On the Dependency of Soccer Scores - A Sparse Bivariate Poisson Model for the UEFA EURO 2016

A. Groll[*] & A. Mayr & T. Kneib & G. Schauberger

[*]Department of Statistics,
Georg-August-University Göttingen

MathSport International 2017 Conference,
Padua, June 28th

Who will celebrate?

Who will cry?

Theoretical Background

Aims

The main aims are to

- find an explicit model for exact numbers of goals
- include covariates
- adjust for possible correlations between numbers of goals

of both competing teams.

Aims

The main aims are to

- find an explicit model for exact numbers of goals
- include covariates
- adjust for possible correlations between numbers of goals

of both competing teams.

$\Rightarrow$ Different approaches for

- EURO 2012 (Groll and Abedieh, 2013)
- World Cup 2014 (Groll, Schauberger and Tutz, 2015)
- EURO 2016

Univariate Model for International Soccer Tournaments

$$y_{ijk}|x_{ik}, x_{jk} \sim Po(\lambda_{ijk}) \quad i,j \in \{1,\ldots,n\},\ i \neq j$$

$$log(\lambda_{ijk}) = \beta_0 + \xi_{ik} - \delta_{jk}$$

n: Number of teams

y_{ijk}: Number of goals scored by team i against opponent j at tournament k

x_{ik}, x_{jk}: Covariate vectors of team i and opponent j varying over tournaments

e.g. **EURO 2012** (Groll and Abedieh, 2013):

$$\xi_{ik} = x_{ik}^T \beta_\xi + b_i$$

$$\delta_{jk} = x_{jk}^T \beta_\delta + b_j$$

Univariate Model for International Soccer Tournaments

$$y_{ijk}|x_{ik}, x_{jk} \sim Po(\lambda_{ijk}) \quad i,j \in \{1,\ldots,n\},\ i \neq j$$

$$log(\lambda_{ijk}) = \beta_0 + \xi_{ik} - \delta_{jk}$$

n: Number of teams

y_{ijk}: Number of goals scored by team i against opponent j at tournament k

x_{ik}, x_{jk}: Covariate vectors of team i and opponent j varying over tournaments

e.g. **World Cup 2014** (Groll, Schauberger and Tutz, 2015):

$$\xi_{ik} = x_{ik}^T \beta + att_i$$

$$\delta_{jk} = x_{jk}^T \beta + def_j$$

Univariate Model for International Soccer Tournaments

$$y_{ijk}|x_{ik}, x_{jk} \sim Po(\lambda_{ijk}) \quad i,j \in \{1,\ldots,n\},\ i \neq j$$

$$log(\lambda_{ijk}) = \beta_0 + \xi_{ik} - \delta_{jk}$$

n: Number of teams

y_{ijk}: Number of goals scored by team i against opponent j at tournament k

x_{ik}, x_{jk}: Covariate vectors of team i and opponent j varying over tournaments

e.g. **World Cup 2014** (Groll, Schauberger and Tutz, 2015):

$$\xi_{ik} = x_{ik}^T \beta + att_i$$

$$\delta_{jk} = x_{jk}^T \beta + def_j$$

$$\Rightarrow\ log(\lambda_{ijk}) = \beta_0 + (x_{ik} - x_{jk})^T \beta + att_i - def_j$$

Correlation between Scores of Both Teams

Dixon and Coles (1997) compared marginal distributions of scores with joint distribution $\Rightarrow$ correlation!

TABLE 2
Estimates of the ratios of the observed joint probability function and the empirical probability function obtained under the assumption of independence between the home and away scores†

