Statistical Methods & Applications (2023) 32:401–423
https://doi.org/10.1007/s10260-022-00656-z

ORIGINAL PAPER

Bayesian GARCH modeling of functional sports data

Patric Dolmeta[1] · Raffaele Argiento[2,3] · Silvia Montagna[3,4]

Accepted: 1 September 2022 / Published online: 16 September 2022

Abstract
The use of statistical methods in sport analytics has gained a rapidly growing interest over the last decade, and nowadays is common practice. In particular, the interest in understanding and predicting an athlete's performance throughout his/her career is motivated by the need to evaluate the efficacy of training programs, anticipate fatigue to prevent injuries and detect unexpected of disproportionate increases in performance that might be indicative of doping. Moreover, fast evolving data gathering technologies require up to date modelling techniques that adapt to the distinctive features of sports data. In this work, we propose a hierarchical Bayesian model for describing and predicting the evolution of performance over time for shot put athletes. We rely both on a smooth functional contribution and on a linear mixed effect model with heteroskedastic errors to represent the athlete-specific trajectories. The resulting model provides an accurate description of the performance trajectories and helps specifying both the intra- and inter-seasonal variability of measurements. Further, the model allows for the prediction of athletes' performance in future sport seasons. We apply our model to an extensive real world data set on performance data of professional shot put athletes recorded at elite competitions.

Keywords Performance analysis · Bayesian functional data analysis · GARCH models · Sport analytics · Latent factor modelling

1 Introduction

Shot put is a track and field event involving throwing ("putting") the shot, a metal ball (7.26 kg/16 lb for men, 4 kg/8.8 lb for women), with one hand as far as possible from a seven-foot diameter (2.135 m) circle. In order for each put to be considered valid, the shot must not drop below the line of the athlete's shoulders and must land inside a designated 35-degree sector. Athletes commonly put four to six times per

Extended author information available on the last page of the article competition, and their best performance is recorded. The complete Technical Rule Book is available at the World Athletics website as attachment C2.1.[1] The same source also provides a short history of the sport, up to date rankings, records and All Time Best lists.

In this work, we are interested in describing (and predicting) the evolution of performance of professional shot put athletes throughout their careers. Figure 1 displays four examples of performance career trajectories of elite athletes. Each point represents the athlete's performance (as measured by throw length in meters) at a professional competition. Figure 1 suggests that as the athlete ages, his/her performance steadily increases (due to improved technique, physique, and practice), plateaus, and then begins to decline. With performance modelling in mind, we notice that different athletes have different career lengths and participate to a different number or professional events each year. Some "local" (i.e., yearly) drops (or peaks) in average performance with respect to the athlete's ongoing trend are also noticeable. For example, Christian Cantwell plays professionally for eighteen years, enjoys several years of consistently good performance, then his average performance drops on the 15th year of career, a year that can be recognised as a year of "worse" performances. Zhang Qi, instead, competes professionally only for thirteen years, generally participates to fewer events each year with respect to Cantwell, but similarly experiences more variable performance on the 7th year of career, with a drop in his average performance that year.

Performance results, such as those displayed in Fig. 1 for shot put, are collected at professional competitions during the sport season, which varies from sport to sport. Indeed, events are traditionally concentrated in some months of the year or across consecutive calendar years, depending on the sport. For example, many football leagues, such as the English Premier League, start in August and end in May of the following year, whereas the Formula 1 championship begins in March and ends at the end of November. Further, weather and environmental conditions may affect the performances or even the practicability of the sport itself (Winter sports as opposed to outdoor Summer sports). Thus, sport performance data have the distinctive characteristic of being time-gathered, in that they are collected only during the sport season. We believe such feature needs to be taken into account when modelling performance to provide an accurate description of career trajectories. Indeed, in the early career seasons an athlete will generally underperform with respect to his/her overall career average performance, and will outperform in later seasons. Shot put events range over the whole calendar year, with indoor competitions held during Winter months and major tournaments, like the Olympics, the Diamond League and the World Championship organised during Summer. Indeed, there are no evident (off-season) "gaps" in data collection in Fig. 1. Since training sessions take place during Winter months in preparation for major tournaments, it is reasonable to say that the shot put season corresponds with the calendar year. Additionally, we have already noted how the synthesis provided by yearly means captures some distinctive features of the temporal evolution of performances. In Fig. 1, vertical lines represent new years' days: the time point at which seasons change.

[1] https://www.worldathletics.org/about-iaaf/documents/book-of-rules.

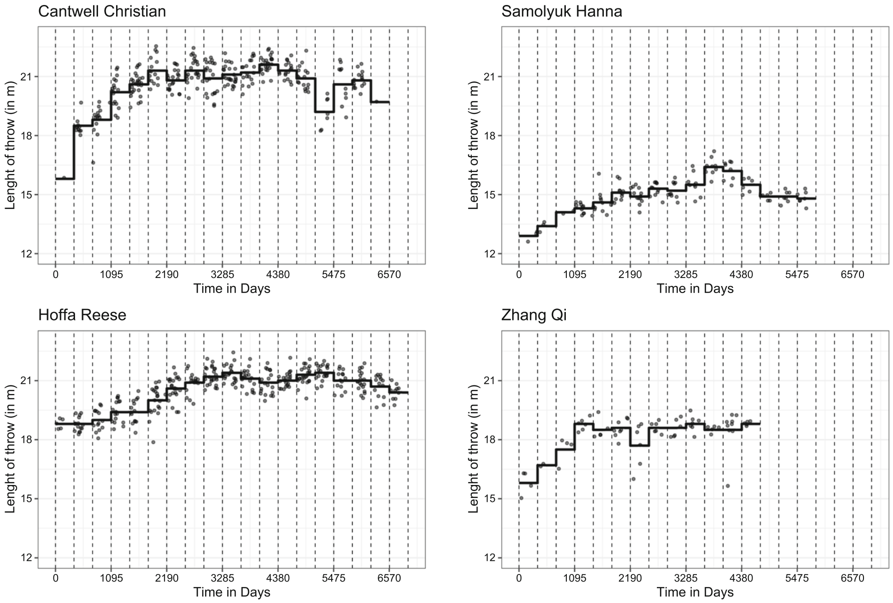

Fig. 1 Each panel displays the performance results of a professional shot put athlete throughout his/her career at elite competitions. Performance is measured in meters (length of the throw) and plotted against the days elapsed since January 1st of each athlete's career starting year. The dark vertical lines represent the sport season changing points, namely new year days, while the step functions show the athlete-specific yearly empirical mean in performance

The literature associated with career performance in sport has grown considerably in the last decade. Wimmer et al. (2011) model performance results in decathlon via a semi-parametric latent variable model. Casals and Martinez (2013) identify variables which may potentially influence player performance in basketball, and rely on mixed effects models to study their relative contribution in explaining points and win score in NBA games. Malcata et al. (2014) derive triathlon athletes' individual quadratic performance trajectories for peak performance prediction via linear mixed models, accounting for covariates. Koulis et al. (2014) use hidden Markov models to assess cricket players' reliability and predict their batting performances. More recently, Lee and Page (2021) study the performance curve, peak and optimal age of a basketball player. They also quantify the influence of each player on the game, and investigate which factors are predictive of a successful professional career. For the different tasks, they rely on several techniques ranging from mixed-effects models, dynamic time warping of performance curves to popular machine learning methods.

Some contributions that bear closer similarities with our modelling strategy (described below) are Scott et al. 1999, Page and Quintana (2015), Vaci et al. (2019), and Montagna and Hopker (2018). Scott et al. (1999) compare players' careers in different sports on the basis of performance. Specifically, they want to compare athletes competing in different eras accounting for the relative difficulty of each year within a sport. The proposed model is an additive model with components for the innate ability of players, the effects of ageing and the differences between years.

