Annals of Operations Research (2023) 325:167–183
https://doi.org/10.1007/s10479-022-04660-0

ORIGINAL RESEARCH

Football tracking data: a copula-based hidden Markov model for classification of tactics in football

Marius Ötting[1] · Dimitris Karlis[2]

Accepted: 14 February 2022 / Published online: 24 April 2022

Abstract

Driven by recent advances in technology, tracking devices allow to collect high-frequency data on the position of players in (association) football matches and in many other sports. Although such data sets are available to every professional team, most teams still rely on time-consuming video analysis when analysing future opponents, for example with regard to how goals were scored or a team's general style of play. In this contribution, we provide a data-driven approach for automated classification of tactics in football. For that purpose, we consider hidden Markov models (HMMs) to analyse high-frequency tracking data, where the underlying states serve for a team's tactic. In particular, as space control in football has been considered a major driver of success, we focus on the effective playing space, which is the convex hull created by the players excluding the goalkeeper. This quantity relates to both playing style and team behavior. Using copula-based HMMs, we model jointly the effective playing space of both teams to account for the competitive nature of the game. Our model thus provides an estimate of a team's playing style at each time point, which can be beneficial for team managers but also of huge interest to football fans.

Keywords Copula · Football · Hidden Markov model · Sports analytics · Tracking data

1 Introduction

Recent advances in technology have produced tremendous changes in several areas of society, and sports is not an exception. In many sports, tracking devices nowadays allow to produce huge amounts of data during a match. Such data sets cover the positions of players and the ball, often sampled at 1–25 Hz, thus producing fine-grained tracking data and trajectories of all players. The analysis of such data can help coaches and scouts in several aspects by producing considerable insights into, *inter alia*, game strategy and tactics, player evaluation,

✉ Dimitris Karlis
karlis@aueb.gr

Marius Ötting
marius.oetting@uni-bielefeld.de

[1] Bielefeld University, Bielefeld, Germany

[2] Athens University of Economics and Business, Athens, Greece

goal analysis, judging referee decision, and talent identification. There has been tremendous increase of interest on such data as they contribute to effectively evaluate the performance at both individual and team level in team sports.

In sports like American football and basketball, tracking data have been considered to investigate teams' strategies (Lopez, 2020) and also to create novel measures thereof (Cervone et al., 2016; Franks et al., 2015). Such investigations require tracking data, as analyses based on play-level data do not include variables that impact on-field behavior very often. Also in (association) football, the analysis of tracking data has increased rapidly in recent years. Corresponding studies investigate passing performance (see, e.g., Kempe et al., 2018; Goes et al., 2019), team formation (see, e.g., Memmert et al., 2019; Frencken et al., 2011), and space creation (Fernandez & Bornn, 2018), to name few. Goes et al. (2021b) provide an overview on the existing literature on football tracking data.

Whereas the existing literature on football tracking data mostly focuses on analysing certain events in a match such as passes or goals, this paper aims at providing a tool for classifying a team's playing style. For that purpose, we do not focus on single events, but instead model an *entire* match. We believe that such tool is useful for managers to analyse future opponents, as most teams rely on time-consuming video analysis when investigating, for example, an opponent's style of play. By considering a data-driven approach instead, teams could benefit from a quantitative and hence more objective analysis of future opponents, which can be carried out much faster compared to video analysis. To provide such an automated classification of a team's playing style, we consider a high-resolution tracking data set. Specifically, we analyse the convex hull spanned by the players of a team, as this metric comprises several tactical aspects. When a team is in offense it is reasonable to increase its convex hull, whereas the opposite seems intuitively plausible when in defense. To analyse our data, we consider hidden Markov models (HMMs), where the states can be interpreted as a team's underlying tactics (such as offensive vs. defensive style of play). The decoded states of the fitted HMM enable a comprehensive analysis of the two teams. In particular, events in a match such as scored goals and clear-cut chances can be linked to the decoded states (i.e. to the underlying tactics), such that managers can very easily analyse whether a goal was scored from, for example, usual build-up play or from a counter attack. Also the time between changing the states can reveal the speed of a team to adjust to the circumstances of the match, and this transition can also characterize the success of the play.

An exploratory analysis of our tracking data is presented in Sect. 2. Section 3 covers the methodological background on HMMs. As a starting point, we consider two separate univariate HMMs for the two teams. To allow for potential interactions between the teams, we build a copula-based bivariate HMM. Section 4 presents the corresponding results. Moreover, with the decoded states enabling an automated classification of a team's playing style, we provide further use cases of our approach in practice, which may be of huge interest to all stakeholders including team managers as well as football fans and consumers.

2 Data

2.1 About the data

The tracking data set considered in this contribution was provided by the company Metrica Sports, who uploaded the data to Github—see https://github.com/metrica-sports/sample-data. Since Metrica Sports anonymised the data, there is no further information given when and in which league the match has taken place. The data set includes the (x, y) positions of all players and the ball, which are sampled with a resolution of 25 Hz. Together with these information, the data also include additional timestamps for events such as goals, shots on goal, and corner kicks. The original data have been preprocessed to remove time intervals where the match stopped (e.g. after the goal for one team when all players celebrate together, or when the match was stopped for a free kick). Since this is not useful information, we removed these observations.

2.2 Effective playing space

Space and its control is recognized as an important element for success in football. For this reason several metrics have been considered to measure the space usage from the teams and how this can affect the outcome (see, e.g., Silva et al., 2014; Ric et al., 2017; Fernandez & Bornn, 2018).

In this contribution, we focus on a specific metric derived from player tracking data, which aims at examining several tactical aspects. Our metric is the convex hull created by the players of a team excluding the goalkeeper. It is also referred as effective playing space (EPS) (Memmert et al., 2019), calculated as the surface area (in square meters) of the convex hull of all players of the team excluding the goalkeepers as a measure of the playing area used by the team in a given situation.

Strategically, a team needs to spread out its players so that the entire field is adequately covered. When a team is in offense it is reasonable to attempt to increase the convex hull, the opposite is intuitively plausible when in defense. The strategy of the attacking team is to have a large hull or at least players not too close such that the defending team needs to increase its size. Aside from such standard offense and defense strategies, the EPS also concerns more involved tactics. In particular, football teams often play a somewhat offensive style of defending, where the ball carrier is attacked and opposing players are marked—such a style of play is often summarised as "pressing". In terms of the EPS, attacking the ball carrier and marking opposing players results in larger EPS values for the defending team compared to a standard defensive style of play. Thus, considering the interplay between the hulls of the offensive and defensive team is important to understand the tactics. Furthermore, we believe that the EPS relates to several other game characteristics:

1. Different team formations result in different convex hull sizes as a result of the positions of the players.
2. The EPS also relates to the way the team attacks. For example, for wing play teams, the corresponding EPS is likely to be larger than for teams which mostly attack through the middle.
3. Observing the EPS as a a time series, as data are available in very small time granularity, can help to investigate the speed of transition from the defense to offense.
4. Further insights can be derived by analysing how the EPS relates to specific events in the match such as goals and goal-scoring chances.

