Skill Importance in Women's Soccer

Matt Heiner
Gilbert W. Fellingham
Camille Thomas

Department of Statistics
Brigham Young University

March 4, 2013

Introduction

According to FIFA, there are 240 million active soccer players worldwide, 20 million of which are women. [1]

Most soccer analytics come from private enterprise and amateur enthusiasts, although there is some work in academia.

Many analyses focus on one of two things:

1. Predicting match outcomes
2. Modeling play sequences spatially or temporally

[1] FIFA (2006)

Introduction

Information useful to coaches:

- Which skills are most important
- Team strengths, weaknesses
- Effective strategies

How coaches can use the information:

- Focus practice efforts
- Know the “percentage plays” during games
- More informed recruiting

Introduction

Our goal is to identify which skills lead to scoring opportunities. To do so we:

1. Model play sequences as absorbing Markov chains with skills as states
2. Estimate transition probabilities in the chain
3. Use the model to estimate the probability of each skill leading to a shot

Data

Play-by-play sequences were recorded for 22 NCAA Division I Women's Soccer matches using a notation system developed by Thomas (2006). The data appear as:

	Skill	Rating	Team	Notes1	Notes2	Notes3	Game
1	PASS	6	HOME	Kickoff			22
2	1-T PASS	0	HOME				22
3	TRAP	5	AWAY				22
4	DRIB	5	AWAY				22
5	DRIB	0	AWAY				22

Data

There are five distinct skills with ratings depending on the outcome. Higher ratings are better outcomes (e.g. 0 is turnover).

Skill	Possible Ratings	Description
SHOT	0-3	
PASS	0-8	
1-T PASS	0-8	First-touch pass
DRIB	0-5	Dribbling touch
TRAP	0-5	Controlling touch

Model

We model the data as a Markov chain with states defined by skill-rating combinations.

Some pass skills are identified as corner kicks, free kicks, goal kicks, kick-offs, and throw-ins. We define these as separate states. We also introduce turnover, bad turnover, fouled, and deflected OB as states.

Markov Chain Review

A Markov chain is a *stochastic process*, or set of random variables indexed by time. Let $X_t \in \{s_1, s_2, \ldots, s_m\}$ for $t = 1, 2, \ldots$ where $\{s_1, s_2, \ldots, s_m\}$ are m states. This process is a *Markov chain* if it satisfies

$$P(X_{t+1} = x | X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) = P(X_{t+1} = x | X_t = x_t).$$

This is called a *transition probability*. We will assume transition probabilities between all the states remain constant. We denote the transition probability from state s_j to state s_k with $p_{j,k}$.

See Lawler (1995)

Markov Chain Review

We can arrange all of our transition probabilities into a matrix

$$\mathbf{P} = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,m} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m,1} & \cdots & \cdots & p_{m,m} \end{pmatrix}$$

where each row consists of the transition probabilities *from* a state and sums to one.

See Lawler (1995)

Markov Chain Review

This notation is especially useful if we are interested in the probability of transitioning from state j to state k in two steps. Conveniently, this is the (j, k) entry of $\mathbf{P}^2$.

Similarly, the probability of transitioning from state j to state k in n steps is the (j, k) entry of $\mathbf{P}^n$.

See Lawler (1995)

Markov Chain Example

Consider this simple Markov chain. The first two states are called *absorbing states* because once the cain enters one of those states, it cannot leave. This is an *absorbing Markov chain* and the non-absorbing states are called *transient states*.

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ .25 & .25 & .15 & .35 \\ 0 & .1 & .7 & .2 \end{pmatrix}$$

Markov Chain Example

$$\mathbf{P}^2 \approx \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ .29 & .32 & .27 & .12 \\ .18 & .30 & .25 & .29 \end{pmatrix} \quad \mathbf{P}^{12} \approx \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ .457 & .536 & .004 & .003 \\ .398 & .592 & .006 & .004 \end{pmatrix}$$

$$\mathbf{P}^5 \approx \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ .41 & .48 & .06 & .05 \\ .33 & .50 & .10 & .07 \end{pmatrix} \quad \mathbf{P}^{\infty} \approx \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ .46 & .54 & 0 & 0 \\ .40 & .60 & 0 & 0 \end{pmatrix}$$

Markov Chain Review

In general, if we have an absorbing Markov chain

$$\mathbf{P} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{B} & \mathbf{Q} \end{pmatrix}$$

where **B** and **Q** contain transition probabilities for transient states, we can solve

$$lim_{n \to \infty} \mathbf{P}^n = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{A} & \mathbf{0} \end{pmatrix}$$

where **A** contains the eventual absorption probabilities and is given by

$$\mathbf{A} = (\mathbf{I} - \mathbf{Q})^{-1}\mathbf{B}. \tag{1}$$

See Lawler (1995)

Model

In our soccer analysis, we model shots and turnovers as absorbing states.

