Missing data patterns in runners' careers: do they matter?

Mattia Stival
Department of Statistical Sciences, University of Padova, Padova, Italy.
E-mail: mattia.stival@unipd.it
Mauro Bernardi
Department of Statistical Sciences, University of Padova, Padova, Italy.
Manuela Cattelan
Department of Statistical Sciences, University of Padova, Padova, Italy.
Petros Dellaportas
Department of Statistical Science, University College London, London, Great Britain.
Department of Statistics, Athens University of Economics and Business, Athens, Greece.

Summary. Predicting the future performance of young runners is an important research issue in experimental sports science and performance analysis. We analyse a data set with annual seasonal best performances of male middle distance runners for a period of 14 years and provide a modelling framework that accounts for both the fact that each runner has typically run in three distance events (800, 1500 and 5000 meters) and the presence of periods of no running activities. We propose a latent class matrix-variate state space model and we empirically demonstrate that accounting for missing data patterns in runners' careers improves the out of sample prediction of their performances over time. In particular, we demonstrate that for this analysis, the missing data patterns provide valuable information for the prediction of runner's performance.

Keywords: Informative missing data; Longitudinal latent class analysis; Matrix-variate state-space model; Sparse mixture model; Sports performance analysis.

1. Introduction

Planning the future career of young runners is a relevant aspect of the work of coaches, whose role is to guide them during training so that they can perform at their best in competitions. Identifying runner's capabilities and future possibilities is important for multiple reasons. It allows the training load to be appropriately allocated over the years, for improving their performances and reducing their risk of injuries. Good planning, along with support during injuries, has been identified as one of the relevant factors that help avoiding drop-outs of runners (Bussmann, 1999). Moreover, good planning is important also from a psychological and emotional point of view, as it allows runners to strive for achievable goals and collect successes over the years. Pleasant emotions, including satisfaction, have been associated with positive outcomes in, e.g. mental health, performance and engagement (Cece et al., 2019). In this context, the identification of possible careers for a runner, in terms of observed personal performance trajectories over

time, is of paramount importance. For example, identifying the period in which runners reach their peaks can help prepare them for the most important events in their career. Similarly, the knowledge of the expected progress of different runners over the years provides an indication of whether the training process has been carried out correctly. The analysis of athletes' trajectory is carried out in various sports. Leroy et al. (2018) study young swimmers' progression using a functional clustering approach, while Boccia et al. (2017) focus on individual careers of Italian long and high jumpers to figure out which characteristics of young jumpers are predictive of good-level results during their careers.

We focus on the analysis of performances of Italian male middle distance runners, born in 1988, in a period ranging from 2006 to 2019. Previous studies on middle distance runners are few or limited to samples with a small number of runners (see, e.g., Weippert et al., 2021). We use a combination of latent class and matrix-variate state space models. Latent class models for time dependent data have been extensively studied in clustering (see, among others, Frühwirth-Schnatter, 2011; Maharaj et al., 2019; Bartolucci and Murphy, 2015) and hidden Markov models (Cappé et al., 2005; Frühwirth-Schnatter, 2006; Bartolucci and Farcomeni, 2015). They allow to capture the heterogeneity in the careers of runners, thereby describing various possible observable scenarios. Combining them with state space models offers additional advantages, including the possibility of building models for multivariate time series in an intuitive manner as well as the possibility of leveraging well-known tools for inference, including the treatment of missing data (Durbin and Koopman, 2012). Unlike other types of runners and sports, middle distance runners have the major feature of competing in different distances, i.e. in the 800, 1500 and 5000 meters distances, as well as in other spurious ones (i.e. the mile, 3000 meters, etc.). The choice of discipline in which to compete is subjective and typically associated with personal attitudes (Mooses et al., 2013). A runner capable of developing higher speed and greater power typically competes in shorter distances, with respect to those with greater endurance who compete in longer distances. As a consequence, observations in different distances are available for each runner over time, but the absence of a particular discipline can be informative on the runner's attitude. Beyond the variability among subjects related to the type of discipline performed, there is also variability in the developing of runners' careers related to both their abilities and histories. Runners that begin their career late in life are less likely to reach high levels; similarly, runners with unsatisfactory careers are likely to end their careers earlier, with respect to those satisfied with their performances (Hernandez et al., 2011). These aspects are related to drop-in and drop-out phenomena, defined as the events where runners enter and exit the observed sample, respectively.

A key interesting and important question that naturally emerges is whether the absence of data is really associated with observed performances. We attempt to shed light in this question by proposing a matrix state space model in which multivariate time series are clustered together on the basis of their observed trend. Matrix-variate state space models have found application in finance and engineering in past years (see, e.g., Choukroun et al., 2006; Wang and West, 2009), but have recently gained additional interest in the statistical literature to analyze problems in which observations over time are matrices (see, e.g., Hsu et al., 2021; Chen et al., 2020, 2021). In this part of model

specification, clustering is achieved via a latent selection matrix which is involved in the measurement equation. We propose to include temporal dynamics that aim to describe missing data patterns using two different processes. First, runner's personal history is described by a three state process, which describes their entry (drop-in) and exit (drop-out) from the sample. Second, different propensities to compete in different distances are considered to describe runner's personal attitude. The probabilities of both the processes are assumed to be dependent on the latent classes stored in the selection matrix previously mentioned, allowing to consider the possible relation of missing data patterns with the observed performances. In this way, clustering is not only achieved on the basis of runners' performances, but the presence and absence of data is considered informative as well. Works on latent class and clustering models for longitudinal data where the presence of missing values may be informative on the latent structure behind the data are few or limited to domain-specific works (see, e.g., Bartolucci and Murphy, 2015; Mikalsen et al., 2018).

Since the seminal work on missing values by Rubin (1976), researchers have wondered if and when it is possible to ignore the presence of missing values in their datasets. In this work, we consider this problem in a pure predictive framework, in which missing values will be considered as informative if having information on their presence and distribution over time helps in predicting runners performances over the years. If so, one could think to a causal relationship, in a Granger's sense, between missing values and observed performances. Although coherent with our model construction, we avoid the use of definition of informativeness of missing data in a causal sense. Indeed, while correlation between missing data and performances is typically expected in sports performance analysis, direct cause-effect relationships between them and their directions are not clearly defined in the sports science literature.

The rest of the paper is organized as follows: Section 2 presents a new publicly available dataset on middle distance runners; Section 3 describes the proposed model; Section 4 discusses the likelihood and the prior specification; Section 5 presents the evaluation strategy of the model; Section 6 shows the results with the real data. Additional details on the data, the algorithms, and the results are reported in the supplementary material accompanying the paper.

2. Data and exploratory analysis

Our data refer to annual seasonal best performances of male Italian runners, born in 1988, on 5000, 1500 and 800 meters distances in a period between 2006 and 2019. They were collected from the annual rankings accessible on the website of the Italian athletics federation (www.fidal.it), which stores results and rankings in competitions since 2005. All runners with at least two observations were selected and the data are illustrated in Figure 1 with the help of a local regression fit that allows us to perceive a U-shaped curve that describes the distribution of sample trajectories across ages. U-shapes are typically observed in the evolution of runners careers (Haugen et al., 2018). However, their shape can be biased by the presence of missing data, related for example to early exit or late entry in the sample.

Indeed, unlike other type of data, the presence of missing values is predominant in

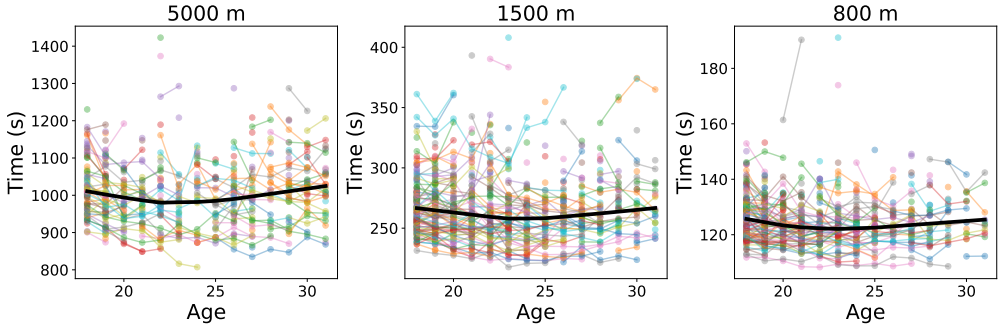

Fig. 1. Seasonal best performances of 369 Italian male middle distance runners in 5000, 1500, and 800 meters distances. Points represent the observed performances. Lines connect the performances in consecutive years of the same runner. The black lines are obtained using local regression.

the careers of these runners. Out of 15498 observable seasonal best performances of $Q = 369$ runners in $P = 3$ distances and $T = 14$ years, only 2411 seasonal best results are observed. A missing value is observed for one runner if the runner does not conclude (and, hence, record) any official competition in a specific discipline during one year. The reasons for not observing any performance can be multiple. The runner can be *not in career* in a specific year, and hence no races are performed during that year. Alternatively, a runner in career can decide not to compete in a specific discipline for several reasons, such as lack of preparation, attitude, technical choices, etc. To understand how missing data patterns differ between runners, Figure 2 shows the observed patterns of missing data for nine distinct runners present in the sample. The patterns shown differ in various features. Some runners have few observations, such as runners 241 and 119. Other runners are characterized by long careers, such as runners 129 and 256. These two runners are interesting because they differ in the type of discipline they run: while runner 129 competes only in the 1500 meters discipline, runner 256 competes in all the distances, recording a different number of observations for each distance. The observed differences are typically associated both with technical choices, but also with different attitudes of the runners, leading them to compete in races of different length, according for example to their endurance and speed abilities. We define the drop-in and drop-out as the runner age in which the first observation in at least one discipline is present and the age after the last performance is observed, respectively. Naturally, the careers of the runners differ both in length and the age they start racing. Based on this definition, runner 129 drops-in at age 19 and drops-out at age 31 and runner 233 drops-in at age 18 and has not dropped-out in the period under examination.

