Mathematics Invades Competitive Sports

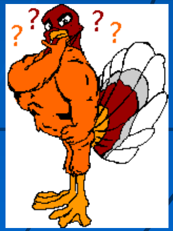
Kenneth Massey
Virginia Tech

Goals

- *Objectively* measure the performance of a team relative to the schedule faced
- Correct for disparate schedules (esp. college sports)
- Predictive vs. Retrodictive
  - Wins, scores, date, stats, homefield, preseason, other ?
- Seed playoffs

Georgia 26 Tennessee 24
Auburn 24 Georgia 17
Syracuse 31 Auburn 14
Georgia Tech 13 Syracuse 7
Virginia 39 Georgia Tech 38
Wisconsin 26 Virginia 17
Michigan St 42 Wisconsin 28
Minnesota 28 Michigan St 19
Toledo 38 Minnesota 7
Ball St 24 Toledo 20
N. Iowa 42 Ball St 39
Illinois St 42 N. Iowa 14
SW Texas 20 Illinois St 13
Nicholls St 33 SW Texas 14
Grambling 37 Nicholls St 28
Alabama St 45 Grambling 38

Alcorn St 20 Alabama St 17
Fort Valley 31 Alcorn St 16
Tuskegee 35 Fort Valley 28
Morehouse 14 Tuskegee 3
Benedict 20 Morehouse 0
Lane 24 Benedict 22
Miles 16 Lane 15
W. Alabama 35 Miles 12
Belhaven 21 W. Alabama 0
Pikeville 30 Belhaven 21
Cumberland KY 34 Pikeville 29
Bethel TN 40 Cumberland KY 27
Westminster MO 24 Bethel TN 21
Greenville 40 Westminster MO 14
Eureka 35 Greenville 28



Prediction: Eureka by 339 points over Tennessee

Challenges

- There is no property of transitivity!
- Disparate schedules

  "Mt. Union Syndrome"

  Separate Divisions
- Strength vs. Performance vs. Results

  Environment

  (venue, homefield, weather, day/night, crowd)

  Teams don't always play at full potential

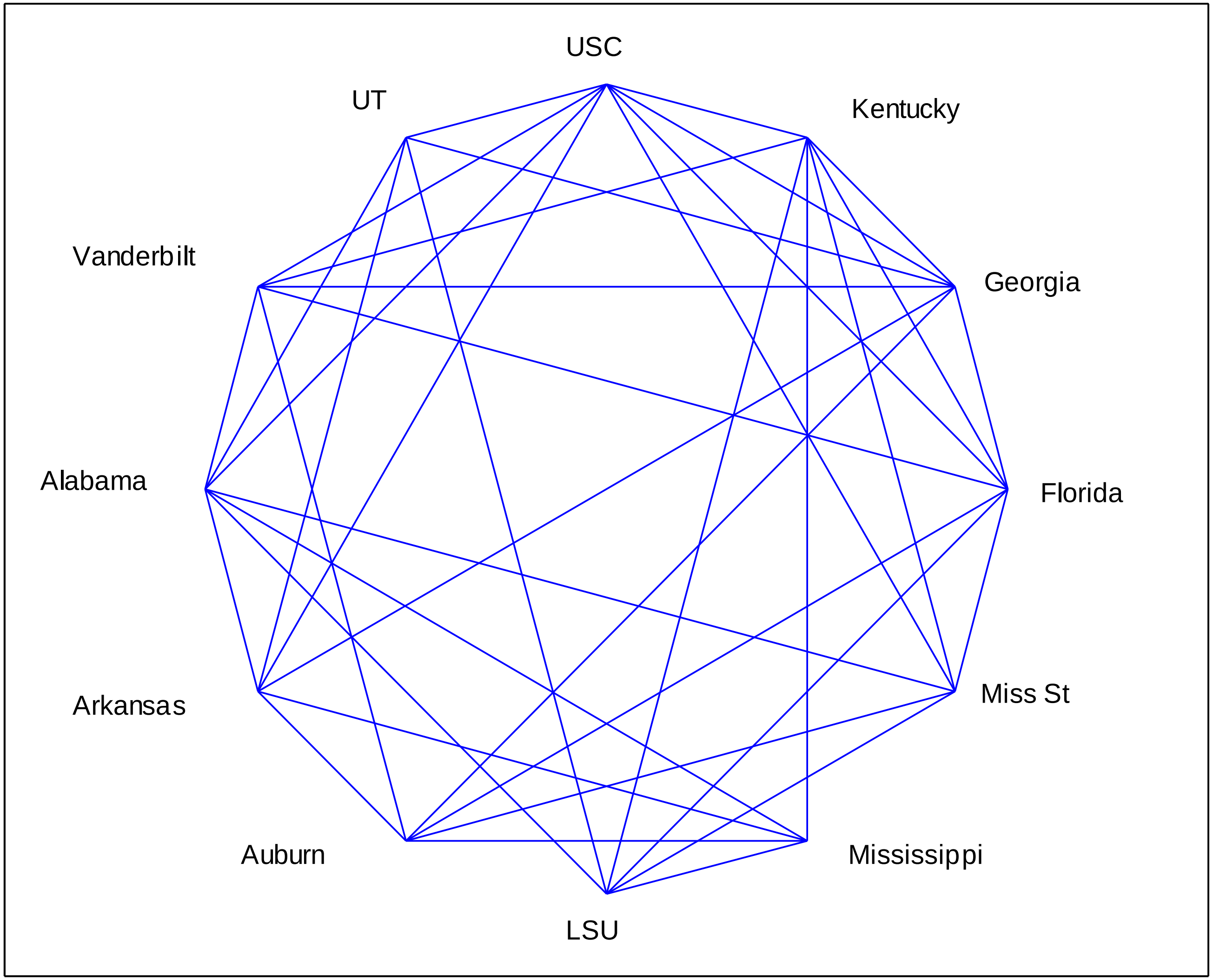
  (injury, unfavorable matchups, intangible, psychological)

