external factors like geography or environment). Within the parabolic pacing strategy, three shapes were characterized, a U-shaped (with a symmetric pacing profile), a J-shaped (where a small pacing decrease is followed by a steep rise in pace) and reverse J-shaped (where a strong pacing decrease is followed by a small rise in pace). Hanley (2015) studied pacing in the IAAF World Half Marathon (21.1 km) championships and found that the top runners maintained a constant speed for 15 km followed by a small decrease from 15 to 20 km and a strong increase in speed for the final 1.1 km whereas other runners had a gradual decrease in speed over the first 20 km followed by an increase for the final 1.1 km. Lima-Silva et al. (2010) did a similar analysis for a 10 km running race and observed similar pacing profiles where the pace was constant or slowly decreasing for the first 9600 m followed by a sudden increase for the final 400 m. Further, March et al. (2011) and Santos-Lozano et al. (2014) found similar pacing profiles in races over the marathon distance.

There have been fewer studies of pacing within ultra marathon events. However, Lambert et al. (2004) showed that in a race over 100 km, the pace of athletes tends to decrease over the duration of the event; they did not observe a strong increase at the end of the race, but this may because their data are aggregated over 10 km intervals.

Therefore, the performance in terms of speed, minimizing resting and avoiding dropping out are dictated by a number of factors for runners in ultra running events of long duration. Thus, it is of great interest to investigate the strategies used and to see how different clusters of runners utilize different strategies during such an event.

This will also allow for comparison of the pacing strategies in 24-h ultra races with races of a shorter duration.

2.2 Data

For each of the $n = 299$ runners we have a record of the lap time for each completed lap until the 24-h period was completed. Overall, 219 of the runners were still running at the end of the 24-h period and 80 runners finished running a significant time period before the end of the 24-h period.

In Table 1 we report some summary statistics for the runners: age at the start of the race, the number of completed laps, number of laps completed in a non-standard way (e.g. resting during the lap), speed per lap and average speed per athlete (km/h). In these summaries, we consider a lap to be completed in a non-standard way when the lap speed is below 4 km/h which indicates that the runner may have stopped during the lap or commenced walking during the lap. Further, 68% of the participants in the race were male and 32% were female. In order to properly read the table, note that the column "Speed" refers to single laps, whereas the column "Av. Speed" is referred to athletes and then it happens that the maximum average speed is lower than the maximum speed and the minimum average speed is higher than the minimum speed.

The only covariates available for the runners are age and gender. Whilst the performance of the runners may depend on many other factors including training, experience and nutrition, this information is not available for modeling purposes. However, the event is one for elite athletes selected by their home nation, so the athletes need to be experienced and well prepared to participate in the event.

The main variables of interest in this study are the runner's speed per lap and a categorical variable that records the runner's behavior (or status) during a lap (i.e. if the subject is running, resting, or leaves the race in a certain lap). The trajectory of the speed and of the categorical variable throughout the race are displayed in Figure 1.

Table 1: Descriptive statistics for the runners, the lap count and the lap speed data for the participants in the race.

	Age	N. laps	N. laps (not running)	Speed (km/h)	Av. speed (km/h)
Min.	21.0	14.000	0.000	4.000	4.133
1st Qu.	39.0	55.000	0.000	8.106	8.312
Median	45.0	80.000	0.000	9.450	8.998
Mean	45.5	74.550	1.087	9.184	8.965
3rd Qu.	51.0	92.000	2.000	10.440	9.708
Max.	72.0	116.000	13.000	13.930	11.390

The overall appearance from Figure 1 is that the speed has a convex shaped behavior with respect to the lap number. Different explanations may be conjectured for this shape. A natural explanation is that runners tend to decrease the speed during the first part of the competition but they increase the speed when the end of the race is getting close. However, there is a drop-out effect due to the race being run for a fixed time (24 h) rather than a fixed distance. Thus, faster runners complete more laps, whereas slower runners complete fewer laps and they are not considered in the computation of the average speed when the lap number is large. Moreover, there is one further form of drop-out due to a runner leaving the race before the end of the competition and this may have a similar effect on the shape of the average lap speed.

We also note that the proportion of subjects still running dramatically decreases after the 50th lap and this is again due the two forms of drop-out mentioned above. On the other hand, we have a parabolic behavior for the proportion of subjects resting in a certain lap with the proportions being very low for low lap numbers and decreasing again for high lap numbers.

Thus, appropriate statistical modeling of the race data will need to account for the features found in this exploratory data analysis; the development of such a model is given in Section 3.

3 The statistical model

For the sample of n runners, let L_i denote the random variable for the number of laps completed by runner i before the end of the race, with $i = 1, \ldots, n$. Moreover, let B_{il} be the discrete random variable for the behavior (or status) of runner i during lap l, with $l = 1, \ldots, l_i$, where l_i is the realization of L_i (the convention of using lower-case letters for realizations of random variables or vectors is used throughout the article). In particular, $B_{il} = 0$ stands for a standard run lap, $B_{il} = 1$ for a lap in which the runner rests, and $B_{il} = 2$ denotes that subject i leaves before the end of the race during lap l. For a runner with a good performance we expect to observe all values of B_{il} equal to 0 (or almost all values equal to 0). Finally, we denote the speed at which runner i completed lap l by Y_{il}. The observed speed, y_{il}, is available for $i = 1, \ldots, n$ and $l = 1, \ldots, l_i$ and when $B_{il} = 0$; when $B_{il} = 1$ the observed speed, y_{il}, is not relevant in our analysis and when $B_{il} = 2$ no speed is observed because the runner is finished running.