The EPS has been considered in previous studies, where it was investigated together with the centroid positions of a team, and used as predictor for the match outcome (Frencken et al., 2011; Goes et al., 2021a). Also its relationship with team formation is examined in Baptista et al. (2020).

With the high-resolution tracking data at hand, we can easily derive the EPS from the (x, y) positions of all players. Figure 1 shows the convex hulls of both teams together with the individual positions of all players for four example situations found in our data. These

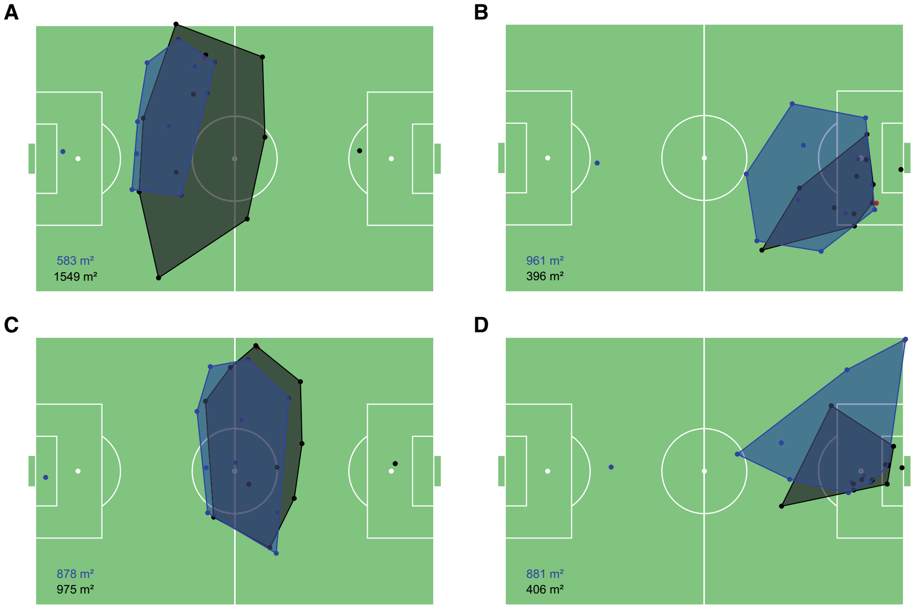

Fig. 1 Positions of players and their corresponding convex hulls for four example situations. (**A**) team B is attacking with a fairly large convex hull, (**B**) team A is attacking and scores a goal, (**C**) team B is in build-up play and team A is pressing, (**D**) team A has a corner kick

examples refer to different match situations, and the EPS also varies substantially across these situations. Panel (A) in Fig. 1 shows a situation where the team shown in black (team B) is attacking with a fairly large convex hull of about 1,500 square meters. In panel (B), the team shown in blue (team A) is attacking, and the player having the ball scores a goal in that situation. Comparing the convex hulls between the two attacking teams in panels (A) and (B), we observe a substantially lower EPS for team A than for team B. In panel (C), team B is in build-up play and team A is pressing, which is different to panel (A), where team A is playing more defensive. Comparing these two situations in more detail—i.e. panels (A) and (C)—we observe that team B has a larger EPS than team A in both situations. For team A, the EPS is increased substantially when pressing (panel C) compared to when playing more defensive (panel A). We observe such an increase in the EPS for team A since team A attacks the ball carrier and marks the opposing players, which is not the case for the situation shown in panel (A). Panel (D) refers to a corner kick from team A.

The examples above reveal that there is a substantial amount of interactions between the EPS of the two teams which can be revealing about their tactics. This is further underlined by Fig. 2, where we observe a negative correlation between the two teams' EPS. When inspecting the EPS when goals were scored (highlighted in Fig. 2 with different symbols), we observe that the goal-scoring team always had a larger EPS than the opposing team—we thus believe that jointly examining the EPS can provide helpful insights into a team's tactics. For the full sample, the means for the EPS are 980 for team A (min: 60, max: 2241) and 931 for team B (min: 60, max: 2524). From the histograms in Fig. 3 we see that values of the EPS between 500 and 1500 are most likely, while fairly large values ($> 1500\text{m}^2$) are more likely for team A than for team B. The resulting bivariate time series is shown in Fig. 4—note that we consider only time points where the ball was in play, totalling in 84,253 observations. For

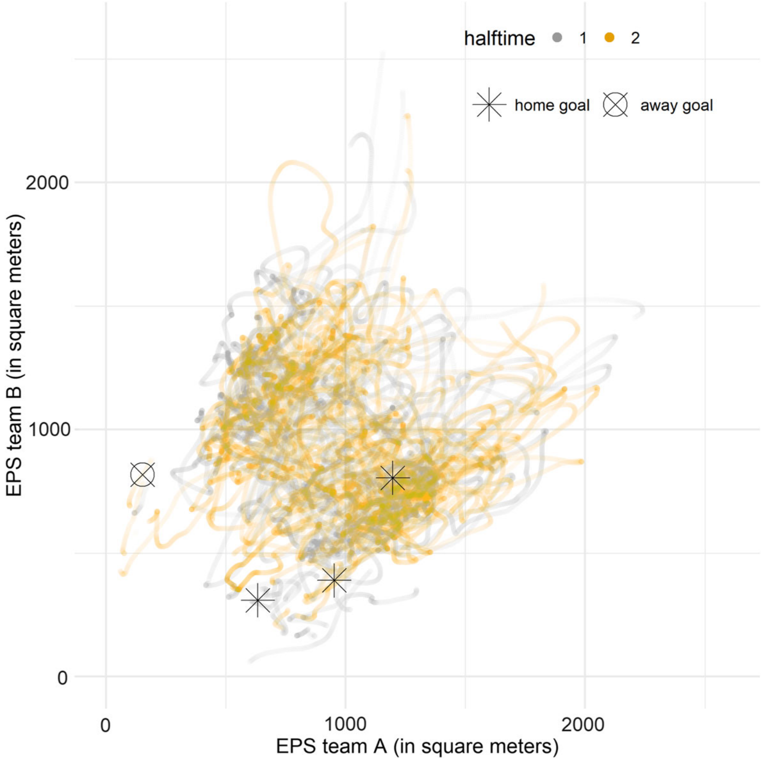

Fig. 2 Scatter plot showing the two teams' EPS. Colours indicate whether observations belong to the first (grey) or second half (yellow), and goals scored are highlighted with different point shapes

example, around the 25th minute we see that for about 2 minutes no values were recorded since there was a short stoppage (potentially due to an injury) at that time.

