Beating the bookie

A look at statistical models for prediction of football matches

Helge Langseth

Norwegian University of Science and Technology

SCAI 2013

NTNU

Building a model for match outcomes

- Suppose we want to build a model to **predict the outcomes** of the English Premier League:
  - **20 teams**, all play each other twice during a season.
  - Each team plays 38 matches, **380 games** per season in total.
- The model should predict the outcome of Team i meeting Team j, based on all games played previously in the season.
- The quality is meassured by the systems ability to **win bets**.
  - A bet (e.g., "*Liverpool to win*") is offered with **odds** ω.
  - The model generates the corresponding **probability** p.
  - A bet is **only rational** whenever the **expected gain is positive**, i.e., $p \cdot \omega \geq 1$.
  - Accurate predictions imply a usefull betting agent, thus our goal is to generate good **probability estimates** for upcoming games based on the history of the **season so far**.

Maher (1982)

An early attempt at building a statistical model:

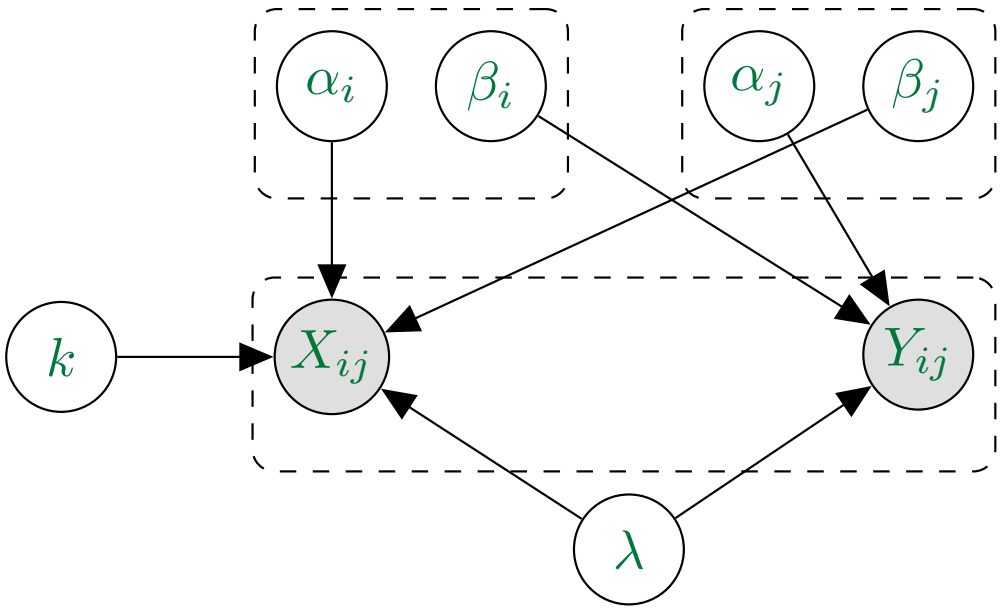
- $X_{ij} \sim \text{Poisson}(k \cdot \lambda \cdot \alpha_i \beta_j)$, where:
  - X_{ij} is no. goals scored by Team i vs. Team j playing at home.
  - k captures the home-team advantage.
  - λ is a normalization constant.
  - α_i is the **attacking** strength of Team i.
  - β_j is the **defending** strength of Team j.
- $Y_{ij} \sim \text{Poisson}(\alpha_j \beta_i)$; Y_{ij} is no. goals scored by Team j.
- Crucially — and surprisingly — he assumes $X_{ij} \perp\!\!\!\perp Y_{ij} | \lambda$.
- The model is under-specified, so he requires

$$\text{avg}_\ell \left(\alpha_\ell \right) = \text{avg}_\ell \left(\beta_\ell \right) = 1.$$

Maher (1982)

An early attempt at building a statistical model:

- $X_{ij} \sim \text{Poisson}(k \cdot \lambda \cdot \alpha_i \beta_j)$, where:
  - X_{ij} is no. goals scored by Team i vs. Team j playing at home.
  - k captures the home-team advantage.
  - λ is a normalization constant.
  - α_i is the **attacking** strength of Team i.
  - β_j is the **defending** strength of Team j.
- $Y_{ij} \sim \text{Poisson}(\alpha_j \beta_i)$; Y_{ij} is no. goals scored by Team j.



