Colley's Bias Free College Football Ranking Method:

The Colley Matrix Explained

Wesley N. Colley

Ph.D., Princeton University

ABSTRACT

Colley's matrix method for ranking college football teams is explained in detail, with many examples and explicit derivations. The method is based on very simple statistical principles, and uses only Div. I-A wins and losses as input — margin of victory does not matter. The scheme adjusts effectively for strength of schedule, in a way that is free of bias toward conference, tradition, or region. Comparison of rankings produced by this method to those produced by the press polls shows that despite its simplicity, the scheme produces common sense results.

Subject headings: methods: mathematical

1. Introduction

The problem of ranking college football teams has a long and intriguing history. For decades, the national championship in college football, and/or the opportunity to play for that championship, has been determined not by playoff, but by rankings. Until recently, those rankings had been strictly the accumulated wisdom of opining press-writers and coaches. Embarrassment occurred when, as in 1990, the principal polls disagreed, and selected different national champions.

A large part of the problem was the conference alignments of the bowl games, where championships were determined. Michigan, for instance, won a national championship in 1997 by playing a team in the Rose Bowl that was not even in the top 5, because the Big 10 champ always played the Pac-10 champ in the Rose Bowl, regardless of national ramifications.

In reaction to a growing demand for a more reliable national championship, the NCAA set up in 1998 the *Bowl Championship Series* (BCS), consisting of an alliance among the Sugar Bowl, the Orange Bowl and the Fiesta Bowl, one of which would *always* pit the #1 and #2 teams in the country against each other to play for the national title (unless one or more were a Big 10 or Pac-10 participant). The question was, how to guarantee best that the true #1 and #2 teams were selected...

By the time of the formation of the BCS (and even long before) many had begun to ask the question, can a machine rank the teams more correctly than the pollsters, with less bias than the humans might have? With the advent of easily accessible computer power in the 1990's, many "computer polls" had emerged, and some even appeared in publications as esteemed as the *New York Times*, *USA Today*, and the *Seattle Times*. In fact, by 1998, many of these computer rankings had matured to the point of some reliability and trustworthiness in the eyes of the public.

As such, the BCS included computer rankings as a part of the ranking that would ultimately determine who played for the title each year. Several computer rankings would be averaged together, and that average would be averaged with the "human" polls (with some other factors) to form the best possible ranking of the teams, and hence determine their eligibility to play in the three BCS bowl games. The somewhat controversial method, despite some implausible circumstances, has worked brilliantly in producing 4 undisputed national champions. With the addition of the Rose Bowl (and its Big 10/Pac-10 alliances) in 2000, the likelihood of a split title seems very small.

Given the importance of the computer rankings in determining the national title game, one must consider the simple question, "Are the computers getting it right?" Fans have doubted occasionally when the computer rankings have seemed to favor local teams, disagreed with one another, or simply disagreed with the party line bandied about by pundits.

Making matters worse is that many of the computer ranking systems have appeared to be byzantine "black boxes," with elaborate details and machinery, but insufficient description to be reproduced. For instance, many of the computer methods have claimed to include a home/away bonus, or "conference strength," or a particular weight to opponent's winning percentage, etc., but without a complete, detailed description, we're left just to trust that all that information is being distilled in some correct way.

With no means of thoroughly understanding or verifying the computer rankings, fans have had little reason to reconsider their doubts. A critical feature, therefore, of the Colley Matrix method is that this paper formally defines *exactly* how the rankings are formed, and shows them to be explicitly bias-free. Fans may check the results during the season to verify that the method is truly without bias.

With luck, I will persuade the reader that the Colley Matrix method:

1. has no bias toward conference, tradition, history, etc., (and, hence, has no pre-season poll),
2. is reproducible,
3. uses a minimum of assumptions,
4. uses no *ad hoc* adjustments,
5. nonetheless adjusts for strength of schedule,
6. ignores runaway scores, and
7. produces common sense results.