3 Methods

3.1 A univariate HMM for the EPS

Figure 4 further underlines that there are periods in the match where both teams' EPS are fairly high (e.g. around minute 80), as well as periods where one team has a fairly low EPS, while we observe a high EPS for the other team (e.g. around minute 50). HMMs thus constitute a natural modelling approach for our time series data, as they accommodate the idea of a match progressing through different phases, with potentially changing tactics of the two teams.

HMMs are a very flexible statistical tool for modelling time series data (Zucchini et al., 2016). In the basic model formulation, HMMs involve two components: an unobserved Markov chain with N possible states, and an observed state-dependent process, whose observations are assumed to be generated by one of N distributions as selected by the Markov chain. Here, the states can be interpreted as the underlying tactics of a team. In the simplest model formulation with two states, the states could, for example, be interpreted as either the

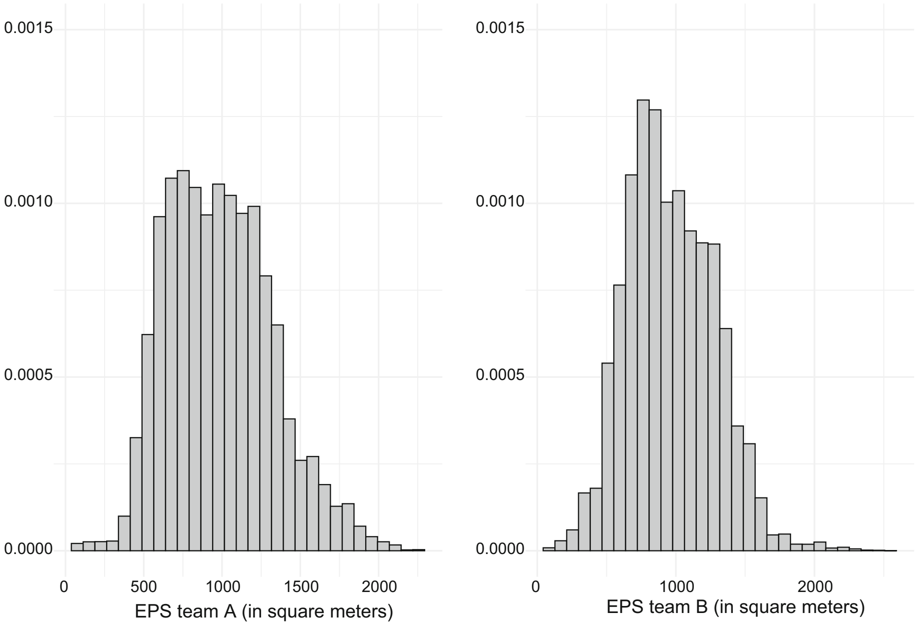

Fig. 3 Histogram of the EPS for team A (left panel) and team B (right panel)

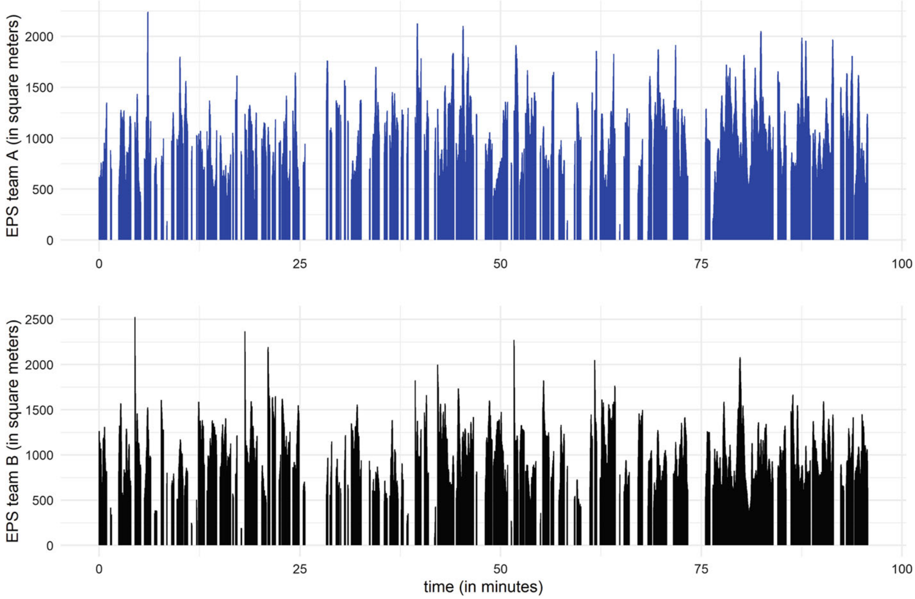

Fig. 4 Time series of the EPS for team A (top) and team B (bottom)

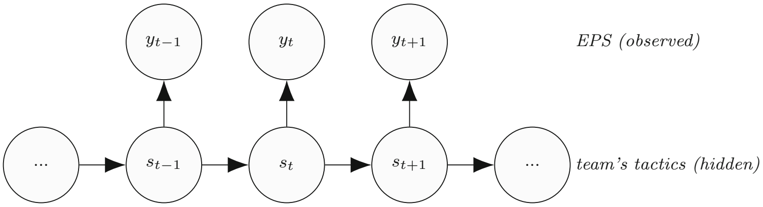

Fig. 5 Dependence structure of the HMM considered: each EPS observation y_t is assumed to be generated by one of N distributions according to the state process s_t

team considered or the opponent having a high level of control over the pitch, as indicated by the EPS. Figure 5 shows the model structure as directed graph.

