Prediction of the FIFA World Cup 2014 by a regularized random effects model

A. Groll and G. Schauberger

Mathematics Institute
Group of Financial Mathematics and Stochastics

Department of Statistics
Seminar for Applied Stochastic

Ludwig-Maximilians-Universität München

04/06/2014

Who will celebrate?

Who will cry?



Prediction of the FIFA World Cup 2014 based on the match outcomes of preceding world cups (2002-2010).

Simplest approach: (binary) Bradley-Terry-Luce model

$$P(Y_{ij} = 1) = \frac{\exp(\alpha_i - \alpha_j)}{1 + \exp(\alpha_i - \alpha_j)}$$

$\Rightarrow$ logistic regression model for the probability that team i beats team j
$\Rightarrow$ α_i represents the strength of team i

Possible extensions/modifications:

- ordered response (including draws)
- Poisson response (number of goals)
- incorporation of covariates

Modified BTL-model

Extensions:

- team-specific random effects $b_i \sim N(0, \sigma_b^2)$ instead of fixed effects α_i
- Poisson response: number of goals scored in single matches of a team
- incorporation of several covariates

Modified BTL-model

Extensions:

- team-specific random effects $b_i \sim N(0, \sigma_b^2)$ instead of fixed effects α_i
- Poisson response: number of goals scored in single matches of a team
- incorporation of several covariates

$\Rightarrow$ **Model equation**:

$$y_{ijk} | \mathbf{x}_{ik}, \mathbf{x}_{jk}, b_i, b_j \sim Pois(\lambda_{ijk})$$

$$log(\lambda_{ijk}) = \mathbf{x}_{ik}^T \beta_{team} + \mathbf{x}_{jk}^T \beta_{oppo} + b_i - b_j \,.$$

- team $i, i = 1, \ldots, n$, n number of teams;
- opponent $j, j = 1, \ldots, n$;
- tournament $k, k = 1, 2, 3$;
- $\mathbf{x}_{ik}, \mathbf{x}_{jk}$ covariates of teams i and j, varying over tournaments
- y_{ijk} number of goals scored by team i against team j in tournament k

Covariates

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

Covariates

- **Economic Factors:**
  GDP per capita, population
- **Sportive Factors:**
  Fairness, ODDSET odds, FIFA rank, confederation

Covariates

- **Economic Factors:**
  GDP per capita, population
- **Sportive Factors:**
  Fairness, ODDSET odds, FIFA rank, confederation
- **Home advantage:**
  host of the world cup, same continent as host

Covariates

- **Economic Factors:**
  GDP per capita, population
- **Sportive Factors:**
  Fairness, ODDSET odds, FIFA rank, confederation
- **Home advantage:**
  host of the world cup, same continent as host
- **Factors describing the team's structure**
  Maximum number of teammates, Second maximum number of teammates, average age, number of CL players, number of Europa League players, number of players abroad

Covariates

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

Covariates

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

All variables are incorporated of both the team whose goals are considered and its opponent!

Data: FIFA World Cups 2002 - 2010

Structure of the data set

goals	team	opponent	age	rank	odds	...	age.o	rank.o	odds.o	...
0	France	Senegal	28.30	1	0.15	...	24.30	42	0.01	...
1	Senegal	France	24.30	42	0.01	...	28.30	1	0.15	...
1	Uruguay	Denmark	25.30	24	0.01	...	27.40	20	0.01	...
2	Denmark	Uruguay	27.40	20	0.01	...	25.30	24	0.01	...
1	Denmark	Senegal	27.40	20	0.01	...	24.30	42	0.01	...
1	Senegal	Denmark	24.30	42	0.01	...	27.40	20	0.01	...
0	France	Uruguay	28.30	1	0.15	...	25.30	24	0.01	...
0	Uruguay	France	25.30	24	0.01	...	28.30	1	0.15	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Structure of the design matrix of the random effects:

France	Senegal	Uruguay	Denmark	Paraguay	...
1	-1	0	0	0	...
-1	1	0	0	0	...
0	0	1	-1	0	...
0	0	-1	1	0	...
0	-1	0	1	0	...
0	1	0	-1	0	...
1	0	-1	0	0	...
-1	0	1	0	0	...
⋮	⋮	⋮	⋮	⋮	⋱

Data: FIFA World Cups 2002 - 2010

Problem: there are two teams at the FIFA World Cup 2014 (Bosnia and Herzegovina, Colombia), which have not participated in any of the 2002 -2010 world cups.

Data: FIFA World Cups 2002 - 2010

Problem: there are two teams at the FIFA World Cup 2014 (Bosnia and Herzegovina, Colombia), which have not participated in any of the 2002 -2010 world cups.

⇒ to obtain nonetheless reasonable random effects estimates for both teams for the prediction of the FIFA World Cup 2014, we collect all teams that have only participated once in the tournaments between 2002 and 2014 in a group called **NEWCOMERS**.

This concerns the following 12 teams:

Angola, China, Czech Republic, Ireland, New Zealand, North Korea, Senegal, Slovakia, Togo, Trinidad & Tobago, Turkey, Ukraine

The **generalized linear mixed model** (GLMM) has the form

$$g(\mu_{it}) = \boldsymbol{x}_{it}^T \beta + \boldsymbol{z}_{it}^T \boldsymbol{b} = \eta_{it},$$

where g is a monotonic differentiable link function.

The **generalized linear mixed model** (GLMM) has the form

$$g(\mu_{it}) = \boldsymbol{x}_{it}^T \beta + \boldsymbol{z}_{it}^T \boldsymbol{b} = \eta_{it},$$

where g is a monotonic differentiable link function.

Notations:

- y_{it} observation t in cluster i, $i = 1, \dots, n$, $t = 1, \dots, T_i$.
- $\boldsymbol{x}_{it}$ covariate vector associated with fixed effects $\beta_0, \beta_1, \dots, \beta_p$.
- $\boldsymbol{z}_{it}^T$ covariate vector associated with random effects.
- $y_{i1}, \dots, y_{iT_i}$ are conditionally independent, with means $\mu_{it} = E(y_{it}|\boldsymbol{x}_{it}, \boldsymbol{z}_{it}, \mathbf{b})$ and variances $var(y_{it}|\mu_{it}) = \phi \upsilon(\mu_{it})$, $\upsilon(.)$ is a known variance function and ϕ is a scale parameter.
- $\mathbf{b}^T = (\mathbf{b}_1^T, \dots, \mathbf{b}_n^T)$ cluster-specific random effects of dimension s, $\boldsymbol{b}_i \sim N(\mathbf{0}, \boldsymbol{Q}(\boldsymbol{\varrho}))$ with covariance matrix $\boldsymbol{Q}(\boldsymbol{\varrho})$.

The conditional density of y_{it}, given the explanatory variables $\boldsymbol{x}_{it}$ and $\boldsymbol{z}_{it}$ and the random effects $\boldsymbol{b}$, is assumed to be of **exponential family type**

$$f(y_{it}|\boldsymbol{x}_{it}, \boldsymbol{z}_{it}, \boldsymbol{b}) = \exp\left\{\frac{(y_{it}^T\theta_{it} - \kappa(\theta_{it}))}{\phi} + c(y_{it}, \phi)\right\},$$

with natural parameter $\theta_{it} = \theta(\mu_{it})$, log normalization constant $c(.)$ and dispersion parameter ϕ.