Nonparametric models are suggested for the ageing functions while hierarchical random curves allow for different ageing effects across athletes. Results are then used to classify athletes also across different eras. Page and Quintana (2015) treat NBA players' performance measurements as error prone observations from underlying curves, which are flexibly modelled via Bayesian penalised B-splines. Further, a hierarchical structure allows borrowing of information across trajectories and induces grouping (clustering) of the athletes guided by the smoothing of the individual curves. Vaci et al. (2019) describe the ageing curves of elite basketball players. The contribution relies on Bayesian structural modelling for the extraction of two latent factors describing development and ageing, respectively. The interaction of these factors provides insights on the rates of development and deterioration of skills over the course of a player's lifetime. Finally, Montagna and Hopker (2018) propose a Bayesian latent factor regression model for detecting the doping status of athletes given their shot put performance results and other covariates. However, the authors limit the analysis to data collected from 2012, whereas our interest is in describing the trajectories in performance over the whole time span available for our data (1996 to 2016). For this time span, a global smoothness assumption for the trajectories could be too restrictive. Indeed, data may exhibit jumps across seasons. See, for example, the distinctive drops and peaks in Cantwell's performance in Fig. 1. The presence of jumps between seasons is even more striking when yearly average performances are considered (step functions in Fig. 1).

In this work, we propose a Bayesian hierarchical additive model to describe the evolution of performance of professional shot put athletes. Our additive model consists of three components. First, we consider a smooth functional component for capturing the overall variability in athletes' performances, following Montagna et al. (2012) and Montagna and Hopker (2018). The second additive component is a random intercept, that quantifies the seasonal mean performance for each athlete as a deviation from a grand mean. This component captures the inter-seasonal variability of the data set, whereas the smooth functional component describes the intra-seasonal evolution of performances. Finally, we complete our model specification accounting for the effect of covariates through a regressive component. We embed our model in a Bayesian framework by proposing suitable prior distributions for all parameters of interest. We believe the presented model represents a flexible tool to analyse evolution of performances in measurable sports, namely, all those disciplines for which results can be summarised by a unique measure (e.g., distance, time or weight). Further, our model can be used to predict the expected performance of an athlete at future competitions within a given season, or in future seasons. To the best of our knowledge, this represents a novel contribution in the sport analytics literature, in particular when considering the attempt to model the seasonal nature of sports data.

The rest of the paper is organised as follows. In Sect. 2, we describe the motivating case study. In Sect. 3, we present the proposed model and briefly discuss possible alternative settings. Moreover, we elicit priors for our Bayesian approach. In Sect. 4, we outline the algorithm for posterior computation. In Sect. 5, we discuss posterior estimates, argue on the performance of the model and interpret the model's parameters form a sport analytics perspective. Conclusions are presented in Sect. 6.

2 The world athletics shot put data set

World Athletics (WA) is the world governing body for track and field athletic sports. It provides standardized rules, competition programs, regulated technical equipment, a list of official world records and verified measurements. The data at our disposal was obtained with permission from an open results database (www.tilastopaja.eu) following institutional ethical approval (Prop_72_2017_18). The data set comprises 56,000 measurements of WA recognized elite shot put competitions for 1115 athletes from 1976 to 2016. For each athlete, the data set reports the date of the event, the best result in meters, the finishing position, an indication of any doping violation during the athlete's career as well as demographic information (athlete's name, WA ID number, date of birth, sex and country of birth).

In this work, we restrict our analysis to results for athletes performing after 1996. Indeed, we pursue consistency of measurement accuracy, and 1996 represents a turning point in anti-doping regulation and fraud detection procedures. The resulting data set is still sufficiently broad for our purposes. It contains 41,033 observations for 653 athletes (309 males and 344 females). The outcome of interest is the shot distance, which ranges from a minimum if 10.6 up to a maximum if 22.56 m, with a mean of 17.30 m.

As shown in Fig. 1 for a selection of athletes, data are collected over time. Hereafter we will denote as t_{ij} the time at which the jth observation for athlete i is recorded. t_{ij} corresponds to the time elapsed from January 1st of each athlete's career starting year to the date of the competition. Accordingly, equal time values for different athletes can refer to different calendar years, but represent the same moment in those athletes' careers (e.g., three years into their careers). Moreover, different athletes will have observations ranging over a large time span, according to the length of their careers. Having described seasons as calendar years, athletes will also compete in a different number of seasons. Figure 2 shows the total number of

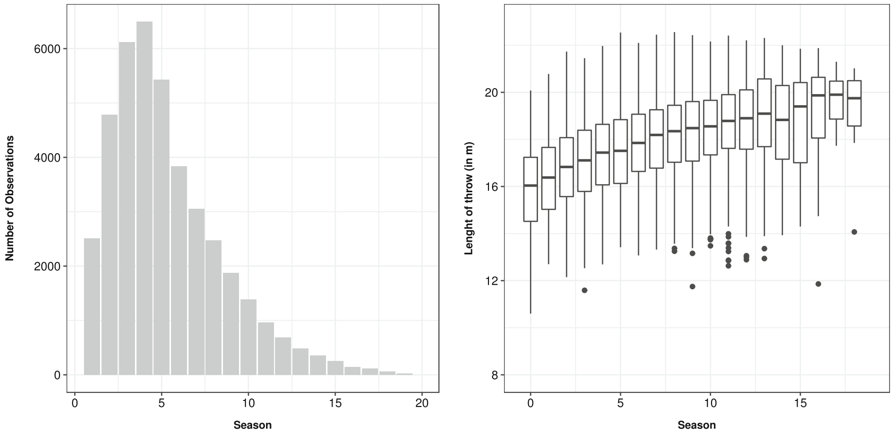

Fig. 2 Left: total number of observations per season for all athletes. Right: each boxplot shows the distribution of the athletes' performances within each season

observations per season as well as boxplots of the distribution of mean performances across the various seasons. We notice that, after season 4, the number of observations across all athletes per season decreases. The decreasing number of observations as seasons increase is indicative of the fact that just a minority of athletes have very long careers, and fewer athletes imply fewer data in later seasons. 19 years is the longest observed career among all shot-putters in the dataset. Only three athletes in our dataset enjoy a 19-years long career, namely, Catarina Andersson, Jon Kalnas, and Reese Hoffa. The median career length (i.e., number of seasons in sport) is 7 years, and the shortest career length is 2 years of professional play. A general increasing trend in performance can be observed as a function of career length (right panel in Fig. 2) or, equivalently, the age of the athlete. In the following, we will discuss two different modeling choices for age, respectively, accounting for its time dependence and considering age as a fixed quantity, namely the age of the athlete at the beginning of his/her career.

Table 1 reports descriptive statistics suggesting how sex, environment and doping have an effect on the average value of the result. We point out that in our dataset we only have 18 athletes who tested positive for doping at some point in their career. Information on the date the test was taken (or if multiple tests were taken) is not available in the data. As expected, performances for men are, on average, higher than for women. Similar effects, despite less evident in magnitude, also hold true for the variable environment, which takes values indoor and outdoor. Regarding environment, a further remark is in order. We already pointed out that major WA events take place outdoor (27,800 observations) during Summer months, whereas less competitive events are held inside between November and March (13,200 observations). Figure 3 displays results in grey when recorded outdoor, and in black otherwise. We can clearly see that field events gather in Summer months, whereas indoor events take place during Winter (solid lines).