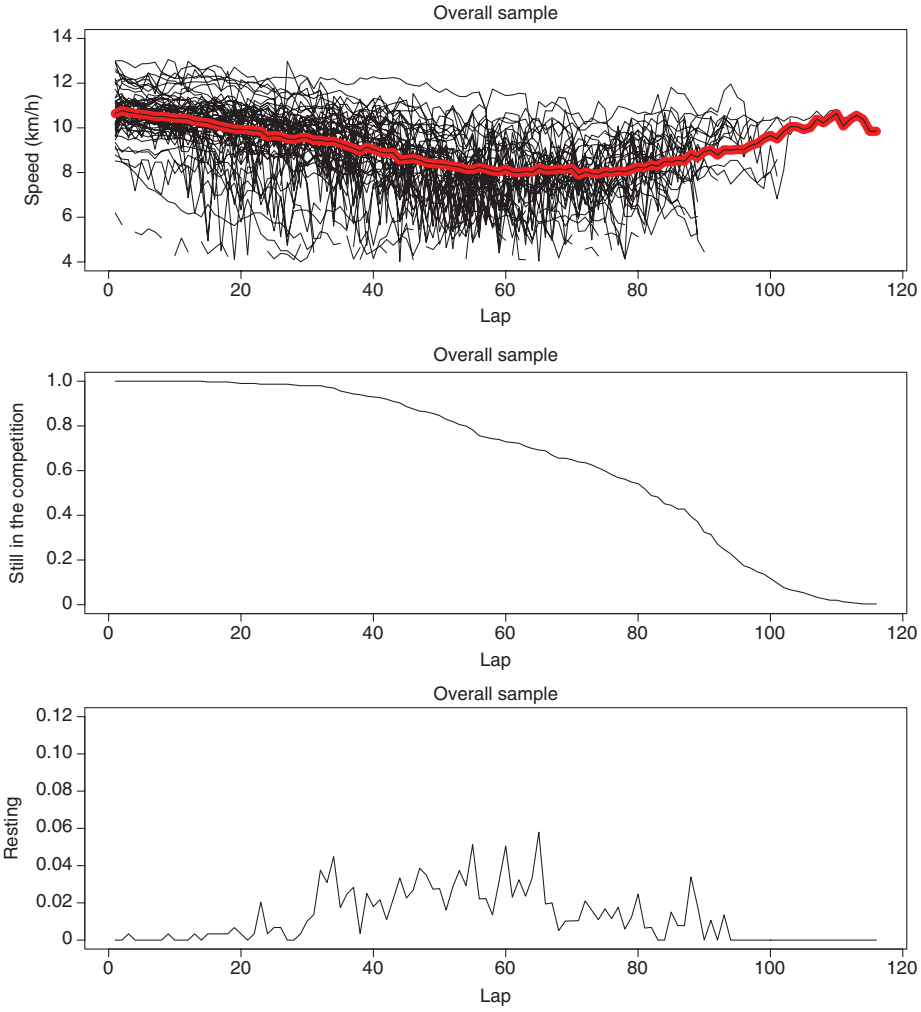

Figure 1: The top panel shows trajectories of speed for the individual runners (one every five) and the average speed per lap in red; the middle panel shows the proportion of subjects still running in a certain lap; the bottom panel shows the proportion of runners resting in a certain lap.

3.1 Model assumptions

We adopt a latent trajectory model (Muthén and Shedden 1999; Roeder et al. 1999; Muthén 2004; Bollen and Curran 2006) that accounts for different possible strategies in running and at the same time facilitates clustering runners according to the adopted strategy by considering that for certain laps they may be running normally ($B_{il} = 0$), also they may have a rest ($B_{il} = 1$), they may finish before the end of the race ($B_{il} = 2$), or they finish because the 24-h time limit is reached. Essentially different lap performances are grouped into a finite number of possible states and different strategies are represented by specific probabilities for these states. Also the runners are clustered in finite number of latent classes according to their overall performance and the *a priori* probabilities to belong to each cluster are allowed to depend on individual covariates.

Let U_i denote a latent variable for the overall performance of runner i and let k denote the number of its possible values, labeled from 1 to k, with the corresponding mass probabilities indicated by $\pi_{iu} = p(u_i = u)$ where π_{iu} may depend on runner-specific covariates. Each of the values (u) identifies a *cluster of runners*. The model is based on the following assumptions for every runner i given that he/she is in cluster u:

– on the first lap ($l = 1$), B_{il} has a generalized Bernoulli distribution with probabilities parametrized on the basis of multinomial logits, that is,

$$
\begin{aligned}
p(B_{il}=0 \mid U_i=u) &= \frac{1}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_{1u})+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_{2u})},\\
p(B_{il}=1 \mid U_i=u) &= \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_{1u})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_{1u})+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_{2u})},\\
p(B_{il}=2 \mid U_i=u) &= \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_{2u})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_{1u})+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_{2u})},
\end{aligned}
\qquad (1)
$$

where $\boldsymbol{x}_l$ is a function of l; in particular, every $\boldsymbol{x}_l$ is a column vector containing the terms of an orthogonal polynomial (Kennedy and Gentle 1980) of order r, which in our application is fixed equal to 3. If $B_{il} = 0$, then we assume the following model for the lap speed:

$$Y_{il} \mid B_{il} = 0,\ U_i = u \sim N(\mu_{lu}, \sigma^2),\ \mu_{lu} = \boldsymbol{x}_l'\boldsymbol{\beta}_u. \tag{2}$$