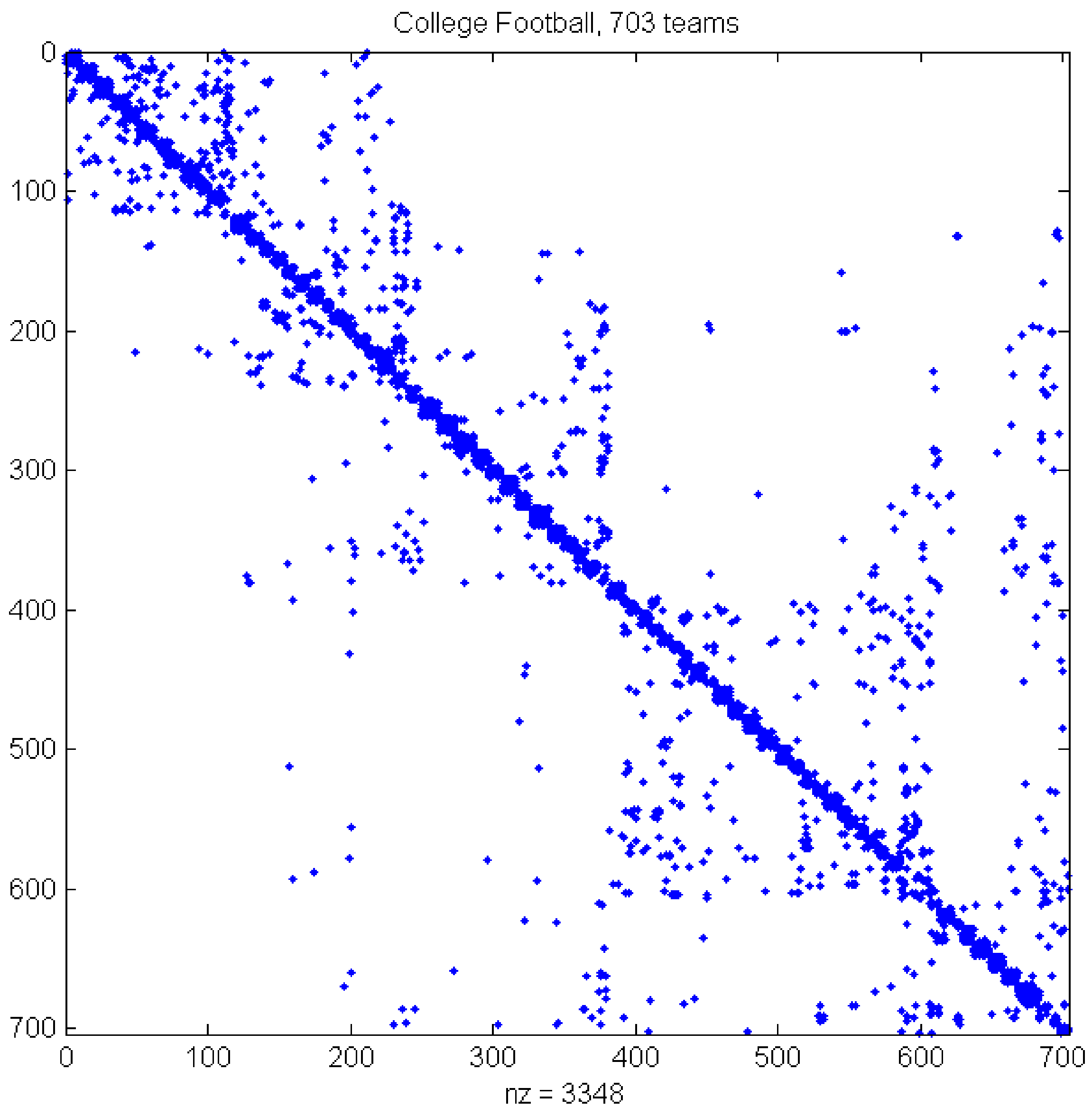
  The score isn't always a good indicator