3 The model

Let n denote the total number of athletes in the study. We assume that shot put performances for athlete i are given by noisy measurements of an underlying function $g_i(t_{ij})$:

Table 1 Performance results conditioned on covariates

	Mean	Sd	Max	Min
Total	17.30	1.78	22.56	10.6
Women	16.09	1.35	21.70	10.6
Men	18.55	1.21	22.56	12.93
Not doped	17.30	1.79	22.56	10.6
Doped	17.77	1.35	20.88	13.55
Indoor	17.17	1.70	22.23	12.15
Outdoor	17.38	1.81	22.56	10.6

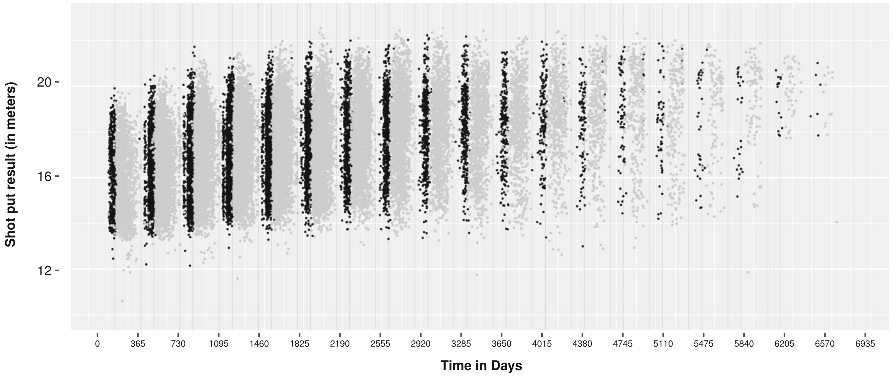

Fig. 3 Performance results displayed according to the variable environment, which takes values indoor (black) and outdoor (gray). Vertical lines corresponding to the label ticks represent season changes, whereas the two enclosing it indicate the boundaries of winter months

$$y_{ij} = g_i(t_{ij}) + \epsilon_{ij} \tag{1}$$

with $\epsilon_{ij} \overset{iid}{\sim} N(0, \psi^2)$ independent errors. Recall t_{ij} is the time at which the jth observation for athlete i is collected, for $j = 1, \ldots, n_i$, where n_i is the total number of measurements available on athlete i.

We further suggest an explicit functional form for $g_i(t_{ij})$:

$$g_i(t_{ij}) = f_i(t_{ij}) + \mu_{is} + \boldsymbol{x}_i(t_{ij})\boldsymbol{\beta} \tag{2}$$

where $f_i(t)$ is a smooth functional component for intra-seasonal variability, μ_{is} a season-specific intercept, and $\boldsymbol{x}_i(t)\boldsymbol{\beta}$ is an additional multiple regression component. Here $s \in \{1, 2, \ldots, S_i\}$ indicates the season in which the shot was recorded. Specifically, $\mu_{is} \equiv \mu_i(t_{ij}) = \sum_{s=1}^{S_i} \mu_{is} \; \mathbb{I}_{(t_i^s, t_i^{s+1})}(t_{ij})$ is an athlete-specific step function taking value μ_{is} for all time points in season s, delimited by t_i^s and t_i^{s+1}. For a complete treatment of the notation used insofar, please refer to Table 2.

Before discussing each of the three terms in Eq. (2) in more detail, we remark that our model can be seen as a generalised additive mixed-effects model (GAMM). Specifically, time is allowed to have a non-linear impact on an athlete's performance via the smooth functional component f_i, there is a linear regression component for the covariates, and the season-specific intercept is itself a functional expansion, taking constant values over seasons. We acknowledge that GAMMs have been proposed before in the sport analytics literature [e.g., Vaci et al. (2019)], but not with the direct intent to capture the major sources of variability of performance data.

3.1 The functional component

The functional component $f_i(t)$ is meant to capture the subject-specific global evolution of the response variable. It explains the global dependence of the data from time. We require that these functions display a smooth behaviour: the latter is assured

Table 2 Mathematical notation

Symbol	Meaning
i	Index identifying the athlete
j	Index identifying a specific observation
n	Total number of athletes
N	Total number of observations
t_{ij}	Time point at which the j'th observation of athlete i is recorded
S_i	Total number of seasons for athlete i
s	The currently considered season
g_{is}	Number of observations in season s for athlete i
r	The number of covariates to be considered
y_{ij}	Response variable at time j for athlete i

by assuming $\{f_i(t)\}_{i=1}^{n}$ are linear combinations of smooth basis functions, $\{b_r(t)\}_{r=1}^{p}$. Note that, both the nature and the number p of these bases are to be determined according to some properties we wish them to satisfy. In particular, we assume:

$$f_i(t) = \sum_{r=1}^{p} \theta_{ir} b_r(t) \tag{3}$$

where $\{b_m(t)\}_{r=1}^{p}$ represent the B-spline basis de Boor (1978) and $\{\theta_{ir}\}_{r=1}^{p}$ are subject-specific coefficients.

We briefly recall that the B-spline basis of degree k on $[L, U]$ is a collection of p polynomials defined recursively on a sequence of points, known as knots, and indicated with $L \equiv t_1 \leq \ldots \leq t_{p+k+1} \equiv U$. We follow the common approach of choosing $k = 3$, leading to cubic splines (see, for instance, Marsden (1974)). Moreover, we assume the knot sequence to be equispaced and $(k + 1)$-open. That is, the first and last $k + 1$ knots are identified with the extremes of the definition interval, whereas the remaining $p - k - 1$ knots divide said interval into sets of the same length. Under these assumptions, each basis function $b_j(t)$ has compact support over $k + 1$ knots, precisely $[t_j, t_{j+k+1}]$. Moreover, together they span the space of piecewise polynomial functions of degree k on $[L, U]$ with breakpoints $\{t_n\}_{n=1}^{p+K+1}$. Finally, such functions are twice continuously differentiable at the breakpoints, de facto eliminating any visible type of discontinuity and providing a smooth result.

Choices for the number and location of the knots are ultimately driven by the desired level of smoothness of the estimated trajectories in a particular application. For simplicity, we have decided to set the interior knots to be equally-spaced, although a different choice would not pose any conceptual or practically challenges. For example, if the data was denser in a particular part of the input space one could choose to place more knots in the corresponding sub-interval to allow for more flexible estimates locally. We did not deem this to be of practical importance in our application. Regarding the number of basis functions, splines have overlapping supports hence, in general, several bases compete in capturing the local behaviour of the function to be approximated. However, if a sufficiently large number of knots is chosen, some splines will have their support completely contained in a given season. In the shot put application, we have 19 seasons in total (Fig. 2). If we want at least one spline to be completely supported in each one of the 19 seasons, we need each season to contain at least 4 knots. Accordingly, we require about 75 internal knots, equivalently 80° degrees of freedom. The majority of the resulting bases will have supports ranging over several seasons, but 19 of them will concentrate uniquely on a particular season, trying to explain the within-season variability. In general, sensitivity analysis is required to choose the number and location of the knots, and changing either possibly requires tuning the other as necessary to find the best combination to fit the data.

With the sake of tractability, a low dimensional representation of the individual curves is of interest. Following the approach by Montagna et al. (2012), we exploit a sparse latent factor model on the basis coefficients:

$$\theta_{ir} = \sum_{l=1}^{k} \lambda_{rl}\eta_{il} + \xi_{ir}, \tag{4}$$

for $r = 1,\ldots,p$, where λ_{rl} are the entries of a $(p \times k)$ factor loading matrix $\mathbf{\Lambda}$, and $\boldsymbol{\eta_i} = [\eta_{i1},\ldots,\eta_{ik}]^\top$ is a vector of k latent factors for subject i. Finally, $\boldsymbol{\xi_i} = [\xi_{i1},\ldots,\xi_{ip}]^\top$ is a residual vector, independent of all other variables in the model. We assume:

$$\boldsymbol{\eta_i} \overset{iid}{\sim} N_k(\boldsymbol{0},\boldsymbol{I}) \tag{5}$$

and the error terms $\boldsymbol{\xi_i}$ are assumed to have normal distribution with diagonal covariance matrix, $\boldsymbol{\xi_i} \overset{iid}{\sim} N_p(\boldsymbol{0}, diag(\sigma_1^{-2},\ldots,\sigma_p^{-2}))$, with $\sigma_j^{-2} \overset{iid}{\sim} Ga(a_\sigma, b_\sigma)$.

