Intelligence in Web and Sports Rating by Least Squares and Maximum Likelihood Estimation

Chee Wei TAN

Ranking Web Pages: Google's PageRank

The Anatomy of a Large-Scale Hypertextual Web Search Engine, Brin and Page, 1998

http://infolab.stanford.edu/~backrub/google.html

Positive and Nonnegative Vectors and Matrices[1]

A matrix or vector is

- positive (or elementwise positive) if all its entries are positive
- nonnegative (or elementwise nonnegative) if all its entries are nonnegative

The notation $x > y$ ($x \geq y$) means $x - y$ is elementwise positive(nonnegative)
If $A \geq 0$ and $z \geq 0$, then we have $Az \geq 0$. Conversely: if for all $z \geq 0$, we have $Az \geq 0$, then we can conclude $A \geq 0$
If $A > 0$ and $z \geq 0$, $z \neq 0$, then $Az > 0$. Conversely, if whenever $z \geq 0, z \neq 0$, we have $Az > 0$, then we can conclude $A > 0$

[1]This and next seven slides from Course Reader of Lecture 17 of EE363: Linear Dynamical Systems by Professor Stephen Boyd, Stanford University, Winter Quarter 2008-09.

Regular Nonnegative Matrices

- Suppose $A \in \mathbf{R}^{n \times n}$, with $A \geq 0$
- A is called regular if for some $k \geq 1, A^k > 0$
- meaning: form directed graph on nodes $1, \ldots, n$, with an arc from j to i whenever $A_{ij} > 0$
- then $(A^k)_{ij} > 0$ if and only if there is a path of length k from j to i
- A is regular if for some k there is a path of length k from every node to every other node

Perron-Frobenius Theorem for Regular Matrices

Suppose $A \in \mathbf{R}^{n \times n}$ is nonnegative and regular, i.e., $A^k > 0$ for some k, then

- there is an eigenvalue $\rho(A)$ of A that is real and positive, with positive left and right eigenvectors
- for any other eigenvalue λ, we have $|\lambda| < \rho(A)$
- the eigenvalue $\rho(A)$ is simple, i.e., has multiplicity one, and corresponds to a 1×1 Jordan block
- This eigenvalue $\rho(A)$ is called the Perron-Frobenius (PF) eigenvalue of A, and the associated positive (left and right) eigenvectors are called the (left and right) Perron-Frobenius eigenvectors (and are unique, up to positive scaling)

Perron-Frobenius Theorem for Nonnegative Matrices

Suppose $A \in \mathbf{R}^{n \times n}$ and $A \geq 0$, then

- there is an eigenvalue $\rho(A)$ of A that is real and nonnegative, with associated nonnegative left and right eigenvectors
- for any other eigenvalue λ of A, we have $|\lambda| \leq \rho(A)$
- $\rho(A)$ is called the Perron-Frobenius (PF) eigenvalue of A, and the associated nonnegative (left and right) eigenvectors are called (left and right) Perron-Frobenius eigenvectors
- The eigenvectors need not be unique, or positive

Markov Chains

Consider stochastic process $X_0, X_1, \ldots$ with values in $\{1, \ldots, n\}$

$$\mathbf{Prob}(X_{t+1} = i | X_t = j) = P_{ij}$$

- P is called the transition probability matrix; clearly $P_{ij} \geq 0$
- Let $p_t \in \mathbf{R}^n$ be the distribution of X_t, i.e., $(p_t)_i = \mathbf{Prob}(X_t = i)$, then we have

$$p_{t+1} = Pp_t$$

  note: standard notation uses transpose of P, and row vectors for probability distributions
- P is a stochastic matrix, i.e., $P \geq 0$ and $\mathbf{1}^T P = \mathbf{1}^T$
- So $\mathbf{1}$ is a left eigenvector with eigenvalue 1, which is in fact the Perron-Frobenius eigenvalue of P