Home goals	*Estimates of ratios for the following numbers of away goals:*					
	0	*1*	*2*	*3*	*4*	*5*
0	111.5 (3.52)	92.0 (2.87)	103.4 (4.18)	82.1 (7.67)	96.4 (15.31)	96.8 (28.12)
1	93.7 (2.43)	105.7 (2.00)	99.3 (3.74)	103.7 (6.31)	86.9 (13.15)	108.3 (19.99)
2	99.6 (2.91)	101.7 (2.11)	99.2 (3.78)	97.4 (7.41)	95.9 (17.4)	106.7 (23.77)
3	100.3 (4.25)	98.5 (3.61)	91.8 (6.51)	116.6 (11.03)	139.8 (23.85)	75.4 (40.5)
4	91.0 (7.07)	93.8 (7.16)	108.6 (10.74)	138.0 (16.31)	111.7 (32.86)	90.4 (55.33)
5	94.1 (13.24)	102.3 (12.28)	114.3 (20.6)	73.3 (31.01)	120.8 (74.71)	130.4 (129.7)
6	139.1 (31.95)	49.1 (23.66)	146.4 (41.33)	45.3 (57.84)	174.1 (122.2)	—

†The numbers are multiplied by 100 for clarity. Standard errors are given in parentheses.

Source: Dixon and Coles (1997)

$\Rightarrow$ Introduction of additional dependence parameter

But: They did not compare conditional distributions!

$\Rightarrow$ Their linear predictors are not independent!

The Bivariate Poisson Distribution

$X_k \overset{ind.}{\sim} Po(\lambda_k),\ k = 1, 2, 3,\ \lambda_k > 0$

$\Rightarrow Y_1 = X_1 + X_3$ and $Y_2 = X_2 + X_3$ follow a joint bivariate Poisson distribution

$$(Y_1, Y_2) \sim Po_2(\lambda_1, \lambda_2, \lambda_3)$$

The Bivariate Poisson Distribution

$X_k \overset{ind.}{\sim} Po(\lambda_k),\ k = 1, 2, 3,\ \lambda_k > 0$

$\Rightarrow Y_1 = X_1 + X_3$ and $Y_2 = X_2 + X_3$ follow a joint bivariate Poisson distribution

$$(Y_1, Y_2) \sim Po_2(\lambda_1, \lambda_2, \lambda_3)$$

Probability function:

$$P_{Y_1,Y_2}(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2)$$

$$= \exp(-(\lambda_1 + \lambda_2 + \lambda_3)) \frac{\lambda_1^{y_1}}{y_1!} \frac{\lambda_2^{y_2}}{y_2!} \sum_{k=0}^{\min(y_1,y_2)} \binom{y_1}{k} \binom{y_2}{k} k! \left(\frac{\lambda_3}{\lambda_1 \lambda_2}\right)^k$$

The Bivariate Poisson Distribution

$X_k \overset{ind.}{\sim} Po(\lambda_k),\ k = 1,2,3,\ \lambda_k > 0$

$\Rightarrow Y_1 = X_1 + X_3$ and $Y_2 = X_2 + X_3$ follow a joint bivariate Poisson distribution

$$(Y_1, Y_2) \sim Po_2(\lambda_1, \lambda_2, \lambda_3)$$

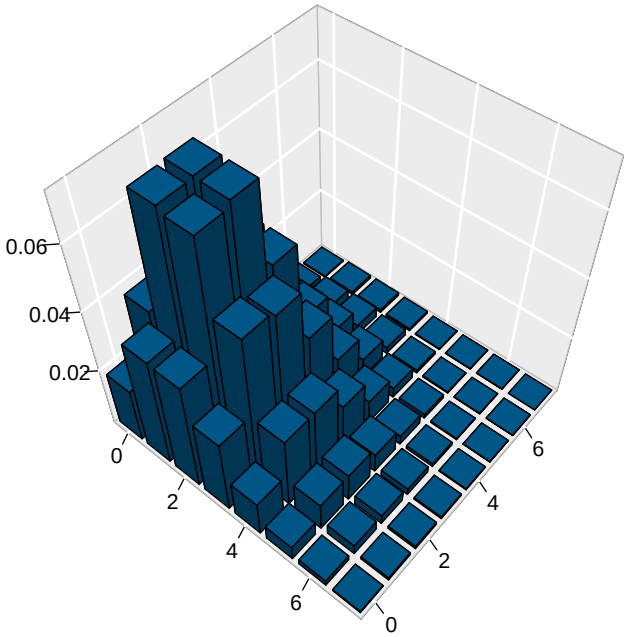
Probability function:

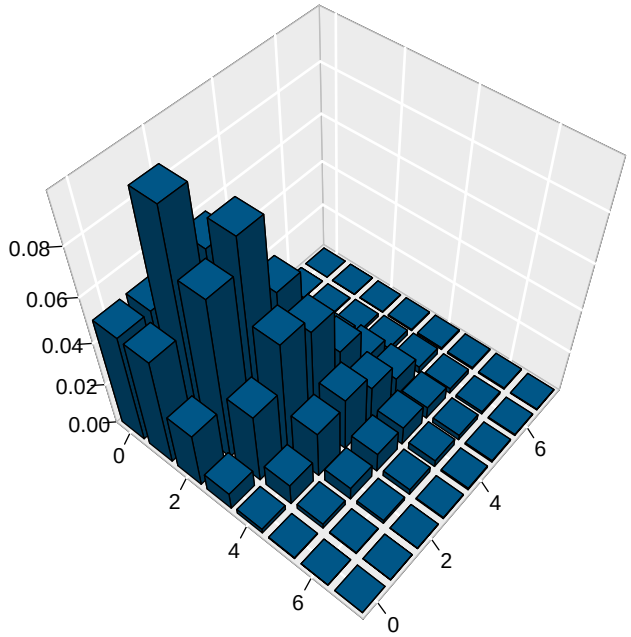
$$P_{Y_1,Y_2}(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2)$$

$$= \exp(-(\lambda_1 + \lambda_2 + \lambda_3)) \frac{\lambda_1^{y_1}}{y_1!} \frac{\lambda_2^{y_2}}{y_2!} \sum_{k=0}^{\min(y_1,y_2)} \binom{y_1}{k} \binom{y_2}{k} k! \left(\frac{\lambda_3}{\lambda_1 \lambda_2}\right)^k$$

- $E(Y_1) = \lambda_1 + \lambda_3$
- $E(Y_2) = \lambda_2 + \lambda_3$
- $cov(Y_1, Y_2) = \lambda_3$

The Bivariate Poisson Distribution





- $\lambda_1 = 2$
- $\lambda_2 = 2$
- $\lambda_3 = 0$

- $\lambda_1 = 1$
- $\lambda_2 = 1$
- $\lambda_3 = 1$

Re-parametrization of bivariate Poisson distribution

Replace $\lambda_1 = \gamma_1\gamma_2$ and $\lambda_2 = \frac{\gamma_1}{\gamma_2}$:

$$
\begin{aligned}
& P_{Y_1,Y_2}(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2) \\
= \quad & \exp(-(\gamma_1(\gamma_2 + \gamma_2^{-1}) + \lambda_3)) \frac{(\gamma_1\gamma_2)^{y_1}}{y_1!} \frac{(\frac{\gamma_1}{\gamma_2})^{y_2}}{y_2!} \sum_{k=0}^{\min(y_1,y_2)} \binom{y_1}{k} \binom{y_2}{k} k! \left(\frac{\lambda_3}{\gamma_1^2}\right)^k
\end{aligned}
$$

$$
\begin{aligned}
\gamma_1 &= \exp(\beta_0) \\
\gamma_2 &= \exp(\tilde{\mathbf{x}}^T \boldsymbol{\beta}) \\
\lambda_3 &= \exp(\alpha_0 + |\tilde{\mathbf{x}}|^T \boldsymbol{\alpha})
\end{aligned}
$$

with $\tilde{\mathbf{x}} = \mathbf{x}_1 - \mathbf{x}_2$.

Re-parametrization of bivariate Poisson distribution

Replace $\lambda_1 = \gamma_1\gamma_2$ and $\lambda_2 = \frac{\gamma_1}{\gamma_2}$:

$$
\begin{aligned}
& P_{Y_1,Y_2}(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2) \\
= \quad & \exp(-(\gamma_1(\gamma_2 + \gamma_2^{-1}) + \lambda_3)) \frac{(\gamma_1\gamma_2)^{y_1}}{y_1!} \frac{(\frac{\gamma_1}{\gamma_2})^{y_2}}{y_2!} \sum_{k=0}^{\min(y_1,y_2)} \binom{y_1}{k} \binom{y_2}{k} k! \left(\frac{\lambda_3}{\gamma_1^2}\right)^k
\end{aligned}
$$