The conditional density of y_{it}, given the explanatory variables $\boldsymbol{x}_{it}$ and $\boldsymbol{z}_{it}$ and the random effects $\boldsymbol{b}$, is assumed to be of **exponential family type**

$$f(y_{it}|\boldsymbol{x}_{it}, \boldsymbol{z}_{it}, \boldsymbol{b}) = \exp\left\{\frac{(y_{it}^T\theta_{it} - \kappa(\theta_{it}))}{\phi} + c(y_{it}, \phi)\right\},$$

with natural parameter $\theta_{it} = \theta(\mu_{it})$, log normalization constant $c(.)$ and dispersion parameter ϕ.

Define the parameter vector $\boldsymbol{\delta}^T = (\beta^T, \mathbf{b}^T)$.

The conditional density of y_{it}, given the explanatory variables $\boldsymbol{x}_{it}$ and $\boldsymbol{z}_{it}$ and the random effects $\boldsymbol{b}$, is assumed to be of **exponential family type**

$$f(y_{it}|\boldsymbol{x}_{it}, \boldsymbol{z}_{it}, \boldsymbol{b}) = \exp\left\{\frac{(y_{it}^T\theta_{it} - \kappa(\theta_{it}))}{\phi} + c(y_{it}, \phi)\right\},$$

with natural parameter $\theta_{it} = \theta(\mu_{it})$, log normalization constant $c(.)$ and dispersion parameter ϕ.

Define the parameter vector $\boldsymbol{\delta}^T = (\beta^T, \mathbf{b}^T)$.

The parameter vector of the covariance structure $\boldsymbol{\varrho}$ together with the dispersion parameter ϕ are collected in $\boldsymbol{\gamma}^T = (\phi, \boldsymbol{\varrho}^T)$.

Likelihood

Exponential Family

The conditional density of y_{it}, given the explanatory variables $\boldsymbol{x}_{it}$ and $\boldsymbol{z}_{it}$ and the random effects $\boldsymbol{b}$, is assumed to be of **exponential family type**

$$f(y_{it}|\boldsymbol{x}_{it}, \boldsymbol{z}_{it}, \boldsymbol{b}) = \exp\left\{\frac{(y_{it}^T\theta_{it} - \kappa(\theta_{it}))}{\phi} + c(y_{it}, \phi)\right\},$$

with natural parameter $\theta_{it} = \theta(\mu_{it})$, log normalization constant $c(.)$ and dispersion parameter ϕ.

Define the parameter vector $\boldsymbol{\delta}^T = (\beta^T, \mathbf{b}^T)$.

The parameter vector of the covariance structure $\boldsymbol{\varrho}$ together with the dispersion parameter ϕ are collected in $\boldsymbol{\gamma}^T = (\phi, \boldsymbol{\varrho}^T)$.

For known γ and given data $\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T$, estimation of $\boldsymbol{\delta}$ is based on maximizing the corresponding log-likelihood.

Log-Likelihood

We consider the following log-likelihood

$$l(\boldsymbol{\delta},\boldsymbol{\gamma}) = \log\left(\int f(\mathbf{y}|\boldsymbol{\delta},\boldsymbol{\gamma})p(\mathbf{b},\boldsymbol{\gamma})d\mathbf{b}\right),$$

where $p(\mathbf{b},\boldsymbol{\gamma})$ denotes the density of the random effects.

Log-Likelihood

We consider the following log-likelihood

$$l(\boldsymbol{\delta},\boldsymbol{\gamma}) = \log\left(\int f(\mathbf{y}|\boldsymbol{\delta},\boldsymbol{\gamma})p(\mathbf{b},\boldsymbol{\gamma})d\mathbf{b}\right),$$

where $p(\mathbf{b},\boldsymbol{\gamma})$ denotes the density of the random effects.

Laplace approximation (Breslow & Clayton, 1993) yields the log-likelihood

$$l^{\mathsf{app}}(\boldsymbol{\delta},\boldsymbol{\gamma}) = \sum_{i=1}^{n}\log f(\mathbf{y}_i|\boldsymbol{\delta},\boldsymbol{\gamma}) - \frac{1}{2}\mathbf{b}^T\mathbf{Q_b}(\varrho)^{-1}\mathbf{b}\,,$$

with $\mathbf{Q_b} = diag(\mathbf{Q},\ldots,\mathbf{Q})$.

We obtain estimates of the parameter vector $\boldsymbol{\delta}$ via Fisher scoring:

$$\hat{\boldsymbol{\delta}}^{(l)} = \hat{\boldsymbol{\delta}}^{(l-1)} + (\mathbf{F}^{\mathsf{app}}(\hat{\boldsymbol{\delta}}^{(l-1)}))^{-1}\mathbf{s}^{\mathsf{app}}(\hat{\boldsymbol{\delta}}^{(l-1)}).$$

The log-likelihood is expanded to include the penalty term $\lambda \sum_{i=1}^{p} |\beta_i|$:

$$l^{\text{pen}}(\boldsymbol{\delta}, \boldsymbol{\gamma}) = l^{\text{app}}(\boldsymbol{\delta}, \boldsymbol{\gamma}) - \lambda \sum_{i=1}^{p} |\beta_i|. \tag{1}$$

For given $\hat{\gamma}$ the optimization problem reduces to

$$\hat{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta}}{\text{argmax}}\ l^{\text{pen}}(\boldsymbol{\delta}, \hat{\gamma}) = \underset{\boldsymbol{\delta}}{\text{argmax}}\ \left[l^{\text{app}}(\boldsymbol{\delta}, \hat{\gamma}) - \lambda \sum_{i=1}^{p} |\beta_i| \right]. \tag{2}$$

We will use a full gradient algorithm based on the algorithm of Goeman (2010).

L_1-penalized GLMMs

The log-likelihood is expanded to include the penalty term $\lambda \sum_{i=1}^{p} |\beta_i|$:

$$l^{\text{pen}}(\boldsymbol{\delta}, \boldsymbol{\gamma}) = l^{\text{app}}(\boldsymbol{\delta}, \boldsymbol{\gamma}) - \lambda \sum_{i=1}^{p} |\beta_i|. \quad (1)$$

For given $\hat{\gamma}$ the optimization problem reduces to

$$\hat{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta}}{\operatorname{argmax}}\ l^{\text{pen}}(\boldsymbol{\delta}, \hat{\gamma}) = \underset{\boldsymbol{\delta}}{\operatorname{argmax}} \left[l^{\text{app}}(\boldsymbol{\delta}, \hat{\gamma}) - \lambda \sum_{i=1}^{p} |\beta_i| \right]. \quad (2)$$

We will use a full gradient algorithm based on the algorithm of Goeman (2010).