For the modelling of the factor loading matrix $\mathbf{\Lambda}$, we follow the approach in Bhattacharya and Dunson (2011) and adopt a multiplicative gamma process shrinkage (MGPS) prior:

$$\begin{aligned}
&\lambda_{rl}|\phi_{rl}^{-1},\tau_l^{-1} \overset{iid}{\sim} N\left(0,\phi_{rl}^{-1}\tau_l^{-1}\right) \qquad \text{with} \\
&\phi_{rl} \sim Ga\left(\frac{\nu_\phi}{2},\frac{\nu_\phi}{2}\right) \qquad \tau_l = \prod_{v=1}^{h} \varpi_v \\
&\varpi_1 \sim Ga(a_1,1) \qquad \varpi_v \sim Ga(a_v,1), \quad v \geq 1
\end{aligned} \tag{6}$$

The impact on shrinkage and on dimensionality reduction deriving from the adoption of a MGPS prior on the loadings is discussed extensively in Montagna et al. (2012), and follows along the same lines here. The number of latent factors, k, is not set a priori but adaptively learnt, as discussed in Bhattacharya and Dunson (2011) and Montagna et al. (2012). We defer to Sect. 1 of the Online Supplementary Material for more details.

3.2 The seasonal component

Early graphical displays and straightforward exploratory analysis suggest a significant variability of the average response across seasons, as displayed in Fig. 1. Namely, performances prove to be gathered over pre-determined time intervals, the seasons (calendar years). However, it is reasonable to expect some degree of dependence for the average performance across seasons. To model such dependence, an autoregressive model for seasonal intercepts can be proposed. The idea behind this choice is to allow for borrowing of information across seasons, in the sense that the seasonal intercept μ_{is} at season s is influenced by the intercept at season $s-1$ through the autoregressive coefficient ρ_i. Specifically:

$$\mu_{is} \mid \rho_i, \sigma_\mu^2 \overset{iid}{\sim} N\left(\rho_i \mu_{i(s-1)}, \sigma_\mu^2\right) \tag{7}$$

However, when we first implemented this model, we noted how residuals presented a pattern which we would like to intercept with a finer model (see Sect. 2 in the Online Supplementary Material). Therefore, we consider a random intercept model with Normal Generalized Autoregressive Conditional Heteroskedastic (GARCH) errors Bollerslev (1986). Specifically,

$$\mu_{is} \mid m, h_{is} \ = m + \zeta_{is} \overset{iid}{\sim} N(m, h_{is}) \tag{8}$$

$$h_{is} \ = \alpha_0 + \alpha_1 \zeta_{is-1}^2 + \upsilon h_{is-1} \tag{9}$$

where $\alpha_0 > 0, \alpha_1 \geq 0$ and $\upsilon \geq 0$ to ensure a positive conditional variance and $\zeta_{is} = \mu_{is} - m$ with $h_{i0} = \zeta_{i0} := 0$ for convenience. The additional assumption of wide-sense stationarity with

$$\begin{aligned} \mathbb{E}(\zeta_t) &= 0 \\ \mathbb{V}ar(\zeta_t) &= \alpha_0 (1 - \alpha_1 - \upsilon)^{-1} \\ \mathbb{C}ov(\zeta_t, \zeta_s) &= 0 \text{ for } t \neq s \end{aligned}$$

is guaranteed by requiring $\alpha_1 + \upsilon < 1$, as proven by Bollerslev (1986).

Three parameters of the seasonal component require prior specification: the overall mean m and the conditional variance parameters, υ and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^\top$. For the autoregressive and heteroskedastic parameters of the GARCH model, we propose non-informative priors satisfying the positivity constraint. For the overall mean parameter, we rely on a more informative Normal prior centred around the mean suggested by posterior analysis of preliminary versions of the model. In particular:

$$\begin{aligned} m &\sim N(\mu_{m_0}, \Sigma_{m_0}) \\ \boldsymbol{\alpha} &\sim N_2(\mu_\alpha, \Sigma_\alpha) \ \mathbb{I}\{\boldsymbol{\alpha} > 0\} \\ \upsilon &\sim N(\mu_\upsilon, \Sigma_\upsilon) \ \mathbb{I}\{\upsilon \geq 0\} \end{aligned} \tag{10}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$ is a bidimensional vector. We complete the model specification assuming that the parameters are statistically independent and noticing that the hypothesis needed for wide-sense stationarity do not translate into actual prior conditions on the parameters. Hence, one of the objects of our analysis becomes to test whether the constraint $\alpha_1 + \upsilon < 1$ holds true.

3.3 Covariates

We consider the effect of three covariates, gender, age and environment, and assume conjugate prior choices for the covariates coefficients:

$$\begin{aligned} \boldsymbol{\beta} &\overset{iid}{\sim} N(\boldsymbol{\beta_0}, \sigma^2_\beta \boldsymbol{I}) \\ \sigma_\beta^{-2} &\sim Ga\left(\frac{\nu_\beta}{2}, \frac{\nu_\beta \sigma^2_\beta}{2}\right) \end{aligned} \tag{11}$$

4 Estimation and inference

4.1 The Bayesian update

Because of the additive nature of the overall sampling model (1)–(2), we are able to exploit a blocked Gibbs sampler grouping together the parameters of the three modelling components described in Sect. 3. Note first that, because of the high dimensionality of the problem, it is computationally convenient to choose conditionally conjugate prior distributions for the parameters. Indeed, conjugacy guarantees analytical tractability of posterior distributions. In some cases, specifically for the conditional variances of GARCH errors, no conjugate model exists and updates rely on an adaptive version of the Metropolis Hastings algorithm for posterior sampling.

Algorithm 1 outlines our sampling scheme, while details are presented in Sect. 3 of the Online Supplementary Material. As far as the parameters of the functional component $\boldsymbol{\theta}_i$ are concerned, we follow Montagna et al. (2012) by choosing conditionally conjugate prior distributions so that the update proceeds via simple Gibbs sampling steps. Analogously, the update of the regression coefficients $\boldsymbol{\beta}$ and the error term ψ proceeds straightforwardly by sampling from their full conditional posterior distributions. Conjugate priors for the GARCH parameters m, ϖ and $\boldsymbol{\alpha}$ are not available, therefore we resort to adaptive Metropolis schemes to draw values from their full conditionals. Specifically, we build an adaptive scale Metropolis such that the covariance matrix of the proposal density adapts at each iteration to achieve an *optimal* acceptance rate (see Haario et al. (2001)).

Further details about the algorithm can be found in Sect. 3 of the Online Supplementary Material, whereas code is available at https://github.com/PatricDolmeta/Bayesian-GARCH-Modeling-of-Functional-Sports-Data.

