Modelling volleyball data using a Bayesian approach

Leonardo Egidi[a], Ioannis Ntzoufras[b]
legidi@units.it, ntzoufras@aueb.gr

July 2nd, 2019

MathSport 2019

[a]Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche *Bruno de Finetti*, Università degli Studi di Trieste, Trieste, Italy

[b]Department of Statistics, Athens University of Economics and Business, Athens, Greece

My only previous experience in volleyball...

Scout guy for a female's volleyball team in my city!

Summary

Three possible models

Data: Italian SuperLega 2017-2018

Results

Posterior checks

References

Modelling volleyball data

- Unlike what happens for other major sports, modelling volleyball match outcomes has not been much addressed by statisticians and mathematicians [Ferrante and Fonseca, 2014].
- Game complexity:
  1. the number of sets is a random variable (from 3 to 5);
  2. the number of points achieved by the two teams in each set varies depending on whether they are in the fifth set or not;
  3. the number of final set points for two competing teams may exceed 25 when both the teams reach 24 points (**24-deuce**).

  Natural **hierarchy** of points within sets.
- In our perspective, the task of modelling volleyball match results should follow a top-down strategy, from the sets to the single points. Challenging from a statistical point of view!

Modelling volleyball data

- We maintain with the idea to replicate the hierarchy of the game into our models.
- Set-by-set negative binomial model for the points achieved by the team **loosing** the single set: the distribution of the points is then conditional to the set result.
- Strengths' difference among the teams: in the Bayesian approach teams' abilities are easily incorporated into the model by use of some **weakly-informative prior distributions** [Gelman et al., 2008].

Three possible models

Modelling volleyball data: models' likelihood

▷ Y_{Ag}, Y_{Bg}: the points for each set $g = 1, 2, \ldots, G$ collected by the two competing teams, A and B.

▷ W_g: the binary indicator for the win of the home team. $W_g = 1$ if team A wins the set, 0 otherwise.

▷ Y_g: the number of points for the team loosing the g-th set, $Y_g = W_g Y_{Bg} + (1 - W_g) Y_{Ag}$:

▷ Three possible models are proposed:

$$\begin{aligned} Y_g &\sim \text{NegBin}(25,\ p_g) I(Y_g \leq 23) && (1) \\ Y_g &\sim \text{Bin}(n,\ 1 - p_g) && (2) \\ Y_g &\sim \text{Pois}(\gamma_g), && (3) \end{aligned}$$

where NegBin, Bin, Pois denote the negative binomial, the binomial and the Poisson distribution, respectively. $I(Y_g \leq 23)$ is the right truncation.

Modelling volleyball data: teams' abilities and priors i

- ▷ $\{\alpha_{A(g)}, \alpha_{B(g)}\}, \{\beta_{A(g)}, \beta_{B(g)}\}$: set teams and point teams abilities for teams $A(g)$ and $B(g)$, respectively.
- ▷ H_s, H_p; μ: set home and point home advantage for the hosting team; common baseline parameter.
- ▷ Set probabilities Team A wins the set with probability ω_g:

$$\begin{aligned} W_g &\sim \text{Bernoulli}(\omega_g), \\ \text{logit}(\omega_g) &= H_s + \alpha_{A(g)} - \alpha_{B(g)}, \end{aligned} \tag{4}$$

- ▷ Point probabilities The logit probability of realizing a point when loosing the set is defined as:

$$\log \frac{1-p_g}{p_g} = \mu + (1 - W_g)H_p + (\beta_{A(g)} - \beta_{B(g)})(1 - 2W_g) \tag{5}$$

Modelling volleyball data: teams' abilities and priors ii

▷ Average points (Poisson model) The logarithm of the average number of points realized by the team loosing the g-th set is

$$\log \gamma_g = \mu + (1 - W_g)H_p + (\beta_{A(g)} - \beta_{B(g)})(1 - 2W_g). \qquad (6)$$

▷ Prior distributions The Bayesian model is completed by assigning some weakly informative priors [Gelman et al., 2008] to the set and point abilities, for each team $t = 1, \ldots,$ nteams:

$$\begin{aligned} \alpha_t, \beta_t &\sim \mathcal{N}(0, 1) \\ \mu, H_p, H_s &\sim \mathcal{N}(0, 10^3) \end{aligned} \qquad (7)$$

▷ Identifiability constraints set and point abilities need to be constrained; in such a framework we impose a sum-to-zero constraint for both the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

Models' extensions i

- Dynamic abilities Teams performance are likely to change over an entire season. An auto-regressive prior distribution for both the point and the set abilities is a common choice for a dynamic assumption of the abilities [Owen, 2011]
- Attacking and defensive abilities Separately model the abilities arising from attack and defense phases, as in football [Karlis and Ntzoufras, 2003].
- Connecting the abilities Joint modelling of both set and point abilities.

