Modeling and Prediction of International Soccer Matches

A. Groll

Faculty of Statistics,
Technical University Dortmund

(joint work with A. Mayr, T. Kneib, G. Schauberger & G. Tutz)

Workshop on Big Data Analytics in Sports,
Brescia, 16th November 2017

technische universität
dortmund

Who will celebrate?

Quellen: youtube.com, EMAJ Magazine, youfrisky.com, Bailiwick Express

Who will cry?

Quellen: youtube.com,pinterest,BBC,theage.com.au

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**

Model for international soccer tournaments

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

$$\lambda_{ijk} = \exp\left(\beta_0 + (x_{ik} - x_{jk})^\top \boldsymbol{\beta} + att_i - def_j\right)$$

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

$\boldsymbol{\beta}$: Parameter vector of covariate effects

att_i: (Additional) attack parameter of team i

def_i: (Additional) defense parameter of opponent j

Model for international soccer tournaments

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

$$\lambda_{ijk} = \exp\left(\beta_0 + (x_{ik} - x_{jk})^\top \boldsymbol{\beta} + att_i - def_j\right)$$

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

$\boldsymbol{\beta}$: Parameter vector of covariate effects

att_i: (Additional) attack parameter of team i

def_i: (Additional) defense parameter of opponent j

$\Longrightarrow$ Univariate and conditionally independent!

Regularized estimation

Maximize penalized log-likelihood

$$l_p(\boldsymbol{\alpha}) = l(\boldsymbol{\alpha}) - \lambda J(\boldsymbol{\alpha}),$$

with penalty term

$$J(\boldsymbol{\alpha}) = \underbrace{\sum_{i=1}^{p} |\beta_i|}_{LASSO} + \underbrace{\sqrt{2} \sum_{i=1}^{n} \sqrt{att_i^2 + def_i^2}}_{Group\ LASSO}$$

where $\boldsymbol{\alpha}$ collects all parameters.

The model is estimated with the R-package `grplasso` (Meier et al., 2008).

Covariates

Data basis: Word Cups 2002–2010

Covariates

Data basis: Word Cups 2002–2010

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

Covariates

Data basis: Word Cups 2002–2010

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

- **Sportive Factors:**
  bookmaker's odds, FIFA rank

Covariates

Data basis: Word Cups 2002–2010

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

- **Sportive Factors:**
  bookmaker's odds, FIFA rank

- **Home advantage:**
  host of the world cup, same continent as host

Covariates

Data basis: Word Cups 2002–2010

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

- **Sportive Factors:**
  bookmaker's odds, FIFA rank

- **Home advantage:**
  host of the world cup, same continent as host

- **Factors describing the team's structure**
  (Second) Maximum number of teammates, average age, number of Champions League & Europa League players, number of players abroad

Covariates

Data basis: Word Cups 2002–2010

- **Economic Factors:**
  GDP per capita, population
- **Sportive Factors:**
  bookmaker's odds, FIFA rank
- **Home advantage:**
  host of the world cup, same continent as host
- **Factors describing the team's structure**
  (Second) Maximum number of teammates, average age, number of Champions League & Europa League players, number of players abroad
- **Factors describing the team's coach**
  age, nationality, tenure

Covariates

Data basis: Word Cups 2002–2010

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

- **Sportive Factors:**
  bookmaker's odds, FIFA rank

- **Home advantage:**
  host of the world cup, same continent as host

- **Factors describing the team's structure**
  (Second) Maximum number of teammates, average age, number of Champions League & Europa League players, number of players abroad

- **Factors describing the team's coach**
  age, nationality, tenure

All variables are incorporated as differences between the team whose goals are considered and its opponent!

Extract of the design matrix

FRA	0:0	URU
URU	1:2	DEN

Team	Age	Rank	Odds	...
France	28.3	1	0.149	...
Uruguay	25.3	24	0.009	...
Denmark	27.4	20	0.012	...
⋮	⋮	⋮	⋮	⋱

Extract of the design matrix

FRA	0:0	URU
URU	1:2	DEN

Team	Age	Rank	Odds	...
France	28.3	1	0.149	...
Uruguay	25.3	24	0.009	...
Denmark	27.4	20	0.012	...
⋮	⋮	⋮	⋮	⋱

Goals	Team	Opponent	Age	Rank	Odds	...
0	France	Uruguay	3.00	-23	0.140	...
0	Uruguay	France	-3.00	23	-0.140	...
1	Uruguay	Denmark	-2.10	4	-0.003	...
2	Denmark	Uruguay	2.10	-4	0.003	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