$$
\begin{aligned}
\gamma_1 = \exp(\beta_0) \quad &\Rightarrow \quad \lambda_1 = exp(\beta_0 + \tilde{\mathbf{x}}^T\beta) \\
\gamma_2 = \exp(\tilde{\mathbf{x}}^T\beta) \quad &\Rightarrow \quad \lambda_2 = exp(\beta_0 - \tilde{\mathbf{x}}^T\beta) \\
\lambda_3 = \exp(\alpha_0 + |\tilde{\mathbf{x}}|^T\boldsymbol{\alpha}) &
\end{aligned}
$$

with $\tilde{\mathbf{x}} = \mathbf{x}_1 - \mathbf{x}_2$.

Bivariate Poisson Model for Football Results

$$(y_{ik}, y_{jk})|x_{ik}, x_{jk} \quad \sim \quad Po_2(\gamma_1, \gamma_{ijk2}, \lambda_{ijk3})$$

- $\gamma_1 = \exp(\beta_0)$
- $\gamma_{ijk2} = \exp((\mathbf{x}_{ik} - \mathbf{x}_{jk})^T \beta)$
- $\lambda_{ijk3} = exp(\alpha_0 + |\mathbf{x}_{ik} - \mathbf{x}_{jk}|^T \boldsymbol{\alpha})$

Bivariate Poisson Model for Football Results

$$(y_{ik}, y_{jk})|x_{ik}, x_{jk} \quad \sim \quad Po_2(\gamma_1, \gamma_{ijk2}, \lambda_{ijk3})$$

- $\gamma_1 = \exp(\beta_0)$
- $\gamma_{ijk2} = \exp((\mathbf{x}_{ik} - \mathbf{x}_{jk})^T \beta)$
- $\lambda_{ijk3} = exp(\alpha_0 + |\mathbf{x}_{ik} - \mathbf{x}_{jk}|^T \boldsymbol{\alpha})$

$\Longrightarrow$ Framework of the so-called Generalized Additive Model for Location, Scale and Shape (GAMLSS; Rigby and Stasinopoulos, 2005)

Boosting for GAMLSS

- `R`-package `gamboostLSS` (Hofner, Mayr and Schmid, 2015)
- Allows for variable selection within GAMLSS framework
- Provides a large number of pre-specified distributions
  - Negative binomial distribution
  - Zero-inflated Poisson distribution
  - ...

Boosting for GAMLSS

- `R`-package `gamboostLSS` (Hofner, Mayr and Schmid, 2015)
- Allows for variable selection within GAMLSS framework
- Provides a large number of pre-specified distributions
  - Negative binomial distribution
  - Zero-inflated Poisson distribution
  - ...
- Mostly restricted to univariate responses, first approach for bivariate normal distribution from Andreas Mayr
- Users can specify new distributions (also bivariate) by providing
  - loss/risk function → neg. log-likelihood
  - neg. gradient of loss function → score function
  - possibly suitable offsets for linear predictors

Boosting for GAMLSS

- R-package gamboostLSS (Hofner, Mayr and Schmid, 2015)
- Allows for variable selection within GAMLSS framework
- Provides a large number of pre-specified distributions
  - Negative binomial distribution
  - Zero-inflated Poisson distribution
  - ...
- Mostly restricted to univariate responses, first approach for bivariate normal distribution from Andreas Mayr
- Users can specify new distributions (also bivariate) by providing
  - loss/risk function → neg. log-likelihood
  - neg. gradient of loss function → score function
  - possibly suitable offsets for linear predictors

⇒ We implemented bivariate Poisson distribution

Application to UEFA Europoean Championship 2016

Covariates

- **Economic Factors:**
  - GDP per capita
  - population

Covariates

- **Economic Factors:**
  - GDP per capita
  - population
- **Sportive Factors:**
  - Home advantage
  - ODDSET odds
  - market value
  - FIFA rank
  - UEFA points