Models' extensions ii

- Extra-points after 25. The three models (1), (2) and (3) may be extended specifying a zero-inflated Poisson (ZIP) model for the extra points collected O_g by the loosing-set team in case of 24-deuce:

$$\begin{aligned} Y_g &= O_g + W_g Y_{Bg} + (1 - W_g) Y_{Ag} \\ O_g &\sim \text{ZIPoisson}(p_{0g}, \lambda_g), \end{aligned} \tag{8}$$

where p_{og} describes the proportion of extra zeros (no 24-deuce) and λ_g is the rate parameter:

$$\begin{aligned} \text{logit}(p_{0g}) =& m + \delta(\alpha_{A(g)} - \alpha_{B(g)}) + \gamma(\beta_{A(g)} - \beta_{B(g)}) \\ \log(\lambda_g) =& \eta \\ \delta, \gamma \sim& \mathcal{N}(0,1);\ \eta \sim \mathcal{N}^+(0, 10^2) \end{aligned} \tag{9}$$

Data: Italian SuperLega 2017-2018

Data: Italian SuperLega 2017-2018

We use the regular season of the Italian SuperLega 2017-2018 dataset to validate and fit models (1), (2) and (3). The dataset consists of the results - considered both at set and points levels - of:

- 182 matches
- 680 sets
- 14 teams

At the end of the regular season, Sir Safety Perugia achieved the greatest number of points (70), whereas BCC castellana Grotte achieved the lowest number of points (10).

Results

Model fitting: model choice

- We fit the model via the `rjags` R package [Plummer, 2018] performing Gibbs sampling (500 MCMC iterations, burn-in period: 100 iterations).
- DIC (Deviance Information Criteria) for each model with the corresponding number of parameters. The ZIP truncated negative binomial model is the best fitted model.

Model distribution	*Additional model details*	*# param.*	*DIC*
1. Neg. binomial	$r = 25$	31	4779.3
2. Poisson	log-linear model for γ_g	31	4574.5
3. Binomial	$n_g \sim \mathsf{Pois}(46)$	31	8031.6
4. ZIP Tr. Neg. bin.		35	4544.1
5. ZIP Tr. Poisson		35	4565.3
6. ZIP Tr. Binomial	$n_g \sim \mathsf{Pois}(46)$	35	8408.7
7. ZIP Tr. Neg. bin.	att $\neq$ def	49	–
8. ZIP Tr. Neg. bin.	α_t, β_t dynamic	737	4721.7

Table 1: Details of the fitted models: Italian SuperLega 2017-2018 season (MCMC sampling, 500 iterations).

Truncated negative binomial: posterior estimates

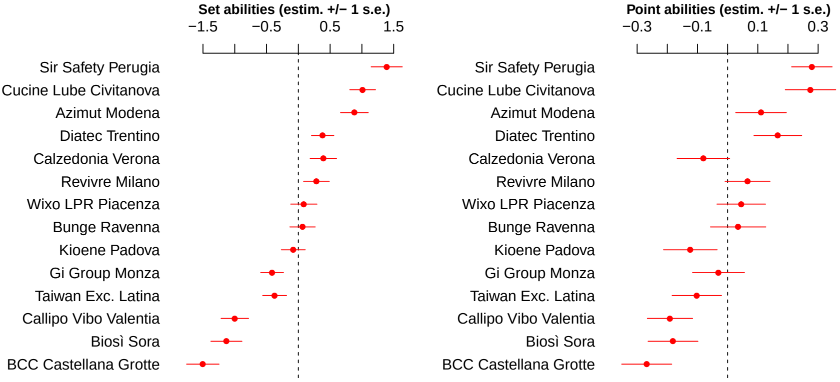
- Posterior estimates for the set home advantage H_s, the point home advantage H_p, the grand intercept μ and the ZIP parameters η, m, δ, γ. There is a clear signal of home advantage both at the set and at the single point level.