Incorporation of categorical predictors: $\sum_{g=1}^{G} \lambda_g ||\boldsymbol{\beta}_{\boldsymbol{I_g}}||_2$ with $\lambda_g = \lambda\sqrt{\text{df}_g}$

1. *Initialization*
   Compute starting values $\hat{\beta}^{(0)}, \hat{\mathbf{b}}^{(0)}, \hat{\gamma}^{(0)}$ and set $\hat{\boldsymbol{\eta}}^{(0)} = \mathbf{X}\hat{\beta}^{(0)} + \mathbf{Z}\hat{\mathbf{b}}^{(0)}$.

2. *Iteration*
   For $l = 1, 2, \ldots$ until convergence:

   a) *Calculation of the log-likelihood gradient for given* $\hat{\gamma}^{(l-1)}$

      With $\mathbf{s}(\boldsymbol{\delta}) = \partial l^{\mathbf{app}}(\boldsymbol{\delta})/\partial\boldsymbol{\delta}$ derive:

$$s^{\mathbf{pen}}_{\beta_{\mathbf{0}}}(\hat{\boldsymbol{\delta}}^{(l-1)}) = s_{\beta_{\mathbf{0}}}(\hat{\boldsymbol{\delta}}^{(l-1)}), \qquad s^{\mathbf{pen}}_{\mathbf{b}_i}(\hat{\boldsymbol{\delta}}^{(l-1)}) = s_{\mathbf{b}_i}(\hat{\boldsymbol{\delta}}^{(l-1)}).$$

1. *Initialization*
   Compute starting values $\hat{\beta}^{(0)}, \hat{\mathbf{b}}^{(0)}, \hat{\gamma}^{(0)}$ and set $\hat{\boldsymbol{\eta}}^{(0)} = \mathbf{X}\hat{\beta}^{(0)} + \mathbf{Z}\hat{\mathbf{b}}^{(0)}$.

2. *Iteration*
   For $l = 1, 2, \ldots$ until convergence:

   a) *Calculation of the log-likelihood gradient for given $\hat{\gamma}^{(l-1)}$*

   With $\mathbf{s}(\boldsymbol{\delta}) = \partial l^{\mathbf{app}}(\boldsymbol{\delta})/\partial\boldsymbol{\delta}$ derive:

$$s^{\mathbf{pen}}_{\beta_{\mathbf{0}}}(\hat{\boldsymbol{\delta}}^{(l-1)}) = s_{\beta_{\mathbf{0}}}(\hat{\boldsymbol{\delta}}^{(l-1)}), \qquad s^{\mathbf{pen}}_{\mathbf{b}_i}(\hat{\boldsymbol{\delta}}^{(l-1)}) = s_{\mathbf{b}_i}(\hat{\boldsymbol{\delta}}^{(l-1)}).$$

   For $\beta_i, i = 1, \ldots, p$ derive:

$$s^{\mathbf{pen}}_{\beta_i}(\hat{\boldsymbol{\delta}}^{(l-1)}) = \begin{cases} s_{\beta_i}(\hat{\boldsymbol{\delta}}^{(l-1)}) - \lambda \,\text{sign}\,(\hat{\beta}_i^{(l-1)}) & \text{if } \hat{\beta}_i^{(l-1)} \neq 0 \\ s_{\beta_i}(\hat{\boldsymbol{\delta}}^{(l-1)}) - \lambda \,\text{sign}\,(s_{\beta_i}(\hat{\boldsymbol{\delta}}^{(l-1)})) & \text{if } \hat{\beta}_i^{(l-1)} = 0, |s_{\beta_i}(\hat{\boldsymbol{\delta}}^{(l-1)})| > \lambda \\ 0 & \text{otherwise.} \end{cases}$$

b) *Calculation of the directional second derivative*

The directional second derivative is given for every $\boldsymbol{\delta}$ and every direction vector $\mathbf{v} \in \mathbb{R}^{p+1+ns}$ by

$$l''_{\text{pen}}(\boldsymbol{\delta}; \mathbf{v}) = \mathbf{v}^T \mathbf{F}^{\text{pen}}(\boldsymbol{\delta})\mathbf{v},$$

b) *Calculation of the directional second derivative*
The directional second derivative is given for every $\boldsymbol{\delta}$ and every direction vector $\mathbf{v} \in \mathbb{R}^{p+1+ns}$ by

$$l''_{\mathbf{pen}}(\boldsymbol{\delta}; \mathbf{v}) = \mathbf{v}^T \mathbf{F}^{\mathbf{pen}}(\boldsymbol{\delta})\mathbf{v},$$

with $\mathbf{F}^{\mathbf{pen}}(\boldsymbol{\delta}) = \mathbf{A}^T\mathbf{W}(\boldsymbol{\delta})\mathbf{A} + \mathbf{K}$ and $\mathbf{A} := [\mathbf{X}, \mathbf{Z}]$, $\mathbf{W}(\boldsymbol{\delta}) = \mathbf{D}(\boldsymbol{\delta})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta})\mathbf{D}(\boldsymbol{\delta})^T$, $\mathbf{K} = diag(0, \ldots, 0, \mathbf{Q}^{-1}, \ldots, \mathbf{Q}^{-1})$, $\mathbf{D}(\boldsymbol{\delta}) = \partial h(\boldsymbol{\eta})/\partial \boldsymbol{\eta}$ and $\boldsymbol{\Sigma}(\boldsymbol{\delta}) = cov(\mathbf{y}|\boldsymbol{\delta})$.

b) *Calculation of the directional second derivative*
The directional second derivative is given for every $\boldsymbol{\delta}$ and every direction vector $\mathbf{v} \in \mathbb{R}^{p+1+ns}$ by

$$l''_{\text{pen}}(\boldsymbol{\delta}; \mathbf{v}) = \mathbf{v}^T \mathbf{F}^{\text{pen}}(\boldsymbol{\delta})\mathbf{v},$$

with $\mathbf{F}^{\text{pen}}(\boldsymbol{\delta}) = \mathbf{A}^T \mathbf{W}(\boldsymbol{\delta})\mathbf{A} + \mathbf{K}$ and $\mathbf{A} := [\mathbf{X}, \mathbf{Z}]$, $\mathbf{W}(\boldsymbol{\delta}) = \mathbf{D}(\boldsymbol{\delta})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta})\mathbf{D}(\boldsymbol{\delta})^T$, $\mathbf{K} = diag(0, \ldots, 0, \mathbf{Q}^{-1}, \ldots, \mathbf{Q}^{-1})$, $\mathbf{D}(\boldsymbol{\delta}) = \partial h(\boldsymbol{\eta})/\partial \boldsymbol{\eta}$ and $\boldsymbol{\Sigma}(\boldsymbol{\delta}) = cov(\mathbf{y}|\boldsymbol{\delta})$.