Algorithm 1: Gibbs Sampler

```
Data: $y_{ij} = (y_{11}, \dots, y_{nn_n})$
Set the required MCMC sample size $G$, the burn-in period $g_0$ and the thinning
    parameter $g_s$.
  Initialise
    $\boldsymbol{\theta}_i^{(0)}, \boldsymbol{\mu}_i^{(0)}, m^{(0)}, \varpi^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}_i^{(0)}, \psi_i^{(0)}$
  For $g = 0, \dots, G$
      Update functional component
          Set partial residuals $y_{ij}^{(1)(g)} = y_{ij} - \mu_{i,s}^{(g)} - \boldsymbol{x_i}(t_{ij})\boldsymbol{\beta}^{(g)}$
          Update $\boldsymbol{\theta}_i^{(g+1)}$ on the base of Section 7.1 of the supplementary material
      Update seasonal component
          Set partial residuals $y_{ij}^{(2)(g)} = y_{ij} - f_i(t_{ij})^{(g)} - \boldsymbol{x_i}(t_{ij})\boldsymbol{\beta}^{(g)}$
          Update $\boldsymbol{\mu}_i^{(g+1)}, m^{(g+1)}, \varpi^{(g+1)}, \boldsymbol{\alpha}^{(g+1)}$ on the base of Section 7.2 of
            the supplementary material
      Update regressive component
          Set partial residuals $y_{ij}^{(3)(g)} = y_{ij} - f_i(t_{ij})^{(g)} - \mu_{i,s}^{(g)}$
          Update $\beta_i^{(g+1)}$ on the base of Section 7.3 of the supplementary material
      Update error term
          Set partial residuals $\epsilon_{ij}^{(g)} = y_{ij} - f_i(t_{ij})^{(g)} - \mu_{i,s}^{(g)} - \boldsymbol{x_i}(t_{ij})\boldsymbol{\beta}^{(g)}$
          Update $\psi^{(g+1)}$ on the base of Section 7.4 of the supplementary
            material
      Return $\boldsymbol{\theta}_i^{(g)}, \boldsymbol{\mu}_i^{(g)}, m^{(g)}, \varpi^{(g)}, \boldsymbol{\alpha}^{(g)}, \boldsymbol{\beta}_i^{(g)}, \psi_i^{(g)}$ for
        $g = g_0, g_0 + g_s, g_0 + 2g_s, \dots, G$
```

4.2 Posterior analysis

The idea of estimating trajectories for athletes' performances is a natural pursuit for the model specification we adopted. Indeed, describing observations as error prone measurements of an unknown underlying function suggests evaluating such function, once retrieved, on any number of points of interest. In practice, we will generate a fine grid of T equispaced time points: $\{t_k\}_{k=1}^T$ between $0 \equiv t_1$ and $1 \equiv t_T$ and evaluate the function on this grid.

In particular, we start by evaluating the athlete-specific functional component by exploiting the basis function representation. Being:

$$\boldsymbol{\Theta}_i = \begin{bmatrix} \theta_{i1}^{(1)} & \theta_{i2}^{(1)} & \dots & \theta_{ip}^{(1)} \\ \theta_{i1}^{(2)} & \theta_{i2}^{(2)} & \dots & \theta_{ip}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{i1}^{(G)} & \theta_{i2}^{(G)} & \dots & \theta_{ip}^{(G)} \end{bmatrix}$$

the matrix of individual-specific spline basis coefficients for all iterations $g = 1, \dots G$ and

$$\boldsymbol{b}^\top = \begin{bmatrix} b_1(t_1) & b_1(t_2) & \dots & b_1(t_k) & \dots & b_1(t_T) \\ b_2(t_1) & b_2(t_2) & \dots & b_2(t_k) & \dots & b_2(t_T) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ b_p(t_1) & b_p(t_2) & \dots & b_p(t_k) & \dots & b_p(t_T) \end{bmatrix}$$

all values of a p-dimensional, degree-3, spline basis on a set of $T+1$ equispaced knots in the unit interval, the estimated contribution of the functional component to the overall trajectory is, at each iteration:

$$f_i^{(g)}(t) = \sum_{r=1}^{p} \theta_{ir}^{(g)} b_r(t) = \Theta_i^{(g)} \boldsymbol{b}_t^\top \qquad \text{for } t = t_1, \dots, t_T,$$

where $\Theta_i^{(g)}$ corresponds to the i-th row of matrix $\boldsymbol{\Theta}_i$.

As for the seasonal linear mixed effect, we modelled it as a piecewise continuous function taking individual- and season-specific values. Hence, when retrieving its estimated effect on any point in the time grid, we need to determine which season it belongs to. As discussed in Sect. 3, time is rescaled so that equal values across individuals indicate the same day of the year, possibly in different years. Therefore, season changes, that occur at new year's days, can be easily computed by straightforward proportions. At this point, the season to which t_k belongs to is obtained by comparison with the season thresholds. In the following Equation, the indicator variable $\chi_{(t \in s)}$ determines to which season each time point belongs to. Accordingly, the estimated contribution of the seasonal component to the overall trajectory is, at each iteration:

$$\mu_i^{(g)}(t) = \sum_{s=1}^{S_i} \mu_{is}^{(g)} \chi_{(t \in s)} \qquad \text{for } t = t_1, \dots, t_T.$$

Lastly, the regressive component has to be taken into account. The estimated contribution of the regressive component to the overall trajectory is, at each iteration:

$$\sum_{l=1}^{r} x_{il}(t) \beta_l^{(g)} = \boldsymbol{x}_i(t) \boldsymbol{\beta}^{(g)} \qquad \text{for } t = t_1, \dots, t_T.$$

Given the three components, the overall estimate of the underlying function is obtained by adding these three components. In particular, the estimated mean trajectory can be written as:

$$\widehat{y_i(t)} = \frac{1}{G} \sum_{g=1}^{G} f_i^{(g)}(t) + \mu_i^{(g)}(t) + \boldsymbol{x}_i(t) \boldsymbol{\beta}^{(g)} \qquad \text{for } t = t_1, \dots, t_T \tag{12}$$

Similarly, 95% credible intervals can be computed to quantify uncertainty around our point estimate.

5 Results

5.1 Fitted model

In this Section, we fit different specifications of our model to the data described in Sect. 2.

In general, we consider the additive structure of the sampling model illustrated in Eq. 2. Table 3 reports an overview on of the six models we compare. In Model M_1, the B-spline basis functions have 80 degrees of freedom, the seasonal component has GARCH errors and three regressors are taken into account: *sex, age and environment*. M_2 represents a slight modification of M_1 given by the fixed-age implementation. Here we consider the covariate *age* not as a time dependent variable, but as a fixed value given by the age at the beginning of each athlete's career. In model M_3 a simpler dependence structure among the seasonal effects is used. Namely, we assume an autoregressive model for μ_{is} (see Eq. 7). For model M_4, we simply consider a larger number of basis functions, i.e. 120, accounting for up to three splines having support in a season and hence meant to better capture the intra-seasonal variability. Finally, models M_5 and M_6 allow for *doping* as additional covariate, both in the case of the time-dependent and time-independent specification of *age*.

Priors were chosen as discussed in Sect. 3, and with hyperparameter choices summarised in Table 4. To argue on the choice of the informative prior for the overall mean parameter m, in Table 5 we also report the results under a slight modification of model M_1, that we denote $M_1^{(2)}$, yielding a vague prior for m.

For all experiments, inference is obtained via posterior samples drawn by the Gibbs sampler introduced in Sect. 4. In particular, we ran 20, 000 iterations with a burn-in period of 60% and a thinning of 5. Performances are compared by means of the logarithm of the pseudo marginal likelihood (LPML) index Geisser and Eddy (1979). This estimator for the log marginal likelihood is based on conditional predictive densities and provides an overall comparison of model fit, with higher values denoting better performing models. The reader may refer to Sect. 4 in the Online Supplementary Material for posterior convergence diagnostics results under model M_1.

