Beating the bookie
A look at statistical models for prediction of football matches

Helge Langseth
Norwegian University of Science and Technology

SCAI 2013

NTNU
Suppose we want to build a model to predict the outcomes of games from 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 quality is measured by the systems ability to win bets.

- A bet (e.g., “Liverpool to win”) is offered with odds $\omega$.
- 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 useful 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.
  - \( \lambda \) is a normalization constant.
  - \( \alpha_i \) is the **attacking** strength of Team \( i \).
  - \( \beta_j \) is the **defending** strength of Team \( j \).

- \( Y_{ij} \sim \text{Poisson}(\lambda \cdot \alpha_j \beta_i) \); \( Y_{ij} \) is no. goals scored by Team \( j \).

- Crucially — and surprisingly — he assumes \( X_{ij} \perp \perp Y_{ij} | \text{Model} \).

- The model is under-specified, so he requires
  \[
  \text{avg}_\ell (\alpha_\ell) = \text{avg}_\ell (\beta_\ell) = 1.
  \]
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.
  - $\lambda$ is a normalization constant.
  - $\alpha_i$ is the attacking strength of Team $i$.
  - $\beta_j$ is the defending strength of Team $j$.

- $Y_{ij} \sim \text{Poisson}(\lambda \cdot \alpha_j \beta_i)$; $Y_{ij}$ is no. goals scored by Team $j$. 
We predict the result of the game between Team $k$ and Team $\ell$ by looking at the probability distributions for $X_{k\ell}$ and $Y_{k\ell}$.

The maximum likelihood parameters for the abilities of the two best teams in the Premier League’s after 11 rounds, are:

<table>
<thead>
<tr>
<th></th>
<th>Attack</th>
<th>Defence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arsenal</td>
<td>1.4</td>
<td>0.9</td>
</tr>
<tr>
<td>Liverpool</td>
<td>1.4</td>
<td>0.8</td>
</tr>
</tbody>
</table>

We can use these parameters (plus $\hat{\kappa}$ and $\hat{\lambda}$) to find, e.g.,

$$P\left(X_{Liv,Ars} > Y_{Liv,Ars}\right).$$
Predictions from the model

- We predict the result of the game between Team $k$ and Team $\ell$ by looking at the probability distributions for $X_{k\ell}$ and $Y_{k\ell}$.
- The **maximum likelihood** parameters for the abilities of the two best teams in the Premier League’s after 11 rounds, are:

<table>
<thead>
<tr>
<th></th>
<th>Attack</th>
<th>Defence</th>
<th>Attack</th>
<th>Defence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arsenal</td>
<td>1.5</td>
<td>1.2</td>
<td>1.4</td>
<td>0.9</td>
</tr>
<tr>
<td>Liverpool</td>
<td>1.0</td>
<td>0.7</td>
<td>1.4</td>
<td>0.8</td>
</tr>
</tbody>
</table>

**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, \( \alpha_i(t) \sim \alpha_i(t+\Delta t) \) is a random walk with st.dev. \( \tau \cdot \Delta t \).
  - Similarly for the defence-strength, \( \beta_i(t) \).

- One HMM/KF-structured model per ability: Latent and time-varying strengths; 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, $\alpha_i^{(t)} \sim \alpha_i^{(t+\Delta t)}$ is a random walk with st.dev. $\tau \cdot \Delta t$.
  - Similarly for the defence-strength, $\beta_i^{(t)}$. 

![Diagram showing the dynamics of attack and defence strengths](attachment:diagram.png)
Small margins can significantly influence the result of a game. This inherent randomness makes the estimation of $\alpha_i^{(t)}$ and $\beta_i^{(t)}$ difficult, as the “signal-to-noise-ratio” is typically small. More data, that “look behind the result”, e.g.,
- No. chances created
- Shot statistics: On target, off target, hitting wood-work
- Passing accuracy
- …

can be useful to uncover the teams’ underlying abilities.
Here we use:

- $\lambda_i^H$: Chance creation rate; home
- $C_{ij}$: Number of chances.
- $F_{ij}$: Number of shots.
- $X_{ij}$: Number of goals.
- $\alpha_i^{(t)}$: The attacking strength.
- $\beta_j^{(t)}$: The defensive strength.
- $\gamma_j^{(t)}$: The goalkeeper strength.
- $\tau$: The scaler in the step-size of the random walk for the abilities.
Money management

- Consider a bet with offered odds $\omega$ and estimated winning probability $p$.
- We require the 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 preferable.

- 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.
## Results

### Premier League 2011-2012

<table>
<thead>
<tr>
<th>Model</th>
<th>Fixed Bet</th>
<th>Fixed Return</th>
<th>Kelly</th>
<th>Var. Adjust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>17.4%</td>
<td>17.4%</td>
<td>23.2%</td>
<td>15.6%</td>
</tr>
<tr>
<td>Dynamic</td>
<td><strong>22.7%</strong></td>
<td>14.3%</td>
<td>21.3%</td>
<td>12.0%</td>
</tr>
<tr>
<td>DataIntensive</td>
<td>20.3%</td>
<td><strong>24.2%</strong></td>
<td>23.0%</td>
<td>14.3%</td>
</tr>
</tbody>
</table>

### Premier League 2012-2013

<table>
<thead>
<tr>
<th>Model</th>
<th>Fixed Bet</th>
<th>Fixed Return</th>
<th>Kelly</th>
<th>Var. Adjust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>-23.7%</td>
<td>-24.9%</td>
<td>-27.8%</td>
<td>-21.2%</td>
</tr>
<tr>
<td>Dynamic</td>
<td>-17.1%</td>
<td>-20.0%</td>
<td>-22.9%</td>
<td>-15.9%</td>
</tr>
<tr>
<td>DataIntensive</td>
<td><strong>-6.3%</strong></td>
<td><strong>-0.7%</strong></td>
<td><strong>-3.4%</strong></td>
<td><strong>0.4%</strong></td>
</tr>
</tbody>
</table>

**2011-2012:** Results are non-conclusive, but **DataIntensive** combined with **FixedReturn** gives the best result.

**2012-2013:** Only **DataIntensive** combined with **Variance Adjustment** beats the bookie.
Although we are looking at betting agents, and not simple classifiers, improving prediction quality 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.

Generate results from more leagues – aiming to understand why some leagues are easier to generate profits from than others.

Replace MCMC simulations with fast approximate Bayesian inference based on variational approximations.