Stationary Distribution

- Let π denote a Perron-Frobenius (right) eigenvector of P, with $\pi \geq 0$ and $\mathbf{1}^T\pi = 1$
- Since $P\pi = \pi$, π corresponds to a stationary distribution or equilibrium distribution of the Markov chain
- Suppose P is regular, which means for some k, $P^k > 0$, since $(P^k)_{ij}$ is $\mathbf{Prob}(X_{t+k} = i | X_t = j)$, this means there is positive probability of transitioning from any state to any other in k steps
- Since P is regular, there is a unique invariant distribution π, which satisfies $\pi > 0$
- the eigenvalue 1 is simple and dominant, so we have $p_t \to \pi$, no matter what the initial distribution p_0, i.e., the distribution of a regular Markov chain always converges to the unique invariant distribution

Convergence Rate to Stationary Distribution

- Rate of convergence to stationary distribution depends on second largest eigenvalue magnitude, i.e.,

$$\mu = \max\{|\lambda_2|, ..., |\lambda_n|\}$$

where λ_i are the eigenvalues of P, and $\lambda_1 = \rho(P) = 1$
(μ is sometimes called the SLEM of the Markov chain)

- The mixing time of the Markov chain is given by

$$T = \frac{1}{\log(1/\mu)}$$

(roughly, number of steps over which deviation from equilibrium distribution decreases by factor e)

Power Method

Assume $A \geq 0$ and regular. Consider

$$x(k+1) = Ax(k),$$

- By Perron-Frobenius theorem, $\rho(A)$ is the unique dominant eigenvalue. Let $v, w > 0$ be the right and left Perron-Frobenius eigenvectors of A, with $\mathbf{1}^T v = 1, w^T v = 1$
- As $k \to \infty, (\rho(A)^{-1}A)^k \to vw^T$ for any $x(0) \geq 0, x(0) \neq 0$, we have

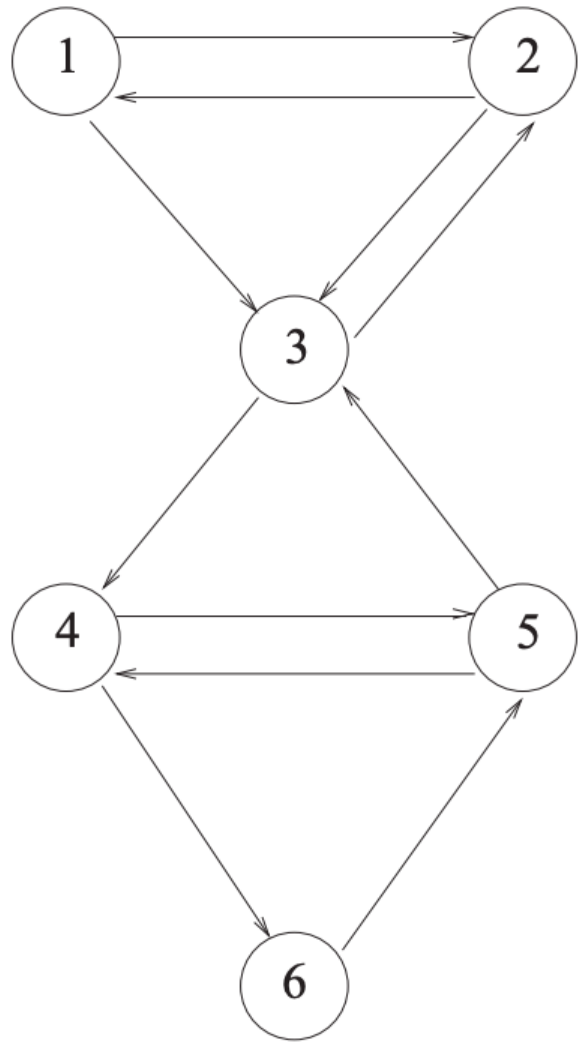
$$\frac{1}{\mathbf{1}^T x(k)} x(k) \to v$$

- As $k \to \infty$, the distribution of $x(k)$ converges to v

Google's PageRank Problem

A search engine wants to rank N webpages. A directed edge $i \to j$ means that Webpage i has a hyperlink pointing to Webpage j. A user surfs from one webpage to another using hyperlinks, and may click on a hyperlink with some probability α or *teleport* to other (possibly not connected) webpage with probability $(1 - \alpha)$. Teleport models a new search (e.g., search & click Google's *I'm Feeling Lucky* button). Let π_i be the rank (PageRank) of Webpage i.