Covariates

- **Economic Factors:**
  - GDP per capita
  - population
- **Sportive Factors:**
  - Home advantage
  - ODDSET odds
  - market value
  - FIFA rank
  - UEFA points
- **Factors describing the team's structure**
  - (Second) maximum number of team-mates
  - Average age
  - Number of CL players
  - Number of Europa League players
  - Age of the national coach
  - Nationality of the national coach
  - Number of players abroad

Structure of Dataset

- Data from 2004, 2008 and 2012
- All covariates are differences between values of both teams

	Team 1	Team 2	Goals 1	Goals 2	Year	Odds	Market Value	⋯
1	Portugal	Greece	1	2	2004	-39.0	7.85	⋯
2	Spain	Russia	1	0	2004	-33.5	7.67	⋯
3	Greece	Spain	1	1	2004	38.5	-7.58	⋯
4	Russia	Portugal	0	2	2004	34.0	-7.94	⋯
5	Spain	Portugal	0	1	2004	0.5	-0.27	⋯
6	Russia	Greece	2	1	2004	-5.0	-0.09	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Parameter Estimates

- $\gamma_1 = \exp(\hat{\beta}_0)$:

  Estimates: $\hat{\beta}_0 = 0.176$,

Parameter Estimates

- $\gamma_1 = \exp(\hat{\beta}_0)$:

  Estimates: $\hat{\beta}_0 = 0.176$,

- $\gamma_{ijk2} = \exp((\mathbf{x}_{ik} - \mathbf{x}_{jk})^T \boldsymbol{\beta})$:

  Estimates: $(\hat{\beta}_{odds}, \hat{\beta}_{marketvalue}, \hat{\beta}_{UEFApoints}) = (-0.120, 0.143, 0.029)$

- $\lambda_{ijk3} = exp(\alpha_0 + |\mathbf{x}_{ik} - \mathbf{x}_{jk}|^T \boldsymbol{\alpha})$:

  Estimates: $\hat{\alpha}_0 = -9.21$, $\hat{\boldsymbol{\alpha}} = \mathbf{0} \Rightarrow \lambda_{ijk3} \approx 0$

Parameter Estimates

- $\gamma_1 = \exp(\hat{\beta}_0)$:

  Estimates: $\hat{\beta}_0 = 0.176$,

- $\gamma_{ijk2} = \exp((\mathbf{x}_{ik} - \mathbf{x}_{jk})^T \boldsymbol{\beta})$:

  Estimates: $(\hat{\beta}_{odds}, \hat{\beta}_{marketvalue}, \hat{\beta}_{UEFApoints}) = (-0.120, 0.143, 0.029)$

- $\lambda_{ijk3} = exp(\alpha_0 + |\mathbf{x}_{ik} - \mathbf{x}_{jk}|^T \boldsymbol{\alpha})$:

  Estimates: $\hat{\alpha}_0 = -9.21$, $\hat{\boldsymbol{\alpha}} = \mathbf{0} \Rightarrow \lambda_{ijk3} \approx 0$

$\Rightarrow$ very simple model

$\Rightarrow$ no (additional) covariance between scores of both teams

Simulation of the tournament progress

- Every single match outcome can be simulated by drawing from the respective Poisson distributions
- Per group, the exact standing after the group stage can be calculated

$\Rightarrow$ Decision on qualification for round of 16 according to UEFA rules

- Draws in knockout stage:
  - Simulate extra time with 1/3 of Poisson parameters
  - Simulate penalty shootout by coin flip