The empirical distribution of runners careers' length, shown in panel (a) of Figure 3, is right skewed, with an average length of 5.04 years. The increased observation at year 14 is due to data censoring. Panel (b) illustrates that around 60% of runners in the sample competes when they are 18 years old, but about 20% of them have already left the competition at the age of 20. A visual exploration that indicates whether these aspects are effectively associated with observed performances are presented in panel (c)

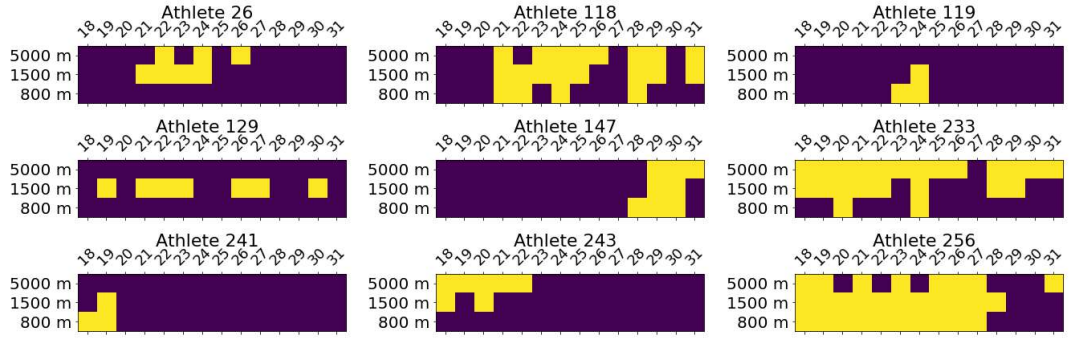

Fig. 2. Missing data patterns for nine runners describing their actual participation in different distances across their ages, shown on top. Yellow squares indicate that the performance is present, blue squares the absence of the observations.

that depicts that performance at drop-in seems to be worse (higher times) if the runners start competing late in life and in panel (d) that shows the distributions of the observed performances, distinguishing between runners with careers longer or shorter than 7 years. In our data runners with longer careers perform typically better than those with shorter careers. The reason for this behavior can be either because runners with unsatisfactory career leave the competitions earlier, or because competing (and, hence, training for it) for a long period is a prerequisite for improving. We refer to the supplementary material for similar plots with other distances.

3. The model

3.1. Clustering longitudinal data with matrix state space model

Let the scalar element $y_{pq,t}$ denote the observation of the performance in discipline p for runner q during year t, for $p = 1, \ldots, P$, $q = 1, \ldots, Q$, and $t = 1, \ldots, T$. To facilitate the exposition, in this Section we assume that the complete set of observations is available in the sense that runners participate in all P distances during the years and that no drop-ins or drop-outs are observed. We assume that runners are divided into G different unobserved groups according to the evolutionary trajectories during their careers. Suppose that runner q belongs to group g, and that their observations over time are described by the following dynamic linear model

$$y_{pq,t} = \mathbf{z}_p^\top \boldsymbol{\alpha}_{p,t}^{(g)} + \varepsilon_{pq,t}, \tag{1}$$

$$\boldsymbol{\alpha}_{p,t+1}^{(g)} = \boldsymbol{T}_p \boldsymbol{\alpha}_{p,t}^{(g)} + \boldsymbol{\xi}_{p,t}^{(g)}, \tag{2}$$

in which $\boldsymbol{\alpha}_{p,1}^{(g)} \sim \mathrm{N}_{F_p}(\widehat{\boldsymbol{\alpha}}_{p,1|0}^{(g)}, \boldsymbol{P}_{p,1|0}^{(g)})$, for $p = 1, \ldots, P$, $t = 1, \ldots, T$, and $\widehat{\boldsymbol{\alpha}}_{p,1|0}^{(g)}$, $\boldsymbol{P}_{p,1|0}^{(g)}$ are fixed mean and variance for the initial state, respectively. The row vector $\mathbf{z}_p^\top$, which has a known structure, links the observation $y_{pq,t}$ to the column vector $\boldsymbol{\alpha}_{p,t}^{(g)}$, which describes the group-specific dynamics of the p–th discipline for all the runners that belong to group g. These dynamics are determined by the state transition equation that describes a

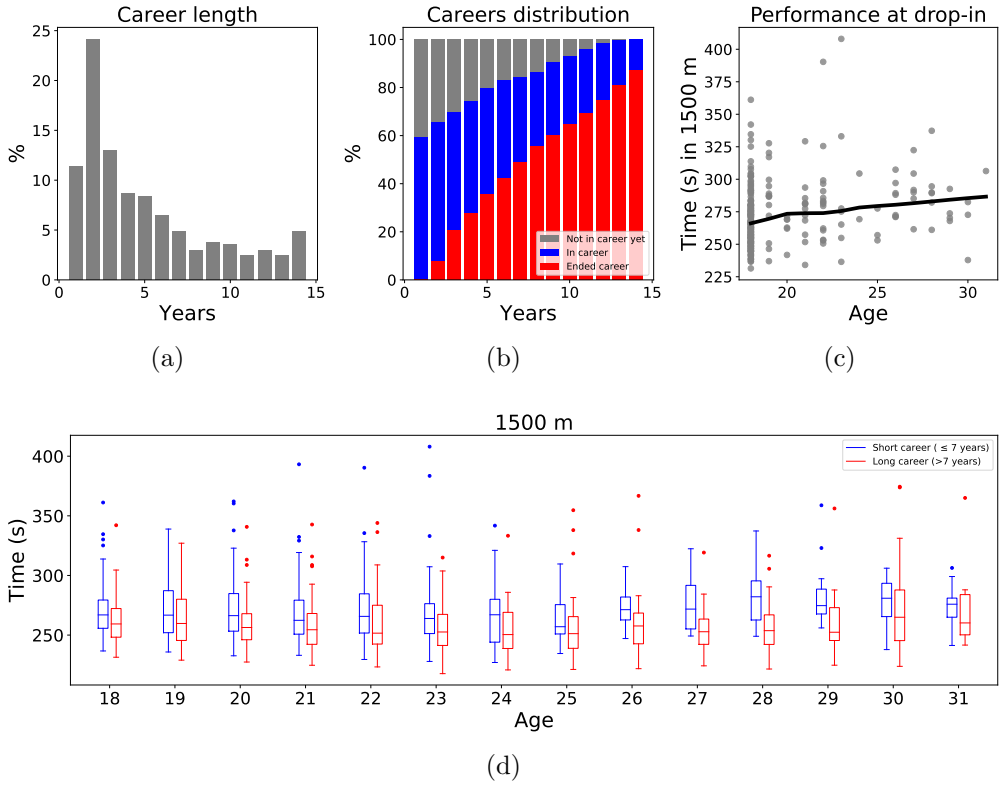

Fig. 3. Panel (a) shows the distribution of runners career length, panel (b) the percentage of runners that are in career (or not) in the different ages considered, and panel (c) the performances at drop-in in 1500 meters. Panel (d) shows the distributions of seasonal best performances over ages in 1500 meters discipline, grouped by the career length of the runners. Blue boxplots represent the performances of runners with a career shorter than 7 years (included), red boxplots the performances of runners with a career longer than 7 years.

first-order autoregressive process with transition matrix $\boldsymbol{T}_p$, which is discipline-specific, known, and shared across all the groups. In this way, for a generic discipline p, we require that the latent states of the different groups are different from each other, but are characterized by the same Markovian dependence induced by $\boldsymbol{T}_p$. Moreover, this dependence is not required to be common across different distances, as $\boldsymbol{T}_p$ may differ from $\boldsymbol{T}_{p'}$ for any $p \neq p'$. The error terms $\varepsilon_{pq,1}, \ldots, \varepsilon_{pq,T}$ are assumed to be Gaussian with zero-mean and variances that can be discipline and subject-specific. They are assumed to be serially independent and independent of both the states $\boldsymbol{\alpha}_{p,1}^{(g)}, \ldots, \boldsymbol{\alpha}_{p,T}^{(g)}$ and the disturbances $\boldsymbol{\xi}_{p,1}^{(g)}, \ldots, \boldsymbol{\xi}_{p,T}^{(g)}$, whose covariance is $\boldsymbol{\Psi}_{pg}$, for $p = 1, \ldots, P$ and $g = 1, \ldots, G$.

Let $\mathbf{y}_{\cdot q,t} = (y_{1q,t}, \ldots, y_{Pq,t})^\top$, $\boldsymbol{\alpha}_t^{(g)} = (\boldsymbol{\alpha}_{1,t}^{(g)\top}, \ldots, \boldsymbol{\alpha}_{P,t}^{(g)\top})^\top$, $\boldsymbol{\varepsilon}_{\cdot q,t} = (\varepsilon_{1q,t}, \ldots, \varepsilon_{Pq,t})^\top$, $\boldsymbol{\xi}_t^{(g)} = (\boldsymbol{\xi}_{1,t}^{(g)\top}, \ldots, \boldsymbol{\xi}_{P,t}^{(g)\top})^\top$, $\widehat{\boldsymbol{\alpha}}_{1|0}^{(g)} = (\widehat{\boldsymbol{\alpha}}_{1,1|0}^{(g)\top}, \ldots, \widehat{\boldsymbol{\alpha}}_{P,1|0}^{(g)\top})^\top$, and $F = \sum_{p=1}^{P} F_p$. Moreover, let $\boldsymbol{T} = \text{blkdiag}(\boldsymbol{T}_1, \ldots, \boldsymbol{T}_P)$, $\boldsymbol{P}_{1|0}^{(g)} = \text{blkdiag}(\boldsymbol{P}_{1,1|0}^{(g)}, \ldots, \boldsymbol{P}_{P,1|0}^{(g)})$, as well as the covariance matrix $\boldsymbol{P}_{1|0} = \text{blkdiag}(\boldsymbol{P}_{1|0}^{(1)}, \ldots, \boldsymbol{P}_{1|0}^{(G)})$ where $\text{blkdiag}(\boldsymbol{X}_a, \ldots, \boldsymbol{X}_z)$ is the block-diagonal operator, creating a block-diagonal matrix with arguments $\boldsymbol{X}_a, \ldots, \boldsymbol{X}_z$ stacked in the main diagonal. Finally, let $\boldsymbol{Z}$ be the $P \times F$ matrix storing, in its p–th row, the row-vector $\mathbf{z}_p^\top$ starting from column $1 + \sum_{j=1}^{p-1} F_j$, and zeros otherwise, and define also the following matrices:

$$\boldsymbol{Y}_t = \begin{bmatrix} \mathbf{y}_{\cdot 1,t} & \ldots & \mathbf{y}_{\cdot Q,t} \end{bmatrix}, \quad \boldsymbol{A}_t = \begin{bmatrix} \boldsymbol{\alpha}_t^{(1)} & \ldots & \boldsymbol{\alpha}_t^{(G)} \end{bmatrix}, \quad \boldsymbol{S}^\top = \begin{bmatrix} \mathbf{s}_{1\cdot}^\top & \ldots & \mathbf{s}_{Q\cdot}^\top \end{bmatrix},$$

$$\boldsymbol{E}_t = \begin{bmatrix} \boldsymbol{\varepsilon}_{\cdot 1,t} & \ldots & \boldsymbol{\varepsilon}_{\cdot Q,t} \end{bmatrix}, \quad \boldsymbol{\Xi}_t = \begin{bmatrix} \boldsymbol{\xi}_t^{(1)} & \ldots & \boldsymbol{\xi}_t^{(G)} \end{bmatrix}, \quad \widehat{\boldsymbol{A}}_{1|0} = \begin{bmatrix} \widehat{\boldsymbol{\alpha}}_{1|0}^{(1)} & \ldots & \widehat{\boldsymbol{\alpha}}_{1|0}^{(G)} \end{bmatrix},$$

where $\mathbf{s}_{q\cdot}^\top = (\mathbb{1}(S_q = 1), \ldots, \mathbb{1}(S_q = G))^\top$ is an allocation vector such that $\mathbb{1}(S_q = g) = 1$ if runner q belongs to group g, and 0 otherwise. Leveraging the previous notation, the model admits a matrix-variate state space representation, in which

$$\boldsymbol{Y}_t = \boldsymbol{Z}\boldsymbol{A}_t\boldsymbol{S}^\top + \boldsymbol{E}_t, \quad \boldsymbol{E}_t \sim \text{MN}_{P,Q}(\mathbf{0}, \boldsymbol{\Sigma}^C \otimes \boldsymbol{\Sigma}^R), \tag{3}$$

$$\boldsymbol{A}_{t+1} = \boldsymbol{T}\boldsymbol{A}_t + \boldsymbol{\Xi}_t, \qquad \boldsymbol{\Xi}_t \sim \text{MN}_{F,G}(\mathbf{0}, \boldsymbol{\Psi}^C \otimes \boldsymbol{\Psi}^R), \tag{4}$$

with $\boldsymbol{A}_1 \sim \text{MN}_{F,G}(\widehat{\boldsymbol{A}}_{1|0}, \boldsymbol{P}_{1|0})$. The matrix $\boldsymbol{S}$ in Equation (3) is a selection matrix, with the role of selecting, for each runner, the columns of states associated with the group the runner belongs to, and silencing the others. The matrices of errors and disturbances are assumed to follow a matrix-variate Normal distribution with covariance matrix decomposed by a Kronecker product (Gupta and Nagar, 2000), which is a typical assumption in models for matrix-variate time series (see, e.g., Wang and West, 2009; Chen et al., 2020). Here, $\boldsymbol{\Sigma}^R$ and $\boldsymbol{\Psi}^R$ are row-covariance matrices with dimensions $P \times P$ and $F \times F$, and measure row-wise dependence of errors and disturbances, respectively. Conversely, the matrices $\boldsymbol{\Sigma}^C$ and $\boldsymbol{\Psi}^C$ are column-covariance matrices with dimensions $Q \times Q$ and $G \times G$ that measure column-wise dependence of errors and disturbances, respectively. Dependent rows or columns are characterized by full covariance matrices, while independent row or columns are characterized by diagonal matrices (Gupta and Nagar, 2000). Thus, the model is general enough to encopass various forms of dependence, while keeping the number of parameters low with respect to alternative full specifications of the covariance matrices.

Although the model is general enough to include a variety of standard state space models (see, e.g. Durbin and Koopman, 2012), in our application we deal with annual-based data describing the careers of different runners, so it seems reasonable to impose the following restrictions: $\boldsymbol{Z} = \boldsymbol{I}_P$, $\boldsymbol{T} = \boldsymbol{I}_P$, $\boldsymbol{\Sigma} = \boldsymbol{I}_Q \otimes \boldsymbol{\Sigma}^R$ and $\boldsymbol{\Psi} = \boldsymbol{I}_G \otimes \boldsymbol{\Psi}^R$. These assumptions imply that the states of different groups describing runners' performance across years are independent of each other and characterized by the same temporal dependence structures, which are those implied by local level models in which the trend of each discipline is a discrete random walk (Durbin and Koopman, 2012). Assuming a priori that performance on discipline p at time $t+1$ is a deviation from that at year t seems reasonable, as a runner is not expected to progress or regress excessively from year to year. Further, setting $\boldsymbol{\Psi}^C$ to be diagonal implies that groups are independent of each other, a typical assumption in clustering. In this framework, the role of the states (i.e. $\boldsymbol{\alpha}_{p,t}^{(g)}$) is to describe various evolution of the performances of the runners. How these states evolve can be considered as the combination of many factors (e.g. individual traits, training, motivation, etc) that lead to unpredictable prior behaviors. Further, conditional on the states and $\boldsymbol{S}$, there is no reason to assume the runners to be dependent between each other. We note also that imposing $\boldsymbol{\Sigma}^C = \boldsymbol{I}_Q$ and $\boldsymbol{\Psi}^C = \boldsymbol{I}_G$ are restrictions even stronger than required, but they help in stabilizing the estimation of the other components given the large number of missing observations present in the data. Removing these restrictions is a delicate aspect in the predictive context in which we fit our model (see Section 5). Indeed, the aim of increasing flexibility struggles with the goal of reducing the variability of the estimates and predictions, due to both the large dimensions of the state space we are considering, but also to the need of adopting a diffuse prior specification and a large number of groups for well capturing the variability of the considered phenomena (see Section 4).

3.2. Missing data inform on clustering structure

The previous section was developed conditional on all data being observed, i.e., when the runners run all P distances during the entire period of observation. However, this is not the case for data that describe the career trajectories of runners, since the lack of data is part of the career itself. To include these factors as informative aspect of runners' careers, we consider two other variables in the model. As first, we consider

$$d_{pq,t} = \begin{cases} 1, & \text{if discipline } p \text{ for runner } q \text{ is observed at time } t, \\ 0, & \text{otherwise,} \end{cases}$$

to describe the presence or absence of the observed discipline for the runners. Then we consider the variable $d^{\star}_{q,t}$ that informs whether the runner q is in career during year t, which is

$$d^{\star}_{q,t} = \begin{cases} 0, & \text{if runner } q \text{ has not started the career before } t \text{ (included),} \\ 1, & \text{if runner } q \text{ is in career during } t, \\ 2, & \text{if runner } q \text{ has finished the career in } t \text{ (included).} \end{cases}$$

The variable $d^{\star}_{q,t}$ is not decreasing in t, and describes the three possible states of runner's career. Moreover, if $d^{\star}_{q,t} \in \{0, 2\}$, then $d_{pq,t} = 0$ with probability 1, for $p = 1, \ldots, P$,

meaning that no distances are observed since the runner is not competing. On the contrary, there might be runners with $d_{pq,t} = 0$, for $p = 1, \ldots, P$, even if $d^{\star}_{q,t} = 1$. This is typical of runners who, despite being in a career, decide not to compete during one specific year, but compete in the following years.

The division into three non-concurrent states allows for the introduction of temporal dynamics within the model of missing data patterns in an easy way. In particular, let $\mathbf{d}^{\star}_q = (d^{\star}_{q,1}, \ldots, d^{\star}_{q,T})$, $\mathbf{d}_{\cdot q,t} = (d_{1q,t}, \ldots, d_{Pq,t})^{\top}$, and $\boldsymbol{D}_q = \begin{bmatrix}\mathbf{d}_{\cdot q,1} & \ldots & \mathbf{d}_{\cdot q,T}\end{bmatrix}$, $\mathcal{D} = \{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_Q\}$, and $\mathcal{D}^{\star} = \{\mathbf{d}^{\star}_1, \ldots, \mathbf{d}^{\star}_Q\}$. First, we make the following independence assumption among different subjects

$$p_{\boldsymbol{\theta}}(\mathcal{D}, \mathcal{D}^{\star} | \boldsymbol{S}) = \prod_{q=1}^{Q} p_{\boldsymbol{\theta}}(\boldsymbol{D}_q, \mathbf{d}^{\star}_q | S_q). \tag{5}$$

As a second step, we let $\mathbf{d}^{\star}_q$ and $\boldsymbol{D}_q$ be dependent on the group $S_q = g$ to which the runner q belongs, and make the following conditional independence assumption

$$\begin{aligned} p_{\boldsymbol{\theta}}(\boldsymbol{D}_q, \mathbf{d}^{\star}_q | S_q) &= p_{\boldsymbol{\theta}}(\boldsymbol{D}_q | \mathbf{d}^{\star}_q, S_q) p_{\boldsymbol{\theta}}(\mathbf{d}^{\star}_q | S_q) \\ &= \prod_{t=1}^{T} \left\{ \prod_{p=1}^{P} p_{\boldsymbol{\theta}}(d_{pq,t} | d^{\star}_{q,t}, S_q) \right\} p_{\boldsymbol{\theta}}(d^{\star}_{q,t} | d^{\star}_{q,t-1}, S_q), \end{aligned} \tag{6}$$

where $p_{\boldsymbol{\theta}}(d^{\star}_{q,1} | d^{\star}_{q,0}, S_q) = \lambda^{\star}_{1g}$ if $d^{\star}_{q,1} = 1$, and $p_{\boldsymbol{\theta}}(d^{\star}_{q,1} | d^{\star}_{q,0}, S_q) = 1 - \lambda^{\star}_{1g}$ if $d^{\star}_{q,1} = 0$. Note that, in Equations (5) and (6), the subscript $\boldsymbol{\theta}$ in $p_{\boldsymbol{\theta}}(A|B)$ denotes conditional dependence of the form $p(A|B, \boldsymbol{\theta})$, for slight abuse of notation, where $\boldsymbol{\theta}$ denotes a set of unknown parameter with finite dimensions (specified later).