	Mean	Median	sd	2.5%	97.5%
H_s	0.16	0.16	0.08	-0.00	0.31
H_p	0.20	0.19	0.07	0.07	0.34
μ	0.36	0.36	0.05	0.26	0.46
η	1.38	1.38	0.07	1.26	1.50
m	2.13	2.10	0.12	1.90	2.39
δ	-0.20	-0.20	0.76	-1.69	1.24
γ	0.09	0.09	0.18	-0.26	0.41

Table 2: ZIP truncated negative binomial model: Posterior estimates for the following parameters: the set home H_s, the point home H_p, the grand intercept μ; η, m, γ, δ (ZIP part).

Set and point abilities

- Estimated set and point abilities (posterior means $\pm$ s. e.), following the actual rank of the Italian SuperLega 2017-2018: the global pattern mirrors almost perfectly the final rank.



Posterior checks

Posterior predictive checks

- To evaluate the overall goodness of fit of our final model, we could draw hypothetical values from the posterior predictive distribution of the model and check how plausible are these replications in comparison with the observed data.
- For each set g, we denote by d_g the set points difference $Y_{Ag} - Y_{Bg}$, and with $\tilde{d}_g^{(s)}, s = 1, \ldots, S$ the corresponding points difference arising from the s-th MCMC replication, $\tilde{y}_{Ag}^{(s)} - \tilde{y}_{Bg}^{(s)}$. Once we replicate new existing values from our model, it is of interest to assess how far they are if compared with the actual data we observed.

Posterior checks: league reconstruction

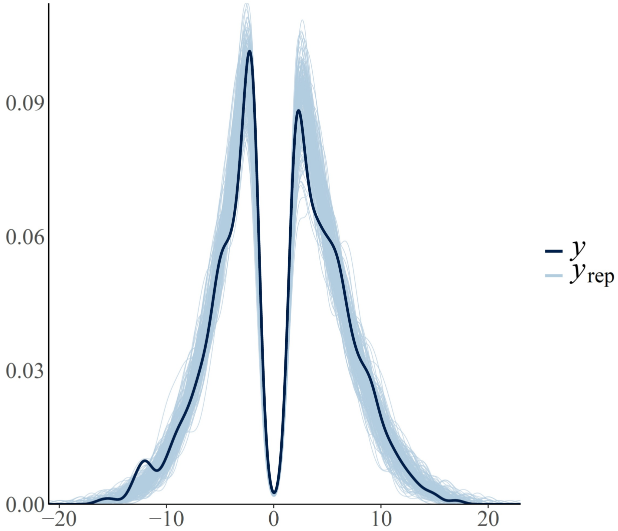
- League reconstruction: replication of hypothetical results from the posterior predictive distribution and league reconstruction. Both simulated rank positions and points are quite close to the observed ones.

Teams	*Exp. Points*	*Actual points*	*Actual rank*
Sir Safety Perugia	69	70	1
Cucine Lube Civitanova	58	64	2
Azimut Modena	51	60	3
Diatec Trentino	47	51	4
Calzedonia Verona	45	50	5
Revivre Milano	41	44	6
Bunge Ravenna	38	41	8
Wixo LPR Piacenza	35	42	7
Kioene Padova	34	35	9
Gi Group Monza	25	28	10
Taiwan Exc. Latina	24	25	11
Biosi Sora	13	13	13
Callipo Vibo Valentia	7	13	12
BCC Castellana Grotte	5	10	14

Table 3: ZIP truncated negative binomial model: final league reconstruction from MCMC sampling along with the actual points and the actual final rank for each team.

Posterior checks: global measure of fit

- Predictive distribution of each $\tilde{d}_g^{(s)}$ (light blue).



Concluding remarks and ongoing work

- The truncated negative binomial model is the best to predict volleyball outcomes.
- The in-sample predictive accuracy is quite good.
- ZIP part for the extra points seems to be a suitable assumption.

Ongoing work:

- Including game's covariates.
- Including points' correlation.
- Predictions on some test set data.
- Other measures of goodness of fit.

Contacts

For further curiosities and analysis related to statistical methods for sports data:

- visit the webpage https://www.leonardoegidi.com/
- write me at legidi@units.it

References

References

Marco Ferrante and Giovanni Fonseca. On the winning probabilities and mean durations of volleyball. *Journal of Quantitative Analysis in Sports*, 10(2):91–98, 2014.

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, Yu-Sung Su, et al. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2 (4):1360–1383, 2008.

Dimitris Karlis and Ioannis Ntzoufras. Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393, 2003.

References ii

Alun Owen. Dynamic bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA Journal of Management Mathematics*, 22(2): 99–113, 2011.

Martyn Plummer. *rjags: Bayesian graphical models using MCMC*, 2018. URL `https://CRAN.R-project.org/package=rjags`. R package version 4-8.