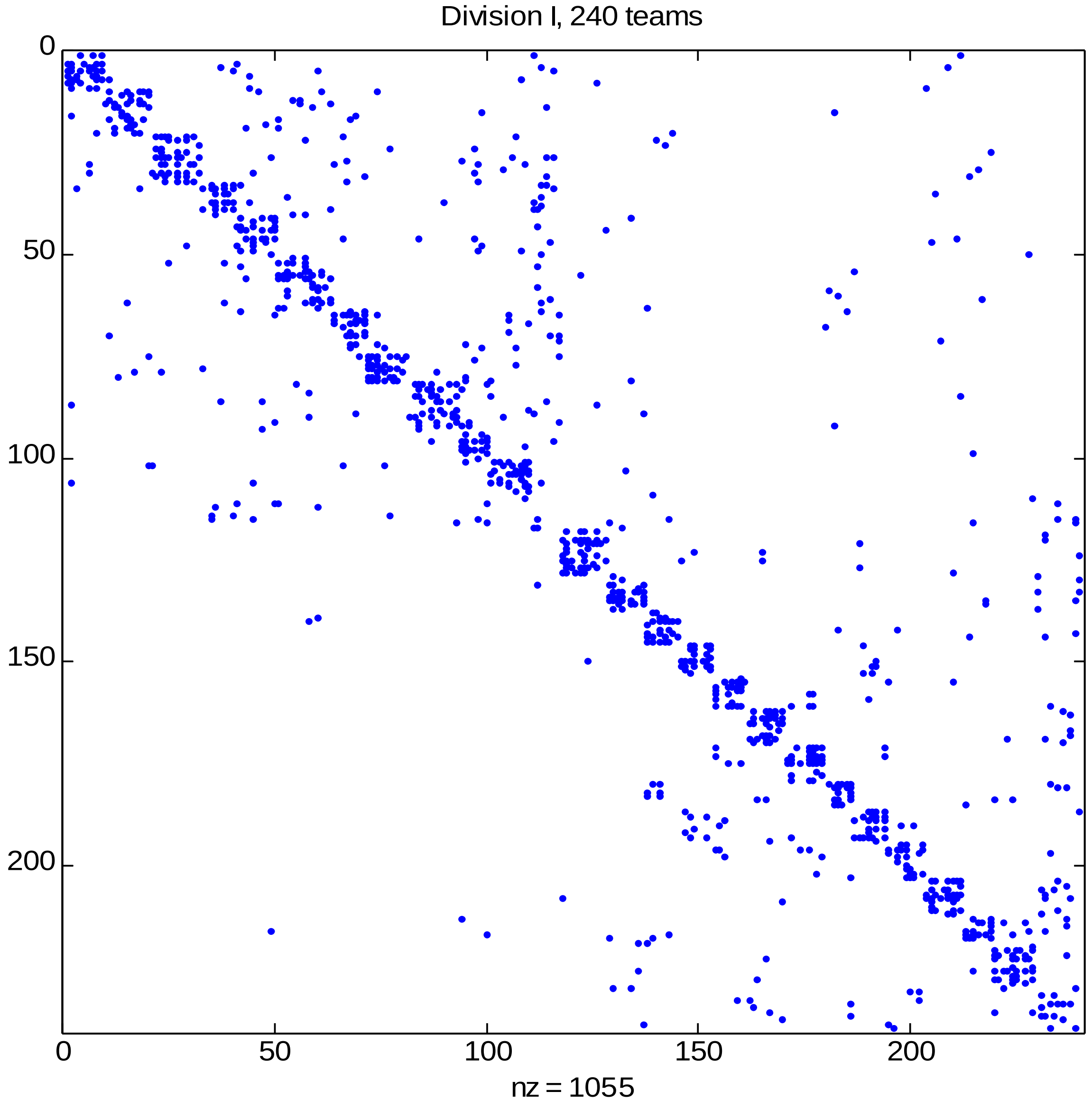
  (coaching philosophy, chaos "bounce of ball")
- Lack of data