We wish to estimate the eventual absorption probability into state `SHOT 3` (Goal) OR `SHOT 2` OR `SHOT 1` from each skill-rating combination, denoted p_α.

Model

First, we estimate the transition probabilities for our 53 transient states by transforming the data into a transition count matrix.

Transition probabilities for impossible transitions (e.g. `PASS 7` to `TURNOVER`) are fixed at zero and are not estimated.

The likelihood for each row of the transition count matrix is multinomial with pdf

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{n}) = \frac{N!}{n_1! n_2! \cdots n_K!} \theta_1^{n_1} \theta_2^{n_2} \cdots \theta_K^{n_K}$$

where $n_1, n_2, \ldots, n_k$ are observed transition counts with $\sum n_i = N$ and $\theta_1, \theta_2, \ldots, \theta_k$ are the transition probabilities we wish to estimate.

See Casella and Berger (2002)

Model

We specify conjugate Dirichlet priors on the transition probabilities $\boldsymbol{\theta}$ for each row

$$\pi(\boldsymbol{\theta};\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\prod_{i=1}^{K}\theta_i^{\alpha_i-1}$$

with $\alpha_i > 0$ and $\sum_{i=1}^{K}\theta_i = 1$.
The posterior distribution on $\boldsymbol{\theta}$ is then

$$\begin{aligned} f(\boldsymbol{\theta}|\mathbf{n}) &\propto \mathcal{L}(\boldsymbol{\theta};\mathbf{n}) \times \pi(\boldsymbol{\theta}) \\ &\propto \theta_1^{n_1+\alpha_1-1}\theta_2^{n_2+\alpha_2-1}\cdots\theta_K^{n_K+\alpha_K-1} \end{aligned} \quad (2)$$

which is distributed Dirichlet with updated $\boldsymbol{\alpha}^* = \boldsymbol{\alpha} + \mathbf{n}$.

See Gelman et al. (2004)

Model

For row j of the transition matrix $\mathbf{P}$, we use prior $\boldsymbol{\alpha}_j = \alpha_j \mathbf{1}$ with α_j values ranging from 0.001 to 0.5 depending on the number of possible transitions from that row. [8]

The posterior distribution on the transition probabilities from row j is then

$$\sim \text{Dir}(\boldsymbol{\alpha}_j + \mathbf{n}_j)$$

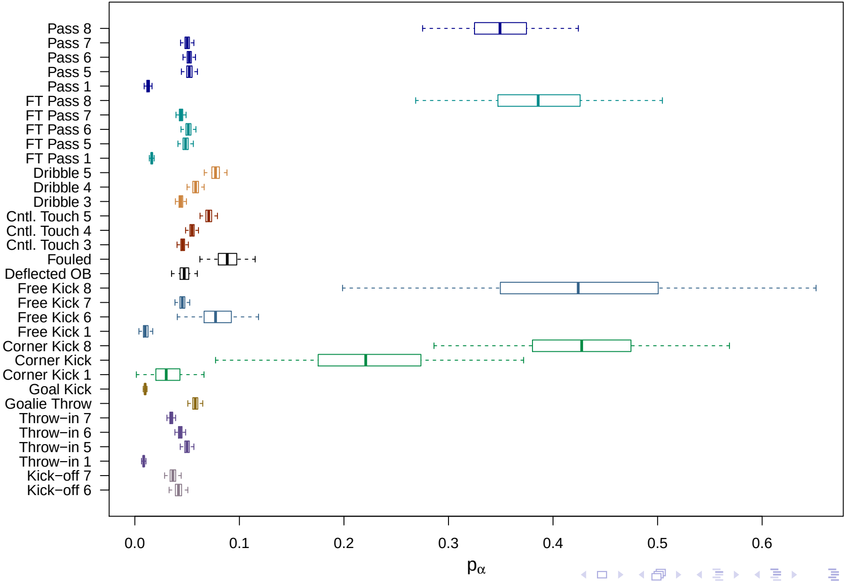
where $\mathbf{n}_j$ is the vector of observed transition counts from skill-rating j.

[8] de Campos and Benavoli (2011)