Example of $N = 6$ Webpages



Google's PageRank Problem

- Intuitively, a webpage's *popularity or importance* (rank) is proportional to the number of webpages each having a hyperlink pointing to it as well as the rank of those webpages
- When the user is at Webpage i, assume the user surfs to Webpage j with probability $P_{ij} = 1/n_i$, where n_i is the number of hyperlinks on Webpage i (i.e., assume the user clicks on any of the hyperlinks on Webpage i equally likely)
- When the user does a new search, assume the user teleports to any other webpage equally likely, i.e., with probability $1/N$. Let $v = (v_1, \ldots, v_N)^T$ where $v_i = 1/N$ for Webpage i
- Allows web surfing to be modeled by a random walk on a Markov chain whose transition probability matrix captures the adjacency matrix of the web graph and also behaviors of the user's web searching (i.e., α) and the search engine (i.e., v)

PageRank Least Squares Solution

Find PageRank $\pi \geq 0$ such that

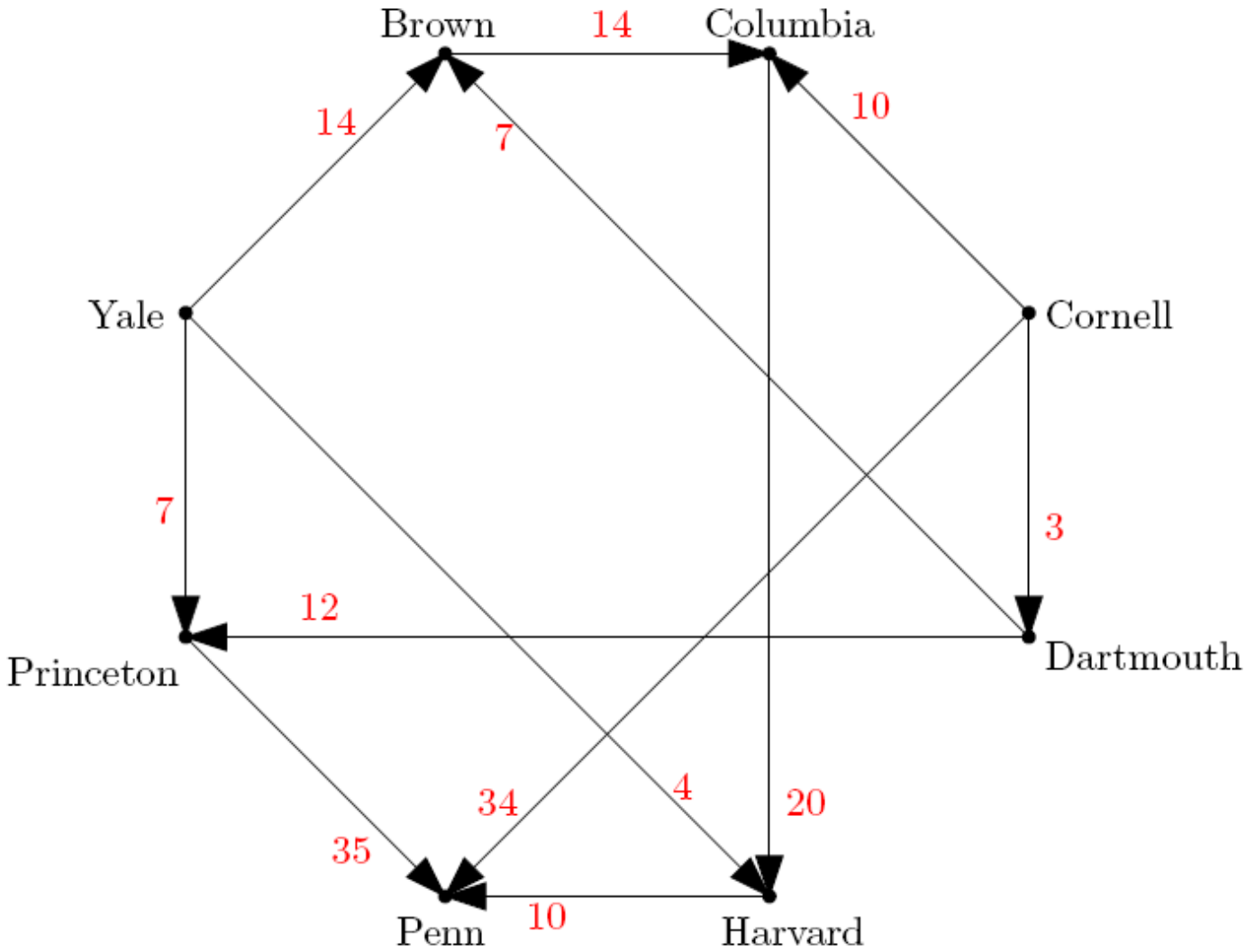
$$\pi = \alpha P\pi + (1-\alpha)v,$$
$$\mathbf{1}^T\pi = 1.$$