2. Wins and Losses Only—Keep it Simple

The most important and most fundamental question when designing a computer ranking system is simply where to start. Usually, in science, one poses a hypothesis and checks it against observation to determine its validity, but in the case of ranking college football teams, there really is no observation—there is no ranking that is an absolute truth, against which to check.

As such, one must form the hypothesis (ranking method), and check it against other rankings systems, such as the press polls, other computer rankings, and, perhaps even common sense, and make sure it seems to be doing something right.

Despite the treachery of checking a scientific method against opinion, we proceed, first by contemplating a methodology. The immediate question becomes what input data to use.

Scores are a good start. One may use score differentials, or score ratios, for instance. One may even invent ways of collapsing runaway scores with mathematical functions like taking the arc-tangent of score ratios, or subtracting the square roots of scores. I even experimented with a histogram equalization method for collapsing runaway scores (which, by the way, produced fairly sensible results).

However, even with considerable mathematical skulduggery, reliance on scores generates some dependence on score margin that surfaces in the rankings at some level. Rightly or wrongly, this dependence has induced teams to curry favor in computer rankings by running up the score against lesser opponents. The situation had degraded to the point in 2001 that the BCS committee instructed its computer rankers either to eliminate score dependence altogether or limit score margins to 21 in their codes.

This is a philosophy I applaud, because using wins and losses only

1. eliminates any bias toward conference, history or tradition,
2. eliminates the need to invoke some *ad hoc* means of deflating runaway scores, and
3. eliminates any other *ad hoc* adjustments, such as home/away tweaks.

By focusing on wins and losses only, we're nearly halfway to accomplishing our goals set out in the Introduction.

A very reasonable question may then be, why can't one just use winning percentages, as do the NFL, NBA, NHL and Major League, to determine standings? The answer is simply that in all those cases, each team plays a very representative fraction of the entire league (more games, fewer teams). In college football, with 117 teams and only 11 games each, there is no way for all teams to play a remotely representative sample. The situation demands some attention to "strength of schedule," and it is herein that lies most of the complication and controversy with the college football computer rankings.

The motivation of the Colley Matrix Method, is, therefore, to use something as closely akin to winning percentage as possible, but that nonetheless corrects efficiently for strength of schedule. The following sections describe exactly how this is accomplished.

3.   The Basic Statistic — Laplace's Method

Note to the reader: In the sections to follow, many mathematical equations will be presented. Many derivations and examples will be based upon principles of probability, integral calculus, and linear algebra. Readers comfortable with those subjects should have no problem with the level of the material.

In forming a rating method based only on wins and losses, the most obvious thing to do is to start with simple winning percentages, the choice of the NFL, NBA, NHL and Major League. But simple winning percentages have some incumbent mathematical nuisances. If nothing else, the fact that a team that hasn't played yet has an undefined winning percentage is unsavory; also a 1-0 team has 100% vs. 0% for an 0-1 team: is the 1-0 team really infinitely better than the 0-1 team?

Therefore, instead of using simple winning percentage (n_w/n_{tot}, with obvious notation), I use a method attributable to the famed mathematician Pierre-Simon Laplace, a method introduced to me by my thesis advisor, Professor J. Richard Gott, III.

The adjustment to simple winning percentage is to add 1 in the numerator and 2 in the de-

nominator to form a new statistic,

$$r \equiv \frac{1 + n_w}{2 + n_{tot}}. \tag{1}$$

All teams at the beginning of the season, when no games have been played, have an equal rating of $1/2$. After winning one game, a team has a $2/3$ rating, while a losing team has a $1/3$ rating, i.e., "twice as good," much more sensible than 100% and 0%, or "infinitely better."

The addition of the 1 and the 2 may seem arbitrary, but there is a precise reason for these numbers; namely, we are equating the win/loss rating problem to the problem of locating a marker on a craps table by trial and error shots of dice. What?