Predictions from the model

- We predict the result of the game between Team k and Team ℓ by looking at the probability distributions for $X_{k\ell}$ and $Y_{k\ell}$.
- The **estimated abilities** of the two best teams in the Premier league after **11 rounds** are:

	After 11 games	
	Attack	Defence
Arsenal	1.42	0.87
Liverpool	1.38	0.83

- Using these parameters we can look at the joint distribution

$$P\left(X_{\text{Liv,Ars}}, Y_{\text{Liv,Ars}}\right).$$

Predictions from the model

- We predict the result of the game between Team k and Team ℓ by looking at the probability distributions for $X_{k\ell}$ and $Y_{k\ell}$.
- The **estimated abilities** of the two best teams in the Premier league after **5 and 11 rounds** are:

	After 5 games		**After 11 games**	
	Attack	Defence	Attack	Defence
Arsenal	1.46	1.23	1.42	0.87
Liverpool	1.02	0.69	1.38	0.83

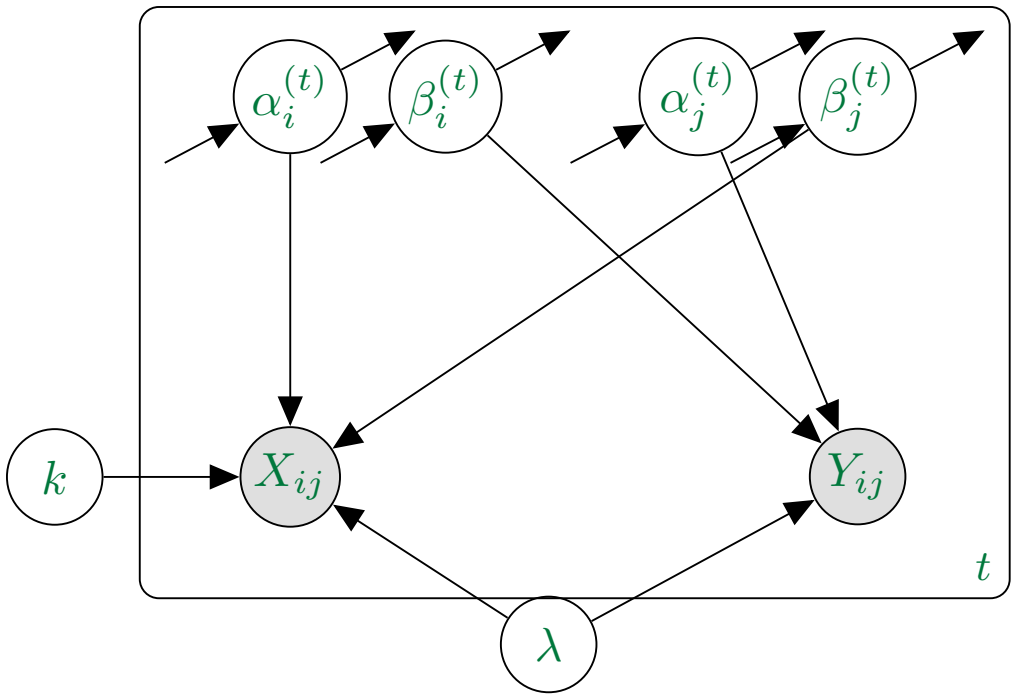
Abilities change over time, so we need a dynamic model!!

Adding dynamics

- We follow, e.g., Rue & Salvesen (2000) and introduce **dynamics** at the **"strength-level"**:
  - Let $\alpha_i^{(t)}$ be the attack-strength for Team i at time t.
  - Then, $\alpha_i^{(t_1)} \left| \left\{ \alpha_i^{(t_2)} \right\} \sim N\left(\alpha_i^{(t_2)}, \frac{|t_1 - t_2|}{\tau} \sigma^2 \right) \right.$
  - Similarly for the defence-strength, $\beta_i^{(t)}$.
- **Hidden Markov** - type model: Unobserved strengths varying over time; partially disclosed through goal-model.
- Assume we observe the result when Team i and Team j :
  - The chains of these teams **get correlated**.
  - Similarly, the strengths of **all teams** Team i and Team j have played previously **get correlated, too!**
- We use **Markov Chain Monte Carlo** to find estimators for the model parameters, and sample results for unseen matches.