Table 3 Models name and description

Symbol	Meaning
M_1	80 df B-splines, GARCH, covariates: sex, age (time dependent), env.
M_2	80 df B-splines, GARCH, covariates: sex, age (time constant), env.
M_3	80 df B-splines, AR, covariates: sex, age (t. dep.), env.
M_4	120 df B-splines, GARCH, covariates: sex, age (t. dep.), env.
M_5	80 df B-splines, GARCH, covariates: sex, age (t. dep.), env., doping
M_6	80 df B-splines, GARCH, covariates: sex, age (t. const.), env., doping

Table 4 Hyperparameter choices. In the first column, we refer to the Equation where the hyperparameter first appears

Ref.	Hyp.	Value	Description
(4)	a_σ	1.0	1st Gamma coeff. of error term in the factor exp.
(4)	b_σ	0.3	2nd Gamma coeff. of error term in the factor exp.
(6)	v_ϕ	9	Gamma coeff.s of local shrink. param. ϕ_{ml}
(6)	a_1	2.1	1st Gamma coeff. of the 1^{st} global shrink. factor δ_1
(6)	b_1	1.0	2nd Gamma coeff. of the 1^{st} global shrink. factor δ_1
(6)	a_l	2.1	1st Gamma coeff. of the lth global shrink. factor δ_l
(6)	b_l	1.0	2nd Gamma coeff. of the lth global shrink. factor δ_l
(10)	μ_{m_0}	-0.2	Mean of the overall mean m
(10)	Σ_{m_0}	0.0001	Variance of the overall mean m
(10)	μ_α	(0.0, 0.0)	Mean vector of the $\boldsymbol{\alpha}$ GARCH coeff.
(10)	Σ_α	$\mathbb{I}_2$	Covariance matrix of the $\boldsymbol{\alpha}$ GARCH coeff.
(10)	μ_ϖ	0.0	Mean of the ϖ GARCH coeff.
(10)	Σ_ϖ	1	Variance of the ϖ GARCH coeff.
(11)	v_β	0.5	1st Gamma coeff.s regression param.
(11)	σ_β	0.5	2nd Gamma coeff.s of regression param.
(1)	μ_ψ	1.0	Mean of the error variance ψ
(1)	σ_ψ	1.0	Variance of the error variance ψ

Table 5 Model and hyperparameter comparison for the models in Table 3

Model	LPML
M_1	$\boldsymbol{-45,943}$
$M_1^{(2)}$	$-46,573$
M_2	$\boldsymbol{-45,472}$
M_3	$-46,544$
M_4	$-46,314$
M_5	$-48,565$
M_6	$-48,122$

Performances for the different models are fairly similar: as a matter of fact, the model specifications do not differ in a significant way. Despite having a slightly lower LPML than the best performing model, M_2, we prefer looking at results for model M_1 with 80° of freedom splines, GARCH errors and three regressors (sex, age, and environment), with the time-dependent age definition because regression parameters prove to be significant in this setting. For the estimation of performance trajectories, we use the method discussed in Sect. 4.2. Figure 4 displays the estimate (with 95% credible bounds) for a random selection of athletes (black) together with one-season-ahead performance prediction (grey). The results are graphically pleasing in terms of model fit, but some comments are of order. First, the amount of variability explained by the random intercepts given the amount of variability explained by the functional component is about 56% of the total variance, thus the random intercepts capture the majority of the variability in the data (refer to Sect. 5 in the Online Supplementary Material for a discussion). Second, the functional component reduces to capture the intra-seasonal variability. Interestingly, the number on non-local bases selected by the adaptive procedure in Bhattacharya and Dunson (2011) is exactly equal to the number of seasons in the data set. This effect seems to be consistent with the choice of degrees of freedom, that limits the support of each spline to a unique season.

To better understand the contribution of the three model components on the overall model fit, we focus on a single athlete (Reese Hoffa, ID 226) and visualise the separate effect of the components in Fig. 5. The first panel (top-left) displays the

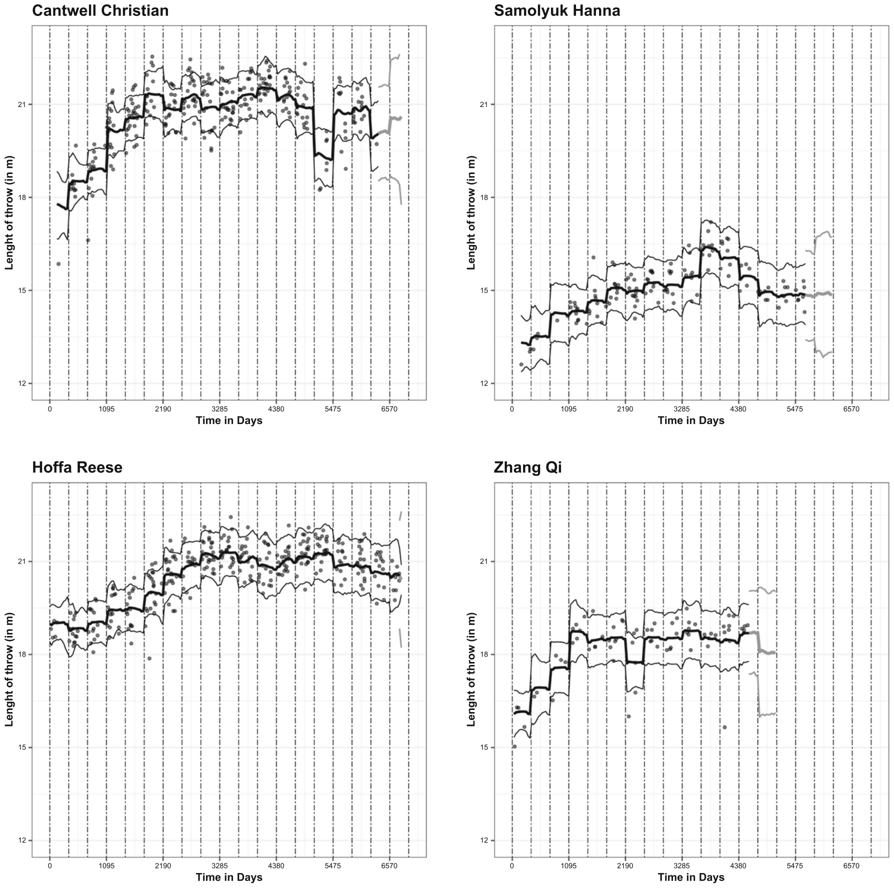

Fig. 4 Performance trajectory estimates for a random selection of athletes. The x-axis denotes the time measured in days from January 1st of the first season of career, whereas on the y-axis there is the length of throw in meters. Vertical lines represent calendar years (seasons in our notation). The final part of each trajectory (grey) for which no observations are available, represents one-season-ahead performance prediction

observed data and Hoffa's overall estimated trajectory. The top-right panel reports the functional contribution (i.e, an estimate of $f_{226}(t)$). The third panel (bottom-left) shows the seasonal contribution (i.e, the estimate of μ_{is}; $i = 226$; $s = 1, \ldots, S_{19}$). Finally, the bottom-right panel shows the effect of the linear regression component. Differently from the other panels, here we visualise the estimated performance associated with different levels of a covariate of interest (environment). In particular, the continuous line represents the trajectory according to Hoffa's observed covariates, while the dashed line represents the estimated performance were all competitions held indoor, all other things being equal. Indoor competitions are