If $B_{il} = 1$ then the distribution of Y_{il} is left unspecified whereas if $B_{il} = 2$ then the process is stopped. The process is also stopped if the distance between the time of the lap (depending on the speed) is close to the end. Parameter vectors $\boldsymbol{\gamma}_{1u}$, $\boldsymbol{\gamma}_{2u}$, and $\boldsymbol{\beta}_u$ and cluster specific, whereas the variance σ^2 is common to all clusters.

– for the following laps ($l > 1$) and provided that the runner is still in the competition, B_{il} and Y_{il} are assumed to have the same distribution as above. Again, if the overall time is close to the end of the race or $B_{il} = 2$, then the precess is stopped as the runner leaves the competition.

Note that there are two forms of drop-out. The first is due to the overall time of the race which is non-informative as it deterministically depends on the previous values of response variables. The second, for the runner leaving the competition before the end, is informative and it is explicitly accounted in by the multinomial logistic regression model (1). Also note that, according to assumption (2), the lap speeds are conditionally independent given latent class and running normally. This assumption needs to be carefully checked on the basis of the corresponding residuals as we will show in Section 4.

Additionally, we allow for individual covariates to affect the distribution of the latent variables U_i. In particular, we adopt a parametrization based on multinomial logits of the following type:

$$\log \frac{p(U_i = u)}{p(U_i = 1)} = \log \frac{\pi_{iu}}{\pi_{i1}} = \boldsymbol{z}_i'\boldsymbol{\delta}_u,\ u = 2, \ldots, k, \tag{3}$$

where $\boldsymbol{z}_i$ is the vector of covariates (including a constant term for the intercept) for individual i, which are considered as fixed and given and $\boldsymbol{\delta}_u$ is the corresponding vector of regression parameters for being in the u-th category instead of the first category. In the context of this race we have the age and gender of each runner available and these are included as covariates.

The labeling on the clusters is arbitrary and thus the model is only identifiable up to a permutation of the cluster labels. This problem is known as the label-switching problem in mixture models (Redner and Walker 1984). When studying the fitted models we label the clusters in terms of increasing race performance so that the results are presented in an intuitive manner.

3.2 Maximum likelihood estimation

In order to express the model likelihood, we have first to express the distribution of the response variables given the latent variables. In particular, for each subject i we observe the sequence $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{il_i})'$; we also observe $\boldsymbol{y}_{i,obs}$ which corresponds to all or a part of the sequence $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{il_i})'$. In particular, if all elements of $\boldsymbol{b}_i$ are equal to 0, then $\boldsymbol{y}_{i,obs}$ and $\boldsymbol{y}_i$ will coincide; if some elements of $\boldsymbol{b}_i$ are equal to 1 or 2, then $\boldsymbol{y}_{i,obs}$ will be a subvector of $\boldsymbol{y}_i$.

Based on the assumptions formulated in the previous section, the distribution of interest has the following density function:

$$f(\boldsymbol{b}_i, \boldsymbol{y}_{i,obs} \mid U_i = u) = \left[\prod_{l=1}^{l_i} p(b_{il} \mid U_i = u)\right]\left[\prod_{l=1:b_{il}=0}^{l_i} \phi(y_{il} \mid U_i = u)\right],$$
$$u = 1, \ldots, k,$$

where $p(b_{il}|U_i = u)$ is defined in (1), the second product is extended to all observed elements of $\boldsymbol{y}_i$, and $\phi(y_{il}|U_i = u)$ denotes the density of the normal distribution defined according to assumption (2). As in a standard finite mixture model, the *manifest distribution* has density that may be obtained as

$$f(\boldsymbol{b}_i, \boldsymbol{y}_{i,obs}) = \sum_{u=1}^{k} \pi_{iu} f(\boldsymbol{b}_i, \boldsymbol{y}_{i,obs} \mid U_i = u).$$

This is the basis for the model log-likelihood, which has expression

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{b}_i, \boldsymbol{y}_{i,obs}),$$

where $\boldsymbol{\theta}$ is a vector containing all model parameters, that is, $\boldsymbol{\beta}_u$, $\boldsymbol{\gamma}_{1u}$, $\boldsymbol{\gamma}_{2u}$, $\boldsymbol{\delta}_u$, for $u = 1, \ldots, k$, and σ^2.

In order to maximize $\ell(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, we rely on the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977). This algorithm has been used extensively for fitting mixture models (see McLachlan and Krishnan 1997; McLachlan and Peel 2000; Fraley and Raftery 2002) in the maximum likelihood framework.