This craps table problem is precisely the one Laplace considered. Imagine a craps table (of unit width) with a marker somewhere on it. We cannot see the marker, but when we cast a die, we are told if our die landed to the left or right of the marker. Our task is to make a good guess as to where that marker is, based on the results of our throws. The analogy to football is that we must make a good guess as to a team's true rating based on wins and losses.

At first, our best guess is that the marker is in the middle, at $r = 1/2$. Mathematically, we are assuming a "flat" distribution, meaning that there is equal probability that the marker is anywhere on the table, since we have no information otherwise—that is to say, a uniform Bayesian prior. The average value within such a flat distribution (shown in Fig. 1 at top left) is $1/2$. Computing that explicitly is called finding the expectation value (or weighted mean, or center of mass). If the probability distribution function of rating $\hat{r}$ is $f(\hat{r})$, then in the case of no games played (no dice thrown), $f(\hat{r}) = 1$, and the expectation value of $\hat{r}$ is

$$r = \langle \hat{r} \rangle = \frac{\int_{r_0}^{r_1} \hat{r} \cdot f(\hat{r}) d\hat{r}}{\int_{r_0}^{r_1} f(\hat{r}) d\hat{r}} = \frac{\int_0^1 \hat{r} d\hat{r}}{\int_0^1 d\hat{r}} = \frac{(\hat{r}^2/2)|_0^1}{\hat{r}|_0^1} = 1/2. \tag{2}$$

Now, if the first die is cast to the left of the divider, the probability density that the marker is at the left wall ($\hat{r} = 0$) has to be zero — you can't throw a die to the left of the left wall. From zero at the left wall, the the probability density must increase to the right. That increase is just linear, because the probability density is just the available space to the left of the marker where your die could have landed; the farther you go to the right, the proportionally more available space there is to the left (see Fig. 1, top right).

The analogy with football here is clear. If you've beaten one team, you cannot be the worst team after one game, and the number of available teams to be worse than yours increases proportionally to your goodness, your rating $\hat{r}$.

The statistical expectation value of the location of the marker (rating of your team) for the one

left throw (one win) case is therefore

$$r = \frac{\int_0^1 \hat{r}^2 d\hat{r}}{\int_0^1 \hat{r} d\hat{r}} = \frac{(\hat{r}^3/3)|_0^1}{(\hat{r}^2/2)|_0^1} = 2/3. \tag{3}$$

If we throw another die to the left, we have not a linear behavior in probability, but parabolic, since the probability densities simply multiply,

$$r = \frac{\int_0^1 \hat{r}^3 d\hat{r}}{\int_0^1 \hat{r}^2 d\hat{r}} = \frac{(\hat{r}^4/4)|_0^1}{(\hat{r}^3/3)|_0^1} = 3/4, \tag{4}$$

as shown at the bottom left in Fig. 1.

However, when a die is thrown to the right, we know that the probability at the right wall has to go to zero, and a term growing linearly from right to left is introduced, $(1 - \hat{r})$ (in exact analogy to the left-thrown die). Therefore, if we have thrown one die to the left and one to the right, we have

$$r = \frac{\int_0^1 (1 - \hat{r})\hat{r}^2 d\hat{r}}{\int_0^1 (1 - \hat{r})\hat{r} d\hat{r}} = \frac{(\hat{r}^3/3 - \hat{r}^4/4)|_0^1}{(\hat{r}^2/2 - \hat{r}^3/3)|_0^1} = 1/2, \tag{5}$$

as shown at the bottom right in Fig. 1.

In general, for n_w wins (left throws of the die) and n_ℓ losses (right throws of the die), the formula is

$$r = \frac{\int_0^1 (1 - \hat{r})^{n_\ell} \hat{r}^{n_w} \hat{r} d\hat{r}}{\int_0^1 (1 - \hat{r})^{n_\ell} \hat{r}^{n_w} d\hat{r}} = \frac{1 + n_w}{2 + n_\ell + n_w}, \tag{6}$$

which recovers equation (1). It is an interesting exercise to check a few more examples.