In Equation (6) we consider the following assumptions: for runner q, the conditional probabilities at time t of transition from state 0 to state 1 is $p_{\boldsymbol{\theta}}(d^{\star}_{q,t} = 1 | d^{\star}_{q,t-1} = 0, S_q = g) = \lambda^{\star}_{1g}$ and from state 1 to state 2 is $p_{\boldsymbol{\theta}}(d^{\star}_{q,t} = 2 | d^{\star}_{q,t-1} = 1, S_q = g) = \lambda^{\star}_{2g}$. Both the probabilities are group dependent but constant over time. By construction, the transitions from state 1 to state 0 or from state 2 to states 0 or 1 are impossible events. Further, for runner q, the conditional probability at time t of observing a value for the generic discipline p is $p_{\boldsymbol{\theta}}(d_{pq,t} = 1 | d^{\star}_{q,t} = 1, S_q = g) = \delta_{pg}$, which is group-dependent, but fixed over time. Although transitions in the prevalence of the type of discipline done in a long career are possible for some runners (e.g. from shorter to longer distances), these transitions are difficult to detect with annual based data—which are summaries of the entire years—since it is enough to compete in only one race of the considered discipline to be included in the discipline-specific ranking lists. Similarly, the assumption of constant probabilities during years used to describe the presence of missing values does not contemplate the possibility that runners would get seriously injured, and, thus, they would not compete in any discipline for more than a year. Although there is no clear indication in the literature about the average duration and severity of an injury in middle distance runners (see, e.g., van Gent et al., 2007), we assume here that severe injuries (injuries that stop competitions for more than one year) are present only in low proportions, leaving open possible investigations on this aspect in the future.

4. Likelihood and posterior distributions

4.1. Likelihood and prior distributions for the proposed model

Let $\mathcal{Y}$ denote the set of observations as if they were fully observed, $\mathcal{Y}^\star$ the set of variables which are effectively observed, and $\widetilde{\mathcal{Y}}$ the completion of $\mathcal{Y}^\star$, i.e. such that $\mathcal{Y} = \mathcal{Y}^\star \cup \widetilde{\mathcal{Y}}$ and $\mathcal{Y}^\star \cap \widetilde{\mathcal{Y}} = \emptyset$. Let also $\mathcal{A} = \{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_T\}$ be the set storing the latent states of the state space model. In order to derive the posterior distribution of the parameters, we present the likelihood of the observed process first, augmented for both the states $\mathcal{A}$, the missing observations $\widetilde{\mathcal{Y}}$, and $\boldsymbol{S}$.

The augmented likelihood is characterized by the following conditional independence structure

$$p_{\boldsymbol{\theta}}(\mathcal{Y}, \mathcal{D}, \mathcal{D}^\star, \mathcal{A}, \boldsymbol{S}) = p_{\boldsymbol{\theta}}(\mathcal{Y}|\mathcal{D}, \mathcal{A}, \boldsymbol{S}) p_{\boldsymbol{\theta}}(\mathcal{D}|\mathcal{D}^\star, \boldsymbol{S}) p_{\boldsymbol{\theta}}(\mathcal{D}^\star|\boldsymbol{S}) p_{\boldsymbol{\theta}}(\boldsymbol{S}) p_{\boldsymbol{\theta}}(\mathcal{A}). \tag{7}$$

In Equation (7), $p_{\boldsymbol{\theta}}(\mathcal{Y}|\mathcal{D}, \mathcal{D}^\star, \mathcal{A}, \boldsymbol{S}) = p_{\boldsymbol{\theta}}(\mathcal{Y}|\mathcal{D}, \mathcal{A}, \boldsymbol{S})$, and is determined by the measurement Equation (3), for which all observations are assumed to be available, and the prior on $\mathcal{A}$ is implicitly determined by the form of the state equation of the state space in Equation (4). However, only $\mathcal{Y}^\star = \{\mathcal{Y}_1^\star, \ldots, \mathcal{Y}_T^\star\}$ is observed, but $p_{\boldsymbol{\theta}}(\mathcal{Y}|\mathcal{D}, \mathcal{A}, \boldsymbol{S})$ can be obtained by conditioning, noting that

$$p_{\boldsymbol{\theta}}(\mathcal{Y}|\mathcal{D}, \mathcal{A}, \boldsymbol{S}) = p_{\boldsymbol{\theta}}(\mathcal{Y}^\star|\mathcal{D}, \mathcal{A}, \boldsymbol{S}) p_{\boldsymbol{\theta}}(\widetilde{\mathcal{E}}|\mathcal{Y}^\star, \mathcal{D}, \boldsymbol{S}),$$

where $\widetilde{\mathcal{E}}$ stores all those entries in $\mathcal{E} = \{\boldsymbol{E}_1, \ldots, \boldsymbol{E}_T\}$ associated with the missing values. To characterize $\boldsymbol{S}$, we make the following independence assumption

$$p_{\boldsymbol{\theta}}(\boldsymbol{S}) = \prod_{q=1}^{Q} p_{\boldsymbol{\theta}}(\mathbf{s}_{q\cdot}) = \prod_{q=1}^{Q} \prod_{g=1}^{G} \pi_g^{\mathbb{1}(S_q = g)}, \tag{8}$$

where $\boldsymbol{\pi} = (\pi_g, \ldots, \pi_G)$ is such that $\pi_g \in (0, 1)$, for $g = 1, \ldots, G$, and $\sum_{g=1}^{G} \pi_g = 1$.

As concerns model parameters, we assume that the prior distribution of $\boldsymbol{\theta}$ factorizes as follows

$$p(\boldsymbol{\theta}) = p(\widehat{\boldsymbol{A}}_{1|0}) p(\boldsymbol{\Sigma}^R) p(\boldsymbol{\Psi}^R) p(\boldsymbol{\pi}) \prod_{g=1}^{G} \left\{ p(\boldsymbol{\lambda}_g^\star) \prod_{p=1}^{P} p(\delta_{pg}) \right\}, \tag{9}$$

where $\boldsymbol{\lambda}_g^\star = (\lambda_{1g}^\star, \lambda_{2g}^\star)$. We further assume the prior distributions of the probabilities driving missing data patterns to be uninformative Beta distributions, such that $\lambda_{1g}^\star \sim \text{Be}(1,1)$, $\lambda_{2g}^\star \sim \text{Be}(1,1)$, and $\delta_{pg} \sim \text{Be}(1,1)$, for any $p = 1, \ldots, P$ and $g = 1, \ldots, G$. Covariance matrices are assumed to be Inverse Wishart distributions, such that $\boldsymbol{\Sigma}^R \sim \text{IW}_P(P+1, \boldsymbol{I}_P)$ and $\boldsymbol{\Psi}^R \sim \text{IW}_P(P+1, \boldsymbol{I}_P)$. For what concerns the state space model, $\widehat{\boldsymbol{A}}_{1|0}$ is assumed to follow a matrix-variate Normal distributions of mean $\bar{\mathbf{y}}_1 \mathbf{1}_G^\top$ and covariance $\boldsymbol{P}_{1|0} = \boldsymbol{I}_G \otimes \boldsymbol{P}_{1|0}^0$, with $\boldsymbol{P}_{1|0}^0 = \text{diag}(p_1^2, \ldots, p_P^2)$, where $\bar{\mathbf{y}}_1$ is the vector storing sample average of observed distances at first time instant and p_p^2 is twice the sample variance of the p–th observed discipline at the first time instant. It is interesting to observe, however, that the number of parameters depends on the number of groups G, which is fixed. We consider an overfitting finite mixture specification of the

model (see, e.g., Malsiner-Walli et al., 2016, 2017), in which G is set to be large, and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G) \sim \text{Dir}_G(e_1, \ldots, e_G)$ with hyper-parameters $e_1 = \ldots = e_G = 1/G$. The prior on the mixture weights favours emptying the extra components, leaving complete symmetry between the different components included in the model. This assumption implies that, during the estimation procedure, the number of filled components may be lower than G, leading to the classical distinction between the number of clusters G^+ (i.e. the number of filled components) and the number of components G included in the model, with $G^+ \leq G$ (see, Malsiner-Walli et al., 2016, 2017; Frühwirth-Schnatter et al., 2021, for an extensive discussion on the topic). Under this prior specification, it is possible to derive a Gibbs sampling algorithm that involves all full conditionals that are conditionally conjugate, see the supplementary material for details. Note that the algorithm allows to explore the $G!$ symmetric modes of the posterior distribution by including a step in which we randomly permute the labels of $\boldsymbol{S}$ (see, e.g. Fruhwirth-Schnatter, 2001; Malsiner-Walli et al., 2016). Draws of the states $\mathcal{A}$ are obtained using the simulation smoothing technique by Durbin and Koopman (2002), applied to a reduced form of the model derived using the reduction by transformation technique (see Jungbacker and Koopman, 2008).