For the EPS data considered in this contribution, the observations and the state process are denoted by y_t and s_t, $t = 1, 2, \dots, T$, respectively. Switches between the state are modelled by the transition probability matrix (t.p.m.) $\mathbf{\Gamma} = \{\gamma_{ij}\}$, where $\gamma_{ij} = \Pr(s_t = j | s_{t-1} = i)$, $i, j = 1, \dots, N$. Since it may very well be the case that a football match does not start in its stationary distribution, we estimate the initial distribution $\boldsymbol{\delta} = \big(\Pr(s_1 = 1), \dots, \Pr(s_1 = N)\big)$, thus resulting in $N - 1$ additional parameters to be estimated. For the model formulation of an HMM to be completed, the number of states N and the class(es) of state-dependent distribution(s) have to be selected. Since y_t (i.e. the EPS) is strictly positive and continuous, we assume a gamma distribution here. The density of the gamma distribution is given by

$$f(y) = \frac{\beta^{\alpha} y^{\alpha-1} e^{-\beta y}}{\Gamma(\alpha)}, \quad y > 0,$$

with scale and shape parameters $\alpha > 0$ and $\beta > 0$, and $\Gamma(\alpha)$ is the gamma function. Since the scale and shape parameters are somewhat cumbersome to interpret, we consider an alternative parameterisation of the gamma distribution with mean μ and standard deviation σ. Specifically, from μ and σ we can obtain α and β as follows:

$$\alpha = \frac{\mu^2}{\sigma^2}, \ \beta = \frac{\sigma^2}{\mu}.$$

The state-dependent densities are contained in the $N \times N$ diagonal matrix $\mathbf{P}(y_t)$, where the i-th diagonal element consists of the density of the observation y_t given state i. The likelihood for our HMM is given by:

$$L = \boldsymbol{\delta}\mathbf{P}(y_1)\mathbf{\Gamma}\mathbf{P}(y_2)\dots\mathbf{\Gamma}\mathbf{P}(y_T)\mathbf{1} \tag{1}$$

with column vector $\mathbf{1} = (1, \dots, 1)' \in \mathbb{R}^N$ (see Zucchini et al., 2016). Calculation of this matrix product expression amounts to running the forward algorithm, which is a recursive tool for efficiently calculating the likelihood of an HMM at low computational cost of $\mathcal{O}(TN^2)$ (see Zucchini et al., 2016). The likelihood is implemented in the statistical software R and maximised numerically (R Core Team, 2021). Below we provide further details on model fitting.

With a fitted HMM at hand, we can decode the underlying states to infer a team's tactics. Specifically, we investigate the most likely trajectory of states using the Viterbi algorithm and seek

$$(s_1^*, \dots, s_T^*) = \underset{s_1,\dots,s_T}{\operatorname{argmax}} \ \Pr(s_1, \dots, s_T | y_1, \dots, y_T),$$

i.e. the most likely state sequence, given the observations of the EPS. Maximising this probability, i.e. finding the optimal of N^T possible state sequences, can be achieved at computational cost $\mathcal{O}(TN^2)$ using the Viterbi algorithm (see Zucchini et al., 2016). The decoded states $s_1^*, \ldots, s_T^*$ are then further investigated and linked to events in the match such as goals.

3.2 A bivariate HMM for the EPS using copulas

In the previous subsection, the two teams' EPS were treated separately which may neglect insightful information on the dependence between the teams' tactics. To overcome these limitations, we model jointly the two teams' EPS $\mathbf{y}_t = (y_{t,\mathrm{A}}, y_{t,\mathrm{B}})$ by assuming a bivariate gamma distribution for each state. This bivariate gamma distribution will be created using copulas, further details about copulas can be found in Joe (2014). Copulas allow to built bivariate (multivariate) models in a flexible way separating the marginal properties to those of dependence. Hence, in our case we can introduce dependence between the two teams' EPS while at the same time we keep control on the marginal properties, which are assumed to be gamma distributions. We aim at using a state-dependent distribution with joint density

$$f(\mathbf{y}_t \mid s_t) = f_1(y_{t,\mathrm{A}} \mid s_t) f_2(y_{t,\mathrm{B}} \mid s_t) c_\theta \Big(F_1(y_{t,\mathrm{A}} \mid s_t), F_2(y_{t,\mathrm{B}} \mid s_t) \Big) \tag{2}$$

where

$$c_\theta(u, v) = \frac{\theta(e^\theta - 1)e^{\theta(1+u+v)}}{\left[e^\theta - e^{(\theta+\theta u)} + e^{\theta(u+v)} - e^{(\theta+\theta v)}\right]^2},$$

i.e. the density of the Frank copula. We further have that $\theta \in \mathcal{R} \setminus \{0\}$. In our case we assume that the marginals are gamma distributions, i.e. F_1, F_2 and f_1, f_2 in Eq. (2) denote the c.d.f. and the p.d.f. of the gamma distribution, respectively, with mean μ_i and σ_i for state i.

As hidden Markov models are special cases of finite mixture models, the model extends the ideas in Kosmidis and Karlis (2016) where copulas were used to built model based clustering with finite mixtures of distributions. HMMs in combination with copulas have previously been used by Härdle et al. (2015), Orfanogiannaki and Karlis (2018), Martino et al. (2020), and Ötting et al. (2021). Here, we have selected the Frank copula for illustration purpose, as in principle any other copula can be used. However, note that since our values are bounded, because of the context, we believe that tail dependence is not present, and the Frank copula is a simple copula that can take care of the dependence in a sufficient way. While using any other copula is possible, optimal copula selection is beyond the scope of the present paper.

To formulate the likelihood for our bivariate HMM, the i-th diagonal element of the $N \times N$ diagonal matrix $\mathbf{P}(y_t)$ in Eq. (1) now consists of the density of the observations given state i as presented in Eq. (2). Since no standard software packages are available for fitting copula-based HMMs, we implement the likelihood in the statistical software `R`. To obtain parameter estimates, the likelihood is numerically maximised using the function `nlm()`. To avoid local maxima, we used 30 sets of starting values for the numerical maximisation by drawing random numbers from uniform distributions. We select the range of starting values for the state-dependent distributions via exploratory analysis. Based on these 30 fitted models, we select the model with the highest likelihood. For a copula-based HMM with $N = 3$ states, model fitting took about 20 minutes on a standard desktop computer.

4 Results

This section starts with the results of the univariate HMM fitted separately to the two teams' tracking data. We first consider the simplest model formulation, that is an HMM with only two states, which can be regarded as a simple baseline model. In the remainder of this section, we will gradually increase the model's complexity to provide more profound insights into the tactics of the two teams.