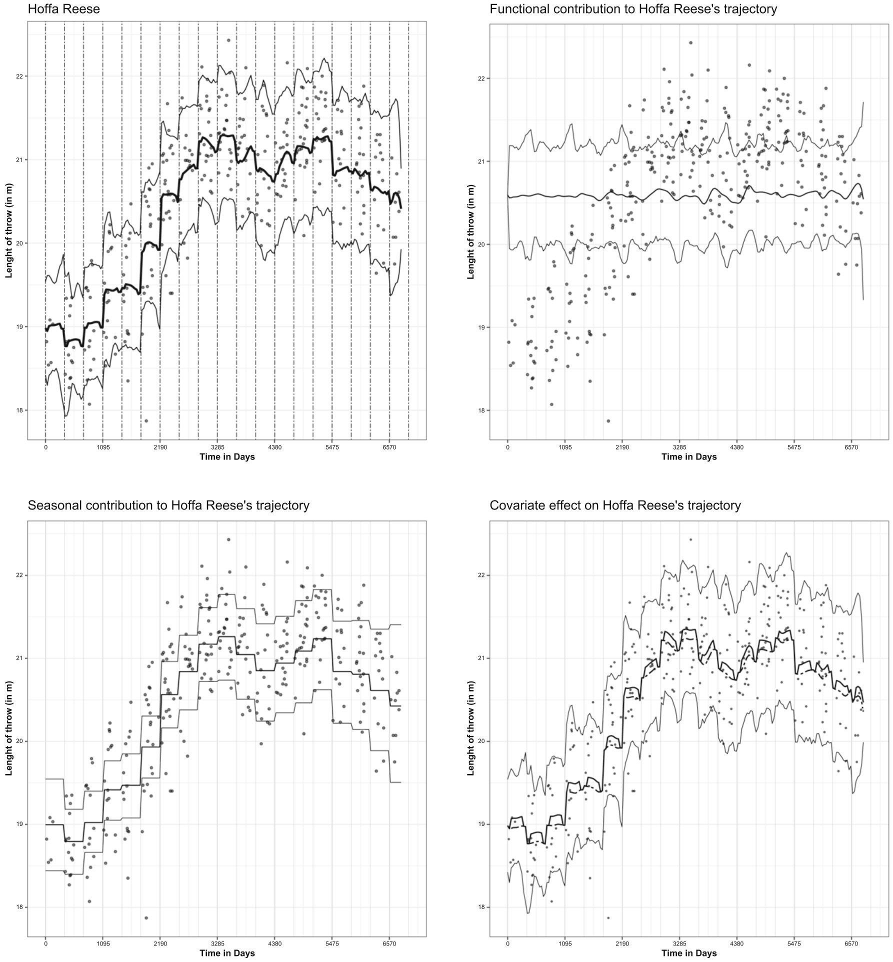

Fig. 5 Single contributions to the whole additive model as in Eq. 12. The first panel is the complete additive model, whereas the second (top-right) displays the functional contribution. The third panel (bottom-left) displays the estimate of the seasonal random intercept, while the bottom-right panel displays the estimated trajectories associated with the different levels of the environment covariate (continuous line for Hoffa's observed covariates, dashed line if all competitions took place indoor). Estimates are embedded into 95% credible bands

typically held in Winter months. We notice an expected drop in performance when competitions take place indoor. Because of the generally small effect in magnitude, it is rather difficult to visualise the effect that different levels of a covariate have on the response, so we omit other plots from the manuscript. We discuss the effect of all other covariates more in detail separately below (Sect. 5.2).

Following a suggestion raised by one reviewer, we investigated further how the different additive components of our model contribute to the overall model fit. Specifically, we performed a model comparison across the included hierarchical levels and components. Please refer to Sect. 6 in the Online Supplementary Material for results.

5.2 Interpretation of the estimated parameters

The regression parameters can be easily interpreted from a sport analytics perspective. It is important to stress that, to improve convergence of the MCMC algorithm, we fitted our model centring athlete-specific data around their average. Accordingly, in our experiments the raw data y_{ij} were substituted by the centered points:

$$\tilde{y}_{ij} = y_{ij} - \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} = y_{ij} - \bar{y}_i \quad \text{for } i = 1, \ldots, n \text{ and } \; j = 1, \ldots, n_i$$

as data-input for the model. We have to take into account this transformation when interpreting the regression parameters, especially when dealing with dummy variables.

We report the posterior mean estimate of the regression coefficients, their standard deviation, the effective sample size (ESS) and the 95% posterior credible bounds for the most interesting models. Table 6 displays the results under model M_1, where covariates are sex (x_1), age (x_2) and environment (x_3). Even if the covariate effect is small in magnitude, we observe that the 95% credible intervals do not contain zero, showing a significant effect. Sex is a binary covariate, taking value equal to one for male athletes. The estimated coefficient for sex is negative ($\widehat{\beta}_{sex} = -0.120$). We remark that, having demeaned the outcome, β_{sex} quantifies the difference in variability of an athlete's performance around his/her average $\bar{y}_i$. Accordingly, males' trajectories express more variability around their average than females. Differently, when using data in the original scale, the estimated gender effect is significant and positive ($\widehat{\beta}_{sex} = 1.243$), suggesting that the length of a throw for a male athlete is expected to be, approximately, 1.2 metres longer than that of a female athlete. The second covariate, age, is time-dependent: as time increases, so will age at a proportional rate, since time is measured in days and age in years. The estimated coefficient for age is positive. This indicates that, generally, performance is expected to improve as the athlete grows older. Finally, environment is a time-dependent, binary covariate taking value equal to one for outdoor events. The estimated effect ($\widehat{\beta}_{env} = 0.0453$) indicates an expected increase in performance when competitions take place outdoor, typically in Summer months, as noted visually in Fig. 5. Parameter estimation under the other models considered in the paper are similar in

Table 6 Posterior mean estimate of the regression coefficients for model M_1(Table 3), together with the standard deviation of their estimate, effective sample size (ESS) with respect to 1600 retained samples, and 95% posterior credible bounds

Coeff	Mean	Sd	ESS	2.5%	97.5%
Sex	-0.120	0.0270	274	-0.175	-0.0675
Age	$6.22e-03$	$9.95e-04$	231	$4.20e-03$	$8.20e-03$
Environment	0.0453	$9.55e-03$	1820	0.0269	0.0643

sign with respect to the ones discussed here. We report complete results in Section 7 of the Online Supplementary Material.

A final comment on results obtained using doping as additional regressor is required. We stressed that the LPML indices for models M_5 and M_6 are quite low, however, this may be due to the fact that the data set is imbalanced, i.e., there are too few doped athletes (18 out of 653). In fact, the ESS of the parameter corresponding to doping (β_4) is very low. Nevertheless, it is interesting to observe that estimates are similar for all common parameters and that the coefficient of the doping regressor is negative (even if the credible intervals contain zero). We conclude that the use of performance-enhancing drugs seems to have a negative effect on the variability of athletes' performances.

5.3 Predictive performance and model validation

Having insofar discussed the model fit to the data, we now turn our attention to performance prediction. Our model can be used for out-of-season prediction of an athlete's performance, that is, to predict his/her expected performance in future seasons. Further, the model can also provide intra-seasonal predictions, that is, the expected evolution within a given season for which partial data is already available. Further, the Bayesian framework our model is embedded into provides us with an automatic mechanism to quantify uncertainty in the predictions. Arguably, major interest from a sport analytics perspective is in predicting the evolution of an athlete's performance in future seasons (e.g., past observed data or out-of-sample), as this quantity provides insights in the future expected development of the athlete (e.g., is the athlete expected to perform better/worse in the future). The estimated value of the athlete's trajectory at time points t_{ij} in season $S_i + 1$, for which no data is yet available, is obtained as:

$$y_i\widehat{(t_{ij})} = \frac{1}{G}\sum_{g=1}^{G} f_i^{(g)}(t_{ij}) + \mu_i^{(g)}(t_{ij}) + \boldsymbol{x}_i(t_{ij})\boldsymbol{\beta}^{(g)}, \tag{13}$$

where g denotes $g-th$ MCMC iteration and G is the total number of posterior samples. Thus, the evaluation (on a future time point) of the functional and regressive components follows the description in Sect. 4.2. One simply needs to evaluate the basis functions on the future time point(s) and needs the future values of the covariates. The new seasonal intercept follows the probabilistic dynamic suggested by the GARCH model:

$$\mu_{iS_i+1} \mid m, h_{iS_i} = m + \zeta_{iS_i} \overset{iid}{\sim} N(m, h_{iS_i}) \tag{14}$$

$$h_{iS_i} = \alpha_0 + \alpha_1 \zeta^2_{iS_i-1} + \varpi h_{iS_i-1} \tag{15}$$

One-season-ahead predictions (with 95% credible bands) are shown, for a random selection of athletes, in Fig. 4. Concerning intra-seasonal predictions, the whole continuous trajectory itself represents the intra-seasonal performance prediction at all those time points for which no data is available. When data is missing completely for a whole season (i.e., the athlete did not compete professionally in a given year to recover from an injury, or when predicting a future season), the task of predicting counterfactual performances is certainly more challenging for the model.