Adding dynamics

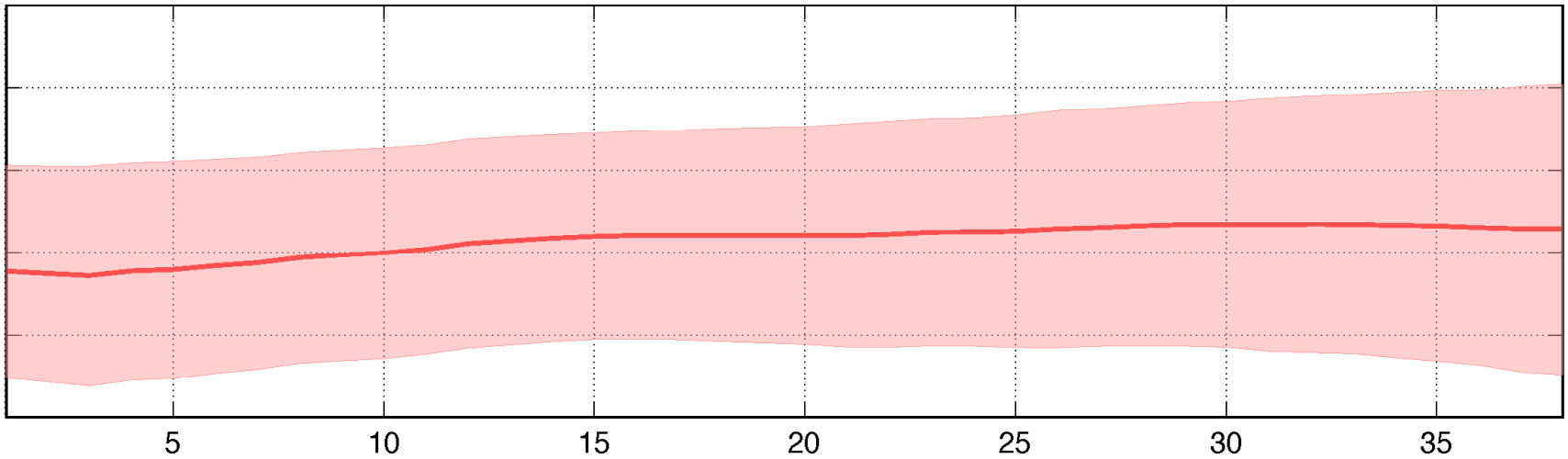
- We follow, e.g., Rue & Salvesen (2000) and introduce **dynamics** at the **"strength-level"**:
  - Let $\alpha_i^{(t)}$ be the attack-strength for Team i at time t.
  - Then, $\alpha_i^{(t_1)} \left| \left\{ \alpha_i^{(t_2)} \right\} \sim N\left(\alpha_i^{(t_2)}, \frac{|t_1 - t_2|}{\tau} \sigma^2 \right) \right.$
  - Similarly for the defence-strength, $\beta_i^{(t)}$.