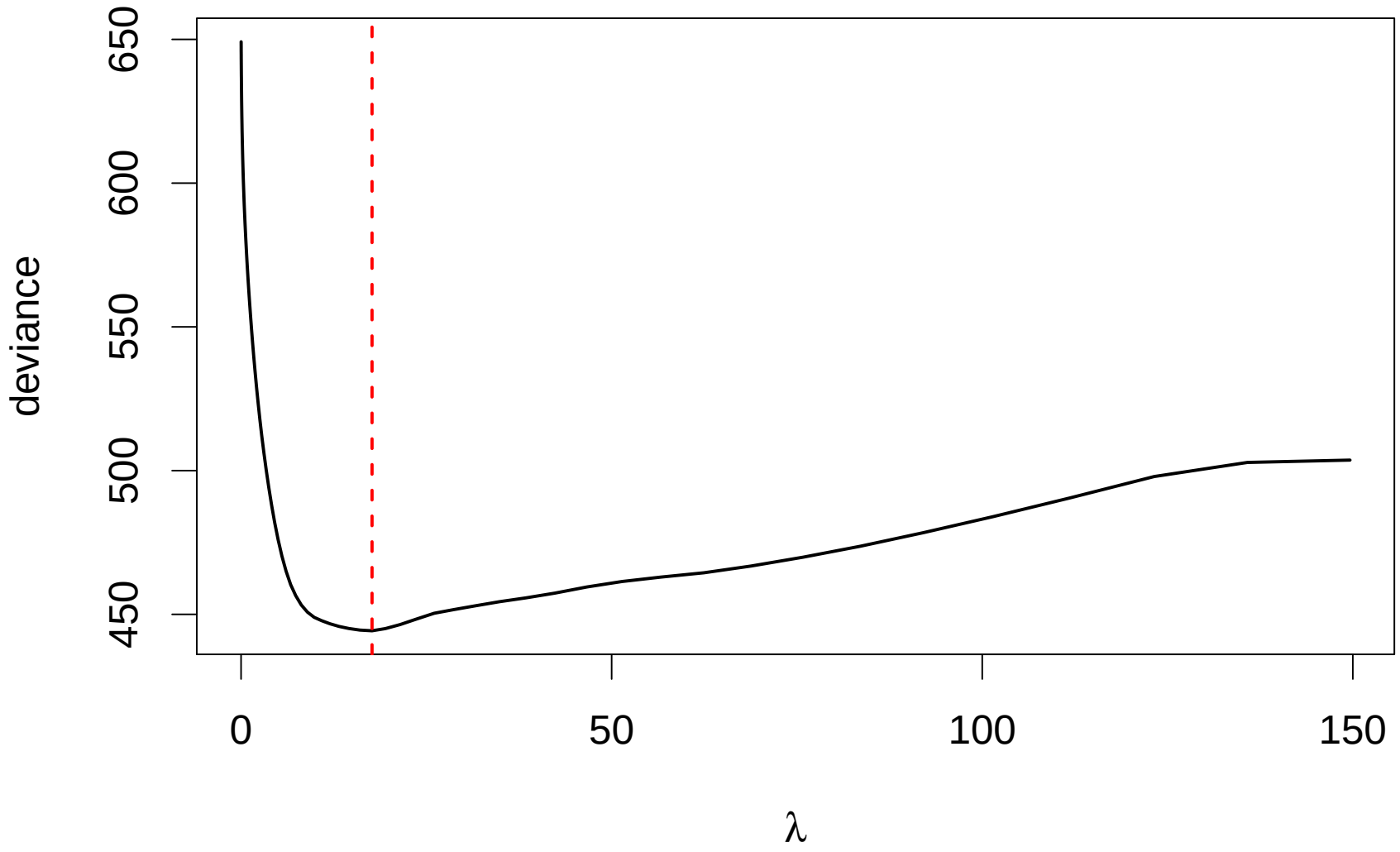
Extract of the design matrix

FRA	0:0	URU
URU	1:2	DEN

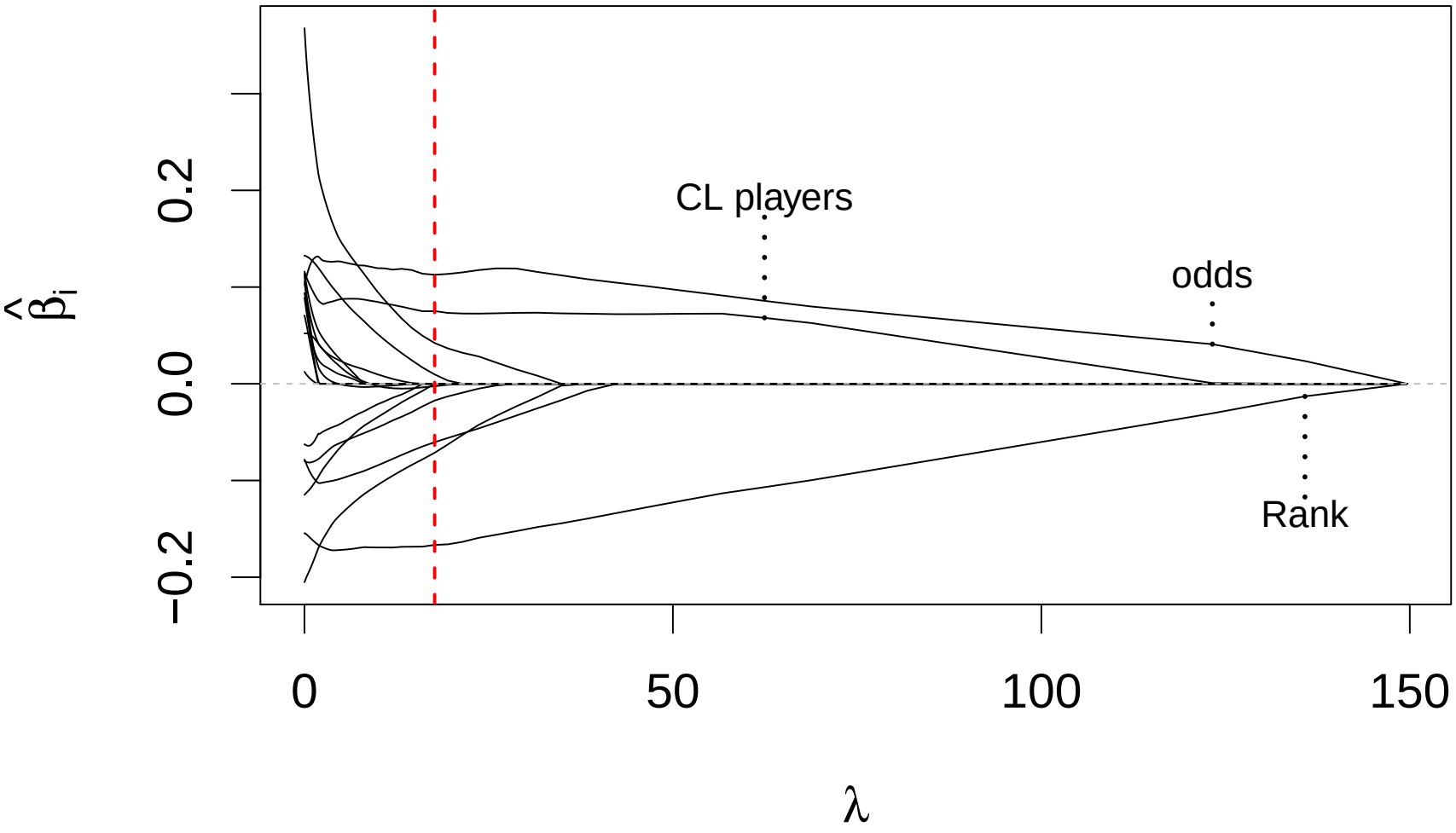
Team	Age	Rank	Odds	...
France	28.3	1	0.149	...
Uruguay	25.3	24	0.009	...
Denmark	27.4	20	0.012	...
⋮	⋮	⋮	⋮	⋱