Model

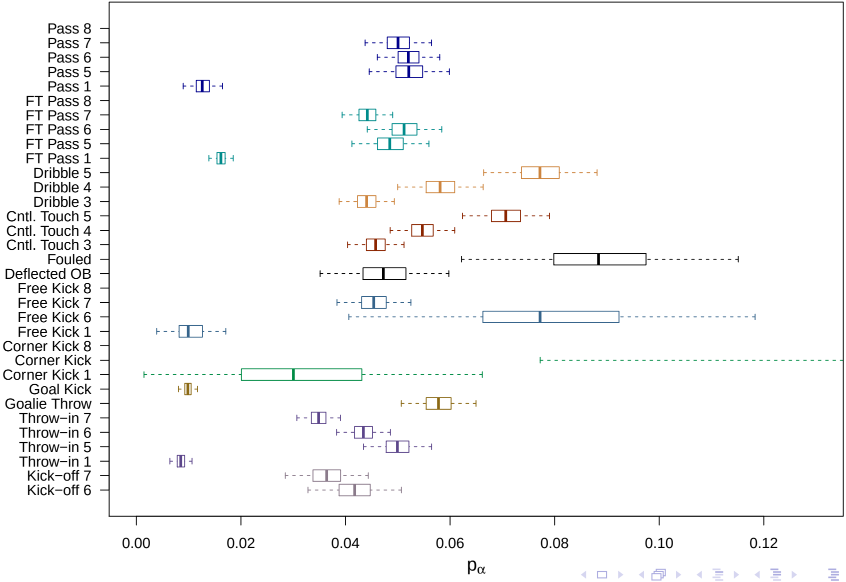
With a posterior distribution on the entire transition probability matrix **P**, we simulate 100,000 posterior transition matrices and for each draw perform the transformation in (1) to get posterior distributions on the eventual absorption probabilities p_α.

Results



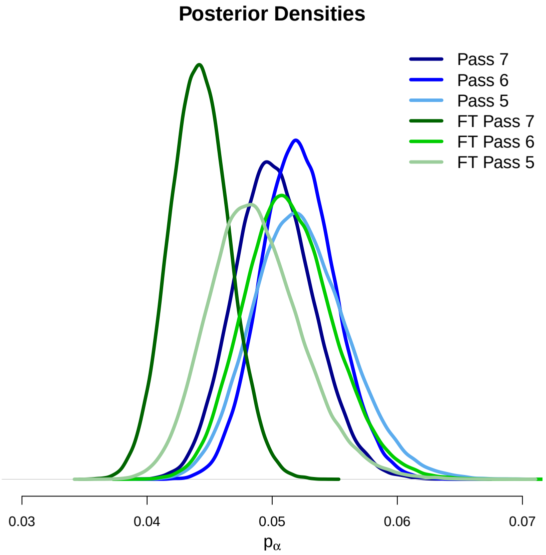
Pass 8
Pass 7
Pass 6
Pass 5
Pass 1
FT Pass 8
FT Pass 7
FT Pass 6
FT Pass 5
FT Pass 1
Dribble 5
Dribble 4
Dribble 3
Cntl. Touch 5
Cntl. Touch 4
Cntl. Touch 3
Fouled
Deflected OB
Free Kick 8
Free Kick 7
Free Kick 6
Free Kick 1
Corner Kick 8
Corner Kick
Corner Kick 1
Goal Kick
Goalie Throw
Throw−in 7
Throw−in 6
Throw−in 5
Throw−in 1
Kick−off 7
Kick−off 6
0.0
0.1
0.2
0.3
0.4
0.5
0.6
p_α

Results



Pass 8
Pass 7
Pass 6
Pass 5
Pass 1
FT Pass 8
FT Pass 7
FT Pass 6
FT Pass 5
FT Pass 1
Dribble 5
Dribble 4
Dribble 3
Cntl. Touch 5
Cntl. Touch 4
Cntl. Touch 3
Fouled
Deflected OB
Free Kick 8
Free Kick 7
Free Kick 6
Free Kick 1
Corner Kick 8
Corner Kick
Corner Kick 1
Goal Kick
Goalie Throw
Throw−in 7
Throw−in 6
Throw−in 5
Throw−in 1
Kick−off 7
Kick−off 6
0.00
0.02
0.04
0.06
0.08
0.10
0.12
p_α

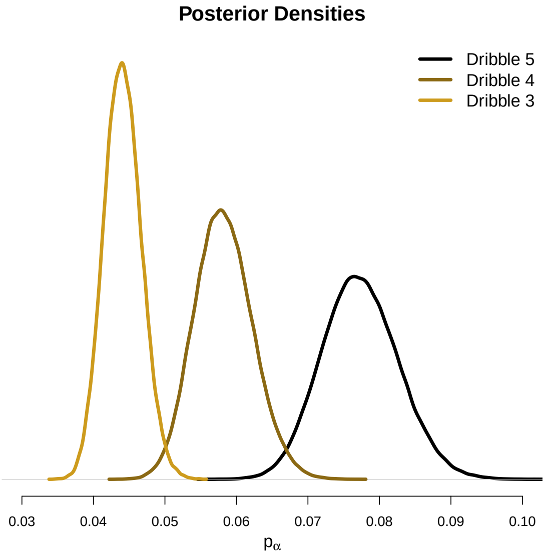
Results



Pass moves are rated:
7-penetrating
6-perpendicular
5-backward

Moves are nearly indistinguishable, indicating that either they are equally valuable, or the notation system would benefit from further distinguishing between pass moves.

