Clustering Athlete Performances in Track and Field Sports

Clustering della Performance degli Atleti di Atletica Leggera

Raffaele Argiento and Alessandro Colombi and Lorenzo Modotti and Silvia Montagna

Abstract This study aims to cluster track and field athletes based on their average seasonal performance. Athletes' performance measurements are treated as random perturbations of an underlying individual step function with season-specific random intercepts. A hierarchical Dirichlet process is used as a nonparametric prior to induce clustering of the observations across seasons and athletes. By linking clusters across seasons, similarities and differences in performance are identified. Using a real-world longitudinal shot put data set, the method is illustrated.

Abstract *L'obiettivo di questo lavoro consiste nel raggruppare atleti di atletica leggera in base alla loro performance stagionale. Le misurazioni di performance degli atleti sono trattate come perturbazioni casuali di una funzione a tratti individuale con intercette casuali stagionali. Si usa un processo nonparametrico di Dirichlet gerarchico a priori per raggruppare le osservazioni tra stagioni e atleti. Unendo i cluster inter-stagionali, si identificano somiglianze e differenze di performance. Il metodo è illustrato utilizzando un dataset reale longitudinale di lancio del peso.*

Key words: Hierarchical Dirichlet process, Longitudinal data analysis, Nonparametric Bayesian modelling, Sports analytics

---

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

Alessandro Colombi
Università degli Studi di Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, Milano, e-mail: a.colombi10@campus.unimib.it

Lorenzo Modotti
Columbia University, 665 West 130th Street, New York, e-mail: lmodotti27@gsb.columbia.edu

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

1 Introduction

Sports analytics employ data and quantitative methods to measure and analyse athletes and teams performance in professional sports, typically with predictive purposes. Specifically, the use of hierarchical Bayesian methods is gaining popularity in sports analytics, as they allow information sharing across time and athletes [1]. The applied motivation for this work comes from a longitudinal dataset of professional shot put athletes, whose performance was measured at each event they competed at, along with their demographic information. This work aims to cluster athletes with a similar average performance within a season and study the evolution in athletes' performance throughout their careers. We rely on a hierarchical Dirichlet process [3] model to generate ties across seasons and athletes. Within each season, athletes sharing the same latent model parameters belong to the same cluster and thus have the same seasonal average performance. The latent parameters are also shared across seasons, thereby also allowing for the global clustering of athletes.

2 Shot put dataset

Shot put is a track and field event in which athletes have to *put* (throw) a heavy spherical ball, the *shot*, as far as they can. Our data set[1] consists of measurements (the throw lengths or marks) recorded during the professional shot put competitions from 1997 to 2016. During these 19 seasons, 41,033 measurements on 653 athletes have been recorded. For each athlete taking part in a competition, a record comprising the mark, the finishing position, information regarding the competition itself and the shot putter is stored. Relevant covariates are the event date, the environment (indoor or outdoor event), sex and age of the athlete.

Since the aim is to model the mean seasonal performance for each shot putter, the season number associated with each observation corresponds to the number of seasons elapsed since the beginning of the athlete's career, also counting seasons in which the athlete did not compete. Thus, season 1 consists of measurements observed during each athlete's first season of activity, regardless of the calendar year in which these have been recorded. In this way, observations result in being grouped in seasons that reflect the athletes' years of experience. Figure 1 depicts the evolution in performance throughout the athlete's career for two randomly selected shot putters from the dataset. Different athletes take part in different competitions, and the length and profile of their performance careers vary. Clearly, performance is expected to vary throughout an athlete's career, but, as evident from this plot, it is rather steady within each season. This suggests that, despite being rough approxima-

[1] Available at `https://github.com/PatricDolmeta/Bayesian-GARCH-Modeling-of-Functional-Sports-Data/blob/main/COMPLETE_DATA.txt`.

tions, the seasonal mean performances capture the essential features of the athletes' careers.