c) *Optimum of directional 2. order Taylor approximation*
Approximate l_{pen} locally from $\boldsymbol{\delta}$ in direction of the gradient:

$$l_{\text{pen}}(\boldsymbol{\delta} + t s^{\text{pen}}(\boldsymbol{\delta})) \approx l_{\text{pen}}(\boldsymbol{\delta}) + t l'_{\text{pen}}(\boldsymbol{\delta}, s^{\text{pen}}(\boldsymbol{\delta})) + \frac{1}{2} t^2 l''_{\text{pen}}(\boldsymbol{\delta}, s^{\text{pen}}(\boldsymbol{\delta}))$$

b) *Calculation of the directional second derivative*
The directional second derivative is given for every $\boldsymbol{\delta}$ and every direction vector $\mathbf{v} \in \mathbb{R}^{p+1+ns}$ by

$$l''_{\mathbf{pen}}(\boldsymbol{\delta}; \mathbf{v}) = \mathbf{v}^T \mathbf{F}^{\mathbf{pen}}(\boldsymbol{\delta})\mathbf{v},$$

with $\mathbf{F}^{\mathbf{pen}}(\boldsymbol{\delta}) = \mathbf{A}^T\mathbf{W}(\boldsymbol{\delta})\mathbf{A} + \mathbf{K}$ and $\mathbf{A} := [\mathbf{X}, \mathbf{Z}]$, $\mathbf{W}(\boldsymbol{\delta}) = \mathbf{D}(\boldsymbol{\delta})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta})\mathbf{D}(\boldsymbol{\delta})^T$, $\mathbf{K} = diag(0, \ldots, 0, \mathbf{Q}^{-1}, \ldots, \mathbf{Q}^{-1})$, $\mathbf{D}(\boldsymbol{\delta}) = \partial h(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$ and $\boldsymbol{\Sigma}(\boldsymbol{\delta}) = cov(\mathbf{y}|\boldsymbol{\delta})$.

c) *Optimum of directional 2. order Taylor approximation*
Approximate $l_{\mathbf{pen}}$ locally from $\boldsymbol{\delta}$ in direction of the gradient:

$$l_{\mathbf{pen}}(\boldsymbol{\delta} + t s^{\mathbf{pen}}(\boldsymbol{\delta})) \approx l_{\mathbf{pen}}(\boldsymbol{\delta}) + t l'_{\mathbf{pen}}(\boldsymbol{\delta}, s^{\mathbf{pen}}(\boldsymbol{\delta})) + \frac{1}{2} t^2 l''_{\mathbf{pen}}(\boldsymbol{\delta}, s^{\mathbf{pen}}(\boldsymbol{\delta}))$$

$$t_{\mathbf{opt}}^{(l-1)} = \frac{||\mathbf{s}^{\mathbf{pen}}(\hat{\boldsymbol{\delta}}^{(l-1)})||_2}{l''_{\mathbf{pen}}(\hat{\boldsymbol{\delta}}^{(l-1)}, \mathbf{s}^{\mathbf{pen}}(\hat{\boldsymbol{\delta}}^{(l-1)}))},$$

$||\cdot||_2$ denoting the L_2-norm.

$$t_{\textbf{edge}}^{(l-1)} = \min_i \left\{ -\frac{\hat{\delta}_i^{(l-1)}}{s_i^{\textbf{pen}}(\hat{\boldsymbol{\delta}}^{(l-1)})} : \text{sign}(\hat{\delta}_i^{(l-1)}) = -\text{sign}[s_i^{\textbf{pen}}(\hat{\boldsymbol{\delta}}^{(l-1)})] \neq 0 \right\}$$

$$t_{\textbf{edge}}^{(l-1)} = \min_i \left\{ -\frac{\hat{\delta}_i^{(l-1)}}{s_i^{\textbf{pen}}(\hat{\boldsymbol{\delta}}^{(l-1)})} : \text{sign}(\hat{\delta}_i^{(l-1)}) = -\text{sign}[s_i^{\textbf{pen}}(\hat{\boldsymbol{\delta}}^{(l-1)})] \neq 0 \right\}$$

d) *Update*

$$\hat{\boldsymbol{\delta}}^{(l)} = \begin{cases} \hat{\boldsymbol{\delta}}^{(l-1)} + t_{\textbf{edge}}^{(l-1)} \mathbf{s}^{\textbf{pen}}(\hat{\boldsymbol{\delta}}^{(l-1)}) & \text{if } t_{\textbf{opt}}^{(l-1)} \geq t_{\textbf{edge}}^{(l-1)} \\ \hat{\boldsymbol{\delta}}^{(l-1)} + t_{\textbf{opt}}^{(l-1)} \mathbf{s}^{\textbf{pen}}(\hat{\boldsymbol{\delta}}^{(l-1)}) & \text{otherwise.} \end{cases}$$

e) *Computation of variance-covariance components*

- Approximate **EM**-type estimate:

$$\hat{\mathbf{Q}}^{(l)} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{V}}_{ii}^{(l)} + \hat{\mathbf{b}}_i^{(l)}(\hat{\mathbf{b}}_i^{(l)})^T), \qquad \text{with} \qquad \hat{\mathbf{V}}_{ii}^{(l)} \in (\mathbf{F}^{\mathbf{pen}}(\hat{\boldsymbol{\delta}}^{(l)}))^{-1}.$$

e) *Computation of variance-covariance components*

- Approximate **EM**-type estimate:

$$\hat{\mathbf{Q}}^{(l)} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{V}}_{ii}^{(l)} + \hat{\mathbf{b}}_i^{(l)}(\hat{\mathbf{b}}_i^{(l)})^T), \qquad \text{with} \qquad \hat{\mathbf{V}}_{ii}^{(l)} \in (\mathbf{F}^{\mathbf{pen}}(\hat{\boldsymbol{\delta}}^{(l)}))^{-1}.$$

- Approximate **REML**-type estimate: maximize

$$\begin{aligned} l(\mathbf{Q_b}) \quad = \quad & -\frac{1}{2}\log(|\mathbf{V}(\hat{\boldsymbol{\delta}}^{(l)})|) - \frac{1}{2}\log(|\mathbf{X}^T\mathbf{V}^{-1}(\hat{\boldsymbol{\delta}}^{(l)})\mathbf{X}|) \\ & -\frac{1}{2}(\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\delta}}^{(l)}) - \mathbf{X}\hat{\beta}^{(l)})^T\mathbf{V}^{-1}(\hat{\boldsymbol{\delta}}^{(l)})(\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\delta}}^{(l)}) - \mathbf{X}\hat{\beta}^{(l)}) \end{aligned}$$

  with respect to $\mathbf{Q_b}$.

e) *Computation of variance-covariance components*

- Approximate **EM**-type estimate:

$$\hat{\mathbf{Q}}^{(l)} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{V}}_{ii}^{(l)} + \hat{\mathbf{b}}_i^{(l)}(\hat{\mathbf{b}}_i^{(l)})^T), \qquad \text{with} \qquad \hat{\mathbf{V}}_{ii}^{(l)} \in (\mathbf{F}^{\mathbf{pen}}(\hat{\boldsymbol{\delta}}^{(l)}))^{-1}.$$