FRA.att	FRA.def	URU.att	URU.def	DEN.att	DEN.def
1	0	0	-1	0	0
0	-1	1	0	0	0
0	0	1	0	0	-1
0	0	0	-1	1	0
⋮	⋮	⋮	⋮	⋮	⋮

Deviance for 10-fold CV



Coefficient paths of the covariate effects



Estimates of the covariate effects

CL players	0.075
UEFA players	0
Age Coach	-0.017
Tenure Coach	-0.071
Legionaires	0
Max. # teammates	0
Sec. max # teammates	0
Age	0
Rank	-0.167
GDP	0.042
Odds	0.113
Population	-0.060
Continent0	0.010
Continent1	-0.003
Nation Coach0	0
Nation Coach1	0
Host0	0
Host1	0

(Unscaled) estimates of team-specific effects

estimated attack parameters				estimated defense parameters			
1.		GER	0.237	1.		SUI	0.599
2.		BRA	0.114	2.		ALG	0.205
3.		URU	0.101	3.		GER	0.181
4.		CRC	0.099	4.		HON	0.065
5.		RSA	0.060	5.		BRA	0.057
6.		BEL	0.042	6.		FRA	0.046
7.		POR	0.019	7.		POR	0.030
8.		ARG	0.000	8.		PAR	0.021
9.		AUS	0.000	9.		ARG	0.000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
28.		USA	0.000	28.		SWE	0.000
29.		SVN	-0.002	29.		USA	0.000
30.		PAR	-0.003	30.		SVN	-0.009
31.		POL	-0.005	31.		POL	-0.012
32.		IRN	-0.040	32.		URU	-0.019
33.		SRB	-0.047	33.		RSA	-0.022
34.		HON	-0.083	34.		BEL	-0.085
35.		FRA	-0.198	35.		CMR	-0.090
36.		SUI	-0.202	36.		IRN	-0.097
37.		CMR	-0.204	37.		SRB	-0.153
38.		TUN	-0.234	38.		TUN	-0.297
39.		ALG	-0.340	39.		CRC	-0.526
40.		KSA	-0.495	40.		KSA	-0.788

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 FIFA (UEFA) rules

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

$\Rightarrow$ 100 000 (1 000 000) simulation runs for the WC 2014 (Euro 2016)

Estimated probabilities (in %) for reaching the different stages in the FIFA World Cup 2014

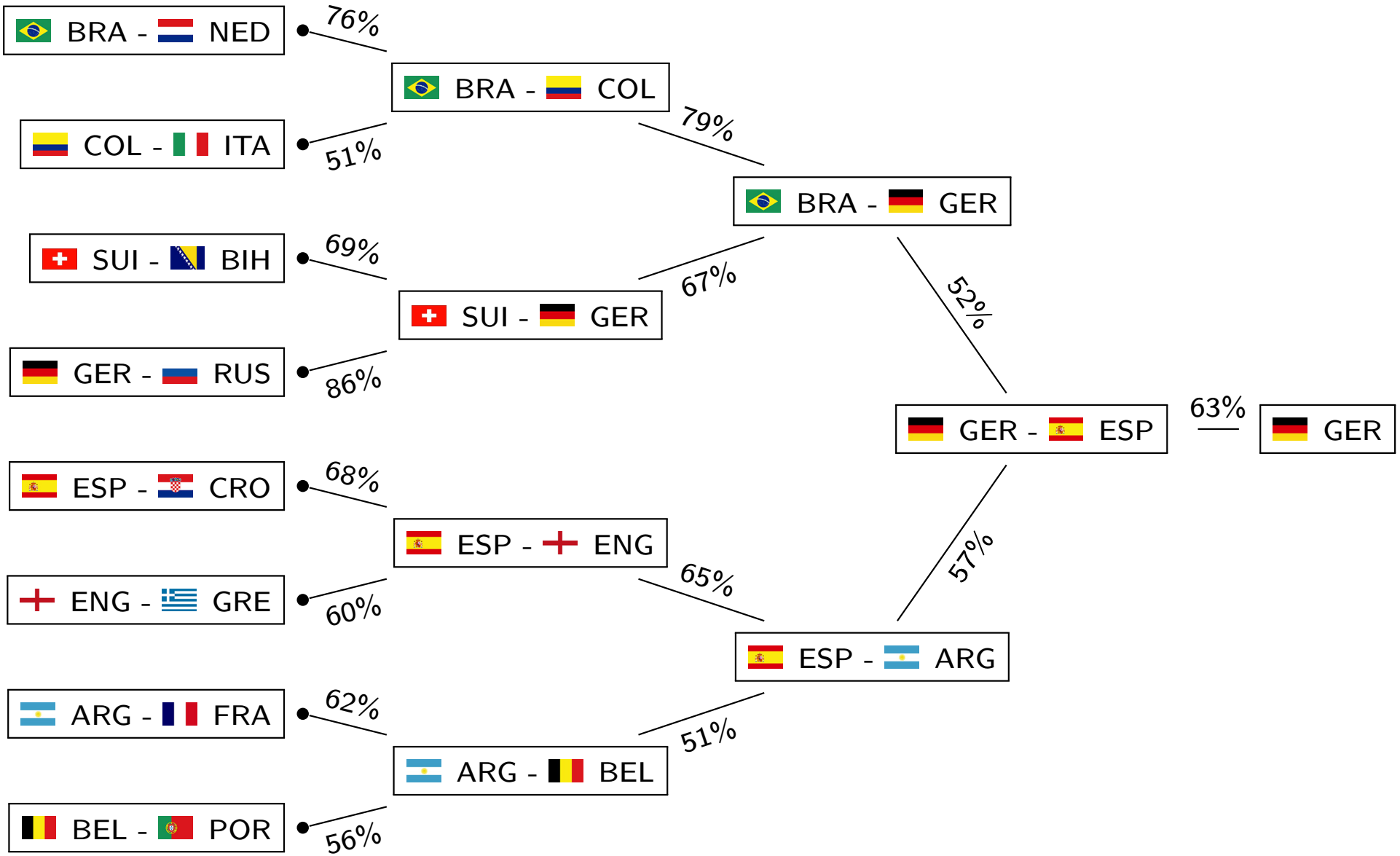
		Round of 16	Quarter finals	Semi finals	Final	World Champion	Bookmaker's Odds
1.	GER	91.4	77.9	57.0	39.2	**27.6**	**14.2**
2.	BRA	91.8	67.9	54.4	30.9	20.0	20.3
3.	SUI	84.2	62.0	35.0	21.6	12.5	0.7
4.	ESP	84.2	52.0	37.8	21.6	12.1	10.9
5.	ARG	90.6	53.2	26.7	15.5	7.3	14.2
6.	POR	60.2	38.6	20.2	10.4	3.6	2.4
7.	BEL	82.5	36.3	19.8	9.3	3.4	5.9
8.	ENG	70.4	41.2	14.7	5.5	1.8	3.5
9.	CRO	58.1	26.1	15.5	5.1	1.6	0.7
10.	FRA	51.2	26.5	9.8	4.6	1.1	3.5
11.	ITA	56.8	31.8	10.8	4.4	1.3	3.5
12.	NED	55.7	21.3	11.8	4.1	1.2	3.5