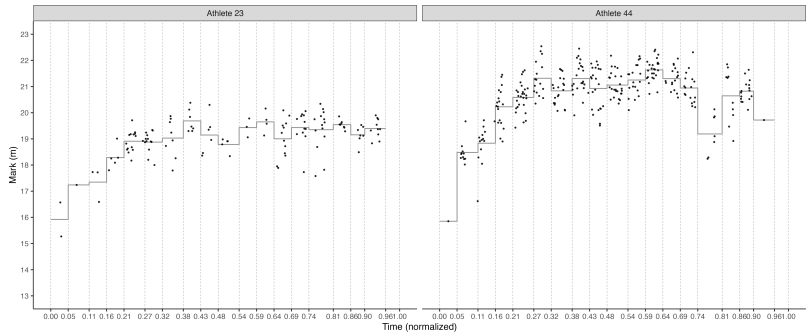

Fig. 1 Shot put measurements collected throughout an athlete's career for two randomly selected athletes with empirical seasonal mean performance (solid grey line). Dotted grey lines delimit seasons.

3 Proposed methodology, prior elicitation and posterior inference

We suppose that n_s athletes compete in season s, with $s \in \{1,\dots,S\}$ and $S = 19$ in our application. Each active athlete i in season s, $i \in \{1,\dots,n_s\}$, takes part in N_{si} events. At each event $j \in \{1,\dots,N_{si}\}$ the athlete's mark Y_{sij} is measured. For simplicity, we rescale time so that measurements are collected at $t_{sij} \in [0,1]$. Moreover, a set of d covariates is available, $\boldsymbol{x}_{sij} := \boldsymbol{x}_i(t_{sij}) = \left[x_{sij}^{(1)},\dots,x_{sij}^{(d)}\right]^\top$. Assuming that observations are noisy measurements of an underlying athlete-specific function, a general model for these data is

$$Y_{sij} = g_i(t_{sij};x_{sij}) + \varepsilon_{sij}, \tag{1}$$

with $\varepsilon_{sij} \overset{\text{iid}}{\sim} \text{N}\left(0, \frac{1}{\tau_{si}}\right)$, where τ_{si} denotes the precision of the distribution. Suppose that the athlete-specific functions are piecewise constant that is

$$g_i(t_{sij}) = \sum_{s=1}^{S} \mu_{si} \mathbb{1}_{[t^{(s)},t^{(s+1)})}(t_{sij}) + \boldsymbol{x}_i(t_{sij})^\top \boldsymbol{\beta}_s, \tag{2}$$

where μ_{si} is a season-specific random intercept, $t^{(s)} := \min_{i,j} t_{sij}$ is the beginning of each season s, $t^{(S+1)} := \max_{i,j} t_{Sij}$ is the end of season S, and $\boldsymbol{\beta}_s$ is a d-dimensional vector of regression parameters, shared among all the athletes in season s. Therefore, within each season s, the athlete's observations are normally distributed as $Y_{sij} \,|\, \mu_{si}, \tau_{si}, \boldsymbol{\beta}_s; \boldsymbol{x}_{sij} \overset{\text{ind}}{\sim} \text{N}\left(\mu_{si} + \boldsymbol{x}_{sij}^\top \boldsymbol{\beta}_s, \frac{1}{\tau_{si}}\right)$. Let $\theta_{si} := (\mu_{si}, \tau_{si})$; the sampling model

(1)-(2) is completed with the hierarchical Dirichlet prior [4] for θ_{si}, i.e.,

$$\theta_{si} \,|\, P_s \overset{\text{iid}}{\sim} P_s; \quad P_s \,|\, \alpha_0, P_0 \overset{\text{iid}}{\sim} \text{DP}(\alpha_0 P_0); \quad P_0 \,|\, \alpha, H \sim \text{DP}(\alpha H),$$