4.1 Univariate HMM for the EPS

For the baseline HMM with two states, the means for the state-dependent gamma distributions are obtained as $\hat{\boldsymbol{\mu}}_{\text{A}} = (1251, 704)$ for team A and $\hat{\boldsymbol{\mu}}_{\text{B}} = (1234, 714)$ for team B, and the estimates for the standard deviations are $\hat{\boldsymbol{\sigma}}_{\text{A}} = (224, 176)$ and $\hat{\boldsymbol{\sigma}}_{\text{B}} = (200, 182)$ for team A and team B, respectively. Both teams have a higher mean in state 1, which can thus be interpreted as the team considered being in an attacking phase, whereas state 2 refers to a more defensive style of play. As we consider a high-resolution data set here, it seems intuitively plausible that we observe a high persistence in both states, with the t.p.m. for team A and team B estimated as

$$\hat{\boldsymbol{\Gamma}}_{\text{A}} = \hat{\boldsymbol{\Gamma}}_{\text{B}} = \begin{pmatrix} 0.998 & 0.002 \\ 0.002 & 0.998 \end{pmatrix}.$$

To provide further insights into the fitted univariate models, Fig. 6 shows the decoded states for both teams which were obtained using the Viterbi algorithm. We see that in most phases where team A is attacking, team B is in defense and vice versa — in fact, the decoded states indicate such pattern in more than 60% of the time. While this finding seems intuitively plausible, the

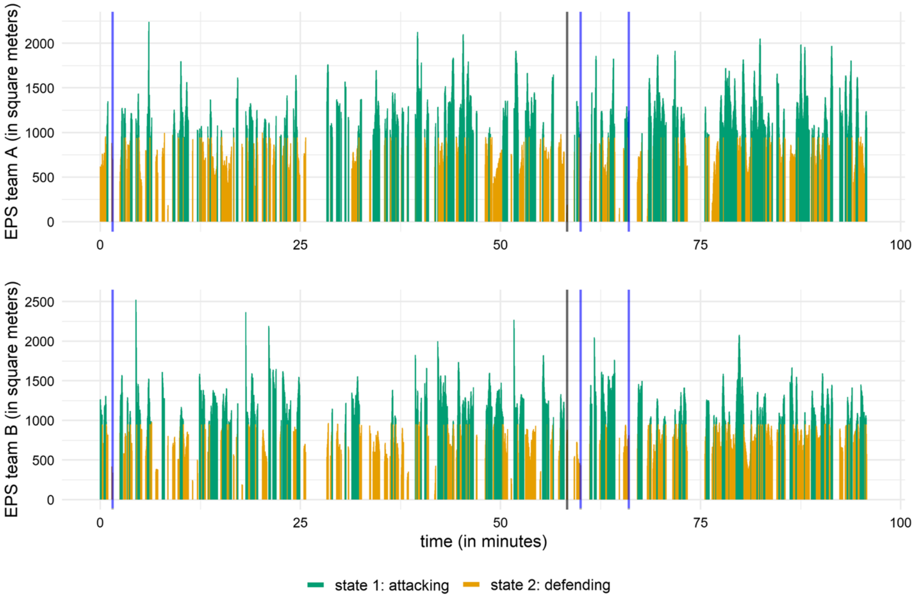

Fig. 6 Decoded sequences for the univariate two-state HMM fitted to the data from team A (top panel) and team B (bottom panel). Blue (black) vertical lines indicate goals scored by the team A (team B)

univariate models do not capture any dependence, since they are fitted separately to the two teams. Hence, to explicitly consider any potential dependence between the two teams' EPS, we consider a bivariate model in the next subsection.

4.2 Bivariate HMM for the EPS

To investigate any potential dependence between the two teams' playing styles, we consider a bivariate HMM as described in Sect. 3. When formulating HMMs for tracking data, it is anything but clear how many states there are in a football match. In the most simple case, there could be two states as considered above for the univariate case, serving for offensive and defensive style of play. Unless analysts have some a priori knowledge on how many different states (and hence tactics) there should be in a match, for example guided by a manager's expertise, the number of states can be chosen via information criteria such as AIC and BIC, but also by inspecting the fitted state-dependent distributions in more detail.

Model selection

For model selection, we fit candidate models with up to five states. Table 1 displays the AIC and BIC values for the candidate models considered as well as the number of estimated parameters. Although both AIC and BIC point to the five-state model, it should be noted that these two information criteria tend to select too complex models—for HMMs, this often results in models with too many states (Pohle et al., 2017). HMMs with many states are usually hard to interpret, and as the number of parameters grows exponentially, parameter estimation may also get unstable. Given such difficulties, Pohle et al. (2017) suggest to closely inspect each candidate model by comparing the estimated state-dependent distributions. Following this approach with a focus on interpretability, Fig. 7 shows the fitted (bivariate) state-dependent distributions for the candidate models with three to five states. Comparing the left and the middle panel in Fig. 7 where a fourth state was added to the model, we see that state 2 of the three-state model is now split up into two states (state 2 and state 3 in the middle panel). Although some complexity is added to the model, we observe only a very limited overlap between the states. However, when adding a fifth state to the model (right panel in Fig. 7), the estimated state-dependent distributions are not clearly distinct as for the model with four states, as we now observe a huge overlap between states 3, 4, and 5. As this hinders interpretation of the model, which is relevant for our application, and since we do not see meaningful additional information in a potential fifth state, from now on we focus on the four-state model.

Table 1 AIC and BIC values for models with 2–5 states as well as the corresponding number of estimated parameters

	# parameters	AIC	BIC
2 states	13	2,308,880	2,309,034
3 states	23	2,250,650	2,250,923
4 states	35	2,207,860	2,208,275
5 states	49	2,174,409	2,174,991

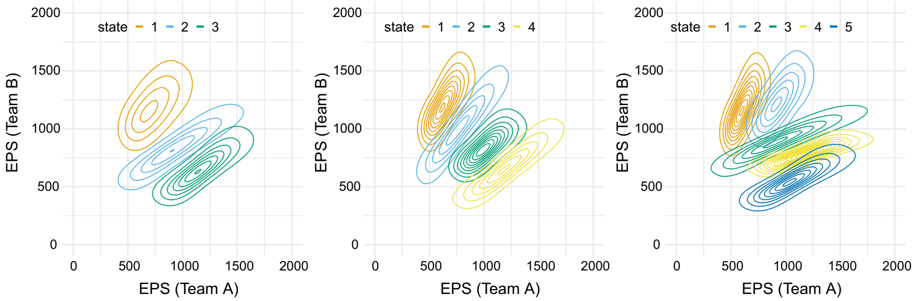

Fig. 7 Fitted state-dependent distributions for the copula-based HMM with three, four, and five states, respectively