Most probable final group standings

Group A 39%	Group B 26%	Group C 15%	Group D 19%
1. BRA	1. ESP	1. COL	1. ENG
2. CRO	2. NED	2. GRE	2. ITA
MEX	CHI	JPN	URU
CMR	AUS	CIV	CRC

Group E 19%	Group F 29%	Group G 38%	Group H 23%
1. SUI	1. ARG	1. GER	1. BEL
2. FRA	2. BIH	2. POR	2. RUS
ECU	NGA	GHA	ALG
HON	IRN	USA	KOR

Most probable course of the knockout stage



Univariate Model for International Soccer Tournaments

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

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

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 tournament k

λ_{ijk}: expected number of goals of team i against opponent j attournament k

ξ_{ik}, δ_{jk}: playing abilities of team i and opponent j

z.B. für **WM 2014** (Groll, Schauberger and Tutz, 2015):

$$\xi_{ik} = \mathbf{x}_{ik}^T\beta + att_i$$
$$\delta_{jk} = \mathbf{x}_{jk}^T\beta + def_j$$

Univariate Model for International Soccer Tournaments

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

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

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 tournament k

λ_{ijk}: expected number of goals of team i against opponent j attournament k

ξ_{ik}, δ_{jk}: playing abilities of team i and opponent j

z.B. für **WM 2014** (Groll, Schauberger and Tutz, 2015):

$$\xi_{ik} = \mathbf{x}_{ik}^T\beta + att_i$$
$$\delta_{jk} = \mathbf{x}_{jk}^T\beta + def_j$$
$$\Rightarrow log(\lambda_{ijk}) = \beta_0 + (\mathbf{x}_{ik} - \mathbf{x}_{jk})^T\beta + att_i - def_j$$

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:

$$\begin{aligned} & P_{Y_1, Y_2}(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2) \\ = \quad & \exp(-(\lambda_1 + \lambda_2 + \lambda_3)) \frac{\lambda_1^{y_1}}{y_1!} \frac{\lambda_2^{y_2}}{y_2!} \sum_{k=0}^{\min(y_1, y_2)} \binom{y_1}{k} \binom{y_2}{k} k! \left(\frac{\lambda_3}{\lambda_1 \lambda_2} \right)^k \end{aligned}$$

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

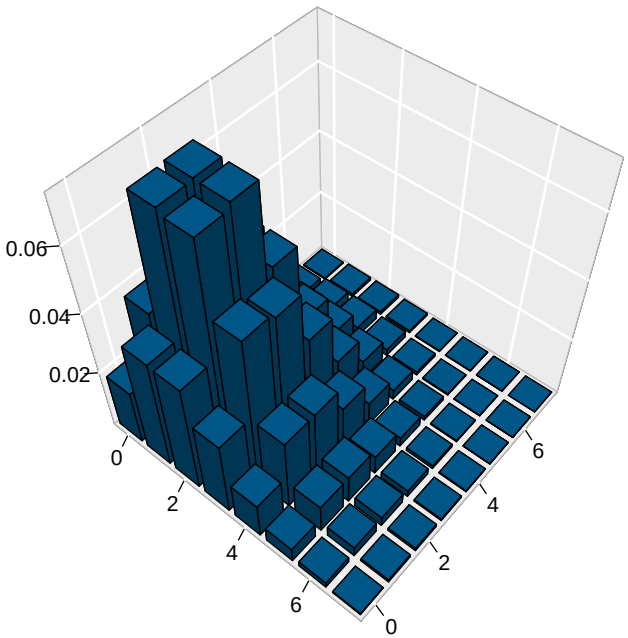
$$(Y_1, Y_2) \sim Po_2(\lambda_1, \lambda_2, \lambda_3)$$