where $\text{DP}(\alpha H)$ denotes a Dirichlet process with concentration parameter α and base distribution H. Assuming conjugacy, the baseline distribution is a Normal-Gamma $H \sim \text{NG}\left(\mu_0, p_0, \frac{\nu_0}{2}, \frac{\nu_0}{2}\xi_0^2\right)$. Finally, we assume $\boldsymbol{\beta}_s \overset{\text{iid}}{\sim} \text{N}_d\left(\boldsymbol{\beta}_0, \Sigma_0\right)$, where N_d denotes the d-dimensional Normal distribution and Σ_0 is its variance-covariance matrix, $\alpha \sim \text{Gamma}(a,b)$, and $\alpha_0 \sim \text{Gamma}(a_0, b_0)$. For posterior computation, we exploit the Chinese restaurant franchise representation of the hierarchical Dirichlet process [3], which allows us to design a Markov chain Monte Carlo sampling scheme for the model above. In this metaphor, customers are represented by parameters θ_{si} and seasons are represented as restaurants. Customers are clustered into tables within each restaurant, and these tables are further clustered into dishes. Observations are clustered across restaurants at the second level of the clustering process restaurants when dishes are associated with tables. One can think that the first customer sitting at each table chooses a dish from a common menu, which is then shared by all subsequent customers at that table. As usual in model-based clustering, we say that two observations, say (s,i) and (l,j), belong to the same cluster if $\theta_{si} = \theta_{lj}$. Under the hierarchical Dirichlet process, the values of the parameters are shared within the seasons, e.g $\theta_{si} = \theta_{sj}$, as well as between the seasons, leading to two-levels, model-based, clustering of the athletes.

4 Results

In this Section, we present the results obtained on the shot put dataset. The hyperparameters of the baseline distribution H were chosen setting $\mu_0 \coloneqq \bar{\boldsymbol{y}} = 0.0$, $p_0 \coloneqq \frac{1}{\text{range}(\boldsymbol{y})^2} = 0.002250395$, $\nu_0 \coloneqq 2$, $\xi_0^2 \coloneqq 0.5$, where $\boldsymbol{y}$ denotes the whole set of observations across athletes and seasons. We included three covariates: sex, age (centered around the global mean) and environment. The hyperparameters of the prior distribution for the multiple regression parameters were set to $\boldsymbol{\beta}_0 \coloneqq [0,0,0]^\top$, $\Sigma_0 \coloneqq I$. The hyperparameters of the prior distributions for the concentration parameters α and α_0 were set to $a = a_0 \coloneqq 1$, $b = b_0 \coloneqq 8$. After a burn-in of $10,000$ iterations, $50,000$ samples have been retained. We examined the estimated posterior distributions of the regression parameters. For the age covariate, it appears that at the beginning of their career, athletes who are older than the mean tend to perform better, probably due to the different stage of physical development. This difference in performance vanishes during their mid-career, but it may become relevant again in the last part. Due to page constraints, we omit other results on covariates from this work as less insightful.

Concerning clustering, 12 global clusters have been found. The estimated locations μ are well dispersed across the range of the observations, while most of the precisions τ are concentrated around 2 and 4. Regarding seasonal clustering, the ac-

tive estimated components within each season are reported in Table 1: the smallest number of clusters per season is 3 (seasons 18 and 19), while the largest is 9 (seasons 7, 8 and 10). Generally, the estimated components are shared across different seasons, as desired. The resulting clustering is illustrated in Figure 2. Despite obser-

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
2	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•			•	
3	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		•	•	•
4	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		
5	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		
6	•		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		
7			•				•					•		•					
8				•	•	•	•		•										
9					•		•	•	•	•	•	•	•	•	•				•
10						•		•									•		
11								•		•	•		•		•				
12										•									
Total	6	5	7	7	8	8	9	9	8	9	8	8	8	8	8	4	6	3	3

Table 1 Active estimated components (rows) per season (columns).