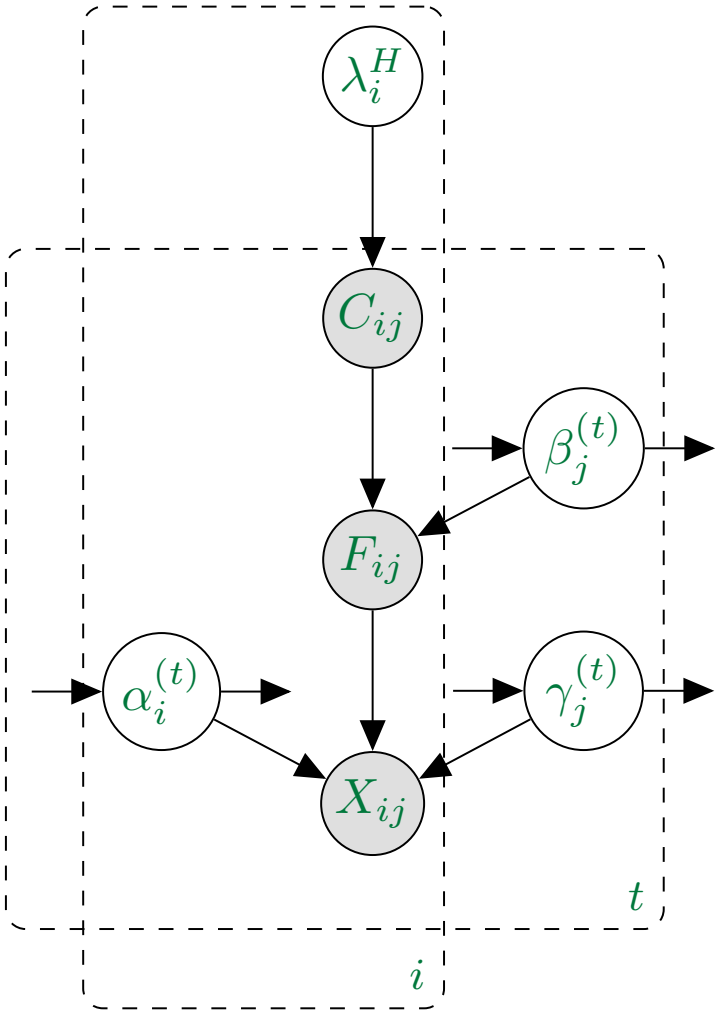
Looking behind the results

- **Small margins** can influence the result of a football match significantly.
- This inherit randomness makes the estimation of $\alpha_i^{(t)}$ and $\beta_i^{(t)}$ difficult, and the **"signal-to-noise-ratio"** is typically small.
- More data, that **"look behind the result"**, e.g.,
  - Shots on goals
  - Possession statistics
  - Passing accuracy

  can be useful to **uncover the teams' underlying abilities**.



Estimated defensive strength for Arsenal over the 2010-11 season.



Here we use:

- λ_i^H: Chance creation rate; home
- C_{ij}: No. *chances*.
- F_{ij}: No. *shots*.
- X_{ij}: No. *goals*.
- $\alpha_\ell^{(t)}$: The *attacking* strength.
- $\beta_\ell^{(t)}$: The *defensive* strength.
- $\gamma_\ell^{(t)}$: The *goalkeeper* strength.

Money management

- Consider a bet with offered **odds** ω and estimated winning **probability** p.
- We require teh **expected gain to be non-negative**, i.e., $p \cdot \omega \geq 1$.
- Consider the two bet-options

  Bet A: $\omega_A = 11.0$, $p_A = 0.1$.
  Bet B: $\omega_B = 1.10$, $p_B = 1.0$.

  Both bets have the same expected return of 1.1 unit per unit staked, but obviously **Bet B** is preferrable.
- It is important to **consider money management** carefully!
- Many strategies exist, we have considered, e.g., `Fixed Bet`, `Fixed Return`, `Kelly's Rule` and `Rue's Variance Adjustment`.

Premier League 2011-2012				
Model	Fixed Bet	Fixed Return	Kelly	Var. Adjust
STATIC	17.4%	17.4%	**23.2%**	**15.6%**
DYNAMIC	**22.7%**	14.3%	21.3%	12.0%
DATAINTENSIVE	20.3%	**24.2%**	23.0%	14.3%