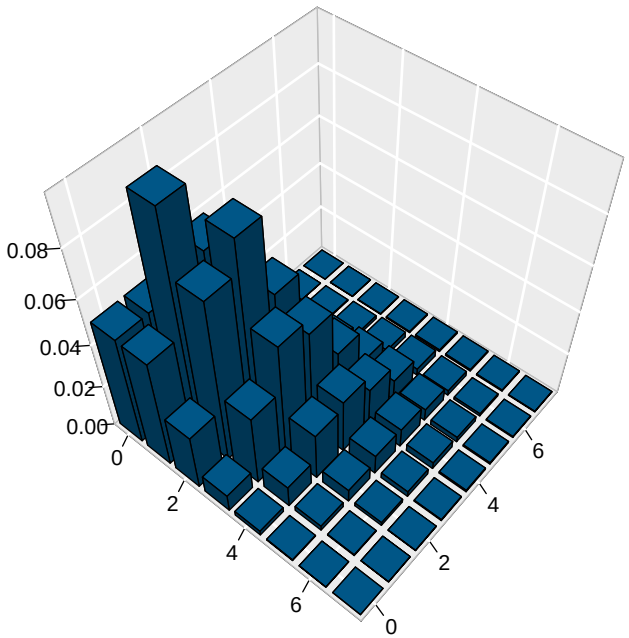
Probability function:

$$\begin{aligned} & P_{Y_1, Y_2}(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2) \\ = \quad & \exp(-(\lambda_1 + \lambda_2 + \lambda_3)) \frac{\lambda_1^{y_1}}{y_1!} \frac{\lambda_2^{y_2}}{y_2!} \sum_{k=0}^{\min(y_1, y_2)} \binom{y_1}{k} \binom{y_2}{k} k! \left(\frac{\lambda_3}{\lambda_1 \lambda_2} \right)^k \end{aligned}$$

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

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 $\lambda_2 = \frac{\gamma_1}{\gamma_2}$:

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

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

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

Re-parametrization of bivariate Poisson distribution

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

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

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

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

$\Longrightarrow$ 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 teammates
  - Average age
  - Number of Champions League players
  - Number of Europa League players
  - Number of players abroad
  - Age of the national coach
  - Nationality of the national coach

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

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

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

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

Probabilities for UEFA European Football Champion 2016

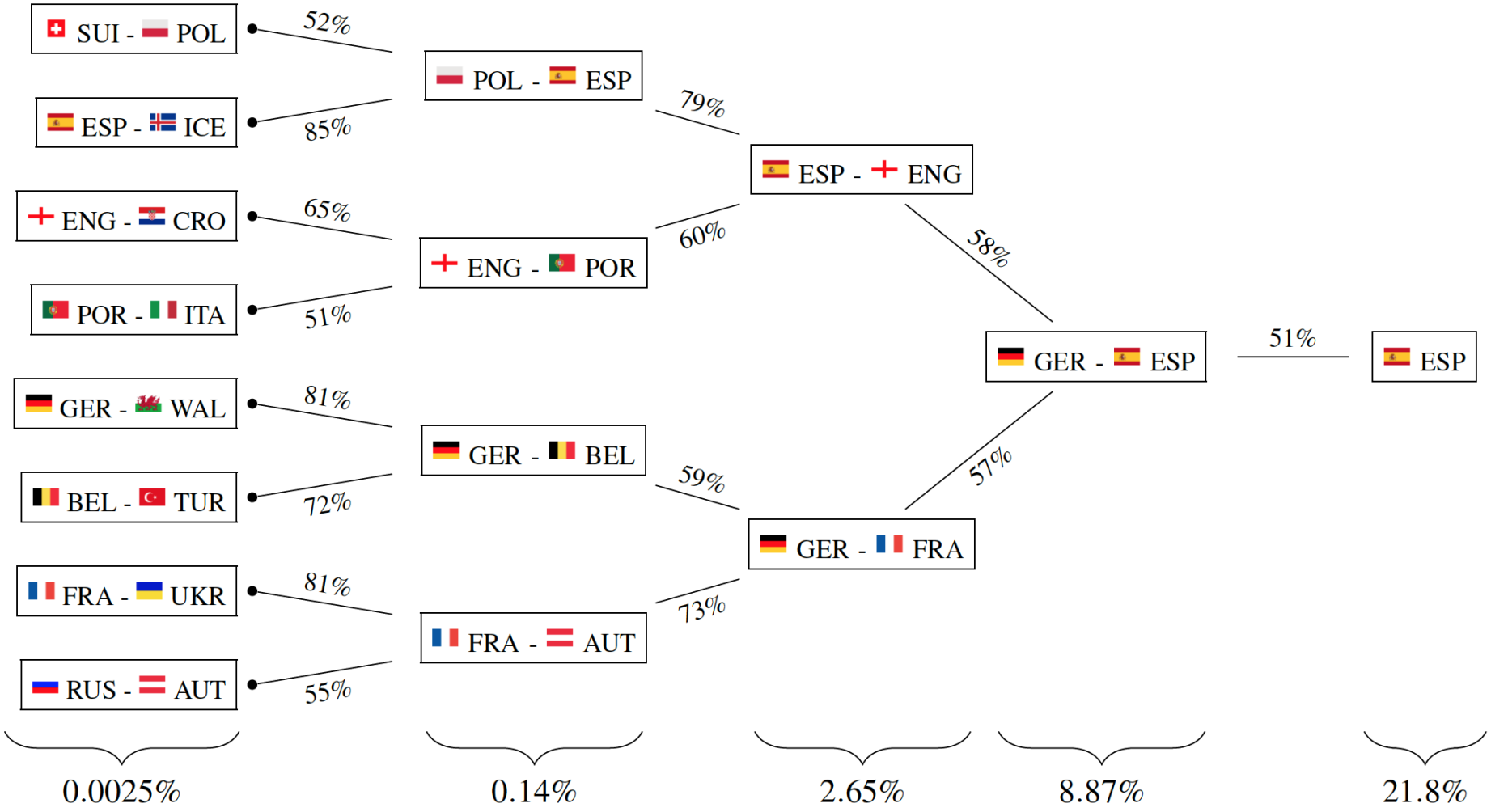
		Round of 16	Quarter Finals	Semi Finals	Final	European Champion	Oddset
1	Spain	95.4	72.9	52.3	35.1	21.8	13.9
2	Germany	99.3	79.5	51.3	34.4	21.0	16.9
3	France	97.5	71.9	48.2	25.8	13.8	18.9
4	England	95.2	69.4	43.4	23.9	12.9	9.2
5	Belgium	93.9	58.7	32.8	18.7	9.5	7.3
6	Portugal	92.5	52.3	27.4	12.6	**5.5**	**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:

- (Conditionally) independent vs. bivariate Poisson model for soccer scores
- Different types of regularization (penalization & boosting)
- Boosting: implementation into framework of GAMLSS via `gamboostLSS`
- Sparse models with very few covariates
- No additional covariance, reduces to two (conditionally) independent Poisson distributions

Prediction Results:

- Survival rates per team and tournament stage
- Most probable course of tournament

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.