- Approximate **REML**-type estimate: maximize

$$\begin{aligned} l(\mathbf{Q_b}) &= -\frac{1}{2}\log(|\mathbf{V}(\hat{\boldsymbol{\delta}}^{(l)})|) - \frac{1}{2}\log(|\mathbf{X}^T\mathbf{V}^{-1}(\hat{\boldsymbol{\delta}}^{(l)})\mathbf{X}|) \\ & -\frac{1}{2}(\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\delta}}^{(l)}) - \mathbf{X}\hat{\beta}^{(l)})^T\mathbf{V}^{-1}(\hat{\boldsymbol{\delta}}^{(l)})(\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\delta}}^{(l)}) - \mathbf{X}\hat{\beta}^{(l)}) \end{aligned}$$

with respect to $\mathbf{Q_b}$.

$$\begin{aligned} &\mathbf{V}(\delta) = \mathbf{W}^{-1}(\delta) + \mathbf{Z}\mathbf{Q_b}\mathbf{Z}^T, && \mathbf{Q_b} = diag(\mathbf{Q}, \ldots, \mathbf{Q}), \\ &\tilde{\boldsymbol{\eta}}(\delta) = [\mathbf{X}, \mathbf{Z}]\,\delta + \mathbf{D}^{-1}(\delta)(\mathbf{y} - \boldsymbol{\mu}(\delta)), && \mathbf{D}(\delta) = \frac{\partial h(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}, \\ &\mathbf{W}(\delta) = \mathbf{D}(\delta)\boldsymbol{\Sigma}^{-1}(\delta)\mathbf{D}(\delta)^T, && \boldsymbol{\Sigma}(\delta) = \text{cov}(\mathbf{y}|\delta). \end{aligned}$$

The overdispersion parameter Φ is estimated by use of Pearson residuals $\hat{r}_{it} = (y_{it} - \hat{\mu}_{it})/(v(\hat{\mu}_{it}))^{\frac{1}{2}}$ as

$$\hat{\Phi} = \frac{1}{N - \mathrm{df}} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \hat{r}_{it}^2, \qquad N = \sum_{i=1}^{n} T_i,$$

where the degrees of freedom (df) correspond to the trace of the hat-matrix.

The overdispersion parameter Φ is estimated by use of Pearson residuals $\hat{r}_{it} = (y_{it} - \hat{\mu}_{it})/(v(\hat{\mu}_{it}))^{\frac{1}{2}}$ as

$$\hat{\Phi} = \frac{1}{N - \text{df}} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \hat{r}_{it}^2, \qquad N = \sum_{i=1}^{n} T_i,$$

where the degrees of freedom (df) correspond to the trace of the hat-matrix.

Call in R using the package glmmLasso:

```
>lasso.obj <- glmmLasso(goals ~ rank + rank.o + continent + continent.o ...,
                        rnd = W, data = soccer, lambda = 100, family = poisson(link = log),
                        control = list(...))
```

The overdispersion parameter Φ is estimated by use of Pearson residuals $\hat{r}_{it} = (y_{it} - \hat{\mu}_{it})/(v(\hat{\mu}_{it}))^{\frac{1}{2}}$ as

$$\hat{\Phi} = \frac{1}{N - \text{df}} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \hat{r}_{it}^2, \qquad N = \sum_{i=1}^{n} T_i,$$

where the degrees of freedom (df) correspond to the trace of the hat-matrix.

Call in R using the package glmmLasso:

```
>lasso.obj <- glmmLasso(goals ~ rank + rank.o + continent + continent.o ...,
                        rnd = W, data = soccer, lambda = 100, family = poisson(link = log),
                        control = list(...))
```

The overdispersion parameter Φ is estimated by use of Pearson residuals $\hat{r}_{it} = (y_{it} - \hat{\mu}_{it})/(v(\hat{\mu}_{it}))^{\frac{1}{2}}$ as

$$\hat{\Phi} = \frac{1}{N - \mathrm{df}} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \hat{r}_{it}^2, \qquad N = \sum_{i=1}^{n} T_i,$$

where the degrees of freedom (df) correspond to the trace of the hat-matrix.

Call in R using the package glmmLasso:

```
>lasso.obj <- glmmLasso(goals ~ rank + rank.o + continent + continent.o ...,
                        rnd = W, data = soccer, lambda = 100, family = poisson(link = log),
                        control = list(...))
```

The overdispersion parameter Φ is estimated by use of Pearson residuals $\hat{r}_{it} = (y_{it} - \hat{\mu}_{it})/(v(\hat{\mu}_{it}))^{\frac{1}{2}}$ as

$$\hat{\Phi} = \frac{1}{N - \text{df}} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \hat{r}_{it}^2, \qquad N = \sum_{i=1}^{n} T_i,$$

where the degrees of freedom (df) correspond to the trace of the hat-matrix.

Call in R using the package glmmLasso:

```
>lasso.obj <- glmmLasso(goals ~ rank + rank.o + continent + continent.o ...,
                        rnd = W, data = soccer, lambda = 100, family = poisson(link = log),
                        control = list(...))
```

The overdispersion parameter Φ is estimated by use of Pearson residuals $\hat{r}_{it} = (y_{it} - \hat{\mu}_{it})/(v(\hat{\mu}_{it}))^{\frac{1}{2}}$ as

$$\hat{\Phi} = \frac{1}{N - \mathrm{df}} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \hat{r}_{it}^2, \qquad N = \sum_{i=1}^{n} T_i,$$

where the degrees of freedom (df) correspond to the trace of the hat-matrix.

Call in R using the package glmmLasso:

```
>lasso.obj <- glmmLasso(goals ~ rank + rank.o + continent + continent.o ...,
                        rnd = W, data = soccer, lambda = 100, family = poisson(link = log),
                        control = list(...))
```

The overdispersion parameter Φ is estimated by use of Pearson residuals $\hat{r}_{it} = (y_{it} - \hat{\mu}_{it})/(v(\hat{\mu}_{it}))^{\frac{1}{2}}$ as

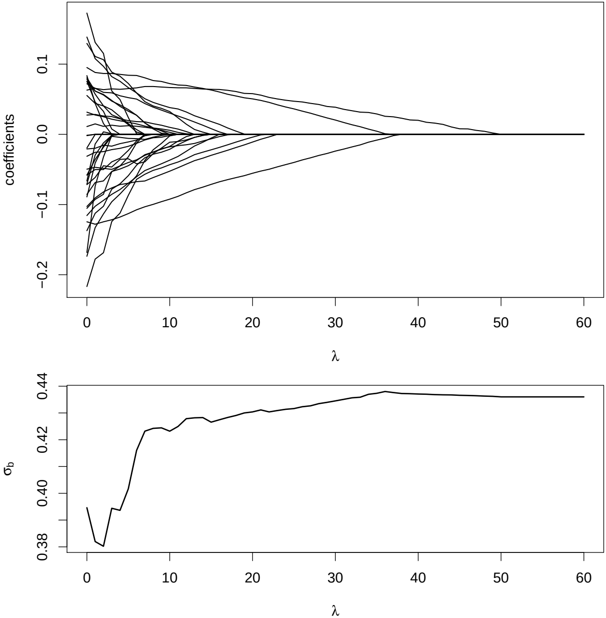
$$\hat{\Phi} = \frac{1}{N - \text{df}} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \hat{r}_{it}^2, \qquad N = \sum_{i=1}^{n} T_i,$$

where the degrees of freedom (df) correspond to the trace of the hat-matrix.

