Bayesian functional mixed effects model for sports data

Modelli funzionali Bayesiani a effetti misti per dati sportivi

Patric Dolmeta, Raffaele Argiento and Silvia Montagna

Abstract The use of statistical methods in sport analytics is common practice nowadays. In this work, we propose a hierarchical Bayesian model for describing and predicting the evolution of performance over time for shot put athletes. We address seasonality and heterogeneity in results by means of a linear mixed effects model with heteroskedastic errors. The model provides an accurate description of the performance trajectories and allows for prediction of athletes' performance in future seasons. We apply our method to an extensive real world data set on performance data of professional shot put athletes recorded at elite competitions.

Abstract *L'impiego di metodi statistici per lo studio dello Sport è ormai largamente diffuso. In questo lavoro, proponiamo un modello Bayesiano gerarchico per la descrizione e la previsione di risultati nel tempo per lanciatori del peso. Grazie a un modello lineare ad effetti misti con errori eteroschedastici, affrontiamo stagionalità e eterogeneità nei risultati. Applichiamo il metodo ad un dataset reale di grandi dimensioni contenente i risultati del lancio del peso in un gran numero di competizioni internazionali e osserviamo una soddisfacente descrizione dei dati e la possibilità di quantificare l'incertezza nelle previsioni di performance future.*

Key words: Performance analysis, Bayesian functional data analysis, GARCH models, Sport analytics

Patric Dolmeta
Bocconi University, Via Roentgen 8, Milano e-mail: patric.dolmeta@unibocconi.it

Raffaele Argiento
Università degli Studi di Bergamo, Via dei Caniana 2, Bergamo e-mail: raffaele.argiento@unibg.it

Silvia Montagna
Università degli Studi di Torino, C.so Unione Sovietica 218/bis, Torino e-mail: silvia.montagna@unito.it

1 Introduction

Shot put is a track and field event involving throwing ("putting") the shot as far as possible. Shot put events range over the whole year, with indoor competitions held during Winter months and major tournaments during the Summer. So, results display some sort of seasonality, as it is common to all sportive competitions. We underline that here the term "seasonality" is not used to indicate a cyclical behaviour of observations over time, as in the literature of time series but, rather, a time dependent gathering of observations. On one hand competitions are traditionally concentrated in some months of the year, and on the other hand weather and environmental conditions may affect the performances or even the practicability of the sport itself. In shot put, it is reasonable to say that seasons coincides with calendar years and that taking seasonality effects into account is necessary to provide an accurate representation of the data.

In this work, we are interested in describing the evolution of performances of professional shot put athletes throughout their careers. We describe results of each athlete as error prone measurements of seasonal means trough a Bayesian mixed effects model, that describes the seasonal mean for each athlete as a deviation from a grand mean.

2 The World Athletics shot put data set

The data was obtained from an open results database (www.tilastopaja.eu) following institutional ethical approval (Prop_72_2017_18). The dataset comprises 41,000 measurements of World Athletics (the world governing body for track and field athletic sports) recognized elite shot put competitions for 653 athletes from 1996 to 2016. For each athlete, the data set reports the date of the event, the shot distance in meters, an indication of doping violation and some demographic information.

The outcome of interest is the shot distance. Data are collected over time: hereafter we will denote as t_{ij} the time at which the j-th 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. Having described seasons as calendar years, athletes will compete in a different number of seasons according to their career length. Figure 1 shows the number of athletes per season as well as boxplots of the distribution of their mean performances across the various seasons. A general increasing trend in performance can be observed as a function of career length (right panel in Figure 1).

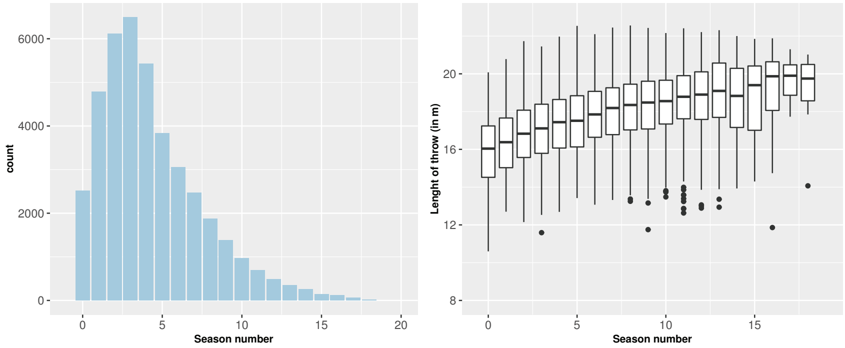

Fig. 1 Left: total number of athletes per season. Right: each boxplot shows the distribution of the athletes' mean performances within each season.