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

Extended references

Regularization:

Groll, A. and G. Tutz (2012): Regularization for generalized additive mixed models by likelihood-based boosting, *Methods of Information in Medicine 51*(2), 168–177.

Tutz, G. and **A. Groll** (2012): Likelihood-based boosting in binary and ordinal random effects models, *Journal of Computational and Graphical Statistics 22*(2), 356–378.

Groll, A. and G. Tutz (2014): Variable selection for generalized linear mixed models by L1-penalized estimation, *Statistics and Computing 24*(2), 137–154.

Groll, A., T. Hastie and G. Tutz, G. (2017): Selection of Effects in Cox Frailty Models by Regularization Methods, *Biometrics*, DOI: 10.1111/biom.12637.

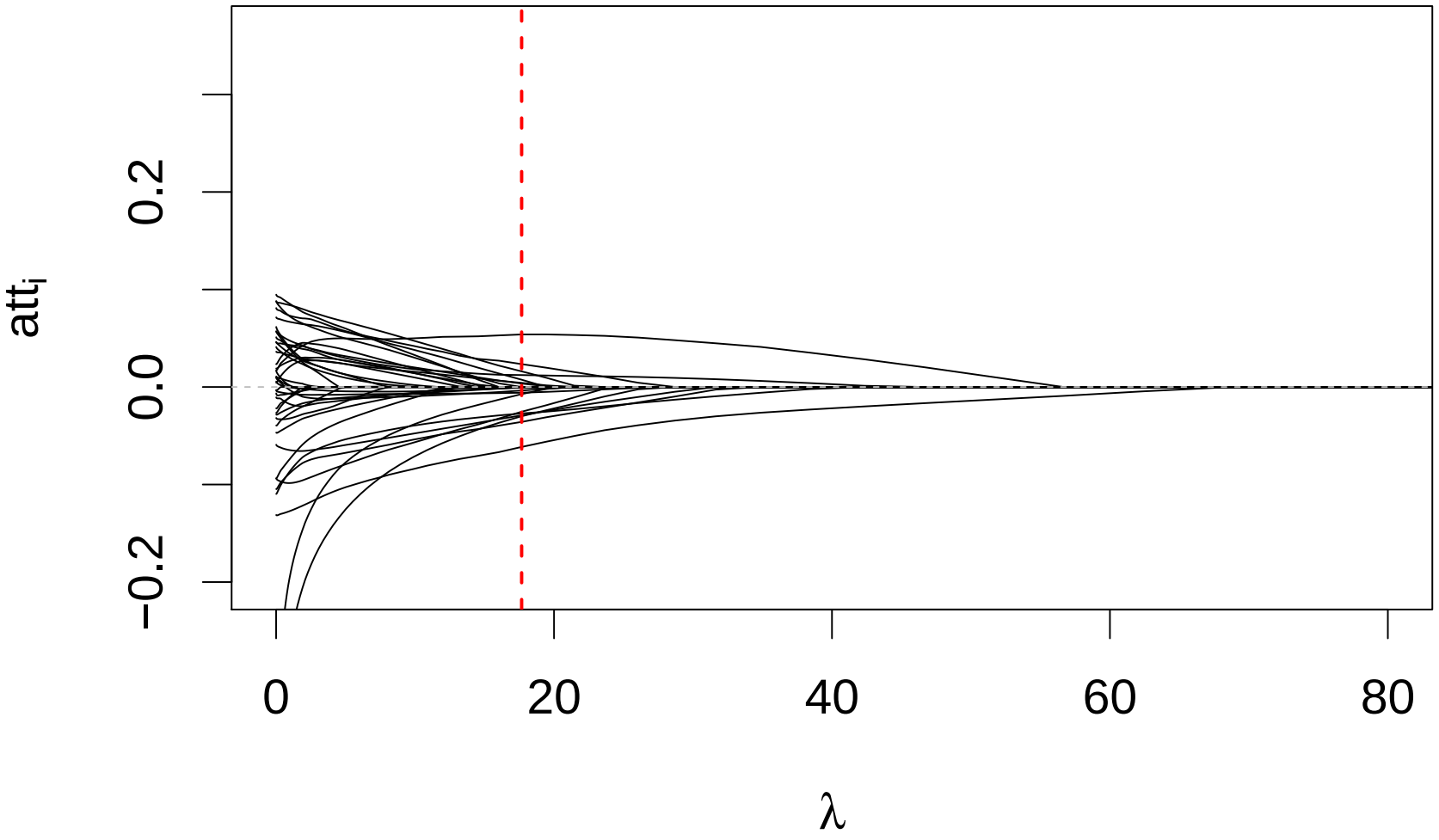
Sports statistics:

Groll, A. and J. Abedieh (2014): A Study on European Football Championships in the GLMM Framework with an Emphasis on UEFA Champions League Experience, In J. R. Bozeman, V. Girardin and C. H. Skiadas (Eds.), *New perspectives on stochastic modeling and data analysis*, Athens: ISAST, 313–321.