4.2. Posterior distribution, alternative specifications and interpretability

The goal of our inferential procedure is to derive quantities of interest (e.g. predictive distributions) from a sample of the posterior distribution

$$p^{\text{C}}(\boldsymbol{\theta}, \mathcal{A}, \boldsymbol{S}, \widetilde{\mathcal{E}} | \mathcal{Y}^\star, \mathcal{D}, \mathcal{D}^\star) \propto p(\boldsymbol{\theta}) p_{\boldsymbol{\theta}}(\mathcal{Y} | \mathcal{D}, \mathcal{A}, \boldsymbol{S}) p_{\boldsymbol{\theta}}(\mathcal{D} | \mathcal{D}^\star, \boldsymbol{S}) p_{\boldsymbol{\theta}}(\mathcal{D}^\star | \boldsymbol{S}) p_{\boldsymbol{\theta}}(\boldsymbol{S}) p_{\boldsymbol{\theta}}(\mathcal{A}). \quad (10)$$

A sample from the posterior distribution can be obtained using a Gibbs sampling scheme, as discussed in the supplementary material. In Equation (10), the superscript $^{\text{C}}$ indicates that the considered posterior is referring to the *complete* model (or model 1), because it assumes that both attitudes and history matter. By restricting the complete model in Equation (10) it is possible to obtain a set of alternative reduced specifications, in which different missing data pattern schemes have different influence for clustering. More specifically, a set of alternative specifications can be derived by dropping the dependence on the selection matrix $\boldsymbol{S}$ in some parts of the model. We consider the following set of alternative specifications:

Model 2: Missing data do not matter:

$$p^{\text{NM}}(\boldsymbol{\theta}, \mathcal{A}, \boldsymbol{S}, \widetilde{\mathcal{E}} | \mathcal{Y}^\star, \mathcal{D}, \mathcal{D}^\star) \propto p(\boldsymbol{\theta}) p_{\boldsymbol{\theta}}(\mathcal{Y} | \mathcal{D}, \mathcal{A}, \boldsymbol{S}) p_{\boldsymbol{\theta}}(\boldsymbol{S}) p_{\boldsymbol{\theta}}(\mathcal{A}); \quad (11)$$

Model 3: Only attitude matters:

$$p^{\text{A}}(\boldsymbol{\theta}, \mathcal{A}, \boldsymbol{S}, \widetilde{\mathcal{E}} | \mathcal{Y}^\star, \mathcal{D}, \mathcal{D}^\star) \propto p(\boldsymbol{\theta}) p_{\boldsymbol{\theta}}(\mathcal{Y} | \mathcal{D}, \mathcal{A}, \boldsymbol{S}) p_{\boldsymbol{\theta}}(\mathcal{D} | \mathcal{D}^\star, \boldsymbol{S}) p_{\boldsymbol{\theta}}(\boldsymbol{S}) p_{\boldsymbol{\theta}}(\mathcal{A}); \quad (12)$$

Model 4: Only history matters:

$$p^{\text{H}}(\boldsymbol{\theta}, \mathcal{A}, \boldsymbol{S}, \widetilde{\mathcal{E}} | \mathcal{Y}^\star, \mathcal{D}, \mathcal{D}^\star) \propto p(\boldsymbol{\theta}) p_{\boldsymbol{\theta}}(\mathcal{Y} | \mathcal{D}, \mathcal{A}, \boldsymbol{S}) p_{\boldsymbol{\theta}}(\mathcal{D}^\star | \boldsymbol{S}) p_{\boldsymbol{\theta}}(\boldsymbol{S}) p_{\boldsymbol{\theta}}(\mathcal{A}). \quad (13)$$

Alternative model specifications in Equations (11)–(13) have meaningful structural interpretations, if compared with the complete model in Equation (10). In model 2 both $p_{\boldsymbol{\theta}}(\mathcal{D}|\mathcal{D}^{\star}, \boldsymbol{S}) = p(\mathcal{D}|\mathcal{D}^{\star})$ and $p_{\boldsymbol{\theta}}(\mathcal{D}^{\star}|\boldsymbol{S}) = p(\mathcal{D}^{\star})$, meaning that neither attitude nor history matter for clustering, and therefore are not correlated with the evolution of performances. In this case, missing data are still considered in the estimation procedure for obtaining $\widetilde{\mathcal{E}}$ and other elements related to the set of completed observations $\mathcal{Y}$ (e.g., $\boldsymbol{\Sigma}^{R}$), but the part of likelihood describing the evolution of $\mathcal{D}$ and $\mathcal{D}^{\star}$ is no longer dependent on $\boldsymbol{\theta}$ and $\mathbf{S}$. Alternatively, in model 3 the attitude matters but history does not. This is obtained by requiring $p_{\boldsymbol{\theta}}(\mathcal{D}^{\star}|\boldsymbol{S}) = p(\mathcal{D}^{\star})$. Note that the dependence of $p_{\boldsymbol{\theta}}(\mathcal{D}|\mathcal{D}^{\star}, \boldsymbol{S})$ on $\mathcal{D}^{\star}$ is preserved, an important aspect because it is involved in the estimation of the parameters δ_{pg} related to runners' attitudes. Finally, in model 4 $p_{\boldsymbol{\theta}}(\mathcal{D}|\mathcal{D}^{\star}, \boldsymbol{S}) = p(\mathcal{D}|\mathcal{D}^{\star})$, leading to a model in which runners' attitudes do not matter for the evolution of the performances. For simplicity, we do not distinguish $p(\boldsymbol{\theta})$ in the different models letting the elements included in $\boldsymbol{\theta}$ under different model specifications differ, based on the single case being considered (e.g. $\boldsymbol{\theta} = \{\hat{\boldsymbol{A}}_{1|0}, \boldsymbol{\Sigma}^{R}, \boldsymbol{\Psi}^{R}, \boldsymbol{\pi}\}$ in model 2). We can provide an interpretation to our model construction from a two-step Bayesian learning perspective. First, different structured priors on $\boldsymbol{S}$ are obtained, which simply reflect different clustering structures that we believe to be relevant for clustering the performances. For example, for the complete model, this structured prior is

$$p^{\mathrm{C}}(\boldsymbol{\theta}, \boldsymbol{S}|\mathcal{D}, \mathcal{D}^{\star}) \propto p(\boldsymbol{\theta}) p_{\boldsymbol{\theta}}(\mathcal{D}|\mathcal{D}^{\star}, \boldsymbol{S}) p_{\boldsymbol{\theta}}(\mathcal{D}^{\star}|\boldsymbol{S}) p_{\boldsymbol{\theta}}(\boldsymbol{S}).$$

Second, the knowledge on the clustering structure is updated by considering the likelihood related to the performances, which is $p_{\boldsymbol{\theta}}(\mathcal{Y}|\mathcal{D}, \mathcal{A}, \boldsymbol{S}) p_{\boldsymbol{\theta}}(\mathcal{A})$. Comparing different models allows to determine which of the four alternative specifications is most credible in explaining the observed variability in runners' performances. Comparisons between models are obtained by assessing the models' abilities to predict the performance of out-of-sample runners, as explained in Section 5.2. We note here that, in our model construction, performances depend directly on missing data patterns, as the term $p_{\boldsymbol{\theta}}(\mathcal{Y}|\mathcal{D}, \mathcal{A}, \boldsymbol{S})$ is considered in Equation (10), and that the opposite direction, in which the performances have a direct influence on the fact that runners remain (or not) in the sample, is not considered. In general, it is easy to imagine that runners with unsatisfactory careers are more likely to leave competitions. Then, the drop-out probability $p_{\boldsymbol{\theta}}(d^{\star}_{q,t} = 2|d^{\star}_{q,t-1} = 1, S_q)$ in Equation (6) should be performance dependent, e.g. $p_{\boldsymbol{\theta}}(d^{\star}_{q,t} = 2|d^{\star}_{q,t-1} = 1, S_q, \mathbf{y}_{\cdot q,t-1})$. Our conjecture is that, conditional on the latent allocation S_q, the decision of competing is the predominant factor for improving the performances, independently of the previous ones, and this is motivated by both physiological and psychological considerations for which defining goals and training for achieving them can help in improving performances as well. Furthermore, although the priors on clustering structure do not depend on the performances, both the posteriors and the posterior probabilities of drop-out do depend. The choice of considering a sufficiently large number of groups G allows to account also for variability present in the data that may be effectively caused by cases in which performances have a direct impact on the choice of leaving the competitions.

Finally, it is interesting to highlight that similar information can be obtained by using any kind of regression model in which $\mathcal{D}$ and $\mathcal{D}^{\star}$ are treated as covariates for explaining the evolution of the performance over time. We discuss these alternative linear models in

Section 6 and Section S.4 of the supplementary material. Our approach differs from these alternatives as it accounts for and quantifies the uncertainty related to $d_{pq,t}$ and $d^{\star}_{q,t}$ and endogenizes the fact that both $y_{pq,t}$, $d_{pq,t}$, and $d^{\star}_{q,t}$ are measures (with errors) of runners abilities, attitudes and histories, respectively, three aspects of runners' careers that are captured by the latent matrix $\mathbf{S}$. The role of the latent matrix $\mathbf{S}$ is indeed to describe the heterogeneity present in the data and to account for the potential correlation between performances and missing data patterns. As a consequence, the latent cluster allocations could represent a meaningful information from a sports science point of view. Despite this, in this paper we do not approach the analysis from the perspective of clustering of runners' performances, preferring a prediction-oriented focus that allows us to highlight if $\mathcal{D}$ and $\mathcal{D}^{\star}$ provide valuable information in determining $\mathcal{Y}^{\star}$. Indeed, when facing the clustering problem with these data, various issues arise, from both a technical and an interpretative point of view. First, it is well-known that model based clustering with symmetric priors suffers from the problem of label switching due to the unidentifiability of the components related to the perfect symmetry of the posterior distribution. Second, in our model the clustering represents a combination of various aspects of $\mathcal{Y}$, $\mathcal{D}$, and $\mathcal{D}^{\star}$, which are mixed together and whose contribution is hard to disentangle to recognize well separated and meaningful clusters. This aspect is even exacerbated by both the fact that we are considering a large number of G, of which $G-G^{+}$ are empty, and by the large state space involving many parameters. Although significant works on relabeling techniques have been proposed in recent years (see, e.g. Wade and Ghahramani, 2018; Malsiner-Walli et al., 2017; Egidi et al., 2018), our high-dimensional model does not leave space for guarantees about interpretable clusters. Thus, our inferential procedure in the next Section is solely based on predictive distributions.