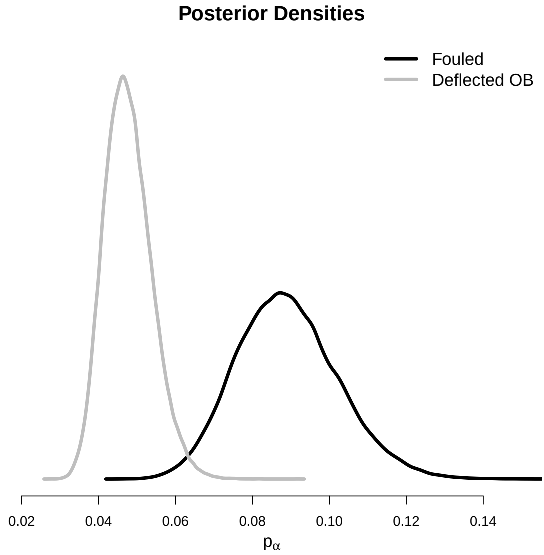
Results



Dribbling touches are rated:
5-penetrating
4-perpendicular
3-backward

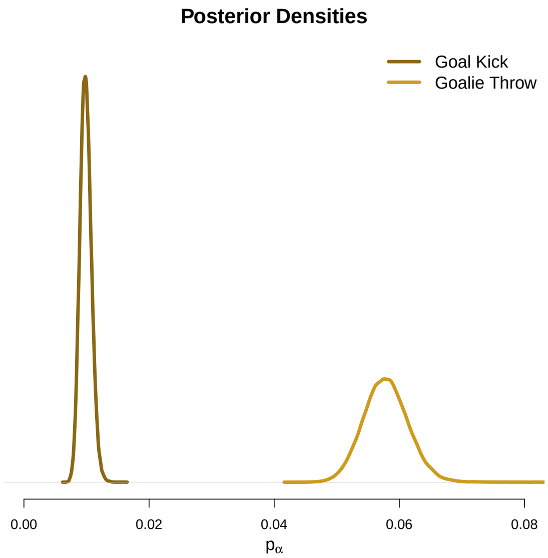
We have a clear distinction, confirming that advancing the ball increases scoring opportunity. We see the same pattern for controlling touches.

Results



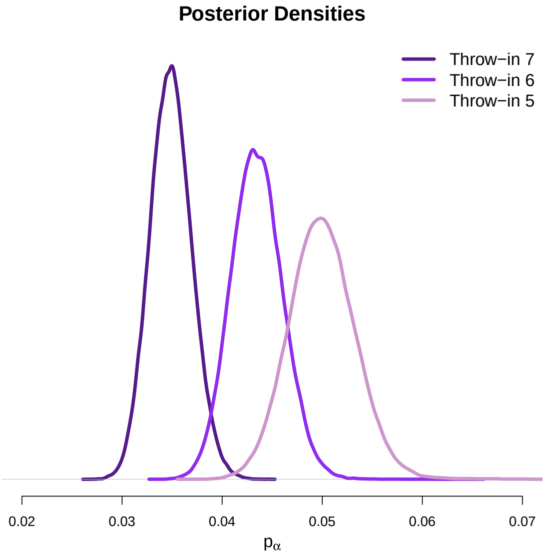
If forced to disrupt a play, it is better for defenders to deflect the ball out of bounds than to foul.

Results



The goalie can give her team a much better chance of taking a shot during the possession by throwing the ball to a teammate than by punting down the field. However, leaving the ball near their own goal introduces greater risk.

Results



Throw-ins are rated:
7-penetrating
6-perpendicular
5-backward

These results are counter-intuitive, but consistent and clear.

Results

These are a few of the insights available through modeling the data as an absorbing Markov chain.

We can also look at the probabilities of skills leading to turnovers and bad turnovers for insight into the risk/reward tradeoffs in an offensive attack.

References I

Casella, G. and Berger, R. L. (2002), *Statistical inference*, Pacific Grove, CA: Duxbury/Thomson Learning, 2nd ed.

de Campos, C. and Benavoli, A. (2011), "Inference with Multinomial Data: Why to Weaken the Prior Strength," in *International Joint Conference on Artificial Intelligence*, retrieved from http://www.aaai.org/ocs/index.php/IJCAI/IJCAI11/paper/view/3292.

FIFA (2006), "Info Plus," Retrieved from http://web.archive.org/web/20060915133001/http://access.fifa.com/infoplus/IP-199_01E_big-count.pdf.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian data analysis*, Boca Raton: CRC Press, 2nd ed.

Lawler, G. F. (1995), *Introduction to stochastic processes*, New York: Chapman & Hall.

Thomas, C. (2006), "Development of a notational analysis system for selected soccer skills of a women's college team," Ph.D. thesis, Brigham Young University.