The EM algorithm is based on alternating the following two steps until convergence in the target function:

– **E-step**: it consists of computing the conditional expected value, given the observed data and the current value of the parameters, of the *complete data* log-likelihood, which is defined as follows:

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{u=1}^{k} z_{iu} \log[\pi_{iu} f(\boldsymbol{b}_i, \boldsymbol{y}_{i,obs} \mid U_i = u)].$$

In the above expression, z_{iu} is an indicator variable equal to 1 if subject i belongs to cluster u (i.e. $U_i = u$), and to 0 otherwise.

– **M-step**: the expected value resulting from the E-step is maximized with respect to $\boldsymbol{\theta}$ and, in this way, this parameter vector is updated.

In practice, the E-step reduces to compute the (conditional) expected value of each indicator variable z_{iu}, denoted by $\hat{z}_{iu}$, by the following simple rule on the basis of the current value of the parameters:

$$\hat{z}_{iu} = \frac{\pi_{iu} f(\boldsymbol{b}_i, \boldsymbol{y}_{i,obs} \mid U_i = u)}{f(\boldsymbol{b}_i, \boldsymbol{y}_{i,obs})}.$$

Regarding the M-step, we can use explicit solutions for the parameter vectors $\boldsymbol{\beta}_u$ and for the common variance σ^2:

$$\boldsymbol{\beta}_u = \left(\sum_{i=1}^{n} \hat{z}_{iu} \sum_{l=1:b_{il}=0}^{l_i} \boldsymbol{x}_l \boldsymbol{x}_l' \right)^{-1} \sum_{i=1}^{n} \hat{z}_{iu} \sum_{l=1:b_{il}=0}^{l_i} y_{il} \boldsymbol{x}_l, \; u = 1, \ldots, k,$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}\sum_{u=1}^{k} \hat{z}_{iu} \sum_{l=1:b_{il}=0}^{l_i} (y_{il} - \mu_{lu})^2}{\sum_{i=1}^{n} o_i},$$

where o_i is the dimension of $\boldsymbol{y}_{i,obs}$, that is, the number of regularly completed laps by runner i. On the other hand, updating the remaining parameters $\boldsymbol{\gamma}_{1u}$ and $\boldsymbol{\gamma}_{2u}$ in (1) requires an iterative algorithm of a Netwon-Raphson type. However, this is a simple algorithm since the objective function being maximized is of the same form as the objective function used when fitting a standard multinomial logit model with weights by maximum likelihood. The same Netwon-Raphson algorithm is also applied to update the parameters $\boldsymbol{\delta}_u$ in (3) that affect the distribution of each latent variables U_i on the basis of the individual covariates. In the case where the π_{iu} probabilities are assumed to be equal for all subjects (i.e. $\pi_{iu} = \pi_u$), we have an explicit solution for the maximization of the expected complete-data log-likelihood with respect to the π_u probabilities:

$$\pi_u = \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{iu}, \; u = 1, \ldots, k.$$

It is important, as for any other iterative algorithm, that the EM algorithm described above is suitably initialized; this amount to guessing starting values for the parameters in $\boldsymbol{\theta}$. We suggest to use both a simple rule providing sensible values for these parameters and a random rule which allows us to properly explore the parameter space. Just to clarify, we choose the starting values for the mass probabilities π_{iu} as $1/k$ for $u = 1, \ldots, k$ under the first rule, which is equivalent to fix the same size for all clusters. The corresponding random starting rule is instead based on first drawing each parameter π_{iu} from a uniform distribution between 0 and 1 and then normalizing these values.

We recall that trying different starting values for the EM algorithm is important to face the problem of multimodality of the likelihood function that may arise in finite mixture models and combining different initialization rules (deterministic and random) is an effective strategy in this regard.

3.3 Model selection

Given that our application is focused on the clustering of individuals in separate groups, the main selection criterion we use for the number of these groups is the Normalized Entropy Criterion (NEC; Celeux and Soromenho 1996; Biernacki, Celeux, and Govaert 1999). This criterion is based on the following index:

$$\text{NEC}_k = \frac{-\sum_{i=1}^{n}\sum_{u=1}^{k} \hat{z}_{iu} \log \hat{z}_{iu}}{\hat{\ell}_k - \hat{\ell}_1}, \; k \geq 2,$$

with $\text{NEC}_1 = 1$, where the numerator corresponds to the entropy and the denominator to the difference in maximum log-likelihood between the model with k classes and with 1 class. According to this approach, the value of k corresponding to the minimum of NEC_k has to be preferred, as it corresponds to the model being the best compromise between separation of the classes (as measured by the entropy) and goodness-of-fit (measured by the log-likelihood).

For completeness, we mention that another important criterion for selecting the number of components of a mixture model is the Bayesian Information Criterion (BIC; Schwartz 1978; Kass and Raftery 1995), which is based on the minimization of the index

$$\text{BIC}_k = -2\hat{\ell}_k + \log(n)(\#\text{par}),$$

where #par is the number of free parameters in the model; for an illustration see McLachlan and Peel (2000), Chapter 6. However, it is known that this criterion typically leads to a less parsimonious model than the model selected with NEC and, in particular, with classes not well separated. Therefore, given the target of our application, we prefer to rely on NEC.

4 Results

The proposed model was fitted, using maximum likelihood, for increasing values of k from 1 to 5. For each model fit, the value of the maximized likelihood, the entropy, NEC and BIC values were computed; the results are shown in Table 2.