To validate the model, we tested its predictive performance against some held-out data. In particular, we performed two analyses. In both cases, we randomly chose 100 athletes from the data set and for these "test set" athletes: (1) we held out and predicted data on a mid-career season ("Mid-career" analysis), and; (2) we held out and predicted data on their last available season ("End of career" analysis). Prediction performance in both cases is quantified in terms of mean squared error between the observed test data and predicted values, and is reported in Table 7. The mid-career prediction error is smaller than the end-of-career's. Likely, this is due to the borrowing of information from both past and future seasons in estimating the current (mid-career) one. The end-of-career predictions resemble the careers's average performance, which might be explained by the centring of the seasonal component around a grand mean. The magnitude of these errors can be explained by the scale of the data, which are centered prior to the analysis, and is comparable to the training MSE (0.273 Mid-career, 0.281 End-career).

Estimated performance trajectories for three randomly selected test set athletes in both the mid-career and end-of-career analyses are displayed in Section 8 of the Online Supplementary Material.

Table 7 Mean squared error between the observed test data and predicted values

Season	MSE
Mid-career	0.304
End-of-career	0.949

Mid-career: data on a mid-career season was held-out and predicted for a random selection of 100 athletes; End-of-career: data on the last season in career was held-out and predicted for a random selection of 100 athletes

6 Discussion

In this paper, we proposed an additive hierarchical Bayesian model for the analysis of athletes' performances in a longitudinal context. Following Montagna and Hopker (2018), we proposed a smooth functional contribution for explaining the overall variability in the data set. The functions are represented by means of a high-dimensional set of pre-specified basis functions, and a factor model on the basis coefficients ensures dimensionality reduction. We enriched the model by allowing for covariates to affect estimates through a linear model regression component. Finally, we addressed the issue of seasonal gathering of sports data by introducing a mixed effects model with GARCH errors which provides evolving random intercepts over different time intervals in the data set. To the best of our knowledge, our additive modelling strategy represents an original contribution to the sport analytics literature, with reasonable predictive performances and statistically significant estimation of performance-driving parameters.

The Bayesian latent factor methodology was originally developed for very sparse longitudinal data, with the purpose of capturing a global trend in subject-specific trajectories. We balanced the model with the requirement of smoothness using a B-spline basis system and adding a seasonal random intercept. However, it is evident that the latter explains the majority of variability in the dataset. Therefore, it might be worth considering a functional basis that batter captures the intra-seasonal variability. Further, we observed that the contribution of the regressive component is consistent across various modelling choices.

We looked at the effect of doping on results, by including it into the linear regression model-component as a covariate. Despite not being significant, the negative effect seems to suggest that performances of doped athletes are less variable. This is in line with previous literature suggesting that doping is more likely used to enhance performances in periods of decreasing fitness than to consolidate already good performances, generally exposed to strict controls. We think this aspect deserves further investigation, considering, for instance, more specific modelling techniques. We remark, however, that investigating doping further by relying on the current data set proves very challenging, both because the data is severely imbalanced (less than 3% of the athletes tested positive at least once throughout their careers), and also because we do not have information on when they tested positive during their career.

Finally, having recognised that the majority of the variability in the data is captured by the seasonal component, alternative modelling choices for the seasonal random intercepts are of interest. A possible direction of research consists in modelling the seasonal intercepts non-parametrically via a hierarchical Dirichlet process mixture model, with the aim of clustering athletes and seasons on the basis of performance. This model will allow to cluster data within a subject's career and across subjects, so to explain the majority of variability in the data set by means of a small number of driving parameters. The non-parametric procedure will further be able to learn the number of required clusters directly from the data.

While the motivation of our work comes from the analysis of shot put data, the methodology presented in this work is applicable to the analysis of performance data collected in all “centimeter-gram“ sports, that is, all those activities where results have the form of a distance, a weight, a speed and, in general, a measurable quantity, collected over time. In conclusion, the attempt of extending existing tools of functional data analysis to the modelling of (shot put) performance data seems promising because of the adaptability of these methodologies to all sorts of performance longitudinal data in measurable sports.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10260-022-00656-z.

Acknowledgements We would like to thank Prof. James Hopker (University of Kent, School of Sport and Exercise Sciences) for sharing the shot put data set, and for the helpful comments.

Funding Open access funding provided by Università Commerciale Luigi Bocconi within the CRUI-CARE Agreement.

Code Availability The code for the Bayesian analysis and the output production (numerical and graphical) is available at https://github.com/PatricDolmeta/Bayesian-GARCH-Modeling-of-Functional-Sports-Data.

Materials Availability The raw data set is available at www.tilastopaja.eu, whereas the data prepared for our analysis at https://github.com/PatricDolmeta/Bayesian-GARCH-Modeling-of-Functional-Sports-Data.

Declarations

Conflict of interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Bhattacharya A, Dunson D (2011) Sparse Bayesian infinite factor models. Biometrika 98:291–306. https://doi.org/10.2307/23076151

Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. J Econom 31(3):307–327

Casals M, Martinez J (2013) Modelling player performance in basketball through mixed models. Int J Perform Anal Sports 13:64–82. https://doi.org/10.1080/24748668.2013.11868632

de Boor C (1978) A practical guide to splines. Springer-Verlag, New York

Geisser S, Eddy WF (1979) A predictive approach to model selection. J Am Stat Assoc 74(365):153–160

Haario H, Saksman E, Tamminen J (2001) An adaptive metropolis algorithm. Bernoulli 7:223–242

Koulis T, Muthukumarana S, Briercliffe C (2014) A bayesian stochastic model for batting performance evaluation in one-day cricket. J Quant Anal Sports 10. https://doi.org/10.1515/jqas-2013-0057

Lee DJ, Page GL (2021) Big data in sports: Predictive models for basketball player’s performance. Math Ind Rep https://doi.org/10.33774/miir-2021-h4x62

Malcata R, Hopkins W, Pearson S (2014) Tracking career performance of successful triathletes. Med Sci Sports Exerc 46:1227–34

Marsden M (1974) Cubic spline interpolation of continuous functions. J Approx Theory 10(2):103–111

Montagna S, Tokdar S, Neelon B, Dunson D (2012) Bayesian latent factor regression for functional and longitudinal data. Biometrics 68:1064–1073. https://doi.org/10.2307/41806025

Montagna S, Hopker J (2018) A bayesian approach for the use of athlete performance data within anti-doping. Front Physiol 9

Page GL, Quintana FA (2015) Predictions based on the clustering of heterogeneous functions via shape and subject-specific covariates. Bayesian Anal 10(2):379–410. https://doi.org/10.1214/14-BA919

Scott MB, Shane CR, Patrick DL (1999) Bridging different eras in sports. J Am Stat Assoc 94(447):661–676. https://doi.org/10.1080/01621459.1999.10474163

Vaci N, Cocić D, Gula B, Bilalic M (2019) Large data and bayesian modeling-aging curves of nba players. Behav Res Methods 51:1544–1564

Wimmer V, Fenske N, Pyrka P, Fahrmeir L (2011) Exploring competition performance in decathlon using semi-parametric latent variable models. J Quant Anal Sports 7:6–6

Authors and Affiliations

Patric Dolmeta[1] · **Raffaele Argiento**[2,3] · **Silvia Montagna**[3,4]

✉ Patric Dolmeta
patric.dolmeta@phd.unibocconi.it

Raffaele Argiento
raffaele.argiento@unibg.it

Silvia Montagna
silvia.montagna@unito.it

[1] Department of Decision Sciences, Università commerciale Luigi Bocconi, Via Guglielmo Röntgen, 1, 20136 Milan, Italy

[2] Department of Economics, Università degli Studi di Bergamo, Via dei Caniana 2, 24127 Bergamo, Italy

[3] Collegio Carlo Alberto, Turin, Italy

[4] ESOMAS Department, Università degli Studi di Torino, C.so Unione Sovietica 218/bis, 10134 Turin, Italy