STATS 700 Project

Latent Variable Model for Paired Comparisons in NCAAM Basketball Scores

Simon Fontaine

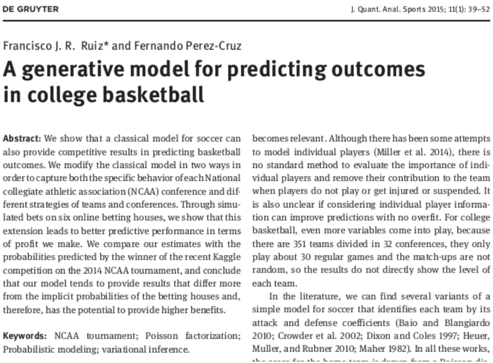
simfont@umich.edu

October 16th, 2020

MOTIVATION

Motivation
Background

- Sports reading group
  - Discussion of a paper on modeling Basketball scores
  - Identification of some possible improvements
- Internship
  - Online video game skill rating
  - Binary outcome
- Current research
  - Closely related to network LVM
  - Binary outcome



DE GRUYTER J. Quant. Anal. Sports 2015; 11(1): 39–52

Francisco J. R. Ruiz* and Fernando Perez-Cruz

A generative model for predicting outcomes in college basketball

Abstract: We show that a classical model for soccer can also provide competitive results in predicting basketball outcomes. We modify the classical model in two ways in order to capture both the specific behavior of each National collegiate athletic association (NCAA) conference and different strategies of teams and conferences. Through simulated bets on six online betting houses, we show that this extension leads to better predictive performance in terms of profit we make. We compare our estimates with the probabilities predicted by the winner of the recent Kaggle competition on the 2014 NCAA tournament, and conclude that our model tends to provide results that differ more from the implicit probabilities of the betting houses and, therefore, has the potential to provide higher benefits.

Keywords: NCAA tournament; Poisson factorization; Probabilistic modeling; variational inference.

becomes relevant. Although there has been some attempts to model individual players (Miller et al. 2014), there is no standard method to evaluate the importance of individual players and remove their contribution to the team when players do not play or get injured or suspended. It is also unclear if considering individual player information can improve predictions with no overfit. For college basketball, even more variables come into play, because there are 351 teams divided in 32 conferences, they only play about 30 regular games and the match-ups are not random, so the results do not directly show the level of each team.

In the literature, we can find several variants of a simple model for soccer that identifies each team by its attack and defense coefficients (Baio and Blangiardo 2010; Crowder et al. 2002; Dixon and Coles 1997; Heuer, Muller, and Rubner 2010; Maher 1982). In all these works,

Figure: The paper discussed in reading group [3]

From a 2018 Kaggle competition [2]:

- All regular-season and post-season outcomes (1985-2017)
- Scores typically between 50 and 100
- $T \approx 350$ teams each season
- $C \approx 30$ conferences each season
- $M \approx 5,000$ matches per season, ≈ 30 matches per season per team
- Conference assignment per season

Pre-processing:

- Consider seasons 2004-2017 (14 seasons)
- Remove all matches that went to overtime

Motivation
Goals

Original model [3]

- Independent Poisson model
- Offensive and Defensive skills
- Non-transitive relationships (multi-dimensional skills)
- Team and conference skills
- “Home-field advantage” modeling

Improvements and differences

- Model correlation between scores
- Gaussian model
- More intuitive team to conference model
- Meaningful “home-field advantage” modeling
- Latent dimension selection
- Independent model through seasons


MODEL
MICHIGAN
2
MICHIGAN
21

Independent models per season:

- m indexes the match in a season
- $t_0(m), t_1(m)$ are the team indices in match m
- $c(t)$ returns the conference index of team t
- $h_0(m), h_1(m)$ identifies if team $i = 0, 1$ is playing at home (1), away (-1) or at a neutral site (0)

[3] considered a single model for 4 consecutive seasons.

Model
Latent variables

All latent variables lie in the same space:

- $C_c^o \in \mathbb{R}^K$: conference c's ability to score (offensive skill)
- $C_c^d \in \mathbb{R}^K$: conference c's ability to prevent score (defensive skill)
- $T_t^o \in \mathbb{R}^K$: team t's ability to score (offensive skill)
- $T_t^d \in \mathbb{R}^K$: team t's ability to prevent score (defensive skill)

Team total skills:

- $S_t^o = T_t^o + \lambda C_{c(t)}^o \in \mathbb{R}^K$
- $S_t^d = T_t^d + \lambda C_{c(t)}^d \in \mathbb{R}^K$
- $\lambda \in \mathbb{R}$ controls the importance of conferences

[3] considered spaces of different dimensions and non-additive combination.

For each match m:

- Compare a team's offense to the opponent's defense with an inner product:

$$M_{m,0} = S_{t_0(m)}^{o\top} S_{t_1(m)}^{d}, \qquad M_{m,1} = S_{t_1(m)}^{o\top} S_{t_0(m)}^{d};$$

  Large inner products associated with larger point production.
- Add the "home-field advantage" $H \in \mathbb{R}$

$$\widetilde{M}_{m,i} = M_{m,i} + H h_i(m), \qquad i = 0, 1.$$

  The effect of playing at home is therefore a skill advantage of $2H$.