3 The model

Let n be the total number of athletes in the study. Shot put performances for athlete i collected over time represent temporal grouped data that we expect to be correlated when associated to the same individual and season. Hence, we propose a mixed-effects model for the shot put performances for athlete i at time point t_{ij} (y_{ij}), inducing a hierarchical structure in the error variance. Accordingly, y_{ij} are deviations from a seasonal mean function $\mu_i(\cdot)$ that takes constant value μ_{is} for any time point in a given season s:

$$y_{ij} = \mu_{is} + \varepsilon_{ij} \tag{1}$$

with $\varepsilon_{ij} \overset{iid}{\sim} N(0, \psi^2)$ independent errors recorded at time t_{ij}, for $j = 1, \dots, n_i$ and n_i is the total number of measurements available on athlete i. We further regard the seasonal- and unit-specific random intercept μ_{is} as the latent effect that quantifies the extent to which performances in season s responds above or below the overall mean m:

$$\mu_{is} = m + \zeta_{i,s}, \quad \text{with} \quad \zeta_{i,s} \overset{iid}{\sim} N(0, h_{i,s}) \tag{2}$$

Again, residuals at this higher level of hierarchy are assumed to be normally distributed, uncorrelated with lower-level residuals, but not uncorrelated among themselves. Indeed, the assumption of homogeneity of variance is inadequate here. Early graphical displays, straightforward exploratory analysis and initial modeling choices suggest a significant variability of the average response and its variance across seasons. In particular, suppose the variance of some athlete's performances during a specific time interval is known, then it will provide insight on future variability. Even further, a history of volatility in results provides information about an athlete's potential more than a background of constant, close to the average, performances. Hence, we consider a random intercept model with Normal Generalized Autoregressive Conditional Heteroskedastic (GARCH) errors [1]. Specifically,

$$\mu_{is} \mid m, h_{is} = m + \zeta_{is} \overset{iid}{\sim} N(m, h_{is}) \tag{3}$$

$$h_{is} = \alpha_0 + \alpha_1 \zeta_{is-1}^2 + \varpi h_{is-1} \tag{4}$$

where $\alpha_0 > 0, \alpha_1 \geq 0$ and $\varpi \geq 0$ 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

$$E(\zeta_t) = 0 \tag{5}$$

$$Var(\zeta_t) = \alpha_0(1 - \alpha_1 - \varpi)^{-1} \tag{6}$$

$$Cov(\zeta_t, \zeta_s) = 0 \text{ for } t \neq s \tag{7}$$

is guaranteed by requiring $\alpha_1 + \varpi < 1$, as proven by [1].
Three parameters of the seasonal component require prior specification: the overall mean m and the conditional variance parameters, ϖ and $\vec{\alpha} = (\alpha_0, \alpha_1)^\top$. 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 centered around the mean suggested by posterior analysis of preliminary versions of the model. In particular:

$$m \sim N(\mu_{m_0}, \Sigma_{m_0}) \tag{8}$$

$$\vec{\alpha} \sim N_2(\mu_\alpha, \Sigma_\alpha)\ I\{\vec{\alpha} > 0\} \tag{9}$$

$$\varpi \sim N(\mu_\varpi, \Sigma_\varpi)\ I\{\varpi \geq 0\} \tag{10}$$

where $\vec{\alpha} = (\alpha_0, \alpha_1)$ 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 $\alpha_1 + \varpi < 1$ holds true.
Because of the recursive definition of the conditional variance, no conjugate model exists for the GARCH parameters. Hence, we rely on an adaptive version of the Metropolis Hastings algorithm for posterior updates. In particular, for parameters m, ϖ and $\vec{\alpha}$ 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 [2]. We ran our sampler for $50{,}000$ iterations, with a burn-in period of 50% and a thinning of 5. We analyzed posterior samples for a variety of function estimates at different time points and for a variety of athletes, and the other model parameters. No issues emerged regarding convergence and mixing of the chains.