- PageRank matrix P is stochastic, high-dimensional and sparse
- The PageRank π is unique. What is its analytical closed form?
- What if we only want accurate approximation of a few entries of PageRank corresponding to top most influential webpages?

The PageRank π is a least squares solution to

$$\begin{array}{ll} \text{minimize} & \|(I-\alpha P)\pi - (1-\alpha)v\|_2 \\ \text{subject to} & \mathbf{1}^T\pi = 1, \pi \geq 0. \end{array}$$

Ranking Ivies' Football Teams



Which is the best team? Is Dartmouth better than Yale?

Ranking Methods in Sports Analytics

Statistical inference methods using optimization techniques and algorithms for problems with graph structures

- First gather data then construct a mathematical model and formulate a convex optimization problem
- Deterministic method
  - Round robin tournament ranking in graph theory
  - Massey's Method of Least Squares
- Probabilistic methods
  - Maximum-likelihood estimation
  - Elo's rating for chess
  - Keener's Method
  - Machine learning-based methods

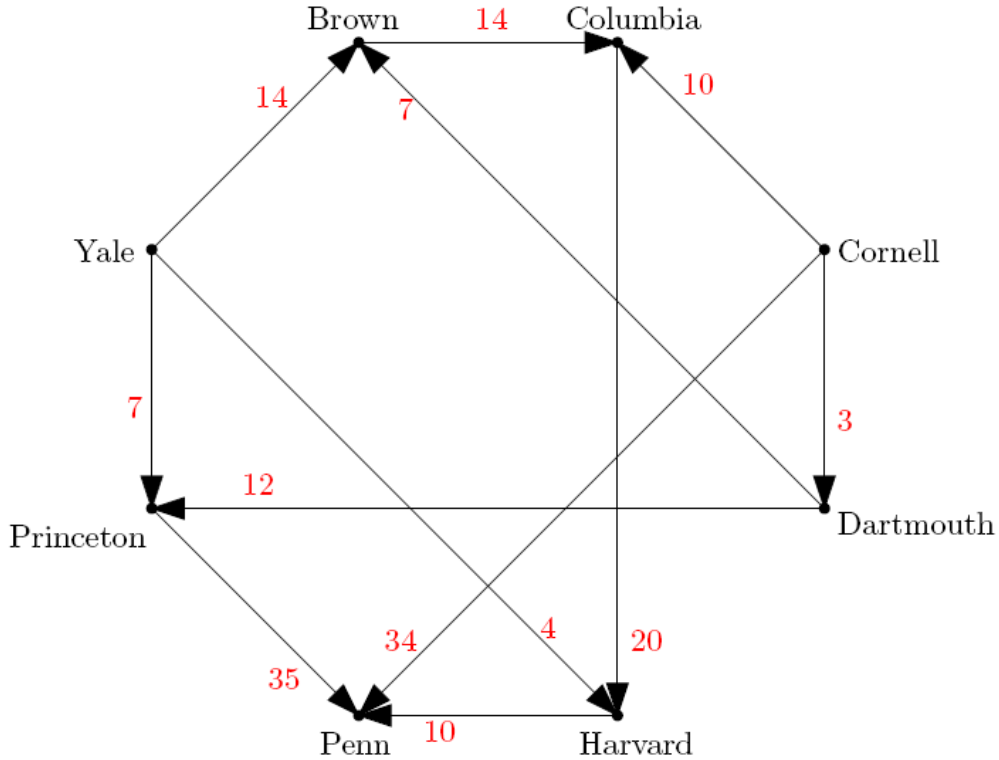
Massey's Problem Formulation

Kenneth Massey's Website: http://www.masseyratings.com

Consider the problem of predicting a rating for all teams. Let r_i be the rating of ith team. If Team i plays against Team j, let the game outcome point difference be $r_i - r_j$, e.g.,

$$r_{Brown} - r_{Yale} = 14$$
$$r_{Columbia} - r_{Brown} = 14$$
$$r_{Columbia} - r_{Cornell} = 10$$

A unique solution for 12 equations and 8 variables?



Massey's Problem Formulation

Let the number of teams and games be n and m respectively

- Consider the $m \times n$ matrix $\mathbf{B}$ where $B_{ki} = 1$, $B_{kj} = -1$, and $B_{kl} = 0$ if $l \neq i, j$ to model that Team i beats Team j in the kth game. We call $\mathbf{B}$ the incidence matrix of the graph.
- Data vector $\mathbf{v} = (v_1, v_2, \ldots, v_m)^T$ where v_k is the game outcome score difference, i.e., margin of victory
- Rating vector $\mathbf{r} = (r_1, r_2, \ldots, r_n)^T$ and we say Team i is better than Team j when $r_i > r_j$

We want to find a rating that is consistent with the game outcome:

$$\mathbf{Br} = \mathbf{v}$$

Massey's Least Squares Problem

Since there might be no solution, let us minimize

$$\|\mathbf{Br} - \mathbf{v}\|_2$$