[3] only added the "home-field advantage" to the team playing at home so the total effect does not cancel compared to neutral-site matches.

Some teams may perform better against certain types of team, but worse against other types.

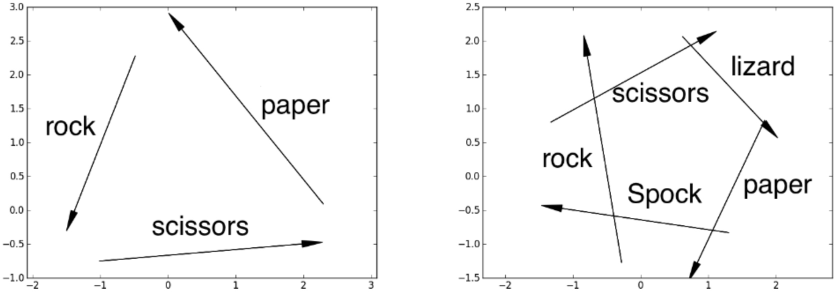


Figure: Multi-dimensional skills and non-transitivity [1].

Denote the vector of in-match ability to produce points:

$$\widetilde{M}_m = \begin{bmatrix} \widetilde{M}_{m,0} \\ \widetilde{M}_{m,1} \end{bmatrix}$$

Gaussian model:

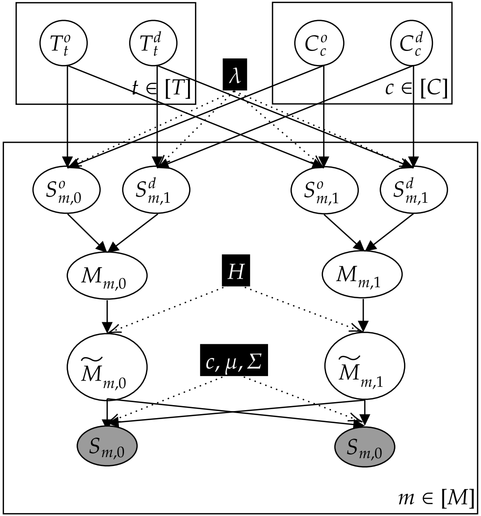
$$S_m \mid \widetilde{M}_m \quad \sim \quad \mathcal{N}_2\left(\mu\mathbf{1} + c\widetilde{M}_m, \Sigma\right),$$

where

- $\mu \in \mathbb{R}$ centers the scores,
- $c \in \mathbb{R}$ scales the in-match ability (N.B. this scales the effect of H),
- $\Sigma \in \mathcal{C}_2^+$ is a PD matrix constrained to

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \qquad \sigma^2 > 0, \rho \in (-1, 1).$$

Model
Graphical representation



ESTIMATION

Constrain each of T^o, T^d, D^o, C^d to be an orthogonal matrix:

- Fixes the scale;
- Produces different components.

Projected (Stochastic) Blockwise Gradient Descent

1. Initialize parameters and latent variables
2. Until convergence of the likelihood:
    - For $v \in \{\text{parameters}, T^o, T^d, D^o, C^d\}$:
        - Sample matches
        - Take a gradient step
        - If $v \neq$ parameters, project onto nearest orthogonal matrix

Details:

- Standard normal priors on T^o, T^d, D^o, C^d
- Mean-field approximation, Gaussian family
- Parameters treated as fixed (uninformative priors)

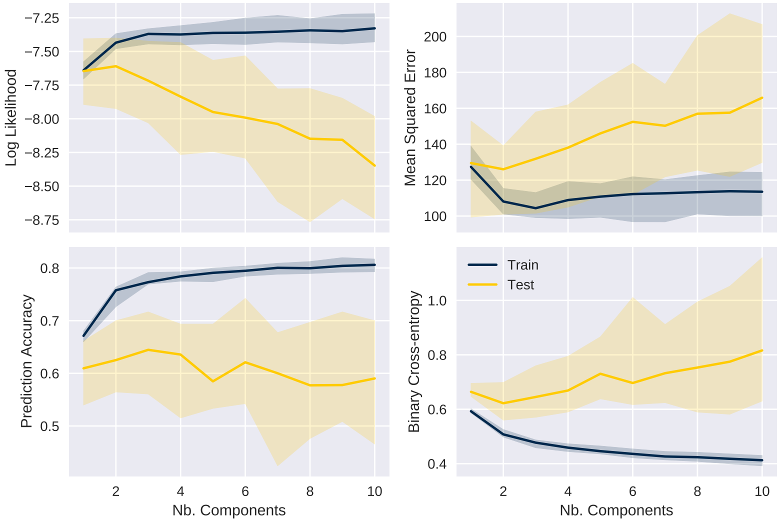
Variational EM with MC expectations

1. Initialize parameters and latent variables
2. Until convergence of the ELBO:
   - Sample matches
   - Sample latent variables using reparameterization trick
   - Compute expected log-likelihood using MC
   - Compute KL term directly
   - Take a gradient step