$\Rightarrow$ 1,000,000 simulation runs for the UEFA European Championship 2016

Probabilities for UEFA European Football Champion 2016

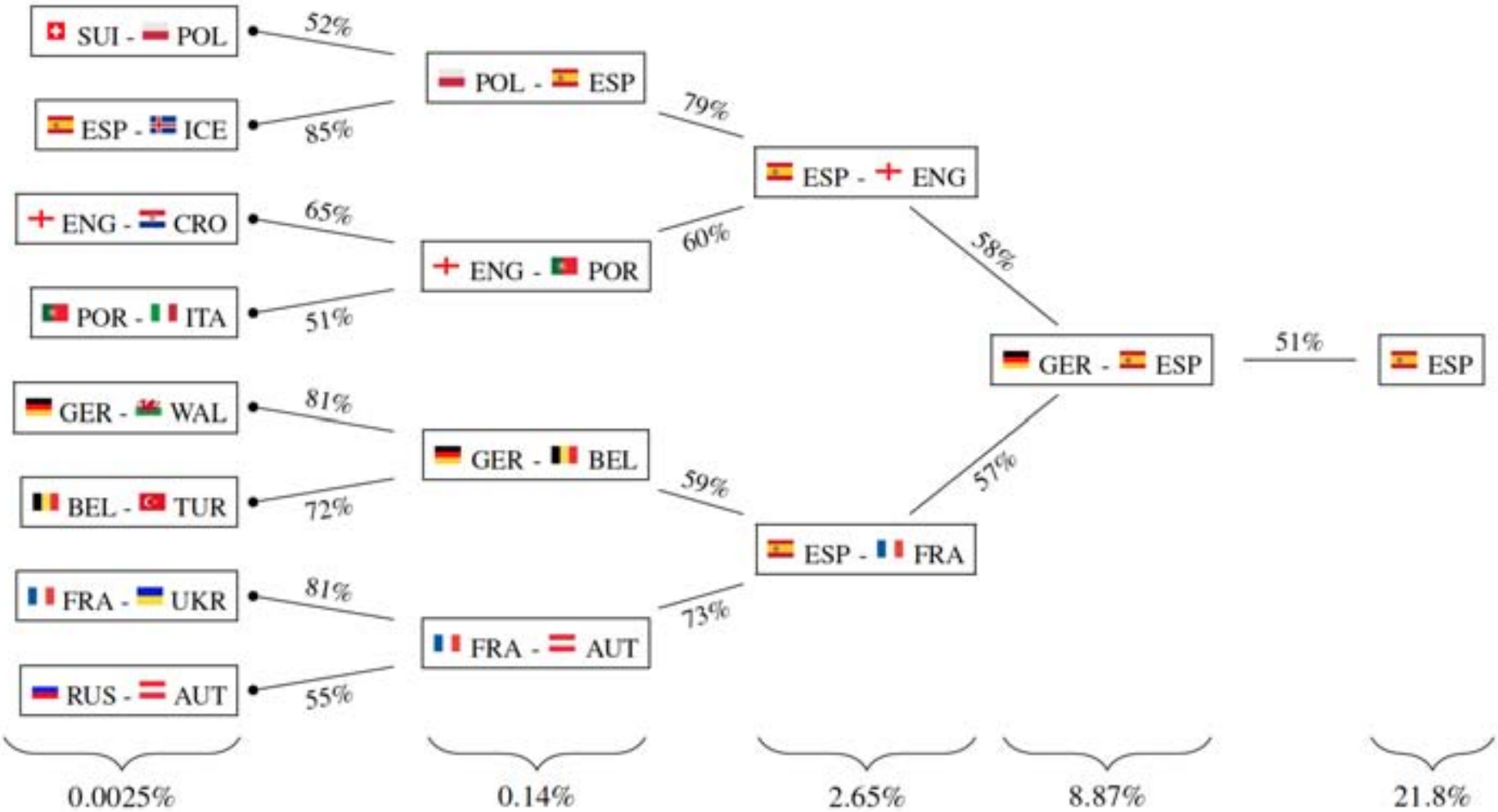
			Round of 16	Quarter Finals	Semi Finals	Final	European Champion	Oddset
1	Spain		95.4	72.9	52.3	35.1	21.8	13.9
2	Germany		99.3	79.5	51.3	34.4	21.0	16.9
3	France		97.5	71.9	48.2	25.8	13.8	18.9
4	England		95.2	69.4	43.4	23.9	12.9	9.2
5	Belgium		93.9	58.7	32.8	18.7	9.5	7.3
6	Portugal		92.5	52.3	27.4	12.6	5.5	4.5
7	Italy		87.7	47.6	23.8	11.4	4.8	5.3
8	Croatia		73.2	35.3	16.8	7.3	2.7	3.2
9	Poland		86.0	42.2	15.6	5.5	1.6	2.0
10	Austria		79.1	34.0	13.4	4.4	1.3	2.7
11	Switzerland		77.9	35.8	13.3	4.3	1.2	1.6
12	Turkey		56.1	21.2	8.3	2.8	0.8	1.6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Most Probable Final Group Standings