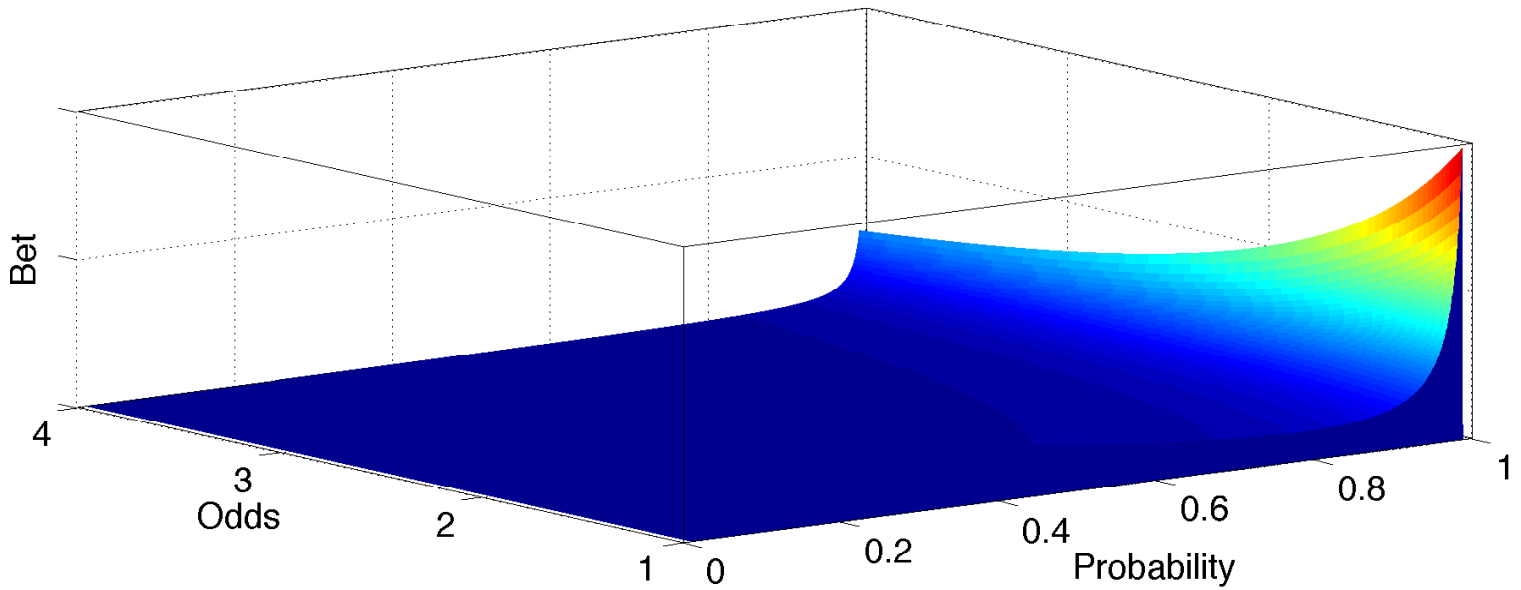
Premier League 2012-2013				
Model	Fixed Bet	Fixed Return	Kelly	Var. Adjust
STATIC	-23.7%	-24.9%	-27.8%	-21.2%
DYNAMIC	-17.1%	-20.0%	-22.9 %	-15.9%
DATAINTENSIVE	**−6.3%**	**−0.7%**	**−3.4%**	**0.4%**

2010-2011: Results similar, but DATAINTENSIVE combined with `FixedReturn` gives best result.

2011-2012: Only DATAINTENSIVE combined with `Var. Adjust` beats the bookie.

The Variance Adjusted strategy

- Rue's Variance Adjustment has been the most robust money management strategy.
- The goal is to minimize the expected profit minus the variance of the profit, leading to bets defined by $C \propto \frac{1}{\omega(1-p)}$.
- In contrast to most other money management strategies, the amount wagered therefore is decreasing in ω.



Future work

- Although we are looking at **betting agents**, and not simple **classifiers**, improving prediction is beneficial:
  - Build models that **incorporate more game-information**; data can be harvested, e.g., from `http://www.whoscored.com`.
  - Combine the **ensemble** of different candidate models into one prediction-engine.
  - Utilize pre-game information about **line-ups** to enhance the predictions.
- Simulate results for **more leagues** – with the aim of understanding why some leagues are easier to generate profits from than others.
- Replace MCMC simulations with fast approximate Bayesian inference based on **variational approximations**.