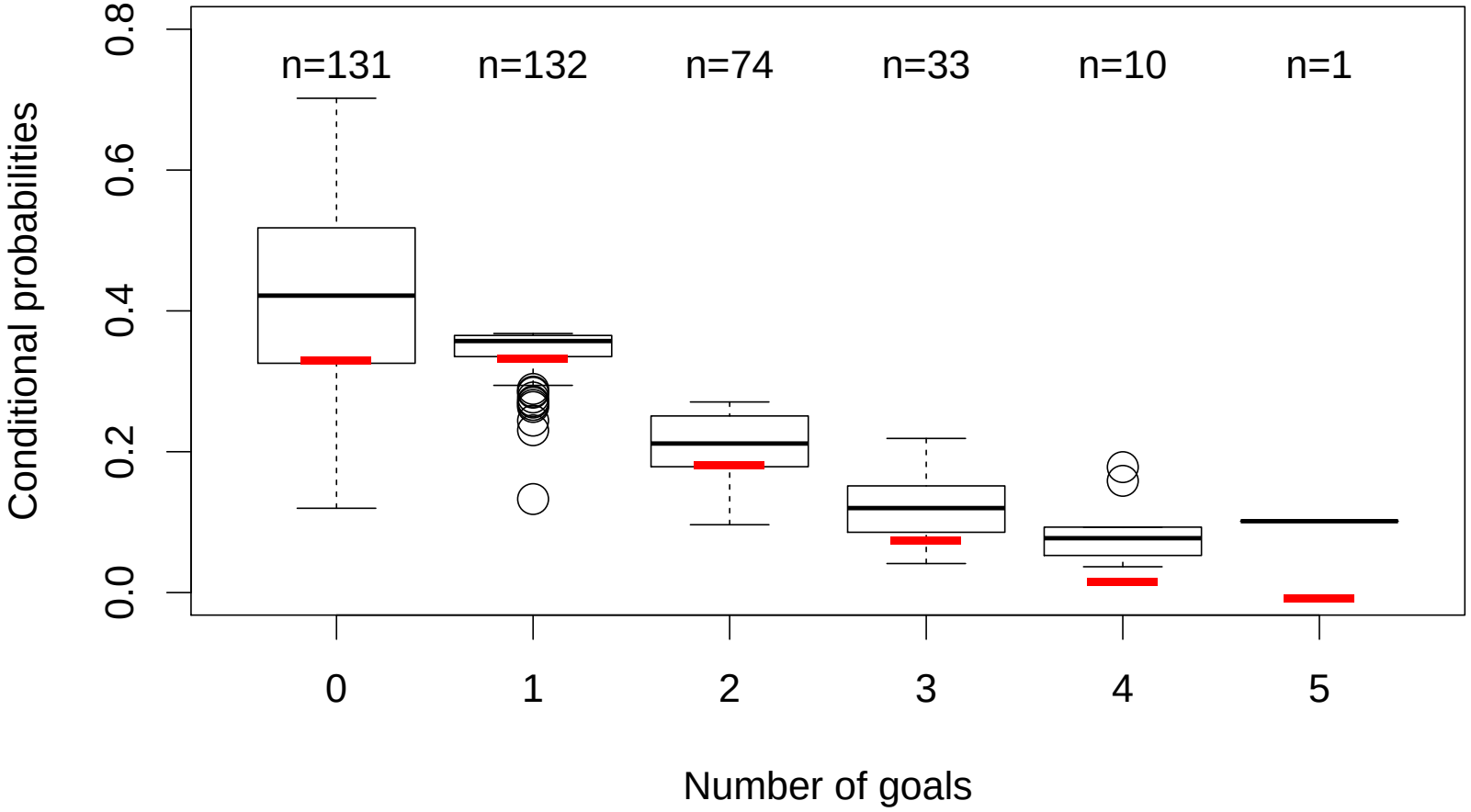
Schauberger, G., **A. Groll** and G. Tutz (2017): Analysis of the importance of on-field covariates in the German Bundesliga. *Journal of Applied Statistics*, DOI: 10.1080/02664763.2017.1383370.

Thank you
for your attention!

Team-specific attack parameters



Goodness-of-fit



Predictive power I

- a possible measurement of the predictive power of the model is to compare it to the "three-way" ODDSET odds for all 64 games of FIFA World Cup 2014

Predictive power I

- a possible measurement of the predictive power of the model is to compare it to the "three-way" ODDSET odds for all 64 games of FIFA World Cup 2014

- compute the three quantities $\tilde{p}_r = 1/\text{odds}_r, r \in \{1, 2, 3\}$, normalizing with $c := \sum_r^3 \tilde{p}_r$ (adjust for the bookmakers' margins)
  $\Longrightarrow$ estimated probabilities $\hat{p}_r = \tilde{p}_r / c$

Predictive power I

- a possible measurement of the predictive power of the model is to compare it to the "three-way" ODDSET odds for all 64 games of FIFA World Cup 2014

- compute the three quantities $\tilde{p}_r = 1/\text{odds}_r, r \in \{1,2,3\}$, normalizing with $c := \sum_r^3 \tilde{p}_r$ (adjust for the bookmakers' margins)
  $\Longrightarrow$ estimated probabilities $\hat{p}_r = \tilde{p}_r/c$

- Use the corresponding (estimated) Poisson distributions $G_{ij} \sim Poisson(\hat{\mu}_{ij})$, $G_{ji} \sim Poisson(\hat{\mu}_{ji})$ to estimate the same probabilities by approximating $\hat{p}_1 = P(G_{ij} > G_{ji}), \hat{p}_2 = P(G_{ij} = G_{ji})$ and $\hat{p}_3 = P(G_{ij} < G_{ji})$ for all 64 games

Predictive power I

- a possible measurement of the predictive power of the model is to compare it to the "three-way" ODDSET odds for all 64 games of FIFA World Cup 2014

- compute the three quantities $\tilde{p}_r = 1/\text{odds}_r, r \in \{1,2,3\}$, normalizing with $c := \sum_r^3 \tilde{p}_r$ (adjust for the bookmakers' margins)
  $\Longrightarrow$ estimated probabilities $\hat{p}_r = \tilde{p}_r/c$

- Use the corresponding (estimated) Poisson distributions $G_{ij} \sim Poisson(\hat{\mu}_{ij})$, $G_{ji} \sim Poisson(\hat{\mu}_{ji})$ to estimate the same probabilities by approximating $\hat{p}_1 = P(G_{ij} > G_{ji}), \hat{p}_2 = P(G_{ij} = G_{ji})$ and $\hat{p}_3 = P(G_{ij} < G_{ji})$ for all 64 games

- For $\omega_m \in \{1,2,3\}, m = 1,\ldots,64$, the average probability of a correct prediction is

$$\bar{p}_{three-way} = \frac{1}{64} \sum_{m=1}^{64} \hat{p}_{1m}^{\delta_{\mathbf{1}\omega_m}} \hat{p}_{2m}^{\delta_{\mathbf{2}\omega_m}} \hat{p}_{3m}^{\delta_{\mathbf{3}\omega_m}}$$