4. Strength of Schedule

The simple statistic developed in the last section would suffice to produce a ranking if we were confident that all teams had played a schedule of similar strength, or for instance a round-robin tournament. While a round-robin with 117 teams would require 6786 games, Division I-A teams play typically a tenth that, so there is absolutely no assurance that the quality of opponents from team to team is close to the same. Contrast this with the NFL, or especially the Major League, where each team plays a very healthy sample of the entire league during the regular season.

This problem is complicated by the addition of still more teams in the form of non-I-A opponents. If one were to use those games as input, he would have to form ratings of all the I-AA teams, which would require ratings of teams in still lesser divisions, since many I-AA teams play

such opponents. Forming sensible ratings which relate Florida State to Emory & Henry is extremely difficult and is frankly beyond the scope of this method. The reason is that my method, in its simplicity, relies on some interconnectedness between opponents, which simply does not exist between a given NAIA squad and a given Division I-A squad—there's barely enough interconnectedness among the I-A teams themselves! Most other computer rankings within the BCS system do endeavor to compute such ratings, and in my opinion, do nearly as good a job as is possible at making sense of such disparate and competitively disconnected teams. To preserve simplicity and total objectivity (no *ad hoc* division adjustment, etc.), my rating system must ignore all games against non-I-A opponents. Therefore, *padding the schedule with I-AA teams contributes absolutely nothing to a team's rating.*

We may then proceed with mathematical adjustments for strength of schedule within Division I-A itself.

The number of wins in equation (1) may be divided into $n_{w,i} = (n_{w,i} - n_{\ell,i})/2 + n_{tot,i}/2$ (which the reader can check). Recognizing that the second term may be written as $\sum^{n_{tot,i}} 1/2$ allows one to identify the sum as that of the ratings of team i's opponents, if those opponents are all random ($r = 1/2$) teams. Instead, then, of using $r = 1/2$ for all opponents, we now use their actual ratings, which gives an obvious correction to $n_{w,i}$.

$$n_{w,i}^{eff} = (n_{w,i} - n_{\ell,i})/2 + \sum_{j=1}^{n_{tot,i}} r_j^i, \tag{7}$$

where r_j^i is the rating of the jth opponent of team i. *The second term (the summation) in equation (7) is the adjustment for strength of schedule.*

Now, the rub. When the teams are not random, but ones which have played other teams, which may or may not have played some teams in common with the first team, etc., how does one possibly figure out simultaneously all the r_j^i's which are inputs to the r_i's, which are themselves r_j^i's for other r_i's, etc.?

5.   The Iterative Scheme

The most obvious way to solve such a problem is a technique called "iteration," which is employed by several of the other computer ranking methods. The way it works is one first computes the ratings, as if all the opponents were random ($r = 1/2$) teams, using equation (1). Next, each team's strength of schedule is computed according to its opponents' ratings, using equation (7). The ratings are re-computed with the new schedule strengths, and then strengths of schedule are re-computed from the new ratings. With luck, the changes to the ratings get smaller and smaller

with each step of these calculations, and after a time, when the changes are negligibly small for any team's rating (a part in a million, say), one calls the list of ratings, at that point, final.