5. Posterior predictive inference and out of sample predictions

5.1. Predictive inference

Let $\mathcal{Y}^{\star}_{[n]}$, $\boldsymbol{D}_{[n]}$, and $\mathbf{d}^{\star}_{[n]}$ denote the random variables describing, respectively, the performances, the participation in the distances and the history of a new runner n, not included in the sample. Let also $\boldsymbol{\Theta} = (\mathcal{A}, \boldsymbol{\theta})$ be the unknown elements which are shared across different runners, characterized by a posterior distribution $p^{j}(\boldsymbol{\Theta}|\mathcal{Y}^{\star}, \mathcal{D}, \mathcal{D}^{\star})$ that can be obtained by means of an MCMC algorithm under model j. We consider the following predictive density:

$$p^{j}(\mathcal{Y}^{\star}_{[n]}|\boldsymbol{D}_{[n]}, \mathbf{d}^{\star}_{[n]}, \mathcal{Y}^{\star}, \mathcal{D}, \mathcal{D}^{\star}) = \int p^{j}(\boldsymbol{\Theta}|\mathcal{Y}^{\star}, \mathcal{D}, \mathcal{D}^{\star}) p^{j}_{\boldsymbol{\Theta}}(\mathcal{Y}^{\star}_{[n]}|\boldsymbol{D}_{[n]}, \mathbf{d}^{\star}_{[n]}) \mathrm{d}\boldsymbol{\Theta}, \tag{14}$$

for $j \in \{\mathrm{C}, \mathrm{NM}, \mathrm{A}, \mathrm{H}\}$ which can be obtained using Monte Carlo estimation. In Equation (14), missing data patterns are supposed to be known and are treated as control variables that potentially have an influence on the predicted performances $\mathcal{Y}^{\star}_{[n]}$. While the posterior distribution $p^{j}(\boldsymbol{\Theta}|\mathcal{Y}^{\star}, \mathcal{D}, \mathcal{D}^{\star})$ is an output of the MCMC algorithms (see supplementary material), the likelihood $p^{j}_{\boldsymbol{\Theta}}(\mathcal{Y}^{\star}_{[n]}|\boldsymbol{D}_{[n]}, \mathbf{d}^{\star}_{[n]})$ for the new individual n can

be obtained by marginalizing over groups as follows:

$$p^j_{\boldsymbol{\Theta}}(\mathcal{Y}^\star_{[n]}|\boldsymbol{D}_{[n]}, \mathbf{d}^\star_{[n]}) = \sum_{g=1}^{G} p^j_{\boldsymbol{\Theta}}(S_{[n]} = g|\boldsymbol{D}_{[n]}, \mathbf{d}^\star_{[n]}) p^j_{\boldsymbol{\Theta}}(\mathcal{Y}^\star_{[n]}|\boldsymbol{D}_{[n]}, \mathbf{d}^\star_{[n]}, S_{[n]} = g).$$

In the equation, the cluster allocation follows a multinomial distribution characterized by weights

$$p^j_{\boldsymbol{\Theta}}(S_{[n]} = g|\boldsymbol{D}_{[n]}, \mathbf{d}^\star_{[n]}) \propto p^j_{\boldsymbol{\Theta}}(S_{[n]} = g) p^j_{\boldsymbol{\Theta}}(\boldsymbol{D}_{[n]}, \mathbf{d}^\star_{[n]}|S_{[n]} = g),$$

that depend on the likelihood related to cluster allocation $p_{\boldsymbol{\Theta}}(S_{[n]} = g)$ but also on the observed missing data patterns, that weigh differently the cluster allocation by means of $p_{\boldsymbol{\Theta}}(\boldsymbol{D}_{[n]}, \mathbf{d}^\star_{[n]}|S_{[n]} = g)$. We note here that the predictive distribution is invariant with respect to permutations of the labels, so that the label switching that usually represents a relevant issue in MCMC-based parameters estimation, in this context, guarantees the exploration of the multiple modes of the posterior distribution. The Monte Carlo procedure for obtaining the predictive distribution is reported in Section S.2 of the supplementary material. Next section details the use of the predictive distribution developed here for evaluating informativeness of missing data patterns under different model specifications.

5.2. Informativeness of missing data: out of sample comparison of alternative specifications

Let $\mathcal{Y}_{[1:N]}$ be a test set, storing the performances in different distances and years of N runners not included in the training sample for model estimation. Let also $y_{p[n],t}$ be the generic scalar element denoting the performance of runner n in discipline p during year t. The quantity

$$\mathbb{Q}^j(y_{p[n],t}) = p^j(y_{p[n],t}|\boldsymbol{D}_{[n]}, \mathbf{d}^\star_{[n]}, \mathcal{Y}^\star, \mathcal{D}, \mathcal{D}^\star),$$

represents the predictive distribution obtained under model j, conditional on missing data patterns $(\boldsymbol{D}_{[n]}, \mathbf{d}^\star_{[n]})$ and the set of available information $(\mathcal{Y}^\star, \mathcal{D}, \mathcal{D}^\star)$. Let also $\widetilde{\mathbb{Q}}^j(y_{p[n],t})$ denote an approximation of $\mathbb{Q}^j(y_{p[n],t})$, given by a set of B particles $\{y^1_{p[n],t}, \ldots, y^B_{p[n],t}\}$. It is possible to use the samples $\widetilde{\mathbb{Q}}^j(y_{p[n],t})$ to evaluate and compare the ability of our proposals in predicting the performances over different distances, for fixed missing data patterns described by $\boldsymbol{D}_{[n]}$, and $\mathbf{d}^\star_{[n]}$. We base our evaluations on the empirical counterpart of the *continuous ranked probability score* (CRPS) and the *interval score* (see, Gneiting and Raftery, 2007; Krüger et al., 2021), preferring models that minimize these scoring rules and that provide adequate prediction interval estimates in term of coverage and interval width. The CRPS is defined as

$$S_1(\mathbb{Q}^j(y_{p[n],t})) = \int_{-\infty}^{\infty} \left\{\mathbb{Q}^j(y_{p[n],t}) - \mathbb{1}(y_{p[n],t} \leq z)\right\}^2 \mathrm{d}z, \tag{15}$$

and the interval score is defined as

$$S_2(\mathbb{Q}^j(y_{p[n],t})) = (u^j_\alpha - l^j_\alpha) + \frac{2}{\alpha}\left\{(l^j_\alpha - y_{p[n],t})\mathbb{1}(y_{p[n],t} < l^j_\alpha) + (y_{p[n],t} - u^j_\alpha)\mathbb{1}(y_{p[n],t} > u^j_\alpha)\right\},$$

where l_α^j and u_α^j are the $\alpha/2$ and $1-\alpha/2$ quantiles for the distribution $\mathbb{Q}^j(y_{p[n],t})$, respectively, and $\alpha \in (0, 1/2)$ is a fixed tolerance. The interval score rewards narrow prediction intervals, and penalizes prediction intervals that do not include the observations. Details on these scores and their computation using sample from the predictive distribution are reported in Krüger et al. (2021) (see, e.g., Equation (9) of their paper). It is relevant to highlight, however, that these scores are scale sensitive, and the scale of the predictions might depend both on the discipline p, the age t, as well as on the specific missing data pattern of runner n we are considering. For this reason, we propose to evaluate the predictions of a reference model j and the prediction of an alternative model j' by means of the following score

$$S_s^{jj'}(\mathcal{Y}_{[1:N]}) = \frac{1}{|\mathcal{Y}_{[1:N]}|} \sum_{n=1}^{N} \sum_{y_{p[n],t} \in \mathcal{Y}_{[n]}} \mathbb{1}(S_s(\mathbb{Q}^j(y_{p[n],t}) < S_s(\mathbb{Q}^{j'}(y_{p[n],t})), \tag{16}$$

for $s \in \{1, 2\}$. In Equation (16), $|\mathcal{Y}_{[1:N]}|$ denotes the number of distinct observations present in the test set $\mathcal{Y}_{[1:N]}$. The scores $S_s^{jj'}(\mathcal{Y}_{[1:N]})$ range in $(0, 1)$, and suggest that model j is overall better than model j' if $S_s^{jj'}(\mathcal{Y}_{[1:N]}) > 0.5$.

6. Real data analysis

6.1. Informativeness of missing data

The models were estimated randomly splitting the runners into a training set, composed of 70% of runners, and a test set with the remaining 30%. Samples from the posterior distributions were obtained on the training set, using the last 2000 of 10000 iterations of the Gibbs sampling for each model (see Section S.3 of the supplementary material for details on chains convergence). The number of components G was fixed to 50. Samples from the predictive distributions in Equation (14) were obtained, conditional on knowing the missing data patterns of runners in the test set (i.e. $\boldsymbol{D}_{[n]}$ and $\mathbf{d}^\star_{[n]}$, for $n = 1, \ldots, N$).

The comparison over different models' predictions was done both graphically and using the scores described in Section 5.2. Results are summarized in Figure 4 and Table 1. In the figures, the real data are represented by black solid lines, while the colored lines delimit 95% quantile-based prediction bands. A model is preferred if: (a) the real data lies within the prediction bands, (b) the predictions bands are narrower. By looking at the results we can claim that model 1 (both attitude and history are considered as informative) and model 3 (only attitude is considered as informative) provide better results in terms of band widths, while including within the bands the real data. As we note, different missing data patterns are represented in the figures. We note the tendency of model 1 and model 3 of yielding lower upper limits of the band, an aspect that is even more evident for runners that participated to many competitions in different years (e.g. runners 100, 63, and 56). Knowing that one runner has competed for many years reduces the uncertainty in the predictions by reducing the probability of observing bad performances. This highlights the effective presence of an association between the abilities of the runners in recording better performances with long histories and their participation in many competitions during the years. Further, by comparing runner 89 with runner 56, for example, it is possible to grasp how entering later in competitions

Table 1. Comparison between models given by $S_s^{jj'}$ with the different scores. In this table, model j (row) is compared to model j' (column) with respect the two scores. Above the diagonal we report the scores for CRPS, below the diagonal we report the scores for IS, computed with $\alpha = 0.05$. A value above 0.5 indicates preference of model j with respect to model j'. Remember that $S_s^{jj'} = 1 - S_s^{j'j}$, for any score considered.

Model	Complete (1)	No missing (2)	Attitude (3)	History (4)
Complete (1)	–	0.574	0.537	0.520
No missing (2)	0.178	–	0.402	0.395
Attitude (3)	0.364	0.757	–	0.511
History (4)	0.233	0.680	0.306	–

leads to more uncertainty in the performance predictions, increasing also the probability of observing worse performances. While the effect of knowing missing data patterns is clear when we compare the upper limits of the predictions bands, the distances between the lower limits produced by the four models appear to be limited and less pronounced. This aspect is interesting because it points out that there are runners that, despite being characterized by short histories, are still able to perform satisfactorily when compared with those with longer histories. These reasonings are conditional on the graphs shown, which are, of course, a selection of the runners in the test set. The complete set of plots is reported in the supplementary material, showing a large number of runners with different missing data patterns and, as a consequence, different behaviors of the bands.