whose optimal solution $\mathbf{r}^\star$ is the rating solution that can be used to predict future game outcome.

- Since this is unconstrained least squares, does one get a solution $\mathbf{r} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{v}$ by solving the normal equations?
  - No! in fact, the Gramian $\mathbf{B}^T\mathbf{B}$ is not invertible (its rank?)
- Normal equations may not have a unique solution and a unique rating is needed. Let's add an equality constraint $\mathbf{r}^T\mathbf{1} = 0$ and consider the constrained least squares problem:

$$\begin{array}{ll} \text{minimize} & \|\mathbf{Br} - \mathbf{v}\|_2 \\ \text{subject to} & \mathbf{r}^T\mathbf{1} = 0. \end{array}$$

Massey's Method to Least Squares Solution

- Exploit the structure of normal equations (Gramian) to derive a unique least squares solution by constraining the rating vector to be orthogonal to the all-ones vector
  - Show the all-ones vector is in the null-space of the Gramian
  - Interpret the rating values that are positive and negative as, respectively, above and below the average rating of zero.
- (Massey's Method) Replace any one row of Gram matrix by the all-ones vector and zero out the corresponding v element
- Can solve the least squares using CVX or the modified normal equations using Jacobi method (check convergence condition).
- Consider the special case of Round Robin Tournament where each team has a match with every other team. What do you observe?

Massey's Method to Least Squares Solution

- Consider a small example: A beats B by 10 points, B beats C by 5 points and C beats A by 1 point. Let r_1, r_2, r_3 be the ratings of A, B and C respectively. Let us minimize

$$\left\| \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix} - \begin{bmatrix} 10 \\ 5 \\ 1 \end{bmatrix} \right\|_2 .$$

- Form the Gram matrix first and then replace any of its row using the equality constraint $\mathbf{r}^T\mathbf{1} = 0$ to obtain, respectively:

$$\begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix} = \begin{bmatrix} 9 \\ -5 \\ -4 \end{bmatrix}, \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix} = \begin{bmatrix} 9 \\ -5 \\ 0 \end{bmatrix}$$

Solving this linear system, we have $r_1^\star = 3$, $r_2^\star = -1.667$, and $r_3^\star = -1.333$.

