MODELLING SOCCER PLAYERS FIELD POSITION VIA MIXTURE OF GAUSSIANS WITH FLEXIBLE WEIGHTS

Marco Berrettini [1], Giuliano Galimberti [1], Thomas Brendan Murphy [2] and Saverio Ranciati [1]

[1] Department of Statistical Sciences , University of Bologna,
(e-mail: marco.berrettini2@unibo.it, giuliano.galimberti@unibo.it, saverio.ranciati2@unibo.it)

[2] School of Mathematics and Statistics, University College Dublin,
(e-mail: brendan.murphy@ucd.ie)

ABSTRACT: An empirical analysis on players' position on the field throughout a soccer match is presented. For this purpose, a Bayesian mixture of experts model is defined, allowing for flexible specification of concomitant covariates on the component weights as smooth functions represented by cubic splines.

KEYWORDS: mixtures of experts models, Gibbs sampling, Bayesian P-splines

1 Introduction

Pettersen et al. (2014) present a dataset of body-sensor traces and corresponding videos from three professional soccer games captured in late 2013 at the Alfheim Stadium in Tromsø, Norway. Tromsø - Stromsogodset is selected for this study, since it is the only one which is valid for the national competition. This game was played on November 3rd, 2013, and it ended with no scores. Player data, including field position, are sampled at 20 Hz using the ZXY Sport Tracking system.
The aim of this analysis is to study how a player's position is affected by a teammate's one and possibily identify a finite number of different phases of the game. Obviously, this relationship depends on many factors, such as the two player's role and which area of the field they are supposed to cover. For this reason, this study focuses on a couple of players playing close to each other.

2 Model specification

The study concentrates on the player covering the right full-back position, identified with tag 9, and assuming that his longitude and latitude (y_1 and y_2,

respectively) can reasonably be approximated by a bivariate Gaussian distribution. Then, the two-dimensional location of the centre-back playing closer to him, Player 13, are taken as concomitant covariates (x_1, x_2). Let $\mathbf{c}$ be a vector of latent variables such that, for each time i, $c_i = g$ if i belongs to cluster g. Conditioning on $\mathbf{c}_i$ and $\mathbf{x}_i$, it is assumed that $\mathbf{y}_i$ follows a Gaussian distribution with vector of means μ_{c_i} and positive definite covariance matrix Σ_{c_i}. Hence, the conditional density of $\mathbf{y}_i$ given $\mathbf{x}_i$ can be written as the following mixture of bivariate Gaussians:

$$f(\mathbf{y}_i|\mathbf{x}_i) = \sum_{g=1}^{G} \pi_g(\mathbf{x}_i) f_{MVN_2}(\mu_g, \Sigma_g), \tag{1}$$

with $f_{MVN_2}(\mu_g, \Sigma_g)$ being the density of a bivariate Gaussian distribution and component weights $\pi_g(\mathbf{x}_i) = \Pr(c_i = g|\mathbf{x}_i) > 0$, so that $\sum_{g=1}^{G} \pi_g(x_i) = 1$, for $i = 1, \ldots, 501$ and $g = 1, 2, \ldots, G$. To allow for flexible specification of such probabilities, a similar methodology to that proposed by Berrettini et al. (2021) for latent class models is adapted to the continuous case. More specifically, prior probabilities are expressed as smooth functions of the covariates represented through Bayesian P-splines (Lang & Brezger, 2004), and estimation is carried out following the data augmentation scheme suggested by Früwirth-Schnatter et al. (2012). Regarding the parameters of the component conditional distributions of the mixture, Gaussian and inverse Wishart priors are respectively assigned to μ_g and Σ_g, as in Marin et al. (2005). The resulting MCMC algorithm does not require any Metropolis-Hastings step.

3 Soccer player positions data

To carry out the analysis, some assumptions are made. In particular, the observations are assumed to be independent across time: to make this assumption more realistic, the data are thinned out to 501 observations over more than 90 minutes of play, leading to a distance of approximately 10 seconds between each pair of consecutive observations. Since between the first and the second half of the game the direction of play changes, preparing this dataset requires a 180° rotation of the locations observed during the second half. The two dimensions of the location of the centre-backs, x_1 and x_2, representing the long and short side of the field, respectively, are assumed to have an additive effect on the log-odds of the component weights. For the analysis, the algorithm is run for fixed G ranging from 1 to 6. The results produced by the best models, in terms of AICM, are selected. Observations are allocated into the G components using the maximum-a-posteriori rule.