Considering that our primary aim is the clustering of runners into distinct states, we rely on the NEC criterion and choose $k = 3$ clusters, corresponding to the minimum of the corresponding index. In any case, we have very good separation (low entropy) between the clusters under any choice of k; thus the model clearly separates runners into clusters and does this in a definitive manner.

For the selected model, with $k = 3$, the parameter estimates together with corresponding standard errors are reported in the Tables 3–5. The clusters have been labeled according to the average speed, with the lowest speed cluster labeled as Cluster 1 and the highest speed cluster labeled as Cluster 3.

In order to interpret the clusters according to the estimated parameters, in Figure 2 we show the mean lap speed and the probability of $B_{il} = 0$ and the conditional probability of $B_{il} = 1$ given $B_{il} > 0$. These curves are based on the estimated parameters, but the represented points are obtained by a Monte Carlo simulation, in order to account for the non-informative drop-out due reaching the race time deadline of 24 h. This procedure amount to randomly draw a large number of trajectories for each cluster and then computing the average trajectory. In practice, each simulated trajectory is obtained as of series of values randomly drawn from the conditional distribution of the response variables Y_{il} and B_{il} given the cluster.

Interestingly, all of the clusters are characterized by a decreasing speed profile but with a rise in speed prior to the end of the race. The rise is particularly strong in Cluster 1 but this can be explained by the fact that a number of runners from this cluster drop out and the remaining runners have a higher average speed than those in the laps prior to when they dropped out. Further, Clusters 2 and 3 initially have a flat speed profile before the profile drops and eventually rises. Thus, the speed profiles of the groups can be characterized as a mix of those outlined in Abbiss and Laursen (2008). Most runners follow an even pacing initial phase for the early laps but this is followed by a reverse-J pacing phase. The athletes in the higher performing groups are able to maintain the even pacing for more laps than the lower performing groups.

It is also worth noting that the effect of the runner covariates, gender and age, as shown in Table 5 are only minor. The possibility of including higher order regression

Table 2: Model summaries for the choice of the appropriate number of strategies (k).

k	Log-likelihood	#par	BIC	Entropy	NEC
1	−42559.12	12	85186.65	0.0000	1.000000
2	−37979.32	27	76112.55	3.0549	0.000667
3	−36073.36	42	72386.13	3.7283	0.000575
4	−35054.75	57	70434.43	5.9694	0.000795
5	−34539.12	72	69488.67	8.2349	0.001027

Table 3: Estimates of the parameters β_u, with standard errors in parentheses.

Power	Cluster (u)		
	1	2	3
0	9.524	9.235	10.014
	(0.209)	(0.042)	(0.016)
1	27.970	2.482	−6.514
	(2.975)	(0.606)	(0.228)
2	29.615	16.135	6.697
	(2.126)	(0.497)	(0.229)
3	5.619	6.702	4.480
	(0.874)	(0.278)	(0.190)

Table 4: Estimates of the parameters γ_{1u} and γ_{2u}.

Cluster (u)		B_{il}	
	Power	1	2
1	0	−12.088	−12.189
		(2.829)	(0.603)
	1	−113.806	−88.901
		(39.343)	(2.426)
	2	−82.106	−71.177
		(25.082)	(3.915)
	3	−18.031	−14.235
		(9.989)	(5.665)
2	0	−6.120	−7.022
		(0.431)	(0.747)
	1	2.824	11.623
		(10.856)	(19.827)
	2	−27.313	−23.202
		(4.749)	(8.282)
	3	2.789	9.233
		(5.842)	(10.780)
3	0	−8.199	−7.644
		(2.163)	(0.863)
	1	33.418	11.216
		(39.297)	(19.556)
	2	−37.100	−24.525
		(21.962)	(9.541)
	3	1.834	3.251
		(16.060)	(11.213)

The standard error of the estimates are given in parentheses.

Table 5: Estimates of the parameters δ_u with standard errors in parentheses.

Covariate	Cluster (k)	
	2	3
Intercept	1.135	0.777
	(0.227)	(0.240)
Gender	−0.298	−0.512
	(0.320)	(0.345)
Age	−0.003	−0.034
	(0.024)	(0.025)

terms was considered but these terms had little effect on the model, so they are omitted. The fitted cluster probabilities can also be seen in Figure 3 where the probability of each cluster membership is shown for males and females and the range of ages of the participants in the race; the probabilities are approximately constant with respect to the covariates. In addition, it is clear that Cluster 2 is the most prevalent, followed by Cluster 3 and Cluster 1. Thus, the cluster of slowest runners is the least prevalent one within the set of competitors.

As already noticed in the entropy calculations from Table 2, the clustering divides the runners into very distinct clusters. In fact, the maximum *a posteriori* probabilities are almost all very close to 1, with a mean value of 0.9933. Thus, the model has effectively clustered the runners into different and distinct strategies.