The predictive scores computed with data of the test set suggest that the complete model is better than the others considering both the scores. However, the interval score suggests this aspect more markedly, highlighting the ability in outperforming the model that does not treat missing values as informative for around 80% of the observations present in the test set. For both the scores, the complete model is better than model 3 (only attitude), which is better than model 4 (only history), which is itself better than model 2 (uninformative missing data patterns). These results highlight how attitude and history (encopassed in the term missing data) seem to be effectively related to performances, giving support to the hypothesis that considering these aspects in the analysis of runners careers is definitely relevant and that, in this context, missing data have to be considered as informative. Note that both the estimates of predictive scores are based on 741 scalar observations which are characterized by different levels of dependence, so that a proper evaluation of uncertainty of these estimates is difficult. In the supplementary material (Sections S.4 and S.5), a three-fold cross validation scheme shows that results seem to be stable. Moreover, we also show how our multivariate modelling approach outperforms simpler models (such as linear regression) and discuss the goodness of fit of the model. Among the limits of the complete model, we highlight the tendency of the model to over-estimate the left tail quantiles (< 0.10) and underestimate the right tail quantiles (> 0.90). In the practical context of interest, although this implies a slight lowering of the marginal predictive coverage, we do not consider this to be an issue if the interest concentrates on the "average" runners, rather than the top performers and high level runners. Note also that our model comparison procedure based on the Interval Score already considers the width of the prediction bands in the calculation, penalizing lower coverage.

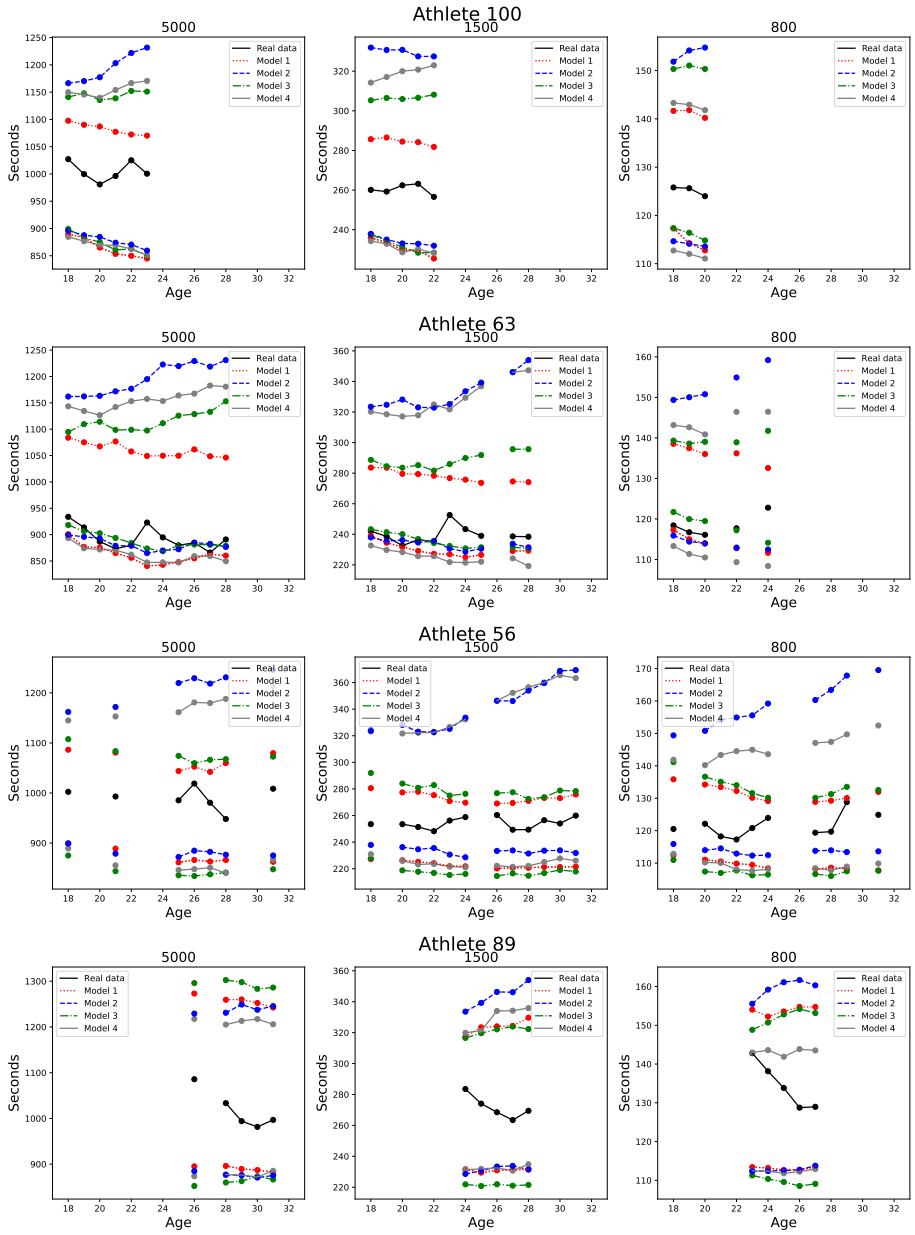

Fig. 4. Quantile-based 95% prediction intervals for the observed distances obtained conditional on knowing the missing data patterns of runners included in the test set. The red-dotted lines represent the intervals for the model that treats missing data as informative, while the blue-dashed lines represent the respective intervals for the model that does not treat them as such. The black lines represent the observed performance of the runners.

6.2. Application

We illustrate how the complete model can be meaningful for sports scientists and coaches, answering to two specific questions, based on samples of size 2000 of the predictive distributions illustrated in Section 5 (see Section S.3 of the supplementary material for details on using multiple chains). The first question is: how do late entry into competitions and early exit from competitions are related to performances? To answer this question, we consider the conditional predictive distribution in Equation (14), in which we vary the age at which the runner enters or exits from the sample, letting the runner participate in both 800 and 1500 discipline for all the years of his career. Comparing the different distributions of the performance allows to catch how the uncertainty related to the predicted performance changes, according to the different histories considered. Figure 5 shows the results of our procedure for the 1500 and 800 meters distances. Solid central lines represent the median of the predictive distributions over different ages. Quantile-based 95% prediction bands are on the contrary represented with different dashed lines, that denote the respective lower an the upper limits. For what concerns drop-in, we note from the left panels that late entry into competition is associated with worse median performances over the years, with lower limits of the predictive confidence bands that worsen only for runners that entry into competition at ages 26 and 30. The upper limits, on the contrary, seem to rise with later drop-in. Based on these results, we can say that for the ideal runners we are considering, later entry in career can still permit to reach good levels, but it is much more likely that their performances will be worse with respect to runners with a longer career (earlier drop-in). A similar reasoning applies for drop-out, in the right panels. The median of the predictive distributions seems indeed to be worse for runners that drop-out earlier in their life, with the upper limits that seem to show more variation with respect to the lower ones. In this case, it is not unlikely to expect runners that drop-out earlier, although their performances seem good, but it is more unlikely that runners that compete for more years record worse results. Similar results for different scenarios can be obtained with an analogous approach and are reported in the supplementary material.

The second question is: how does competing in different disciplines impact the performances? In this case, we consider runners with a complete observed career, and that compete, every year, in the distances with different strategy: the first runner competes only in the 1500 meters discipline; the second runner competes in 1500m and 5000m distances; the third one competes in 1500m and 800m distances; the fourth runner competes in all distances. Results are shown in Figure 6, in terms of their respective predictive distributions in 1500 meters discipline. By comparing the results, we see clearly how, in the training sample, worse performances appear to be associated with the choice of competing only in one discipline over the years. On the contrary, being a runner that competes in more than one discipline seems to be associated with better performances, with differences between the respective predictive distributions which are less evident. More specifically, the greater differences can be seen comparing runners that compete in two distances with the one that competes in all three. Indeed, both the limits of the bands and the median of the predictive distribution appear to be shifted up for the runner that competes in all distances, implying a slightly worse performance for these kind of runners. Based on our results, competing in all distances seems not to be an optimal

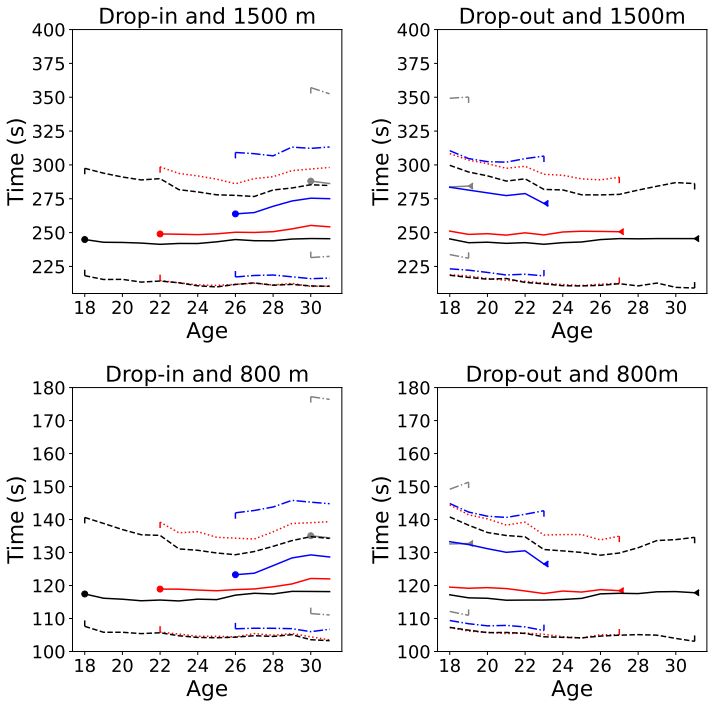

Fig. 5. The association between drop-in and drop-out with performances in the 1500 meters discipline, for an ideal runner that competes regularly in 1500 (top) and 800 meters (bottom). On the left, the runner drops-in at ages 18, 22, 26, 30, respectively. On the right, the runner drops-out at age 20, 24, 28 or after 32 years old. Central solid line indicate the median of the predictive distributions. External lines indicate the 95% confidence bands based on symmetric quantiles of the predictive distributions.