	A	B	C
1	France	England	Germany
2	Switzerland	Wales	Poland
3	Romania	Russia	Ukraine
4	Albania	Slovakia	Nor. Ireland
	21.2%	15.1%	37.6%

	D	E	F
1	Spain	Belgium	Portugal
2	Croatia	Italy	Austria
3	Turkey	Sweden	Iceland
4	Czech Rep.	Ireland	Hungary
	17.7%	17.5%	16.9%

Most Probable Course of Knockout Stage



Summary

Theoretical Results:

- Bivariate Poisson model for scores of both teams
- Implementation into framework of GAMLSS via `gamboostLSS`
- Very sparse model
- No additional covariance, reduces to two (conditionally) independent Poisson distributions

Prediction Results:

- Survival rates per team and tournament stage
- Most probable course of tournament
- Spain favorite team followed by Germany and France

References

Groll, A. and J. Abedieh (2013). Spain retains its title and sets a new record - generalized linear mixed models on European football championships. *Journal of Quantitative Analysis in Sports 9*(1), 51–66.

Groll, A., G. Schauberger, and G. Tutz (2015). Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: An application to the FIFA World Cup 2014. *Journal of Quantitative Analysis in Sports 11*(2), 97–115.

Hofner, B., A. Mayr, and M. Schmid (2015). gamboostLSS: An R package for model building and variable selection in the GAMLSS framework. *Journal of Statistical Software 74*(1), 1–31.

Karlis, D. and I. Ntzoufras (2003). Analysis of sports data by using bivariate poisson models. *The Statistician 52*, 381–393.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 54*(3), 507–554.

Stasinopoulos, D. M. and R. A. Rigby (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software 23*(7), 1–46.

First Idea for Bivariate Model

$$(y_{ik}, y_{jk})|x_{ik}, x_{jk} \quad \sim \quad Po_2(\lambda_{ik1}, \lambda_{jk2}, \lambda_{ijk3})$$

- $log(\lambda_{ik1}) = \beta_0 + \mathbf{x}_{ik}^T \beta$
- $log(\lambda_{jk2}) = \beta_0 + \mathbf{x}_{jk}^T \beta$
- $log(\lambda_{ijk3}) = \alpha_0 + |\mathbf{x}_{ik} - \mathbf{x}_{jk}|^T \boldsymbol{\alpha}$