The trajectories of the runners in each cluster are reported in Figure 4. The plot of the trajectories and the mean trajectory shows that the model fits the data very well and the differences between the clusters are highlighted. In particular, the speed trajectory for each cluster is quite similar but the runners in different clusters are running

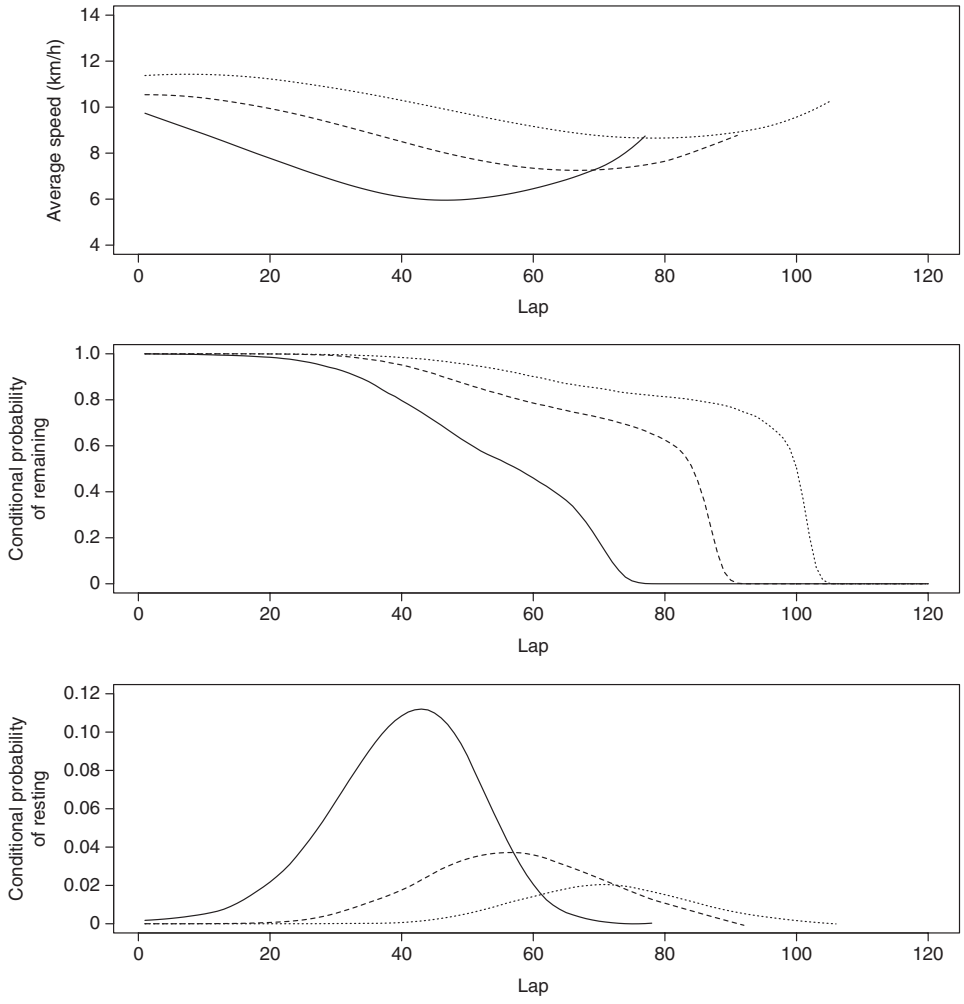

Figure 2: The top panel shows the estimated mean trajectories given the cluster; the middle panel shows the trajectories of the conditional probability that a subject is still running in a certain lap; the lowest panel shows the proportion of a subject resting in a certain lap. The solid line corresponds to the first group, the dashed line to the second group, and the dotted line to the third group.

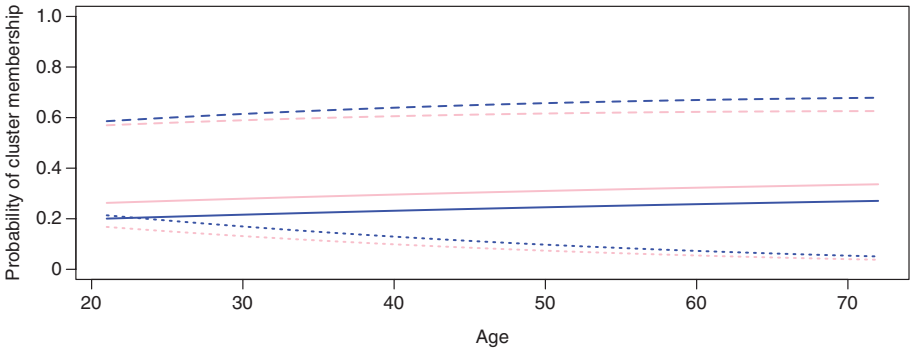

Figure 3: Probability of belonging to each cluster for each gender and age. The pink lines show the probabilities for females and the blue lines for males. The probability of belonging to Cluster 1 is shown as solid line, Cluster 2 as a dashed line and Cluster 3 as a dotted line.

at different average speeds (increasing from Cluster 1 to Cluster 3). Further, the clusters are also characterized by the rate of resting and dropping out with these behaviors being less prevalent and later as the cluster number goes from 1 to 3.