Table 2 Parameter estimates for the state-dependent distributions of the final model, i.e. the Frank-copula HMM with four states

	State 1 *Team A: defense* *Team B: attacking*	State 2 *Team A: pressing* *Team B: attacking*	State 3 *Team A: attacking* *Team B: pressing*	State 4 *Team A: attacking* *Team B: defense*
EPS (Team A)	$\hat{\mu} = 651, \hat{\sigma} = 130$	$\hat{\mu} = 851, \hat{\sigma} = 283$	$\hat{\mu} = 1031, \hat{\sigma} = 174$	$\hat{\mu} = 1277, \hat{\sigma} = 320$
EPS (Team B)	$\hat{\mu} = 1234, \hat{\sigma} = 211$	$\hat{\mu} = 1096, \hat{\sigma} = 341$	$\hat{\mu} = 853, \hat{\sigma} = 154$	$\hat{\mu} = 729, \hat{\sigma} = 250$
Dependence	$\hat{\theta} = 5.326$	$\hat{\theta} = 12.65$	$\hat{\theta} = 5.648$	$\hat{\theta} = 11.41$

Final model

For the model selected in the previous subsection, i.e. the copula-based HMM with four states, Table 1 displays the estimated parameters of the state-dependent distributions. Due to the high-resolution data considered, we obtain a t.p.m. with very high persistence ($\hat{\gamma}_{11} = 0.996$, $\hat{\gamma}_{22} = 0.994$, $\hat{\gamma}_{33} = 0.994$, $\hat{\gamma}_{44} = 0.998$). For states 1 and 4, we observe a fairly large EPS for one team while the other team exhibits a rather low EPS. In Sect. 2, we have seen that such values of the EPS refer to one team being in offense (the one with a large EPS) while the other is in defense. Thus, in state 1 team B is in offense while in state 4 team A is in offense. For states 2 and 3, the exploratory analysis presented in Sect. 2 again facilitates interpretation of the fitted model. As teams tend to have a larger EPS when playing pressing (compared to when defending close to their own goal), states 2 and 3 may refer to situations where the team with the larger convex hull is in offense, while the one with the smaller convex hull is pressing. Such pattern was also underlined by panel (C) in Fig. 1 and will further be further investigated below.

Whereas we observe a overall negative correlation between the two teams' EPS (cf. Fig. 2), the estimated dependence parameter is positive in each state. We consider this an intuitive finding, because whenever one team (usually the attacking team) spreads out its players, which in turn increases the EPS, the other (defending) team has to cover the opponent's players, which also results in an increased EPS.

Decoded states

With each state of our fitted model belonging to a certain match situation, we now further investigate typical tactical patterns by calculating the most likely state sequence given the

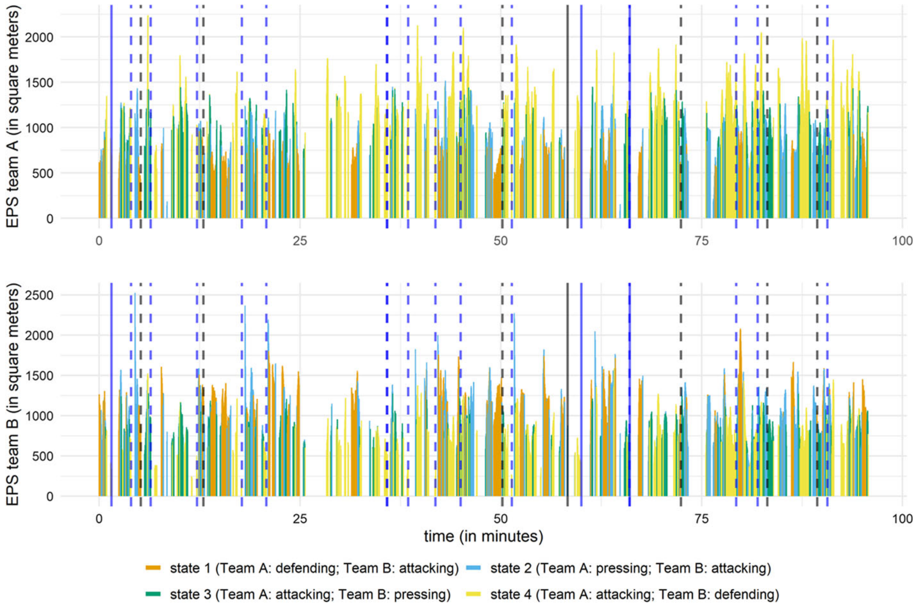

Fig. 8 Decoded sequences for the copula-based HMM with three states. The blue dashed lines indicate goals scores by the home team

observations. For that, we again use the Viterbi algorithm, whose complexity does not increase compared to the univariate model—calculating the most likely state sequence takes less than a minute on a usual desktop computer. Figure 8 shows the resulting decoded bivariate time series, with the solid and dashed lines denoting goals scored and shots on goal, respectively. We can thus very easily infer how goals were scored by inspecting the decoded state at that time. For example, for team A we see that goals were scored when being in state 4, i.e. when in attacking play and team B was defending. The result of the match (team A won 3:1) is also to some extent mirrored by the decoded states, as state 4 was active most of the time with about 19 minutes (corresponding to 34% of the time points).

To further interpret the fitted states, we investigate the position of the players in each state. For that purpose, Fig. 9 shows the geometric median (Vardi & Zhang, 2000) of players' positions for each decoded observation, separated for the four states. From the figures in the first row we can see that if the match is in state 1, the teams are in the half of team A most of the time, whereas in state 2 there are several instances where both teams are in the half of team B. The geometric medians thus further aid the interpretation of our states, as the players of team A are playing more offensive in state 2 compared to state 1, which is in line with the interpretation of team A playing pressing in state 2.

4.3 Use cases

While Figs. 7, 8, and 9 already provide insights by connecting the decoded states to (shots on) goal and players' positions, the decoded states enable further analyses of a team's tactics. This subsection illustrates further use cases of our model, which can aid decision making in practice.

Table 3 Mean number of consecutive seconds spent in each state (according to the decoded sequences shown in Fig. 8)

Mean number of consecutive seconds spent in state 1:	13.10
Mean number of consecutive seconds spent in state 2:	6.782
Mean number of consecutive seconds spent in state 3:	6.325
Mean number of consecutive seconds spent in state 4:	16.60

Use case I: How effective are a team's attacking phases?