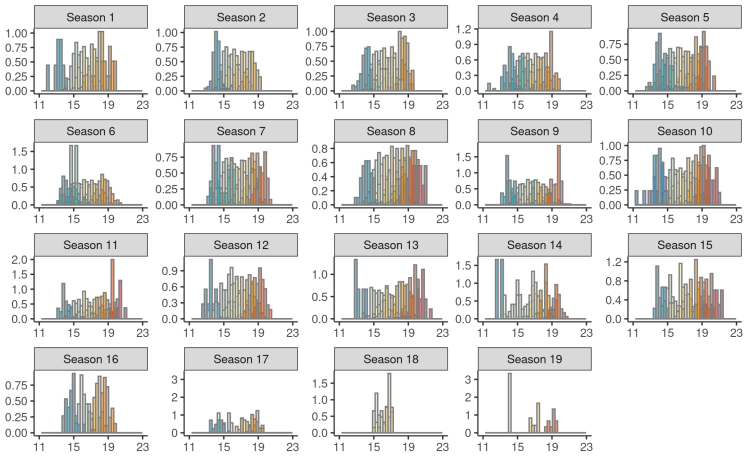

Fig. 2 Histograms of the shot put observations for all athletes split in the 15 seasons and coloured according to the estimated clusters' memberships.

vations are numerous and overlapping within each season, the model has been able to retrieve a reasonable clustering. Figure 3 provides evidence for the goodness of fit at the individual level. For most seasons, the two athletes obtained similar results and indeed they have been clustered together. In seasons 8, 12 and 13 the two shot putters have been assigned to different clusters. By inspecting the plot it seems that athlete 55 performed slightly better than 94 in these seasons, on average. Overall, the model fits the data well. Both local and global clustering of the athletes are reasonable, and the estimated number of clusters allows to model precisely the mean seasonal performance of the shot putters, without overfitting the data.

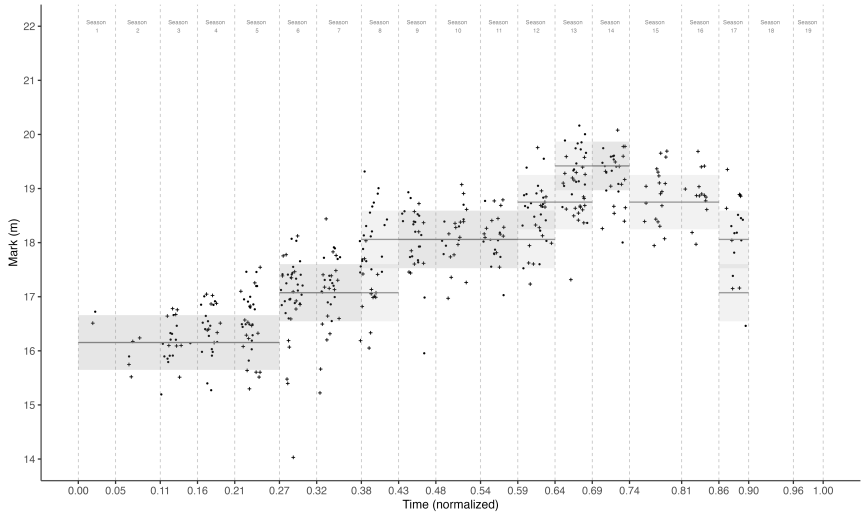

Fig. 3 Shot put measurements for athletes 55 (dot) and 94 (plus) with estimated mean (solid line) of the assigned clusters. The shaded areas comprise values within one estimated standard deviation from the estimated mean.

5 Conclusions

In this work, local and global clustering of longitudinal sports data have been investigated by employing a hierarchical Dirichlet process mixture model. Several extensions of this model may be investigated, for example the hierarchical Dirichlet process could be replaced by more general nonparametric priors. Further, the entire longitudinal curve could be used to induce clustering instead as in [2].

References

1. Baio G., Blangiardo M.: Bayesian hierarchical model for the prediction of football results. Journal of Applied Statistics. **37**, 253–264 (2010)
2. Page, G.L., Quintana, F.A.: Predictions Based on the Clustering of Heterogeneous Functions via Shape and Subject-Specific Covariates. Bayesian Analysis **10**, 379–410 (2015)
3. Teh Y.W., Jordan M.I., Beal M.J., Blei D.M.: Hierarchical Dirichlet Processes. Journal of the American Statistical Association. **101**, 1566–1581 (2006)
4. Teh Y.W., Jordan M.I.: Hierarchical Bayesian nonparametric models with applications. In: Hjort, N.L., Holmes, C., Müller, P., Walker, S.G. (eds.) Bayesian Nonparametrics, pp. 158–207. Cambridge University Press (2010)