Probabilistic Models for Rating

- Model the skill of Team j by parameter $\theta_j > 0, j = 1, 2, ..., k$
- Thurstone-Mosteller Model (1951): $X_i - X_j \backsim$ normal distribution

$$\textbf{Prob}(\text{Team } i \text{ beats } Teamj) = \textbf{Prob}(X_i > X_j) = \Phi\left(\frac{\theta_i - \theta_j}{\sqrt{2\sigma^2}}\right),$$

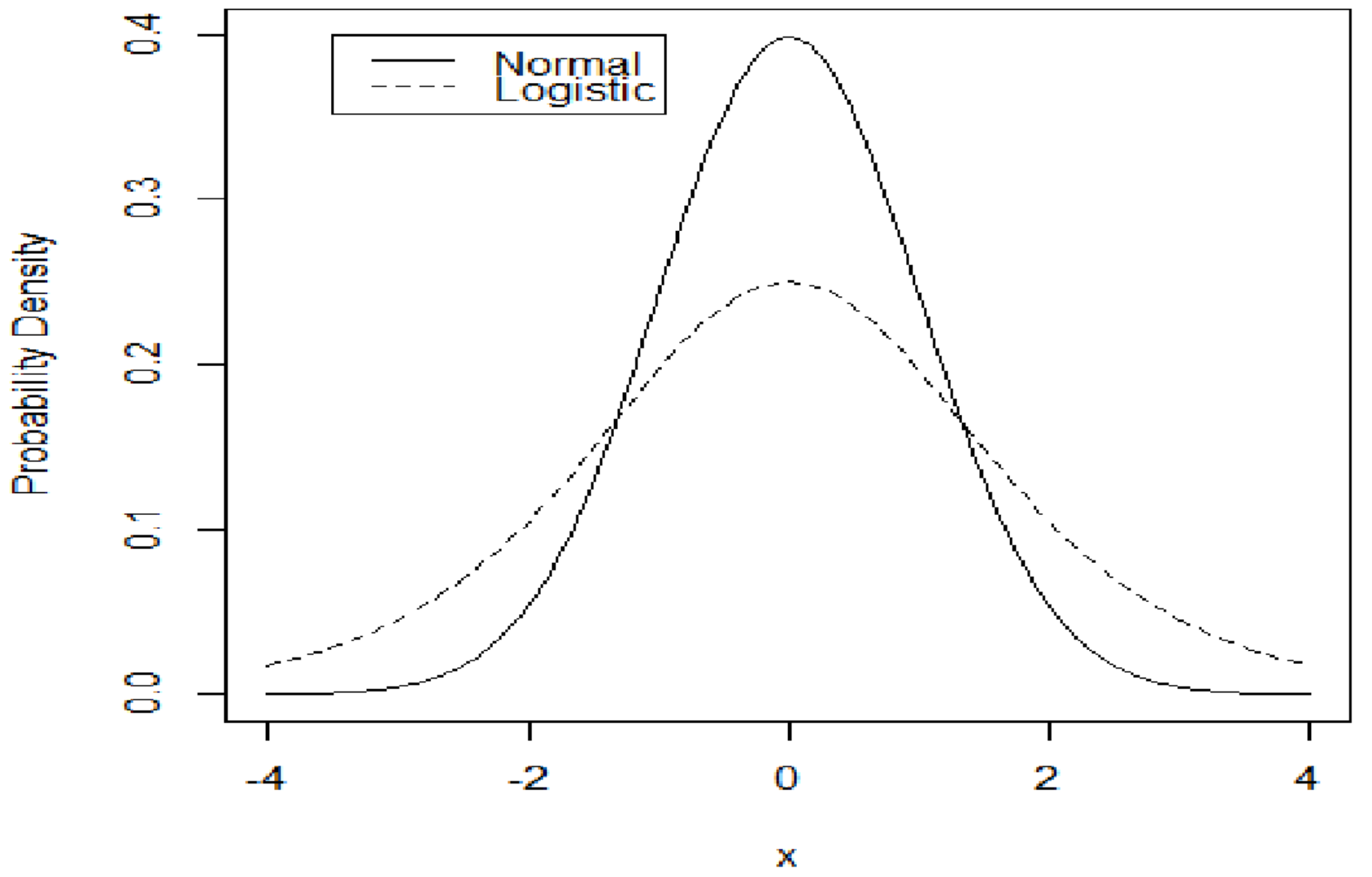
where Φ is cumulative distribution function of the standard Gaussian distribution $N(0, 1)$.