GAMLSS

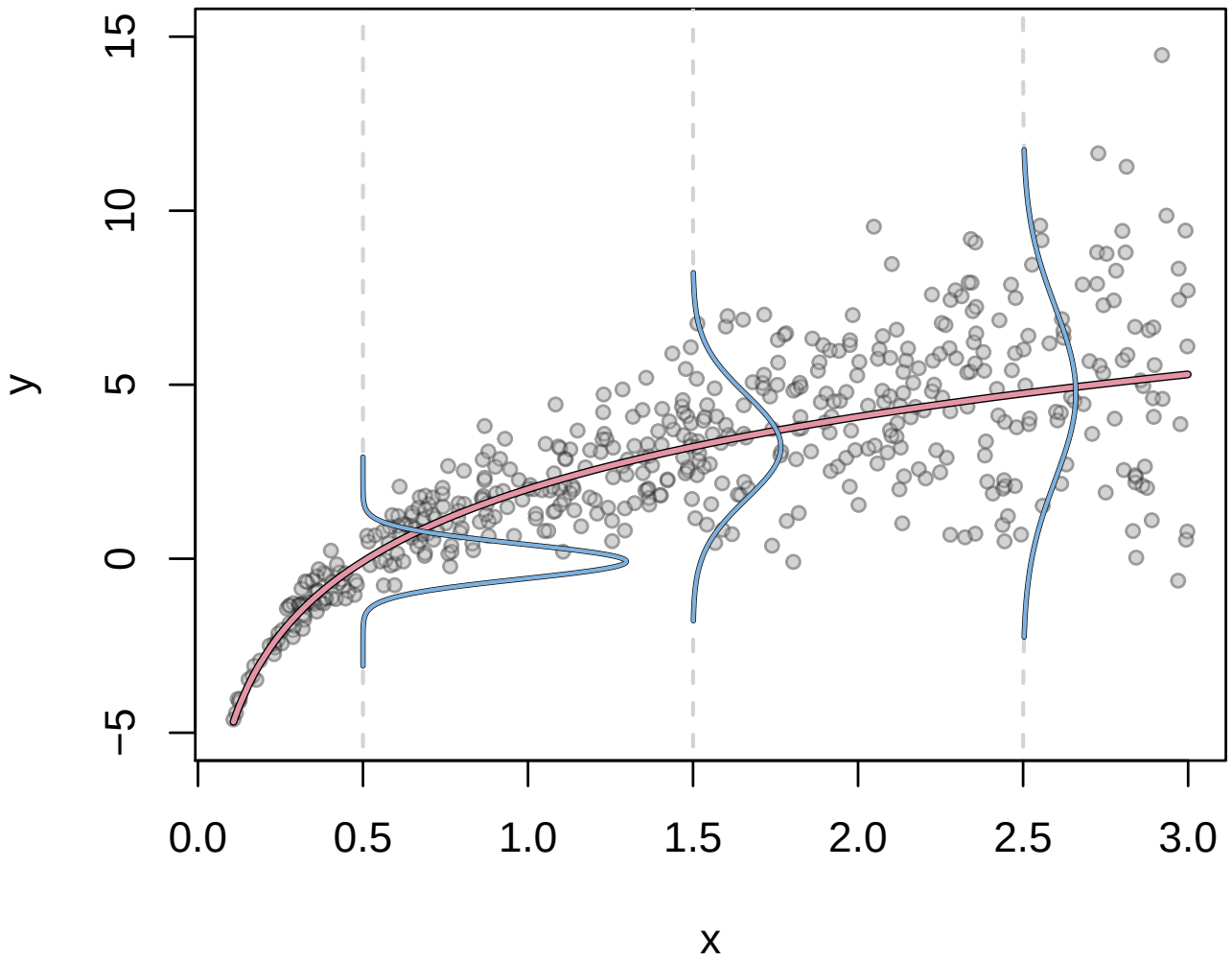
Generalized Additive Models for Location, Scale and Shape

$$
\begin{aligned}
g_1(\mu) &= \eta_\mu = \beta_{0\mu} + \sum_{j=1}^{p_1} f_{j\mu}(x_j) && \text{"location"} \\
g_2(\sigma) &= \eta_\sigma = \beta_{0\sigma} + \sum_{j=1}^{p_2} f_{j\sigma}(x_j) && \text{"scale"} \\
&\vdots \qquad \vdots
\end{aligned}
$$

- Proposed by Rigby and Stasinopoulos (2005)
- Extension of generalized additive models (GAMs)
- The distribution parameters are modeled by specific predictors and associated link functions $g_k(\cdot)$.

Example for GAMLSS

$$Y \sim N(\mu = \beta_{0\mu} + f_\mu(x),\ \sigma = \exp(\beta_{0\sigma} + f_\sigma(x)))$$



First Idea for Bivariate Model

$$(y_{ik}, y_{jk})|x_{ik}, x_{jk} \quad \sim \quad Po_2(\lambda_{ik1}, \lambda_{jk2}, \lambda_{ijk3})$$

- $log(\lambda_{ik1}) = \beta_0 + \mathbf{x}_{ik}^T\beta$
- $log(\lambda_{jk2}) = \beta_0 + \mathbf{x}_{jk}^T\beta$
- $log(\lambda_{ijk3}) = \alpha_0 + |\mathbf{x}_{ik} - \mathbf{x}_{jk}|^T\boldsymbol{\alpha}$

In general, in GAMLSS effects for predictors differ across different components

$\Rightarrow$ Restrictions for parameters would become necessary!

$\Rightarrow$ Solution:

- Re-parametrize bivariate Poisson distribution
- Use differences between covariates