Call in R using the package glmmLasso:

```
>lasso.obj <- glmmLasso(goals ~ rank + rank.o + continent + continent.o ...,
                        rnd = W, data = soccer, lambda = 100, family = poisson(link = log),
                        control = list(...))
```

Coefficient paths for the FIFA World Cup data



Selection of tuning parameter

Idea: **leave-one-out cross validation**, especially designed for soccer data

- for the prediction of single matches we are mainly interested in the general match outcome (win, draw, defeat)
- single matches are treated as observations
- for the omitted observation/match the probabilities for win, draw and defeat are estimated

Selection of tuning parameter

Idea: **leave-one-out cross validation**, especially designed for soccer data

- for the prediction of single matches we are mainly interested in the general match outcome (win, draw, defeat)
- single matches are treated as observations
- for the omitted observation/match the probabilities for win, draw and defeat are estimated

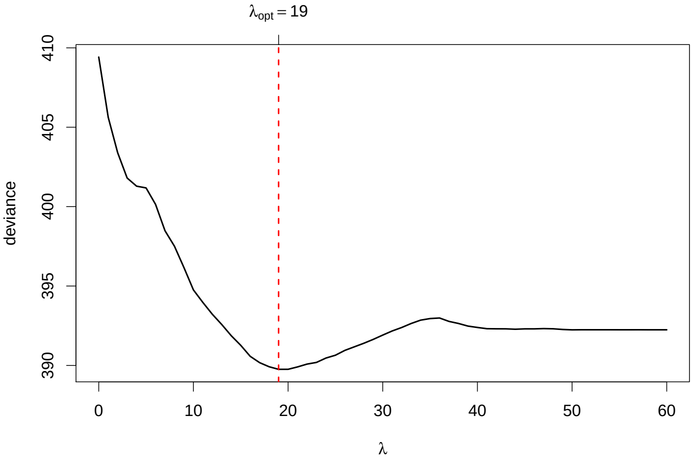
Let the omitted observation/match be a match of team k against team l:

- predictions $\hat{\mu}_k$ and $\hat{\mu}_l$ for a given model
  $\Rightarrow \hat{p}_1 = P(Y_k > Y_l), \hat{p}_2 = P(Y_k = Y_l)$ and $\hat{p}_3 = P(Y_k < Y_l)$
- for a given match outcome $\omega \in \{1, 2, 3\}$, the quantity

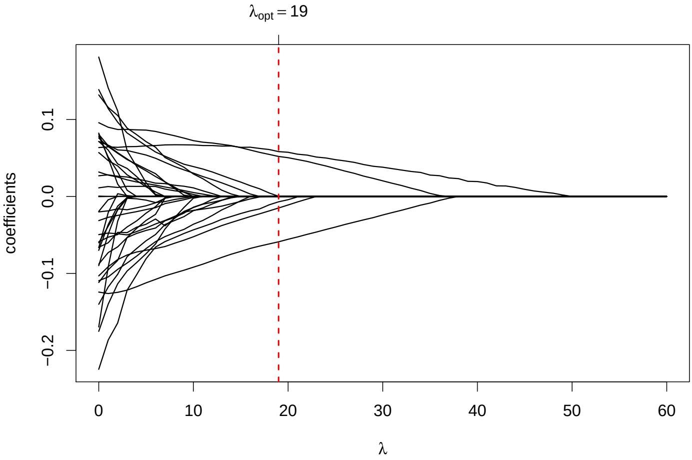
$$\hat{p}_1^{\delta_{1\omega}} \hat{p}_2^{\delta_{2\omega}} \hat{p}_3^{\delta_{3\omega}}$$

  represents the likelihood score of the corresponding match
- $\Rightarrow$ deviance for multinomial logit model

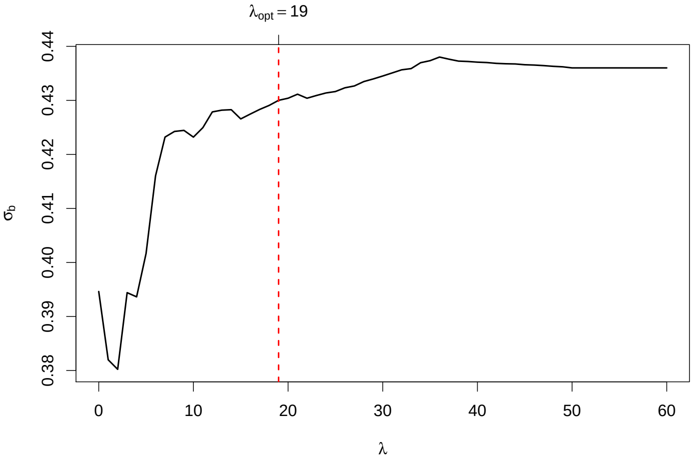
Leave-one-out cross validation



Coefficient paths



Standard deviation of the random effects



Final model based on LOOCV

Team covariates

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

Final model based on LOOCV

Opponent covariates

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

Estimation results

(Intercept)	Nation.Coach	Population	Rank.oppo
0.0169	0.0004	-0.0151	0.0521

GDP.oppo	continent.oppo	Population.oppo	σ_b
-0.0074	-0.0589	0.0583	**0.430**

Estimation results

(Intercept)	Nation.Coach	Population	Rank.oppo
0.0169	0.0004	-0.0151	0.0521

GDP.oppo	continent.oppo	Population.oppo	σ_b
-0.0074	-0.0589	0.0583	**0.430**

1. BRA 0.74	11. URU 0.24	21. RSA 0.00	31. POL -0.28
2. GER 0.72	12. SWE 0.15	22. CRO 0.00	32. SVN -0.31
3. ARG 0.56	13. MEX 0.15	23. AUS -0.01	33. SRB -0.33
4. NED 0.44	14. FRA 0.14	24. JPN -0.05	34. RUS -0.35
5. POR 0.44	15. PAR 0.12	25. NGA -0.06	35. CRC -0.35
6. ESP 0.37	16. ECU 0.10	26. DEN -0.12	36. CMR -0.36
7. ITA 0.30	17. KOR 0.10	27. NEW -0.13	37. ALG -0.37
8. ENG 0.29	18. GHA 0.09	28. HON -0.19	38. IRN -0.41
9. SUI 0.27	19. CHI 0.06	29. BEL -0.20	39. TUN -0.73
10. CIV 0.25	20. USA 0.05	30. GRE -0.27	40. KSA -1.09

Goodness-of-fit (in sample)

- a possible goodness-of-fit criterion is based on the "three-way" ODDSET odds for all 192 games of the FIFA World Cups 2002 - 2010

Goodness-of-fit (in sample)