By means of the decoded states, we can further investigate the effectiveness of a team by inspecting their attacking phases in combination with their shots on goal. From Fig. 8 we can see that around minute 30 team A had a relatively long attacking phase, which was followed by an attacking phase of team B. However, neither of the two teams made a shot on goal in these two phases. Looking at such sequences as a whole, we consider 19 attacking phases for team A—that is, observations where either state 3 or 4 was active for at least 20 consecutive seconds—and also 19 of such phases for team B. Out of these 19 phases, in six phases team A made at least one shot on goal, while team B made only a single shot on goal during their attacking phases. Although team A made substantially more shots on goal than team B, our analysis based on the decoded states reveals that both teams had several attacking phases where they made no shots on goal, thus indicating that such phases could have been used more efficiently by both teams.

Use case II: How quickly does a team attack?

When analysing attacking phases, it is not only of interest whether an attack results in a (shot on) goal, but also how a team attacks. In principle, one can distinguish between quick attacks with relatively high pressure on goal, and less goal-oriented tactics, where the focus is more on controlling the match by keeping possession of the ball. To distinguish between these playing styles, we calculate the number of consecutive seconds spent in each state (according to the decoded states). The mean number of these consecutive seconds is displayed in Table 3. We see that, for example, the consecutive time spent in state 1 (where team A is in defense and team B is attacking) is lower than the time spent in state 4 (where team A is attacking and team B is in defense). One reason for that could be that if team A is attacking, they first take control over the match by keeping possession of the ball, whereas team B is likely to play more directly to the goal.

Use case III: How do teams behave when pressing?

While the previous two use cases considered a further investigation of the attacking play, the third use case focuses on defense play, in particular on teams' behavior when pressing. While there exist different styles of pressing in general, such as a fairly offensive style of pressing for which Jürgen Klopp is well-known, some teams may also play a less offensive style of pressing. With our data and model at hand, we can gain some further insights into pressing tactics. In particular, similar to the teams' geometric medians shown in Fig. 9, we can compute the average formation line, which is simply given by the mean of the players' x-values. When team A is pressing (i.e. when state 2 is active), the average formation line is 38.8 meters away from their own goal. For team B, the average formation line is 46.5 meters away from their own goal in such situations (i.e. when state 3 is active). Thus, team

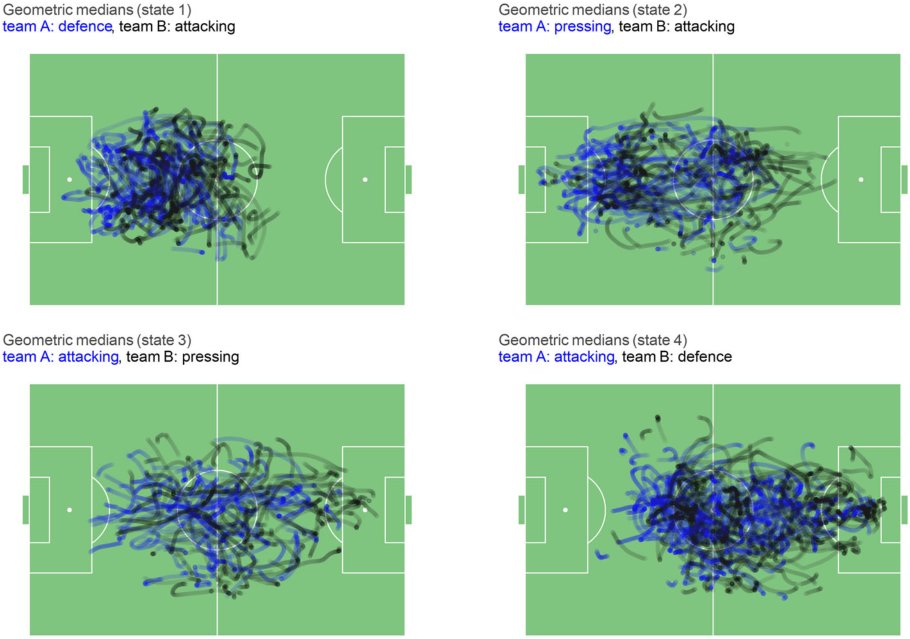

Fig. 9 Geometric medians for state 1, state 2, state 3, and state 4. For simplicity, the data have been transformed such that team A (team B) always plays from left (right) to right (left)

B is playing a more offensive style of pressing than team A, which in turn constitutes a more adventurous style of play. However, this result may also be driven by the fact that team B was trailing most of the time—this aspect will be further discussed in Sect. 5.

5 Discussion

In football, there is a wide interest about analysing tactics and playing styles among managers, media, bookmakers, and sports fans. Unique tracking data sets as considered in this contribution enable an automated classification of such tactics and playing styles. To that end, we provide an approach using copula-based HMMs, which relate the observed EPS to latent states, serving for teams' tactics. A key strength of the proposed approach is that we model an entire match instead of focusing only on certain situations such as set plays. Thus, the proposed approach could be helpful for managers in practice when analysing future opponents, as it provides information on a team's main characteristics such as how shots on goal were made.

By analysing a single tracking data set as a case study, we demonstrate that HMMs provide interpretable states, which correspond to different tactics. The decoded states enable several further analyses, such as to investigate how quickly a team attacks and how teams behave in pressing play. However, some of the patterns found—such as few shots on goal for team A while in attacking phase—may be related to the score of the match, which was 3:1 for team A. In particular, if a team has a clear lead, they may not play with much pressure on goal, but instead aim at taking control over the match by keeping possession of the ball. The score of the match may have also driven team B’s tactics, as their adventurous style of pressing may be driven by the fact that they were trailing most of the time.

Considering that we analyse data from only one match, this contribution is to be regarded as a simple case study of how tracking data sets can be used to detect playing styles in football. Thus, the current analysis has certain limitations mainly due to the fact that we have used only data from one match. If more matches were available, a series of further interesting questions arise. First, as already briefly discussed above, with more data one can relate the EPS to some particular events in the match, like goals or shots on goal in a more detailed manner. Results from one match can simply be regarded as a case study and do provide only limited insights. Second, the strength of the competing teams may affect the EPS of a team. It seems intuitively plausible that teams adjust to their opponent and hence we expect to see such changes. Third, different team formations relate to different EPS. Hence, one can detect from the data such changes in formation and understand the playing style in a more quantitative way. Fourth, since changes in the EPS can be interpreted as changes in the tactic, such in-game modelling could be used by bookmakers. Modelling the EPS as considered here can serve as a basis to see whether more precise estimations of teams’ winning probabilities can be obtained.