Here is a very simple example of the iterative technique, after only one week of play, where a team that has played is either 0-1 against a 1-0 team, or vice versa. Before any iterations, the basic Laplace statistic from equation (1) is computed for each team. Letting $r_{W,0}$ be the initial rating of a winning team, and $r_{L,0}$ be the initial rating of the losing team, one finds that equation (1) initially gives

$$\text{Initial ratings}$$
$$r_{W,0} = (1+1)/(2+1) = 2/3 \approx 0.6667 \tag{8}$$
$$r_{L,0} = (1+0)/(2+1) = 1/3 \approx 0.3333.$$

The first adjustment for strength of schedule (there's been only one game, so schedule strength is just the rating of the one opponent) is made by computing n_w^{eff} for each team, using equation (7):

$$\text{First Correction}$$
$$n_{w,W,1}^{eff} = (1-0)/2 + 1/3 = 5/6 \tag{9}$$
$$n_{w,L,1}^{eff} = (0-1)/2 + 2/3 = 1/6.$$

Because the 1-0 team beat a 0-1 team, worse than an average team, the 1-0 team is punished, and given only $5/6$ of a win, whereas the losing team lost to a 1-0 team, better than an average team, and is rewarded by suffering only $5/6$ of a loss. One can see how the method explicitly gives to one team only by taking from another.

The next step is to re-compute the ratings, given the new n_w^{eff} values. Plugging back into equation (1) yields:

$$\text{Ratings After First Iteration}$$
$$r_{W,1} = (1+5/6)/(2+1) = 11/18 \approx 0.6111 \tag{10}$$
$$r_{L,1} = (1+1/6)/(2+1) = 7/18 \approx 0.3889.$$

Let's look at just one more iteration.

$$\text{Second Correction}$$
$$n_{w,W,2}^{eff} = (1-0)/2 + 7/18 = 8/9 \tag{11}$$
$$n_{w,L,2}^{eff} = (0-1)/2 + 11/18 = 1/9.$$

$$\text{Ratings After Second Iteration}$$
$$r_{W,2} = (1+8/9)/(2+1) = 17/27 \approx 0.6296 \tag{12}$$
$$r_{L,2} = (1+1/9)/(2+1) = 10/27 \approx 0.3704.$$

If one examines the ratings of the winning team after the zeroth, first and second iterations, one finds that the values $r_{W,\{0,1,2\}} \approx \{0.6667, 0.6111, 0.6296\}$, show first a correction down, then

a correction up, by a lesser amount. Corrections that alternate in sign, and shrink in magnitude are hallmarks of *convergence*, meaning that with each iteration, the scheme is closer to finding a final, consistent value. Table 1 shows how these numbers converge to a part in a million after 11 iterations.

In fact, one can demonstrate that the final ratings in this simple case are explicitly the sums of converging series (compare to Table 1),

$$
\begin{aligned}
r_L &= \tfrac{1}{2}\left[1 - \tfrac{1}{3} + \tfrac{1}{9} - \tfrac{1}{27} + \cdots\right] \\
&= \tfrac{1}{2}\sum_{n=0}^{\infty}(-1/3)^n \\
&= \tfrac{1}{2}\cdot\tfrac{1}{1+1/3} = \tfrac{3}{8},
\end{aligned}
\tag{13}
$$

where the last line is the standard formula for the sum of a geometric series. In this simple case, the iterative method converges rapidly and stably, as a classic alternating geometric series.

Also note that the results converge to an average rating of 1/2, which is the same average as if there had been no game played at all; average rating has been conserved.

The ratings may converge nicely, but how can one know that these are the *right* answers? Furthermore, is the method extensible to the prodigiously more complicated case of 117 teams having played 11 or 12 games each?

6. The Colley Matrix Method

The previous section showed how an iterative correction for strength of schedule could provide consistent results that make intuitive sense for the simple one game case, but left us with the question of how do we know that the result is really right?

Let us return, then, to the example of the two teams 1-0, and 0-1 after their first game. Referring to equations (1) and (7), we have

$$
\begin{aligned}
r_W &= \tfrac{1+1/2+r_L}{2+1} \\
r_L &= \tfrac{1-1/2+r_W}{2+1}.
\end{aligned}
\tag{14}
$$

A simple rearrangement gives

$$
\begin{aligned}
3r_W - r_L &= 3/2 \\
-r_W + 3r_L &= 1/2,
\end{aligned}
\tag{15}
$$

a simple two-variable linear system. Plugging in the results from the iterative technique (Table 1), one discovers that indeed $r_W = 5/8$ and $r_L = 3/8$ work exactly.