  Connectedness
- Undefeated / winless teams

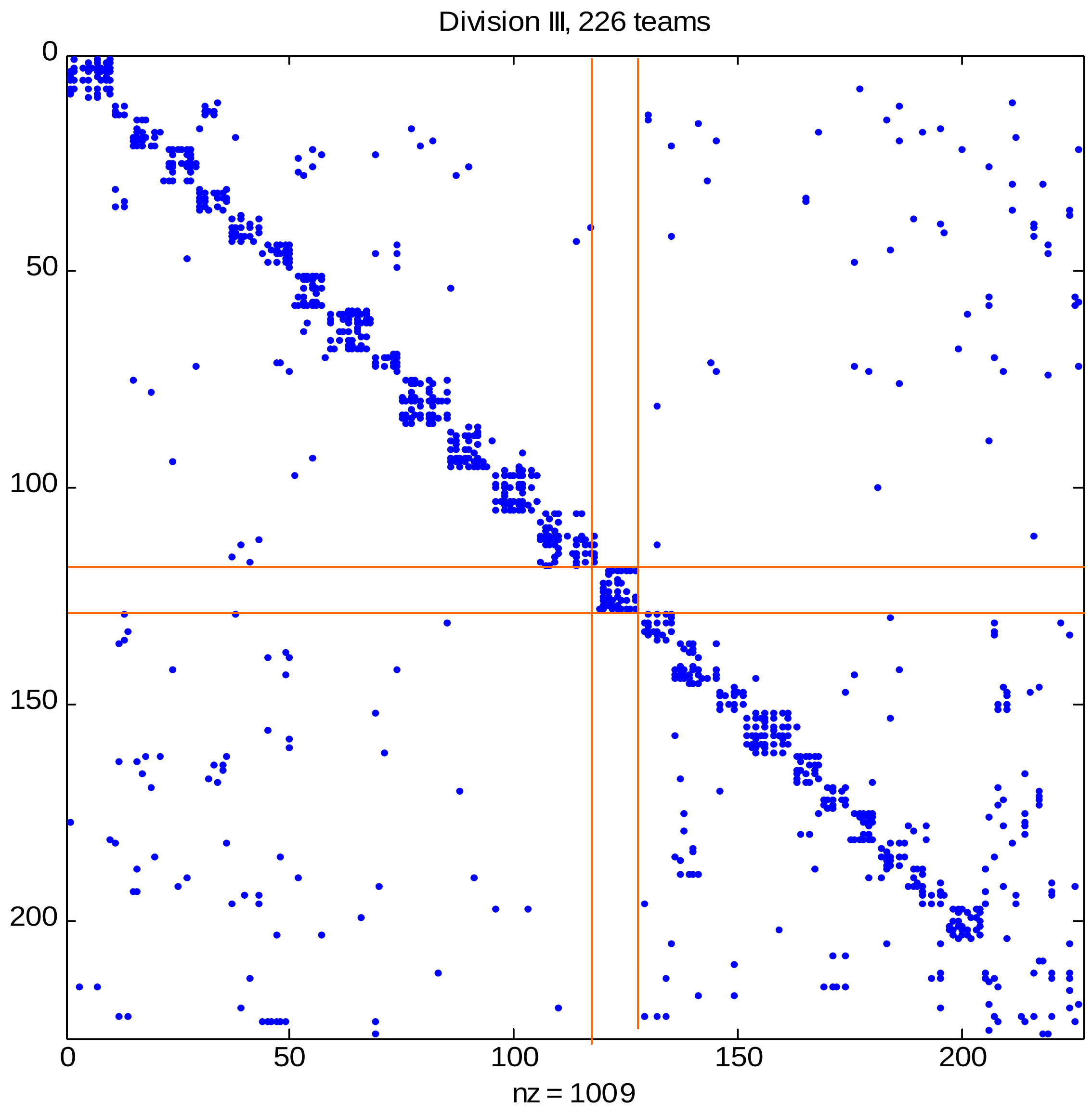
SEC Schedule Graph




College Football, 703 teams
0
100
200
300
400
500
600
700
nz = 3348

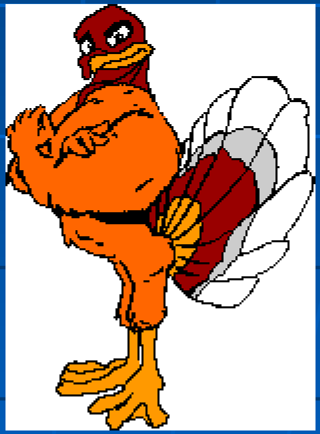


Division I, 240 teams
0
50
100
150
200
0
50
100
150
200
nz = 1055


Division III, 226 teams
0
50
100
150
200
0
50
100
150
200
nz = 1009

Types of Ratings

- Standings / WL% / Points
- Polls Tabulated votes, subjective, time sensitive, no corrections, incomplete analysis
- Formula (RPI)
- Update (Elo chess)
- Least Squares
- MLE
- Matrix (Markov)
- Other (Neural nets)



Bowl Championship Series (BCS)

- Polls
- Computers
- Schedule
- Losses
- Quality Wins

} redundant

Schedule Ratings

- Average rating of opponents (corrected for homefield)
- A good team prefers a less distributed schedule;
  a bad team prefers a more distributed schedule.