As mentioned above, space control is extremely important in football, and studying quantities like the EPS can help to understand a team’s tactic. Related quantities that attempt to measure space control can be also of interest. For example, using the geometric median as in Sect. 4.2, one can understand the position of the team on the pitch and how this may change over time and how it is related to the score, the competition etc. Other measures can be also of value. Gonçalves et al. (2018) considered smaller subgroups of players and calculated the EPS for them, this may also help to consider different tactics and how well they have been implemented. However, the EPS helped to understand the behaviour of a team as a whole which is important in team sports. In a recent paper Bueno et al. (2021) provide some other shape descriptors to evaluate the organisation of football teams on the pitch. Our HMM approach is applicable to such measures with minor modifications.

As tracking data sets become more and more popular, the EPS could also be investigated in other team sports. The proposed modelling framework of copula-based HMMs could thus also be transferred to other sports to provide automated classification of team’s tactics. In particular, sports with team sizes of about 5 or more individuals are best suitable for our approach, as it may very well be the case that otherwise we do not observe substantial variation in the EPS throughout a match. For readers interested in analysing their own tracking data using the approach proposed in this manuscript, we provide data and code via the electronic supplement of this article.

Acknowledgements The authors would like to thank the Associate Editor and the referees for helpful comments that helped improve the paper. Marius Ötting received financial support from the German Research Foundation (DFG), grant number 431536450, which is gratefully acknowledged.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Baptista, J., Travassos, B., Gonçalves, B., Mourão, P., Viana, J. L., & Sampaio, J. (2020). Exploring the effects of playing formations on tactical behavior and external workload during football small-sided games. *The Journal of Strength & Conditioning Research, 34*(7), 2024–2030.

Bueno, MJd. O., Silva, M., Cunha, S. A., Torres, Rd. S., & Moura, F. A. (2021). Multiscale fractal dimension applied to tactical analysis in football: A novel approach to evaluate the shapes of team organization on the pitch. *PlOS One, 16*(9), e0256771.

Cervone, D., D'Amour, A., Bornn, L., & Goldsberry, K. (2016). A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of the American Statistical Association, 111*(514), 585–599.

Fernandez, J., & Bornn, L. (2018). Wide open spaces: A statistical technique for measuring space creation in professional soccer. In: Sloan Sports Analytics Conference.

Franks, A., Miller, A., Bornn, L., Goldsberry, K., et al. (2015). Characterizing the spatial structure of defensive skill in professional basketball. *Annals of Applied Statistics, 9*(1), 94–121.

Frencken, W., Lemmink, K., Delleman, N., & Visscher, C. (2011). Oscillations of centroid position and surface area of soccer teams in small-sided games. *European Journal of Sport Science, 11*(4), 215–223.

Goes, F., Kempe, M., van Norel, J., & Lemmink, K. (2021). Modelling team performance in soccer using tactical features derived from position tracking data. *IMA Journal of Management Mathematics, 32*(4), 519–533.

Goes, F., Meerhoff, L., Bueno, M., Rodrigues, D., Moura, F., Brink, M., Elferink-Gemser, M., Knobbe, A., Cunha, S., Torres, R., et al. (2021). Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *European Journal of Sport Science, 21*(4), 481–496.

Goes, F. R., Kempe, M., Meerhoff, L. A., & Lemmink, K. A. (2019). Not every pass can be an assist: A data-driven model to measure pass effectiveness in professional soccer matches. *Big Data, 7*(1), 57–70.

Gonçalves, B., Folgado, H., Coutinho, D., Marcelino, R., Wong, D., Leite, N., & Sampaio, J. (2018). Changes in effective playing space when considering sub-groups of 3 to 10 players in professional soccer matches. *Journal of Human Kinetics, 62*, 145.

Härdle, W. K., Okhrin, O., & Wang, W. (2015). Hidden Markov structures for dynamic copulae. *Econometric Theory, 31*(5), 981–1015.

Joe, H. (2014). *Dependence modeling with copulas*. CRC Press.

Kempe, M., Goes, F.R., & Lemmink, K.A. (2018). Smart data scouting in professional soccer: Evaluating passing performance based on position tracking data. In *2018 IEEE 14th International Conference on e-Science, IEEE*, pp 409–410.

Kosmidis, I., & Karlis, D. (2016). Model-based clustering using copulas with applications. *Statistics and Computing, 26*(5), 1079–1099.

Lopez, M. J. (2020). Bigger data, better questions, and a return to fourth down behavior: An introduction to a special issue on tracking datain the National Football League. *Journal of Quantitative Analysis in Sports, 16*(2), 73–79.

Martino, A., Guatteri, G., & Paganoni, A. M. (2020). Multivariate hidden Markov models for disease progression. *Statistical Analysis and Data Mining, 13*(5), 499–507.

Memmert, D., Raabe, D., Schwab, S., & Rein, R. (2019). A tactical comparison of the 4-2-3-1 and 3-5-2 formation in soccer: A theory-oriented, experimental approach based on positional data in an 11 vs 11 game set-up. *PlOS One, 14*(1), e0210191.

Orfanogiannaki, K., & Karlis, D. (2018). Multivariate Poisson hidden Markov models with a case study of modelling seismicity. *Australian & New Zealand Journal of Statistics, 60*(3), 301–322.

Ötting, M., Langrock, R., & Maruotti, A. (2021). A copula-based multivariate hidden Markov model for modelling momentum in football. *AStA Advances in Statistical Analysis* pp 1–19.

Pohle, J., Langrock, R., van Beest, F. M., & Schmidt, N. M. (2017). Selecting the number of states in hidden Markov models: Pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics, 22*(3), 270–293.

R Core Team. (2021). R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria, https://www.R-project.org/

Ric, A., Torrents, C., Gonçalves, B., Torres-Ronda, L., Sampaio, J., & Hristovski, R. (2017). Dynamics of tactical behaviour in association football when manipulating players' space of interaction. *PlOS One, 12*(7), e0180773.

Silva, P., Aguiar, P., Duarte, R., Davids, K., Araújo, D., & Garganta, J. (2014). Effects of pitch size and skill level on tactical behaviours of association football players during small-sided and conditioned games. *International Journal of Sports Science & Coaching, 9*(5), 993–1006.

Vardi, Y., & Zhang, C. H. (2000). The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences, 97*(4), 1423–1426.

Zucchini, W., MacDonald, I. L., & Langrock, R. (2016). *Hidden Markov Models for Time Series: An Introduction Using R*. Boca Raton: Chapman & Hall/CRC.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.