This exercise illustrates that linear methods can be used for two teams, but begs the question, can the ratings of many teams, after many games, be computed by simple linear methods?

6.1. The Matrix Solution

Returning to equations (1) and (7), using the same definitions for r_i and r_j^i, one finds that equations (1) and (7) can be rearranged in the form:

$$(2 + n_{tot,i})r_i - \sum_{j=1}^{n_{tot,i}} r_j^i = 1 + (n_{w,i} - n_{\ell,i})/2, \tag{16}$$

which is a system of N linear equations with N variables.

It is convenient at this point to switch to matrix form by rewriting equation (16) as follows,

$$C\vec{r} = \vec{b}, \tag{17}$$

where $\vec{r}$ is a column-vector of all the ratings r_i, and $\vec{b}$ is a column-vector of the right-hand-side of equation (16):

$$b_i = 1 + (n_{w,i} - n_{\ell,i})/2. \tag{18}$$

The matrix C is just slightly more complicated. The ith row of matrix C has as its ith entry $2 + n_{tot,i}$, and a negative entry of the number of games played against each opponent j. In other words,

$$\begin{aligned} c_{ii} &= 2 + n_{tot,i} \\ c_{ij} &= -n_{j,i}, \end{aligned} \tag{19}$$

where $n_{j,i}$ is the number of times team i has played team j.

The matrix C is defined as the **Colley Matrix**. *Solving equations (17)–(19) is the method for rating the teams.* In practice, the matrix equation is solved in double precision by Cholesky decomposition and back-substitution (faster and more stable than Gauss-Jordan inversion, for instance [e.g., Press et al. 1992]). The Cholesky method is available, because the matrices are not only (obviously) symmetric and real, but are also positive definite, which will be discussed in the next section.

6.2. Equivalence of the Matrix and Iterative Methods

The matrix method has been shown to agree with the iterative method in the simple one game case. The question is whether the agreement extends to more complex situations. In a word, "yes," but why?

There is no dazzlingly elegant answer here. If the iterative scheme converges, then equations (1) and (7) are more nearly mutually satisfied with every iteration; otherwise the ratings would have to diverge at some point. When the ratings have converged, and iterating no longer introduces any changes to the ratings, the equations themselves have become simultaneously satisfied—the convergent ratings values have solved equations (1) and (7). Because those equations are identically the same ones solved by the matrix method, the matrix and iterative solutions must be identical as long as the iterative method remains convergent.

The question then becomes shifted to one of the convergence itself, which has been discussed only by example to this point. The convergence is due principally to $n + 2$ denominator in equation (1). I shall not give a rigorous proof as to *exactly* why this is so, but rather motivate the idea in a less rigorous way.

The initial ratings are the final ratings plus some error.

$$\vec{r} = \vec{\rho} + \vec{\delta}, \tag{20}$$

where $\vec{\rho}$ is the vector of the true (final) ratings, and $\vec{\delta}$ the vector of errors. Let us consider the simple case where $\vec{\delta}$ has only one non-zero component, say $\delta_1 \neq 0$. Assuming a round-robin schedule (the slowest to converge), the iterations would proceed as in Table 2. The convergence in Table 2 is slow, with the errors decreasing by $\sim n/(n + 2)$ in each iteration. In practice, the college football schedule is not round-robin outside of each conference, so the convergence factor is more like $\sim (n - 2)/(n + 2) \approx 10/14 \approx 0.71$, so convergence to a part in 10^7 occurs in about 48 iterations. In the 2000 season, for instance, the number of iterations required before median ratings correction fell below 10^7 was 60, so this very simple estimate is correct to 20% for a typical case.

Of course, in reality there are errors in more than one of the ratings, but, because the equations are linear, the principle of superposition applies, and the above calculation changes very little.