Finally, to check the conditional independence assumption in equation (2), we obtained the residuals for each athlete and lap as the difference between the observed speed and the predicted speed given the latent class assignment of the athlete. We then computed the

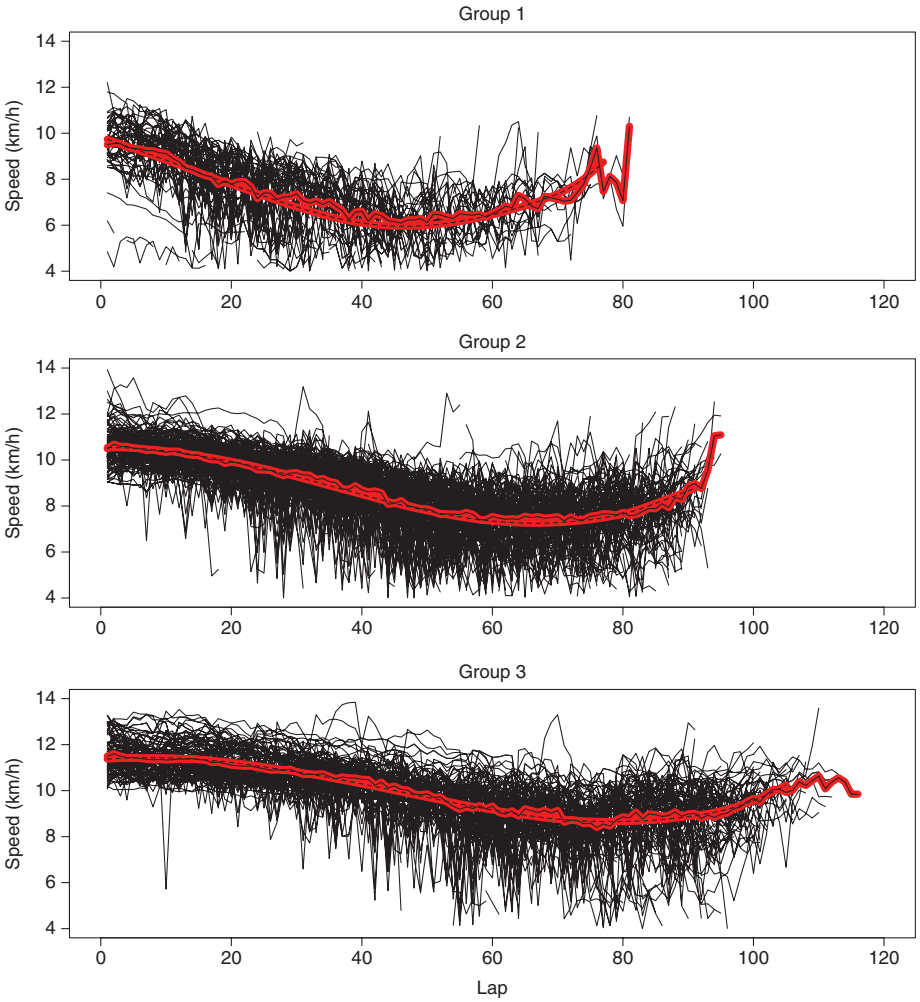

Figure 4: Trajectories of the runners assigned to each cluster with corresponding mean (solid) and estimated mean on the basis of the parameters (dashed).

Table 6: Descriptive statistics about the athlete-specific autocorrelations coefficients between residuals.

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
−0.3625	0.3091	0.4929	0.4681	0.6585	0.9583

autocorrelation of the residuals for each athlete given that he/she is running normally, obtaining in this way n autocorrelation coefficients. Table 6 shows the main descriptive statistics for these autocorrelations. These results indicate that the aforementioned assumption of conditional independence between lap speeds may be too restrictive. To overcome this limitation we suggest some possible extension in Section 5.

5 Discussion and conclusions

The running strategies of the runners in the 2013 International Association of Ultrarunners 24-Hour World Championships has been investigated using a latent trajectory model. The model was constructed to capture the changing speeds of the runners in the race and to facilitate modeling runners who rest or stop during the race duration.

The modeling strategy established that there were three distinct clusters of runners who differed in both their running speed and their prevalence to rest or stop running completely. In all clusters, the runners exhibited a gradual decrease in pace throughout the race; this is similar to the pacing observed by Lambert et al. (2004) in a 100 km race. However, interestingly, in all clusters the average speed of runners increased when the end of the race became closer; this is similar to what has been previously observed in a wide range of race distances as outlined in Section 2.

Further, the propensity to stop shows a peak towards the middle of the race only. The group of best performing runners had very little tendency to stop at any point during the race. The cluster membership was not strongly influenced by either gender or age. The middle speed runners form the largest cluster, faster runners forming the next largest cluster and slower runners forming the smallest cluster.

Likelihood-based inference for this model was achieved using the EM algorithm combined with model selection using the normalized entropy criterion, so that clearly separated clusters yielded.

Limitations of the proposed approach are mainly due to the structure of the data and, in particular, to the reduced number of covariates that are available. In fact, it would be of interest to dispose of more details about the athletes, such as previous performances in similar races. However, if available, this information may be easily included among the individual covariates affecting the probability of belonging to each cluster. Similarly, time-varying covariates related, for instance, to the temporary weather conditions, could be included in the model, but this would require a suitable data manipulation to take into account that the outcomes are referred to each lap run by every athlete and the same lap number may correspond to different moments of the race for different athletes. This is mainly due to the variability of the performances in terms of lap speed.

The proposed approach assumes that, given the latent class, the probability of running normally at any time is independent of any other time. In addition, given the latent class and running normally, the speed at a particular time is conditionally independent of that at any other times. In particular, the diagnostic analysis illustrated at the end of previous section indicates that the second assumption is restrictive for the data at issue. A possibility to relax this assumption is to assume a mixed-effects model based on random intercepts and/or regression coefficients as in the approach of Muthén and Shedden (1999). This approach may result in more precise inferences and reduced bias by addressing the dependence between consecutive laps.