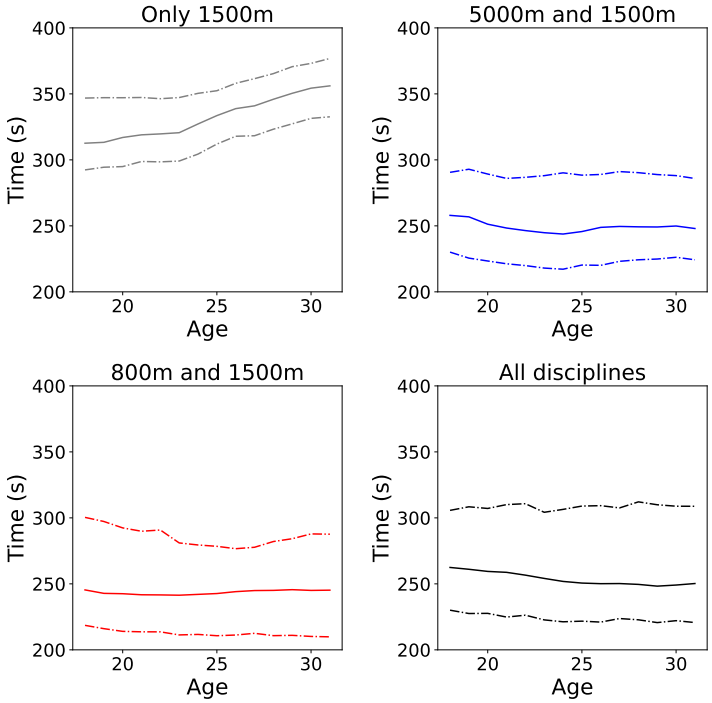

Fig. 6. Predictive distributions of performances in the 1500 meters discipline, obtained for four different runners with different ways of participation in other distances during the years.

choice to achieve better results in the 1500 meters discipline. On the contrary, runners that specialize in 1500 and 5000 meters or in 1500 and 800 meters distances seem to have better performances overall, especially for the latter type of runners. While the answers to the first question can generally be obtained by considering only univariate models, the second one can be addressed only by fitting a multivariate model which allows to understand how competing in different disciplines interacts with performances. Further analyses with other distances are in the supplementary material.

7. Conclusions

This paper investigates whether prediction of the runners' performance is improved by an accurate assessment of the presence of missing data patterns. Our analysis has provided strong evidence that for our data, missing data patterns are informative in predicting performances and they constitute a structural part of the signal explaining the observed variability of the runners' performances.

The statistical analysis took place via a matrix-variate state space model, in which the observed trends were clustered by employing a selection matrix involved in the measurement equation, and by storing the unknown cluster allocations of the runners. To include observed missing data patterns as informative on the clustering structure, two distinct processes were included in the model. The first included the runner's history

as potentially informative by considering when a runner starts or stops competing. The second aimed to include as potentially informative the runner's attitude by considering in which distances the runner mostly participates.

Our results based on out of sample comparisons suggest that it is important to consider both these processes when describing runners' performances whereas considering only attitude is better than considering only history. There is evidence for a deterioration in performance when one runner starts competing later or finishes earlier and for improvement for runners that compete regularly in more distances compared to runners that compete only in one distance. Finally, competing in all three distances does not seem to be associated with better performances with respect to competing in two adjacent distances, such as 800 and 1500 meters or 1500 and 5000 meters.

Our key message is to illustrate the usefulness of considering missing data when describing runners' performances. Our modelling framework can be straightforwardly applied to other athletic disciplines, such as sprinting and hurdling, or even to multidisciplinary competitions such as heptathlon and decathlon. It would be also interesting to investigate whether these findings differ in female runners or in countries with possibly different coaching methodologies.

Data availability statement

The authors confirm that the data supporting the findings of this study are available as supplementary material. Data are openly accessible from the website of Italian athletics federation (FIDAL): `www.fidal.it`.

Acknowledgment

This research was supported by funding from the University of Padova Research Grant 2019-2020, under grant agreement BIRD203991.

References

Bartolucci, F. and Farcomeni, A. (2015) A discrete time event-history approach to informative drop-out in mixed latent markov models with covariates. *Biometrics*, **71**, 80–89.

Bartolucci, F. and Murphy, T. B. (2015) A finite mixture latent trajectory model for modeling ultrarunners' behavior in a 24-hour race. *Journal of Quantitative Analysis in Sports*, **11**, 193–203.

Boccia, G., Moisè, P., Franceschi, A., Trova, F., Panero, D., La Torre, A., Rainoldi, A., Schena, F. and Cardinale, M. (2017) Career performance trajectories in track and field jumping events from youth to senior success: The importance of learning and development. *PLOS ONE*, **12**, 1–15. Publisher: Public Library of Science.

Bussmann, G. (1999) How to prevent "dropout" in competitive sport. *IAAF New Studies in Athletics*, **14**, S. 23–29.

Cappé, O., Moulines, E. and Rydén, T. (2005) *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York.

Cece, V., Guillet-Descas, E., Nicaise, V., Lienhart, N. and Martinent, G. (2019) Longitudinal trajectories of emotions among young athletes involving in intense training centres: Do emotional intelligence and emotional regulation matter? *Psychology of Sport and Exercise*, **43**, 128–136.

Chen, E. Y., Tsay, R. S. and Chen, R. (2020) Constrained factor models for high-dimensional matrix-variate time series. *Journal of the American Statistical Association*, **115**, 775–793.

Chen, R., Xiao, H. and Yang, D. (2021) Autoregressive models for matrix-valued time series. *Journal of Econometrics*, **222**, 539–560.

Choukroun, D., Weiss, H., Bar-Itzhack, I. Y. and Oshman, Y. (2006) Kalman filtering for matrix estimation. *IEEE Transactions on Aerospace and Electronic Systems*, **42**, 147–159.

Durbin, J. and Koopman, S. J. (2002) A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, **89**, 603–615.

— (2012) *Time series analysis by state space methods*, vol. 38 of *Oxford Statistical Science Series*. Oxford University Press, Oxford, second edn.

Egidi, L., Pappada, R., Pauli, F. and Torelli, N. (2018) Relabelling in bayesian mixture models by pivotal units. *Statistics and Computing*, **28**, 957–969.

Fruhwirth-Schnatter, S. (2001) Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, **96**, 194–209.

Frühwirth-Schnatter, S. (2006) *Finite mixture and Markov switching models*. Springer Series in Statistics. Springer, New York.

Frühwirth-Schnatter, S. (2011) Panel data analysis: a survey on model-based clustering of time series. *Advances in Data Analysis and Classification*, **5**, 251–280.

Frühwirth-Schnatter, S., Malsiner-Walli, G. and Grün, B. (2021) Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Analysis*, **16**, 1279 – 1307.

van Gent, R. N., Siem, D., van Middelkoop, M., van Os, A. G., Bierma-Zeinstra, S. M. A. and Koes, B. W. (2007) Incidence and determinants of lower extremity running injuries in long distance runners: a systematic review. *British Journal of Sports Medicine*, **41**, 469–480.

Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378.

Gupta, A. K. and Nagar, D. K. (2000) *Matrix variate distributions*. Chapman and Hall/CRC.

Haugen, T. A., Solberg, P. A., Foster, C., Morán-Navarro, R., Breitschädel, F. and Hopkins, W. G. (2018) Peak age and performance progression in world-class track-and-field athletes. *International Journal of Sports Physiology and Performance*, **13**, 1122 – 1129.

Hernandez, A. E., Mattarella-Micke, A., Redding, R. W., Woods, E. O. and Beilock, S. (2011) Age of acquisition in sport: Starting early matters. *The American Journal of Psychology*, **124**, 253.

Hsu, N.-J., Huang, H.-C. and Tsay, R. S. (2021) Matrix autoregressive spatio-temporal models. *Journal of Computational and Graphical Statistics*, **30**, 1143–1155.

Jungbacker, B. and Koopman, S. J. (2008) Likelihood-based analysis for dynamic factor models. *Tech. rep.*, Tinbergen Institute Discussion Paper.

Krüger, F., Lerch, S., Thorarinsdottir, T. and Gneiting, T. (2021) Predictive inference based on markov chain monte carlo output. *International Statistical Review*, **89**, 274–301.

Leroy, A., Marc, A., Dupas, O., Rey, J. L. and Gey, S. (2018) Functional data analysis in sport science: Example of swimmers' progression curves clustering. *Applied Sciences*, **8**, 1766.

Maharaj, E. A., D'Urso, P. and Caiado, J. (2019) *Time series clustering and classification*. Chapman and Hall/CRC.

Malsiner-Walli, G., Frühwirth-Schnatter, S. and Grün, B. (2016) Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, **26**, 303–324.

Malsiner-Walli, G., Frühwirth-Schnatter, S. and Grün, B. (2017) Identifying mixtures of mixtures using bayesian estimation. *Journal of Computational and Graphical Statistics*, **26**, 285–295.

Mikalsen, K. Ø., Bianchi, F. M., Soguero-Ruiz, C. and Jenssen, R. (2018) Time series cluster kernel for learning similarities between multivariate time series with missing data. *Pattern Recognition*, **76**, 569–581.

Mooses, M., Jürimäe, J., Mäestu, J., Purge, P., Mooses, K. and Jürimäe, T. (2013) Anthropometric and physiological determinants of running performance in middle-and long-distance runners. *Kinesiology*, **45**, 154–162.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.

Wade, S. and Ghahramani, Z. (2018) Bayesian cluster analysis: point estimation and credible balls (with Discussion). *Bayesian Analysis*, **13**, 559 – 626.

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

Weippert, M., Petelczyc, M., Thürkow, C., Behrens, M. and Bruhn, S. (2021) Individual performance progression of german elite female and male middle-distance runners. *European Journal of Sport Science*, **21**, 293–299.