Predictive power I

- a possible measurement of the predictive power of the model is to compare it to the "three-way" ODDSET odds for all 64 games of FIFA World Cup 2014

- compute the three quantities $\tilde{p}_r = 1/\text{odds}_r, r \in \{1,2,3\}$, normalizing with $c := \sum_r^3 \tilde{p}_r$ (adjust for the bookmakers' margins)
  $\Longrightarrow$ estimated probabilities $\hat{p}_r = \tilde{p}_r/c$

- Use the corresponding (estimated) Poisson distributions $G_{ij} \sim Poisson(\hat{\mu}_{ij})$, $G_{ji} \sim Poisson(\hat{\mu}_{ji})$ to estimate the same probabilities by approximating $\hat{p}_1 = P(G_{ij} > G_{ji}), \hat{p}_2 = P(G_{ij} = G_{ji})$ and $\hat{p}_3 = P(G_{ij} < G_{ji})$ for all 64 games

- For $\omega_m \in \{1,2,3\}, m = 1,\ldots,64$, the average probability of a correct prediction is

$$\bar{p}_{three-way} = \frac{1}{64} \sum_{m=1}^{64} \hat{p}_{1m}^{\delta_{\mathbf{1}\omega_m}} \hat{p}_{2m}^{\delta_{\mathbf{2}\omega_m}} \hat{p}_{3m}^{\delta_{\mathbf{3}\omega_m}}$$

WC2002	bookmakers' odds
40.33%	41.45%

Predictive power II

The model and its predicted probabilities served the authors as basis for the following (fortunately successful!) betting strategy:

- bet on 3-way outcome with the highest expected return (given the model holds)
- the highest expected return is calculated as the product of the model's predicted probability and 3-way odd of the bookmakers
- only bet on matches with positive expected return

$\Rightarrow$ placing equal sized bets yielded a return of 33.52%

Correlation between Scores of Both Teams

Dixon and Coles (1997) compared marginal distributions of scores with joint distribution ⇒ 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)

⇒ Introduction of additional dependence parameter

But: They did not compare conditional distributions!

⇒ Their linear predictors are not independent!

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

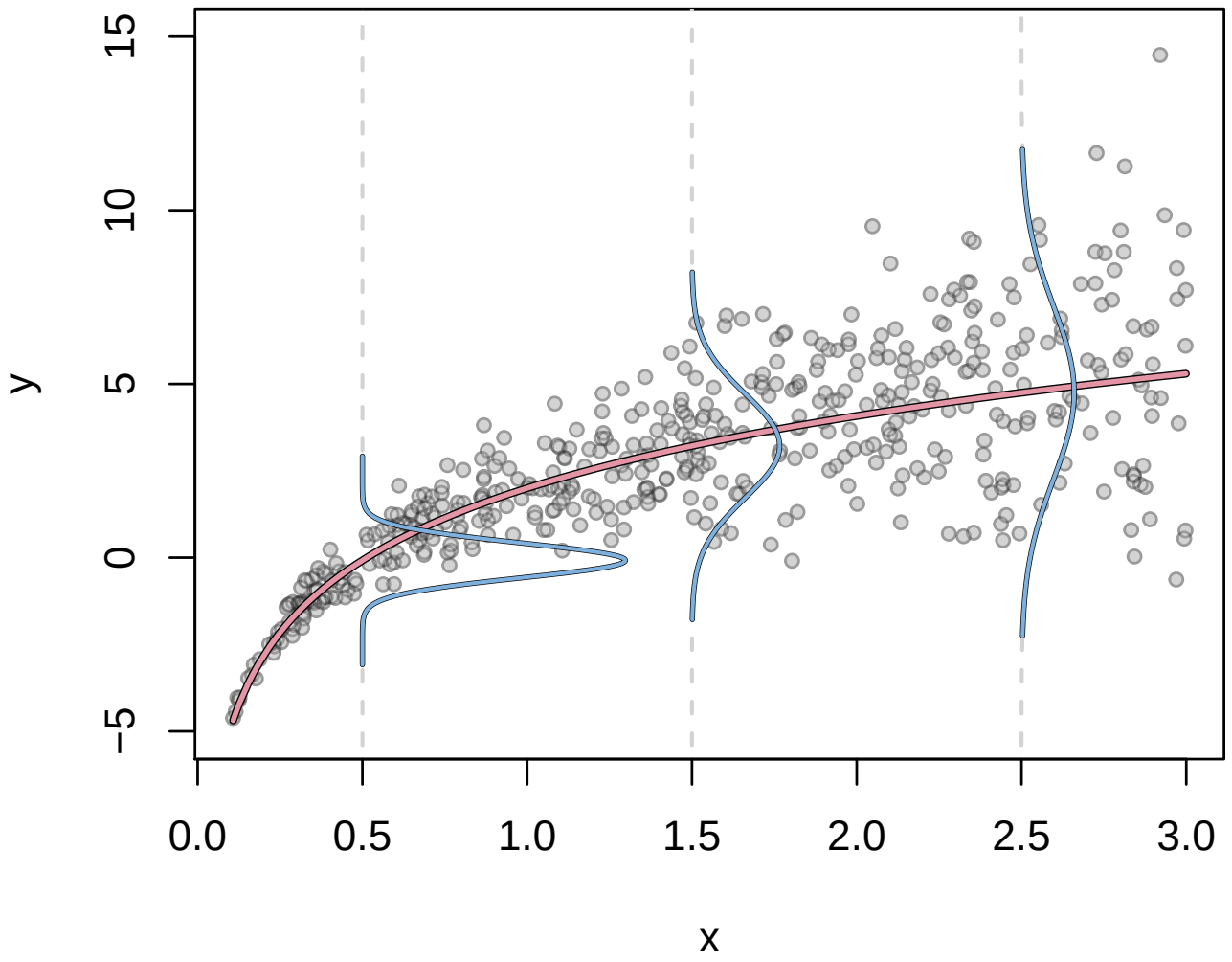
Generalized Additive Models for Location, Scale and Shape

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

- 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

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



First Idea for Bivariate Model

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

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

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