  For example (Florida, Vanderbilt) vs. (Alabama, Arkansas)

BCS Computers

Anderson / Hester	formula
Billingsley	update
Colley	matrix
Massey	MLE (Gaussian)
Matthews	matrix
Rothman	MLE (logistic)
Sagarin	MLE (logistic)
Wolfe / Baker	least squares

Least Squares Model

Assume the expected outcome b_k of a game is a linear function of the rating vector x.

$$E[b_k] = a_k^T x$$

Example: define $b_k = s_i - s_j$ = margin of victory (MOV) for team i over team j

suppose x = r contains the ratings for each team

$$a_k = e_i - e_j$$

$$E[b_k] = r_i - r_j$$

Least Squares Model

Assume there are n teams and $t \geq n$ rating parameters.

Suppose there are m observed game results. Let

$$A = \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_m^T \end{bmatrix} \in \Re^{m \times t} \qquad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \in \Re^{m}$$

We find x to solve the least squares problem:

$$\min_x ||Ax - b||$$

Rank Deficiency

If ker(A) is nontrivial, then there is no unique solution to the least squares problem.

- Additive scale invariance of the model
  Impose additional constraints, such as

$$\sum_{i=1}^{n} r_i = 1^T r = 0$$

- The schedule matrix is not connected
  Compute the minimal norm solution (SVD).
  Impose the constraint on each "group."
  Solve the problem for each group separately.

LS Example

We will use four rating parameters per team

- offense
- defense
- home advantage offense
- home advantage defense

Note that there might not be enough data to warrant such a complex model!

There are two observations per game: s_i and s_j

$$s_i = r_i^o - r_j^d + h_k(h_i^o + h_j^d)$$

$$s_j = r_j^o - r_i^d - h_k(h_j^o + h_i^d)$$

$$x = \begin{bmatrix} r^o \\ r^d \\ h^o \\ h^d \end{bmatrix} \qquad a_k = \begin{bmatrix} e_i & e_j \\ -e_j & -e_i \\ h_k e_i & -h_k e_i \\ h_k e_j & -h_k e_j \end{bmatrix} \qquad b_k = \begin{bmatrix} s_i \\ s_j \end{bmatrix}$$

$$a_k^T x = b_k$$

LS Example

Collecting all of the observations, we have the coefficient matrix:

$$A = [A_o \; A_d \; H_o \; H_d] \in \Re^{m \times 4n}$$

We combine and rearrange the equations with the help of the matrices

$$P = \begin{bmatrix} I & I & & \\ & & I & I \\ I & -I & & \\ & & I & -I \end{bmatrix} \qquad P^{-1} = \frac{1}{2} \begin{bmatrix} I & & I & \\ I & & -I & \\ & I & & I \\ & I & & -I \end{bmatrix}$$

With x = P^{-1}y, the least squares problem becomes

$$\min_y ||AP^{-1}y - b||$$

LS Example

The new coefficient matrix

$$AP^{-1} = \frac{1}{2}[(A_o + A_d)\ (H_o + H_d)\ (A_o - A_d)\ (H_o - H_d)]$$

has the property that the first two blocks of columns are orthogonal to the last two blocks. Therefore the problem decouples into:

$$\min_{y_1} \|\frac{1}{2}[(A_o + A_d)\ (H_o + H_d)]y_1 - b\|$$
$$\min_{y_2} \|\frac{1}{2}[(A_o - A_d)\ (H_o - H_d)]y_2 - b\|$$

where $y = [y_1\ y_2]^T$

In fact, since y = Px,

$$y_1 = \begin{bmatrix} r_o + r_d \\ h_o + h_d \end{bmatrix} \qquad y_2 = \begin{bmatrix} r_o - r_d \\ h_o - h_d \end{bmatrix}$$

We interpret this as first solving for the total rating y1, then determining the offense and defense parts from y2.

More LS Model Features

- Preseason ratings

  Augment observation matrix with $Ix = b_0$

  - Provide reasonable ratings despite lack of data
  - Insure unique solution to least squares problem
  - Pull team values toward the average.

  $$h_i = \frac{1}{n}\sum_k h_k$$

- Weighting $\min_x \|W(Ax-b)\|_2$ or $\min_x \|Ax-b\|_{W*W}$

- Choice of GOF
  - WIF $b_k = \text{sign}(s_i - s_j)$
  - BOMB $b_k = s_i - s_j$

LS Notes

Once necessary constraints, preseason observations, weightings, and change of variables have been applied, the least squares problem:

$$\min_x ||Ax - b||$$

should be full rank, and may be solved by the standard methods.

Note that the least squares solution may be interpreted as a MLE (statistical linear regression).

Also, it can be shown that the LS solution satisfies the expected = actual condition for the GOF.

Any linearly scaled ratings may be divided into off /def using LS.