- Bradley-Terry Model (BT 1952, Zermelo 1929): $X_i - X_j \backsim$ logistic distribution

$$\textbf{Prob}(\text{Team } i \text{ beats } Teamj) = \textbf{Prob}(X_i > X_j) = \frac{\theta_i}{\theta_i + \theta_j}$$

- Can we infer skills from observed performance? Can these probabilistic models predict outcomes of future games?

Normal versus Logistics Distributions



Maximum Likelihood Estimation

- Let the number of times that Team i beats Team j be r_{ij}.
- The likelihood of the outcome of a tournament of five games involving three teams is given by

$$\begin{aligned}&\textbf{Prob}(r_{12}=1, r_{21}=2, r_{31}=2)\\ &\quad=\left(\frac{\theta_1}{\theta_1+\theta_2}\right)\left(\frac{\theta_2}{\theta_2+\theta_1}\right)^2\left(\frac{\theta_3}{\theta_3+\theta_1}\right)^2\end{aligned}$$

- Estimate skill parameters θ subject to $\theta_i > 0$, $\sum_i \theta_i = 1$
- Let the number of teams be n. Given data $\{w_{ij}\}$, $\forall\, i, j$. Consider the Maximum Likelihood Estimation (MLE) problem:

$$\begin{aligned}\text{maximize} \quad & \prod_{i=1, i\neq j}^{n}\left(\frac{\theta_i}{\theta_i+\theta_j}\right)^{w_{ij}}\\ \text{subject to} \quad & \mathbf{1}^T\boldsymbol{\theta}=1,\ \boldsymbol{\theta}\geq\mathbf{0}.\end{aligned}$$

Maximum Likelihood Estimation is Non-Convex

The observations of the outcomes of previous comparisons are assumed to be independent. Taking the logarithm of the objective function, we obtain a log-likelihood function in terms of the parameter vector $\boldsymbol{\theta} = \theta_1, \ldots \theta_n$ as

$$L(\boldsymbol{\theta}) = \sum_{i}^{n} \sum_{j, j \neq i}^{n} w_{ij} \log \theta_i - w_{ij} \log(\theta_i + \theta_j).$$

Denote the number of comparisons "won" by i as W_i, and the number of comparisions made between i and j as N_{ij}, i.e., $N_{ij} = w_{ij} + w_{ji}$ (assuming $w_{ii} = 0$ by convention). This non-convex optimization problem can surprisingly be solved using an iterative algorithm.

Algorithm to Solve the Maximum Likelihood Estimation

Start from an arbitrary positive $\boldsymbol{\theta}(0)$ and repeat until convergence.
Step 1: At the kth iteration, the algorithm updates:

$$\theta_i(k+1) = W_i \left(\sum_{j \neq i} \frac{N_{ij}}{\theta_i(k) + \theta_j(k)} \right)^{-1}$$

for all i.
Step 2: Renormalize all the iterates:

$$\theta_i(k+1) \leftarrow \frac{\theta_i(k+1)}{\sum_{j=1}^{n} \theta_j(k+1)}.$$

This algorithm increases the log-likelihood at every iteration, and converges to a unique solution asymptotically.

Maximum Likelihood Estimation Example

An example of $n = 2$ Teams

	NCCU	NTU
NCCU		8
NTU	2	

$\Rightarrow (\theta_1, \theta_2) = (0.8, 0.2)$

- If $r_{12} \leftarrow 9$, then $(\theta_1, \theta_2) = (0.82, 0.18)$
  θ_1 increases 0.02
  (beat a weaker team, gains little in θ
- If $r_{21} \leftarrow 3$, then $(\theta_1, \theta_2) = (0.73, 0.27)$
  θ_1 decreases 0.07
  (beaten by a weaker team, loses more in θ