- a possible goodness-of-fit criterion is based on the "three-way" ODDSET odds for all 192 games of the FIFA World Cups 2002 - 2010
- compute the three quantities $\tilde{p}_i = 1/\text{odds}_i, i \in I := \{1, 2, 3\}$, normalizing with $c := \sum_{i \in I} \tilde{p}_i$ (adjust for the bookmakers' margins)
  $\Longrightarrow$ estimated probabilities $\hat{p}_i = \tilde{p}_i / c$

Goodness-of-fit (in sample)

- a possible goodness-of-fit criterion is based on the "three-way" ODDSET odds for all 192 games of the FIFA World Cups 2002 - 2010
- compute the three quantities $\tilde{p}_i = 1/\text{odds}_i, i \in I := \{1, 2, 3\}$, normalizing with $c := \sum_{i \in I} \tilde{p}_i$ (adjust for the bookmakers' margins)
  $\Longrightarrow$ estimated probabilities $\hat{p}_i = \tilde{p}_i / c$
- Use the corresponding (estimated) Poisson distributions $Y_k \sim Poisson(\hat{\mu}_k)$, $Y_l \sim Poisson(\hat{\mu}_l)$ to estimate the same probabilities by approximating $\hat{p}_1 = P(Y_k > Y_l), \hat{p}_2 = P(Y_k = Y_l)$ and $\hat{p}_3 = P(Y_k < Y_l)$ for all 192 games

Goodness-of-fit (in sample)

- a possible goodness-of-fit criterion is based on the "three-way" ODDSET odds for all 192 games of the FIFA World Cups 2002 - 2010
- compute the three quantities $\tilde{p}_i = 1/\text{odds}_i, i \in I := \{1, 2, 3\}$, normalizing with $c := \sum_{i \in I} \tilde{p}_i$ (adjust for the bookmakers' margins)
  $\Longrightarrow$ estimated probabilities $\hat{p}_i = \tilde{p}_i / c$
- Use the corresponding (estimated) Poisson distributions $Y_k \sim Poisson(\hat{\mu}_k)$, $Y_l \sim Poisson(\hat{\mu}_l)$ to estimate the same probabilities by approximating $\hat{p}_1 = P(Y_k > Y_l), \hat{p}_2 = P(Y_k = Y_l)$ and $\hat{p}_3 = P(Y_k < Y_l)$ for all 192 games
- For $\omega_j \in I, j = 1, \ldots, 192$, the likelihood is given by the product

$$L(\omega_1, \ldots, \omega_{192}) = \prod_{j=1}^{192} \hat{p}_{1j}^{\delta_{1\omega_j}} \hat{p}_{2j}^{\delta_{2\omega_j}} \hat{p}_{3j}^{\delta_{3\omega_j}}$$

Goodness-of-fit (in sample)

- a possible goodness-of-fit criterion is based on the "three-way" ODDSET odds for all 192 games of the FIFA World Cups 2002 - 2010
- compute the three quantities $\tilde{p}_i = 1/\text{odds}_i, i \in I := \{1,2,3\}$, normalizing with $c := \sum_{i \in I} \tilde{p}_i$ (adjust for the bookmakers' margins)
  $\Longrightarrow$ estimated probabilities $\hat{p}_i = \tilde{p}_i / c$
- Use the corresponding (estimated) Poisson distributions $Y_k \sim Poisson(\hat{\mu}_k)$, $Y_l \sim Poisson(\hat{\mu}_l)$ to estimate the same probabilities by approximating $\hat{p}_1 = P(Y_k > Y_l), \hat{p}_2 = P(Y_k = Y_l)$ and $\hat{p}_3 = P(Y_k < Y_l)$ for all 192 games
- For $\omega_j \in I, j = 1, \ldots, 192$, the likelihood is given by the product

$$L(\omega_1, \ldots, \omega_{192}) = \prod_{j=1}^{192} \hat{p}_{1j}^{\delta_{1\omega_j}} \hat{p}_{2j}^{\delta_{2\omega_j}} \hat{p}_{3j}^{\delta_{3\omega_j}}$$

Corresponding **deviance scores** for `glmmLasso`, `grplasso` and ODDSET odds:

`glmmLasso`		`grplasso`	ODDSET odds
AIC	LOOCV		
334.3549	333.1851	336.7894	365.4724

Simulation of the tournament progress

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

⇒ decision on qualification for round of 16 according to FIFA rules

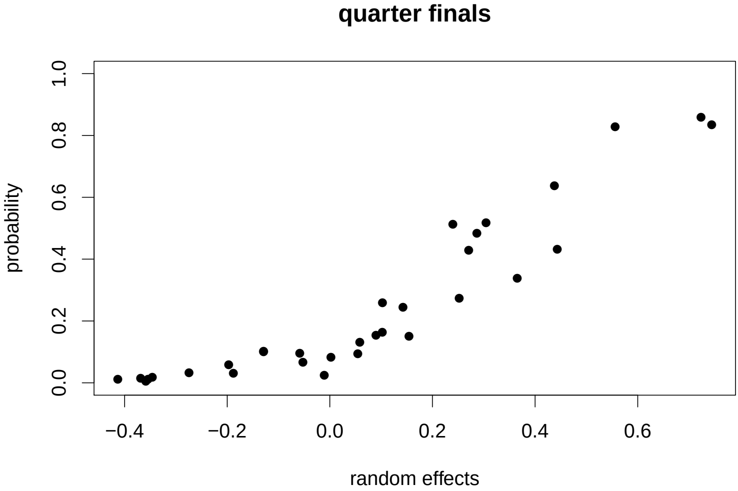
- draws in kockout stage → team with higher probability wins

⇒ 10000 simulation runs for FIFA World Cup 2014

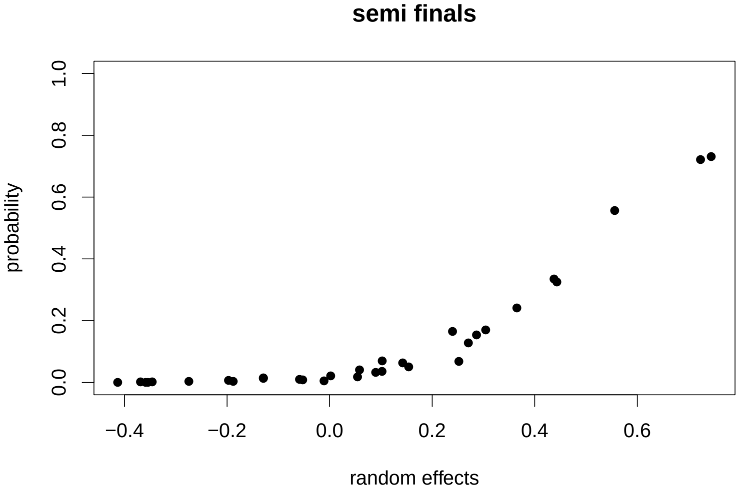
Probabilities for FIFA World Cup Winner 2014