Maximum Likelihood Estimator (MLE) Method

- Optimization Problem

  Choose ratings to maximize the probability of reproducing the observed results

- Game Outcome Function $0 \leq g \leq 1$

  Measures the result of a particular game

- Game Likelihood Function $0 \leq p \leq 1$

  The probability of a given result given a set of ratings

Game Outcome Function

- Win Indicator Function $g(s_i, s_j) = \alpha \ \mathrm{sign}(s_i - s_j)$ $\alpha = 1$
- Score Ratio $g(s_i, s_j) = \frac{s_i + \beta}{s_i + s_j + \beta}$ $\beta = 10$
- Rothman $g(s_i, s_j) = \alpha + (1 - \alpha) p(s_i - s_j)$ $\alpha = 0.5$
- Sagarin ? $g(s_i, s_j) = 1 - \exp\left(c_1 + \frac{s_i - s_j}{c_2}\right)$ $c_1 = -0.9$ $c_2 = -20$
- Massey $g(s_i, s_j) = p\left(\frac{s_i - s_j}{\sqrt{c_1 \sqrt{\frac{s_i + s_j}{c_2}}}}\right)$ $c_1 = 200$ $c_2 = 50$

Note: g could depend on other input, such as stats, or even ratings

GOF Values

s_i	s_j	WIF	SR	Roth	Sag	Mas
21	20	1	.6078	.7650	.6133	.5296
27	20	1	.6491	.8492	.7135	.6924
30	14	1	.7407	.9361	.8173	.8786
42	7	1	.8814	.9926	.9293	.9936
49	7	1	.8939	.9968	.9502	.9981
10	0	1	1	.8843	.7534	.8548
52	42	1	.5962	.8843	.7534	.7270

GOF Bias

Theorem:

Let z represent the measured performance of the favored team.

Suppose that g(z) satisfies the following properties:

g(z) is monotone increasing, $0 \leq g(z) \leq 1$, and $g(z) + g(-z) = 1$

Let f(z) be the p.d.f. for the z and suppose $f(z) < f(-z) \text{ if } z \leq 0$

Then the expected value of g cannot exceed the probability that the favorite will win. Furthermore, equality occurs if and only if g(z) is the win indicator function.

$$E[g(z)] \leq \int_0^\infty f(z)dz$$

Proof

$$
\begin{aligned}
E[g(z)] &= \int_{-\infty}^{\infty} g(z)f(z)dz \\
&= \int_{-\infty}^{0} g(z)f(z)dz + \int_{0}^{\infty} (1 - g(-z))f(z)dz \\
&= \int_{-\infty}^{0} g(z)(f(z) - f(-z))dz + \int_{0}^{\infty} f(z)dz \\
&\leq \int_{0}^{\infty} f(z)dz
\end{aligned}
$$

Clearly equality holds if and only if $g(z) = 0$ for $z < 0$.

Game Likelihood Functions

- Gaussian $p(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(\frac{-z^2}{2}\right) dz$ $p'(z) = \frac{1}{\sqrt{2\pi}} \exp(\frac{-z^2}{2})$