4 Results

Our goal is in estimating trajectories for athletes' performances. To this end, we generate a fine grid of equispaced time points, $\{t_k\}_{k=1}^T$, over our time span and evaluate the trajectory in performance on this grid.
Since we modelled the seasonal mixed effects function as a piecewise continuous function taking individual- and season-specific values, when estimating such function on any point in the time grid, we need to determine which season the time point belongs to. As discussed in Section 2, 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. In the following Equation, the indicator variable $\chi_{(t\in s)}$ determines to which season each time point t belongs to and index g ranges over the total number of iterations G:

$$\widehat{y_i(t)} = \frac{1}{G}\sum_{g=1}^{G}\sum_{s=1}^{S_i}\mu_{is}^{(g)}\chi_{(t\in s)} \quad \text{for } t = t_1,\ldots,t_T. \tag{11}$$

Equation (11) represents the point estimate of athlete i's performance at time t. Similarly, 95% credible intervals can be computed to quantify uncertainty around our point estimate. Estimated trajectories, credible bands and one season ahead prediction are displayed in Figure 2 for a random selection of shut put athletes.

5 Discussion

We proposed a hierarchical Bayesian model for the analysis of athletes' performances in a longitudinal context. We addressed the issue of seasonal gathering of sports data with a mixed effects model with GARCH errors, providing evolving random intercepts over different time intervals in the data set. While the motivation of our work comes from the analysis of shot put performance data, the methodology presented in this work is applicable to the analysis of performance data collected in all measurable sports.
The comprehensive nature of the data set suggests exploiting it further, possibly by including the contribution of covariates on the response. In particular, we believe there is potential for a better understanding of the effects of doping, not only on single performances, but on the overall evolution of a career. Further, we hope also doping detection might be targeted. Additional future developments include more sophisticated modeling choices both for the intraseasonal variability and the seasonal intercepts themselves. A nonparametric Bayesian approach to the hierarchy with the intent of clustering observation both across athletes and seasons is already forthcoming. This way we hope to recognize common patterns in similarly evolving careers for enhanced prediction purposes.

References

1. Bollerslev, T.: Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics **31**(3), 307 – 327 (1986)
2. Haario, H., Saksman, E., Tamminen, J.: An adaptive metropolis algorithm. Bernoulli **7**, 223–242 (2001)

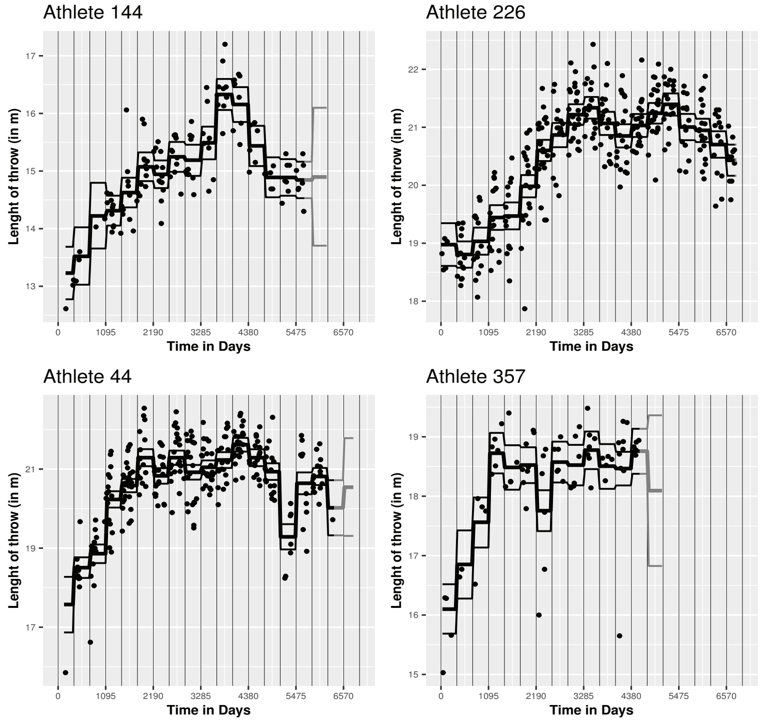

Fig. 2 Performance trajectory estimates for a random selection of athletes. The *x*-axis denotes the time measured in days from January 1st of the first season of career, whereas on the y-axis there is the length of throw in meters. Vertical lines represent calendar years (seasons in our notation). The final part of each trajectory (grey) for which no observations are available, represents one-season-ahead performance prediction.