RESULTS

Results

Selecting the latent dimension (MLE)

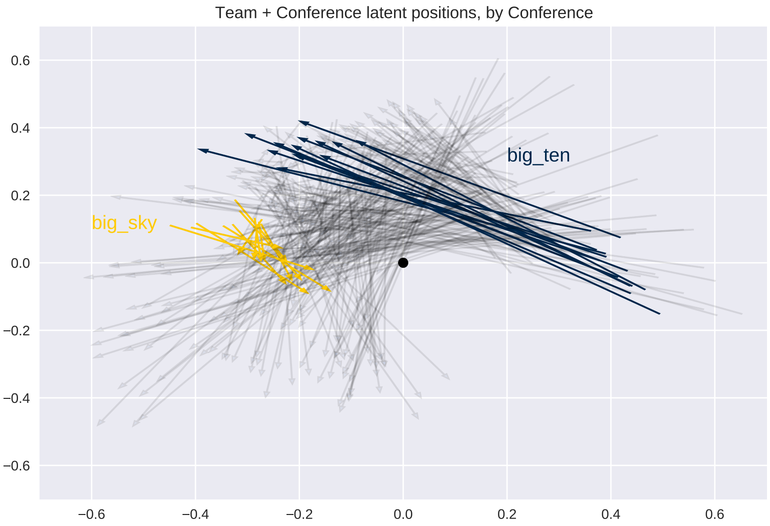


Results

Estimated parameters over time (MLE, $K = 2$)

Results

Team and conference weighting (MLE, $K = 2$)

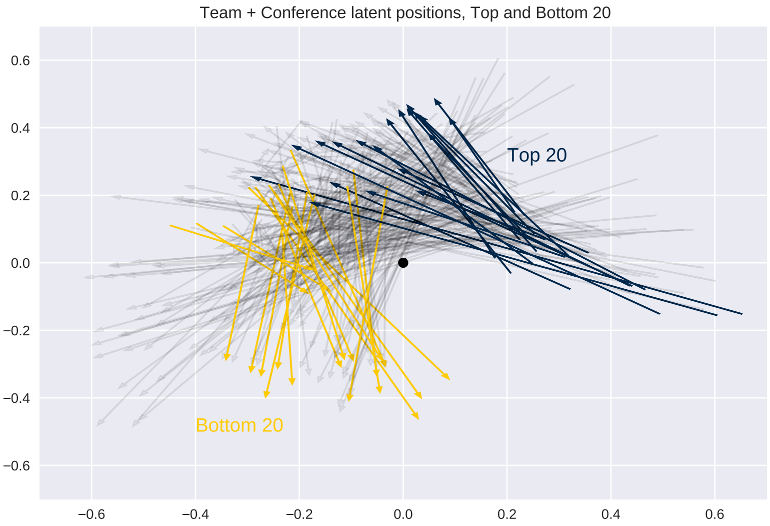


- Make every team play each other once;
- Compute the number of wins.

2017 Rankings

	Team	Proj. wins
1	Villanova	349
2	Purdue	348
3	Florida	344
4	Wisconsin	344
5	North Carolina	343
6	SMU	343
7	Michigan	343
...	...	...
347	NC A&T	6
348	St Francis NY	4
349	Alabama A&M	3
350	Alabama St	2
351	Ark Pine Bluff	0

Results
Best and worst teams (MLE, $K = 2$)



CONCLUSION

Discussion

- Model the dependency between years:
  - Player changes from year to year, but team performance is relatively constant;
  - Constrain/penalize the difference in skills between years

$$T_t^o[y+1] \mid T_t^o[y] \sim \mathcal{N}(T_t^o[y], \tau^2), \qquad \text{LLK} + (T_t^o[y+1] - T_t^o[y])^2$$

  - Component interpretation common across years.
- Variational EM:
  - Similar performance for small K;
  - No improvement as K increases.

THANK YOU!

References

[1] Shuo Chen and Thorsten Joachims. "Modeling Intransitivity in Matchup and Comparison Data". In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. WSDM '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 227–236. ISBN: 9781450337168. DOI: 10.1145/2835776.2835787. URL: https://doi-org.proxy.lib.umich.edu/10.1145/2835776.2835787.

[2] Kaggle. *Google Cloud NCAA® ML Competition 2018-Men's*. 2018. URL: https://www.kaggle.com/c/mens-machine-learning-competition-2018.

[3] Francisco J. R. Ruiz and Fernando Perez-Cruz. "A generative model for predicting outcomes in college basketball". In: *Journal of Quantitative Analysis in Sports* 11.1 (2015), pp. 39–52. DOI: 10.1515/jqas-2014-0055. URL: https://www.degruyter.com/view/journals/jqas/11/1/article-p39.xml.

QUESTIONS?