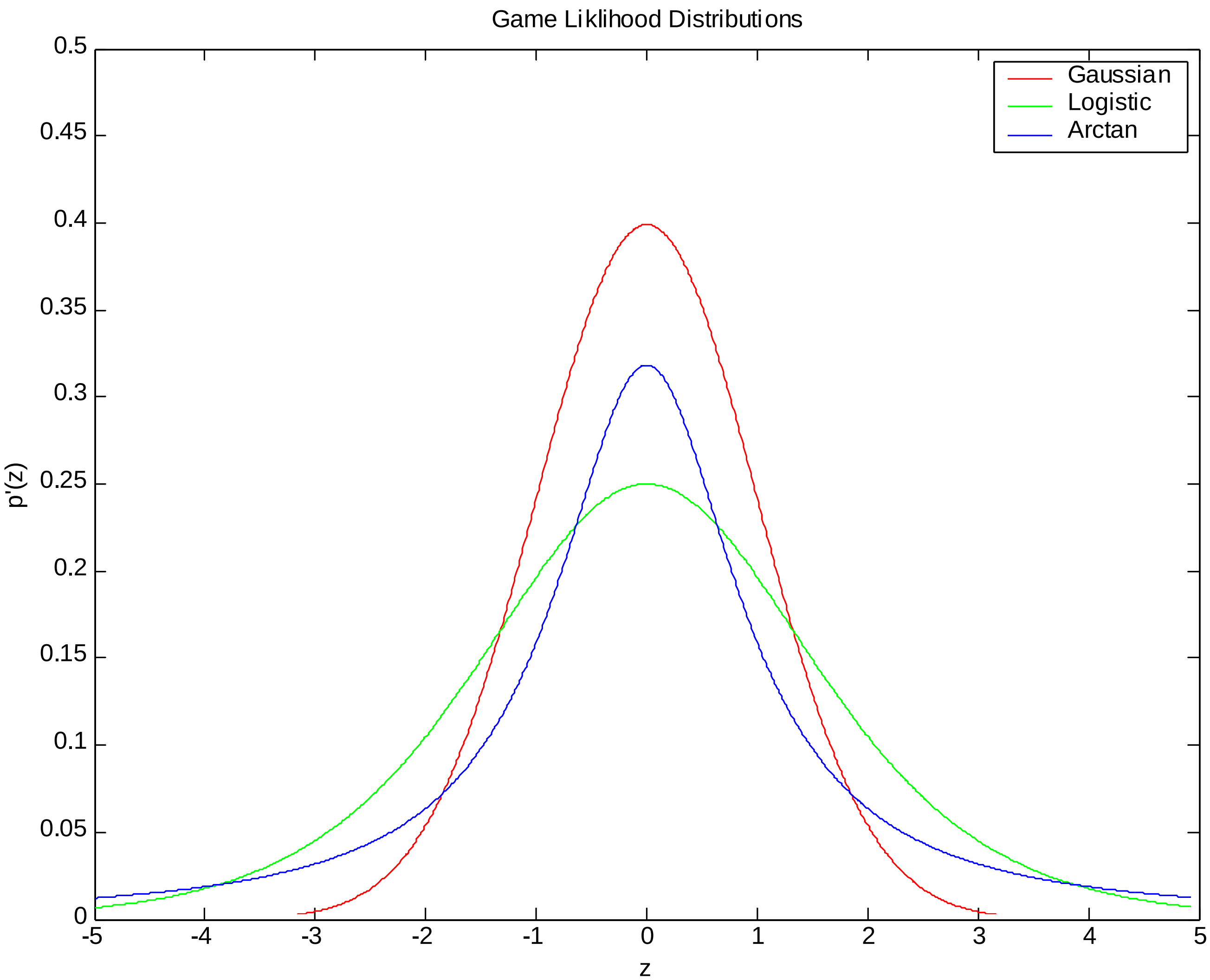
- Logistic $p(z) = \frac{1}{1+e^{-z}}$ $p'(z) = \frac{e^{-z}}{(1+e^{-z})^2} = p(z)(1-p(z))$

Equivalent to $p(r_i, r_j) = \frac{r_i}{r_i + r_j} \quad r_i, r_j \in [0, \infty)$ or $p(r_i, r_j) = \frac{r_i(1-r_j)}{r_i(1-r_j) + r_j(1-r_i)} \quad r_i, r_j \in [0, 1]$

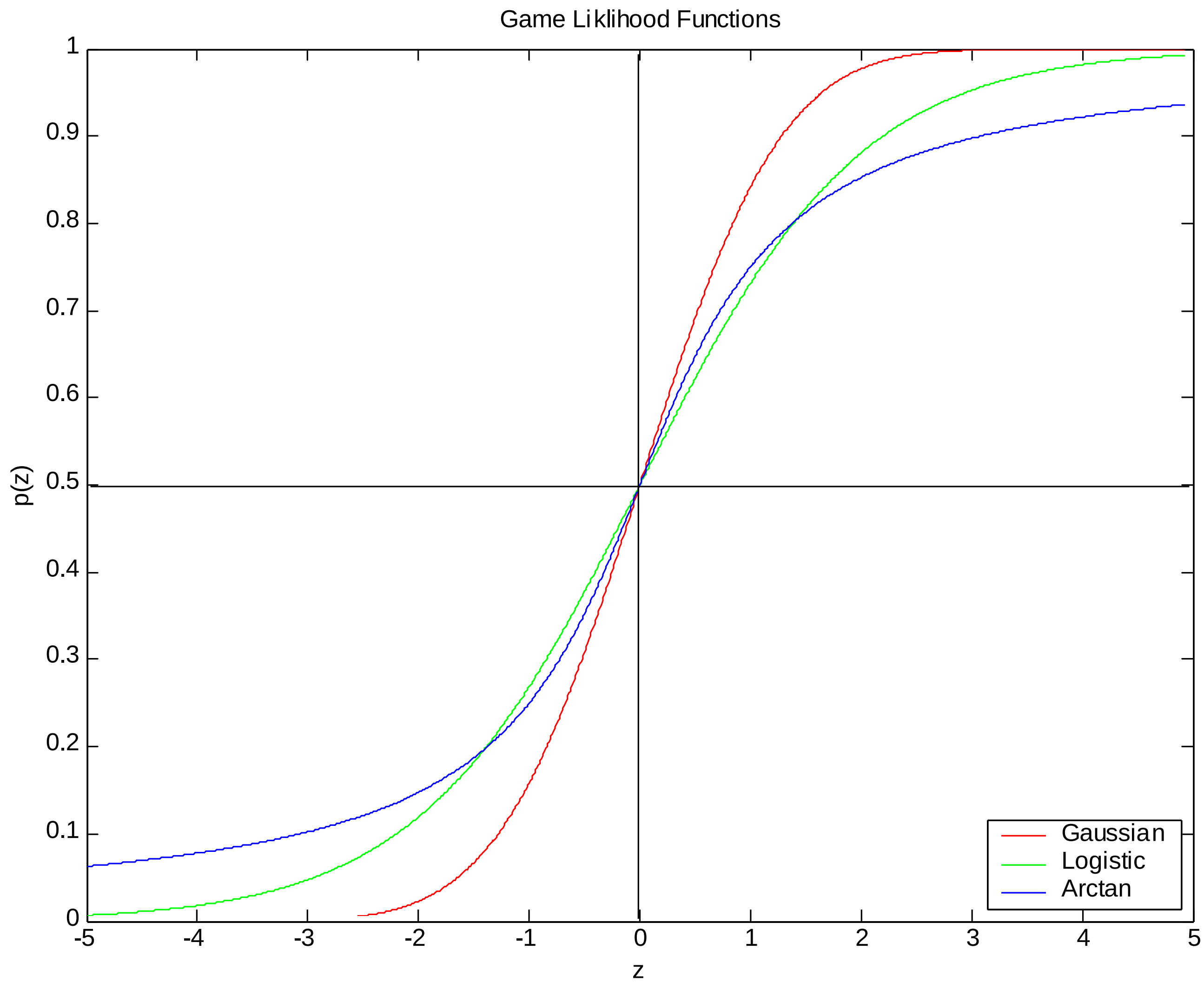
- Arctan $p(z) = \frac{1}{2} + \frac{1}{\pi}\arctan(z)$ $p'(z) = \frac{1}{\pi(1+z^2)}$