team		$\hat{p}_{\text{glmmLasso}}$	$\hat{p}_{\text{Oddset}}$	team		$\hat{p}_{\text{glmmLasso}}$	$\hat{p}_{\text{Oddset}}$
1.	BRA	0.4325	0.2028	17.	GHA	0.0007	0.0071
2.	GER	0.2640	0.1420	18.	USA	0.0003	0.0071
3.	ARG	0.1472	0.1420	19.	BIH	0.0002	0.0047
4.	POR	0.0518	0.0237	20.	AUS	0.0002	0.0024
5.	NED	0.0311	0.0355	21.	BEL	0.0001	0.0592
6.	ESP	0.0214	0.1092	22.	CRO	0.0001	0.0071
7.	URU	0.0129	0.0284	23.	COL	0	0.0394
8.	ITA	0.0104	0.0355	24.	NGA	0	0.0035
9.	ENG	0.0101	0.0355	25.	JPN	0	0.0047
10.	SUI	0.0069	0.0071	26.	HON	0	0.0005
11.	CIV	0.0028	0.0071	27.	GRE	0	0.0071
12.	ECU	0.0024	0.0071	28.	ALG	0	0.0071
13.	FRA	0.0017	0.0355	29.	RUS	0	0.0118
14.	MEX	0.0014	0.0071	30.	CRC	0	0.0071
15.	CHI	0.0010	0.0203	31.	IRN	0	0.0005
16.	KOR	0.0008	0.0024	32.	CMR	0	0.0024

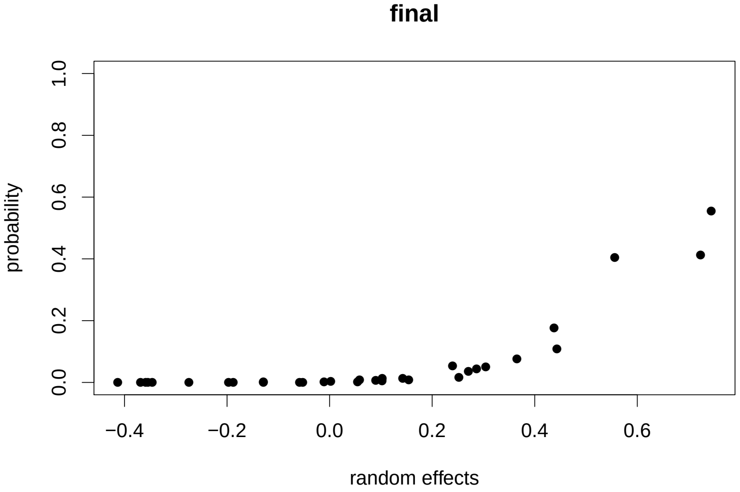
Influence of random effects on tournament success



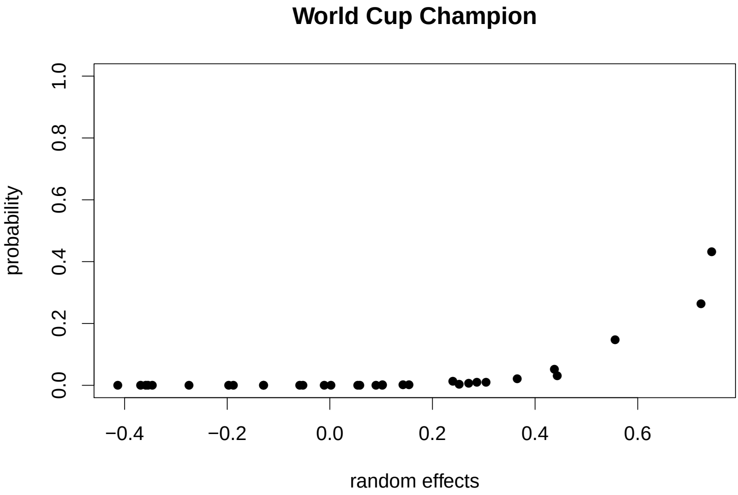
Influence of random effects on tournament success



Influence of random effects on tournament success



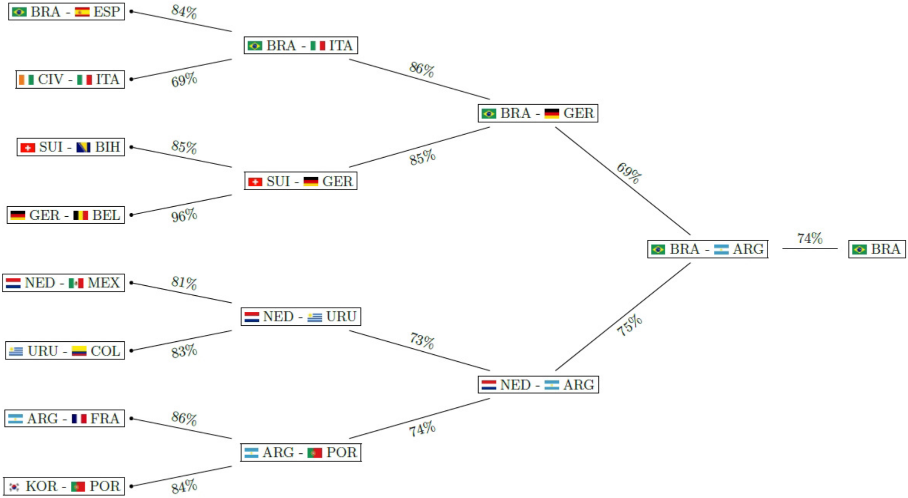
Influence of random effects on tournament success



Most probable tournament outcome: group standings

Group A 46%	Group B 25%	Group C 24%	Group D 17%	Group E 19%	Group F 38%	Group G 42%	Group H 27%
1. BRA	1. NED	1. CIV	1. KOR	1. SUI	1. ARG	1. GER	1. URU
2. MEX	2. ESP	2. COL	2. BEL	2. FRA	2. BIH	2. POR	2. ITA
CRO	CHI	JPN	RUS	ECU	NGA	GHA	ENG
CMR	AUS	GRE	ALG	HON	IRN	USA	CRC

Most probable tournament outcome: knockout stage



Summary

Model:

- the glmmLasso algorithm allows to include complex random effects structures, e.g. for paired comparison data;
- simulations and applications show that the proposed procedure provides **convincing estimators**.

Summary

Model:

- the `glmmLasso` algorithm allows to include complex random effects structures, e.g. for paired comparison data;
- simulations and applications show that the proposed procedure provides **convincing estimators**.

FIFA World Cup application:

- huge impact of team-specific random effects $\Rightarrow$ most variables concerning teams strength are not selected;
- unexpected set of selected variables;
- three top favorites (not including Spain!).

References & Software

Goeman, J. J. (2012). *penalized.* R package version 0.9-42.

Goeman, J. J. (2010). L_1 Penalized Estimation in the Cox Proportional Hazards Model. *Biometrical Journal 52*, 70–84.

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

Groll, A. & G. Tutz (2014): Variable Selection for Generalized Linear Mixed Models by L_1-Penalized Estimation. *Statistics and Computing* 24(2): 137-154

Groll, A. (2014). *glmmLasso: Variable selection for generalized linear mixed models by L1-penalized estimation.* R package version 1.3.0.

Die Hoffnung stirbt zuletzt!