Finally, it is also important recalling that a basic assumption of the proposed model is the independence between athletes in terms of behavior during the race. This assumption rules out possible interactions between runners which would be of interest to study. In particular, there might exist particular "group" strategies that lead to an improvement of the performance of certain athletes. This again would require a more complex data structure and, in particular, a much more sophisticated model having elements of a model for social networks that could be the object of future research.

Acknowledgments: We would like to thank the editors and referees who made constructive suggestions that greatly improved this paper. Francesco Bartolucci acknowledges the financial support from award RBFR12SHVV of the Italian Government (MIUR) (FIRB "Mixture and latent variable models for causal inference and analysis of socio-economic data," 2012). Brendan Murphy was supported by the Science Foundation Ireland funded Insight Research Centre (SFI/12/RC/2289).

References

Abbiss, C. R. and P. B. Laursen. 2008. "Describing and Understanding Pacing Strategies During Athletic Competition." *Sports Medicine* 38:239–252.

Biernacki, C., G. Celeux, and G. Govaert. 1999. "An Improvement of the NEC Criterion for Assessing the Number of Clusters in a Mixture Model." *Pattern Recognition Letters* 20:267–272.

Bollen, K. A. and P. J. Curran. 2006. *Latent Curve Models: A Structural Equation Perspective*. Hoboken, NJ: Wiley.

Celeux, G. and G. Soromenho. 1996. "An Entropy Criterion for Assessing The Number of Clusters in a Mixture Model." *Journal of Classification* 13:195–212.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion)." *Journal of the Royal Statistical Society, Series B* 39:1–38.

Fraley, C. and A. E. Raftery. 2002. "Model-Based Clustering, Discriminant Analysis, and Density Estimation." *Journal of the American Statistical Association* 97:611–631.

Hanley, B. 2015. "Pacing Profiles and Pack Running at the IAAF World Half Marathon Championships." *Journal of Sports Sciences* 33:1189–1195.

Kao, W.-F., C.-L. Shyu, X.-W. Yang, T.-F. Hsu, J.-J. Chen, W.-C. Kao, Polun-Chang, Y.-J. Huang, F.-C. Kuo, C.-I. Huang, and C.-H. Lee. 2008. "Athletic Performance and Serial Weight Changes During 12- and 24-hour Ultra-Marathons." *Clinical Journal of Sport Medicine* 18:155–158.

Kass, R. E. and A. E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90:773–795.

Kennedy, W. J. J. and J. E. Gentle. 1980. *Statistical Computing*. New York: Marcel Dekker.

Lambert, M. I., J. P. Dugas, M. C. Kirkman, G. G. Mokone, and M. R. Waldeck. 2004. "Changes in Running Speeds in a 100 km Ultra-Marathon Race." *Journal of Sports Science and Medicine* 3:167–173.

Lima-Silva, A. E., R. C. Bertuzzi, F. O. Pires, R. V. Barros, J. F. Gagliardi, J. Hammond, M. A. Kiss, and D. J. Bishop. 2010. "Effect of Performance Level on Pacing Strategy During a 10-km Running Race." *European Journal of Applied Physiology* 108:1045–1053.

March, D. S., P. M. Vanderburgh, P. J. Titlebaum, and M. L. Hoops. 2011. "Age, Sex and Finish Time as Determinants of Pacing in the Marathon." *Journal of Strength and Conditioning* 25: 386–391.

McLachlan, G. J. and T. Krishnan. 1997. *The EM Algorithm and Extensions*. New York: John Wiley & Sons Inc.

McLachlan, G. and D. Peel. 2000. *Finite Mixture Models*. New York: John Wiley & Sons Inc.

Muthén, B. 2004. "Latent Variable Analysis: Growth Mixture Modeling and Related Techniques for Longitudinal Data." in *Handbook of Quantitative Methodology for the Social Sciences*, edited by D. Kaplan, Pp. 345–368, Newbury Park, CA: Sage.

Muthén, B. and K. Shedden. 1999. "Finite Mixture Modeling with Mixture Outcomes using the EM Algorithm." *Biometrics* 55:463–469.

Redner, R. A. and H. F. Walker. 1984. "Mixture Densities, Maximum Likelihood and the EM Algorithm." *SIAM Review* 26:195–202.

Roeder, K., K. G. Lynch, and D. S. Nagin. 1999. "Modeling Uncertainty in Latent Class Membership: A Case Study in Criminology." *Journal of the American Statistical Association* 94:766–776.

Santos-Lozano, A., P. Collado, C. Foster, A. Lucia, and N. Garatachea. 2014. "Influence of Sex and Level on Marathon Pacing Strategy. Insights From the New York City Race." *International Journal of Sports Medicine* 35:933–938.

Schwartz, G. 1978. Estimating the Dimension of a Model. *The Annals of Statistics* 6:461–464.

Zingg, M., C. A. Rüst, R. Lepers, T. Rosemann, and B. Knechtle. 2013. "Master Runners Dominate 24-h Ultramarathons Worldwide – A Retrospective Data Analysis from 1998 to 2011." *Extreme Physiology and Medicine* 2:21.