The variable z is a typically a linear function of the rating parameters.

$$z = r_i - r_j + H(h_i + h_j) \in \Re$$



Game Liklihood Distributions
Gaussian
Logistic
Arctan
p'(z)
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
z


Game Liklihood Functions
p(z)
z
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
Gaussian
Logistic
Arctan

MLE Function

For a particular game k, model the probability of the observed result given a set of ratings, x, as:

$$\hat{f}_k(x) = p(z_k)^{g_k}(1 - p(z_k))^{1-g_k}$$

Example: $0.8^{0.6}\ 0.2^{0.4}$

Define the MLE function to be the weighted product of all game probabilities:

$$\hat{f}(x) = \prod_k [f_k(x)]^{-w_k}$$

For computational purposes, we minimize

$$f(x) = \log \hat{f}(x) = \sum_k -w_k f_k(x)$$

where

$$f_k(x) = \log \hat{f}_k(x) = g_k \log p(z_k) + (1 - g_k)\log(1 - p(z_k))$$

MLE Optimization (Logistic Model)

Taking the partial derivatives with respect to a rating parameter x_i

$$\frac{\partial f}{\partial x_i} = \sum_k -w_k \left(\frac{g_k}{p_k} - \frac{1-g_k}{1-p_k} \right) \frac{dp}{dz_k} \frac{\partial z_k}{\partial x_i} = \sum_k \frac{w_k(p_k - g_k)}{p_k(1-p_k)} \frac{dp}{dz_k} \frac{\partial z_k}{\partial x_i}$$

In the logistic model, we have that: $\frac{dp}{dz} = p(1-p)$

Therefore the derivatives reduce to:

$$\frac{\partial f}{\partial x_i} = \sum_k w_k (p_k - g_k) \frac{\partial z_k}{\partial x_i}$$

If we choose z_k so that the coefficient of x_i is always positive, setting the derivative to zero yields the (expected = actual) property:

$$\sum_k w_k p_k = \sum_k w_k g_k$$

The second derivatives are also easy to calculate:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \sum_k w_k p_k (1-p_k) \frac{\partial z_k}{\partial x_i} \frac{\partial z_k}{\partial x_j}$$

MLE Issues

- Non-uniqueness
- Ideally, wins help and losses hurt
- Undefeated / Winless Teams
  - Don't use the WIF.
  - Use prior distribution (Bayesian)
  - Use a linear approximation to f

$$\min_x [\alpha f(x) + (1-\alpha) f'(x)(Gx - x_0)]$$

- Update ratings based on linearization

  (time dependent, n large)

$$F(s,x) = 0 \qquad dx \approx -F_x^{-1} F_s ds$$

Ratings on the Web

- Massey Ratings

  http://www.masseyratings.com
- College Football Rankings
  http://www.cae.wisc.edu/~dwilson/rsfc/rate/index.html
- Bowl Championship Series

  http://www.collegebcs.com