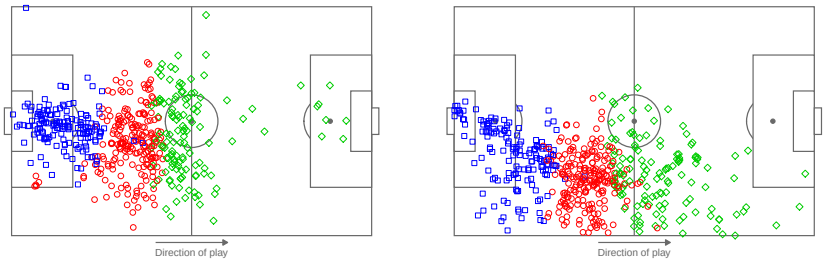

Figure 1. *Locations of Player 13 (left plot) and Player 9 (right plot). Different colors and dot symbols correspond to different clusters.*

4 Results

The best model according to AICM has $G = 3$ components. Figure 1 shows the locations of the two players during the game, allocated according to the 3-component ME model. The clusters does not seem well separated. Indeed, without considering the position of Player 13, the best finite mixture of Gaussians with constant component weights suggests the presence of a single component. These clusters may be interpreted as phases of the game: in particular, the blue dots identify the defensive phase, the green triangles the offensive one, while the red square indicate an intermediate phase. The intermediated phase, originally associated to the first component (in red), is taken as the reference to define the log-odds of mixture weights. The splines' coefficients are transformed accordingly, and, due to space limitations, only the estimated effect of the location of Player 13 on the probability of the defensive phase of Player 9 is reported in Figure 2. The clusters differ mainly with respect to the long side (x_1) of the field, while the location on the short side seems to be less impactful. Lower values of the longitude for Player 13 seem to lead to a higher probability that Player 9 is in the defensive phase, implying him covering the backfield too. This probability drops as x_1 grows, increasing the odds of the offensive phase, characterized by a higher longitude and variability. A huge amount of variability of the estimated effects can be noticed in the plots, especially when the functions reach large absolute values that correspond to 0 or 1 on the scale of the probability. This might be also due to the fact that the locations of the players are not uniformly distributed along the field. It is worth mentioning that this uneven distribution of the observations seems coherent with the specific roles of the two players considered in this analysis.

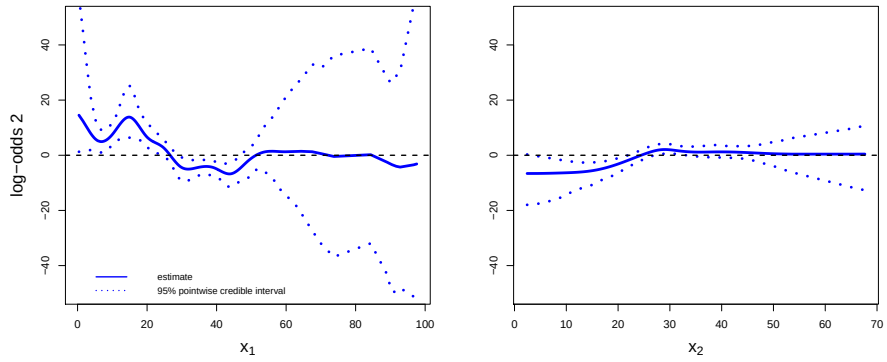

Figure 2. *Estimated effect (and 95% pointwise credible interval) of the location of Player 13,* (x_1, x_2) *on the log-odds of the mixture weights, for Cluster 2 .*

References

BERRETTINI, MARCO, GALIMBERTI, GIULIANO, RANCIATI, SAVERIO, & MURPHY, THOMAS BRENDAN. 2021. Flexible Bayesian modelling of concomitant covariate effects in mixture models. *arXiv preprint arXiv:2105.12852*.

FRÜHWIRTH-SCHNATTER, S., PAMMINGER, C., WEBER, A, & WINTER-EBMER, R. 2012. Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts Markov chain clustering. *Journal of Applied Econometrics*, **27**, 1116–1137.

LANG, STEFAN, & BREZGER, ANDREAS. 2004. Bayesian P-splines. *Journal of computational and graphical statistics*, **13**(1), 183–212.

MARIN, JEAN-MICHEL, MENGERSEN, KERRIE, & ROBERT, CHRISTIAN P. 2005. Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics*, **25**, 459–507.

PETTERSEN, SVEIN ARNE, JOHANSEN, DAG, JOHANSEN, HÅVARD, BERG-JOHANSEN, VEGARD, GADDAM, VAMSIDHAR REDDY, MORTENSEN, ASGEIR, LANGSETH, RAGNAR, GRIWODZ, CARSTEN, STENSLAND, HÅKON KVALE, & HALVORSEN, PÅL. 2014. Soccer video and player position dataset. *Pages 18–23 of: Proceedings of the 5th ACM Multimedia Systems Conference*.