While the preceding is no proof that the scheme is always convergent, the round-robin case is the slowest to converge, and even in that case, we have shown that any error in a single rating does vanish over time, and superposition extends that to errors in multiple ratings. In the sparser case of an actual college football season the convergence can be slightly faster for some teams.

It should be noted that if it weren't for the $+2$ in the denominator of equation (1), the convergence would not occur, since, in the round-robin case, the error decrement would be by a factor of $\sim n/n = 1$, so if this method were based strictly on winning percentages, rather than Laplace's formula, it would fail.

The iterative scheme produces convergent ratings, which upon convergence simultaneously satisfy equations (1) and (7), which are exactly those that the matrix method solves; therefore, the iterative and matrix solutions must be equivalent (and, in practice, are equivalent).

6.3.   Examples of the Colley Matrix Method

We now consider two examples to illustrate the matrix method in action. First, let us return to our friend, the simple two team, one game case. There, we discovered that the ratings could be determined from the linear system

$$
\begin{aligned}
3r_W - r_L &= 3/2 \\
-r_W + 3r_L &= 1/2,
\end{aligned}
\tag{21}
$$

Rewriting this in matrix form,

$$
\begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} r_W \\ r_L \end{bmatrix} = \begin{bmatrix} 3/2 \\ 1/2 \end{bmatrix}
\tag{22}
$$

we recognize that r_W and r_L can be determined by simply inverting the matrix and multiplying by the solution vector on the right hand side.

$$
\begin{bmatrix} 3/8 & 1/8 \\ 1/8 & 3/8 \end{bmatrix} \begin{bmatrix} 3/2 \\ 1/2 \end{bmatrix} = \begin{bmatrix} 5/8 \\ 3/8 \end{bmatrix},
\tag{23}
$$

verifying the iterative and linear solutions.

But let us now consider a more complex example, a five team field, teams a–e. Their records against each other are shown thus (the initial and final ratings are listed for reference, the latter of which will be solved for forthwith):

team	a	b	c	d	e	record	initial rating	final rating
a	∘	∘	W	L	L	1-2	0.4	0.4130
b	∘	∘	L	∘	W	1-1	0.5	0.5217
c	L	W	∘	W	L	2-2	0.5	0.5000
d	W	∘	L	∘	∘	1-1	0.5	0.4783
e	W	L	W	∘	∘	2-1	0.6	0.5870.

The matrix equation, according to equations (17)–(19) is written

$$
\begin{bmatrix}
5 & 0 & -1 & -1 & -1 \\
0 & 4 & -1 & 0 & -1 \\
-1 & -1 & 6 & -1 & -1 \\
-1 & 0 & -1 & 4 & 0 \\
-1 & -1 & -1 & 0 & 5
\end{bmatrix}
\begin{bmatrix} r_a \\ r_b \\ r_c \\ r_d \\ r_e \end{bmatrix}
=
\begin{bmatrix} 1/2 \\ 1 \\ 1 \\ 1 \\ 3/2 \end{bmatrix},
\tag{24}
$$

Solving the matrix equation yields sensible results:

$$\vec{r} = \{19, 24, 23, 22, 27\}/46 \approx \{0.413, 0.522, 0.500, 0.478, 0.587\} \tag{25}$$

As with the two team case, the ratings average to exactly 1/2 (23/46). Notice that team *b*, having played a 2-1 team and a 2-2 team, is rated higher than team *d*, having played a 1-2 team and a 2-2 team, despite identical 1-1 records: an example of how strength of schedule comes into play. Finally, there is consistency in that team *c* has a rating of exactly 1/2, because that team is 2-2 against teams whose ratings average to exactly 1/2.

7.   Comments on the Colley Matrix Method

There has been discussion of the fact that the matrix method (and the iterative method) conserves an average ranking of 1/2. The reason I emphasize that point is that, as such, the ratings herein require no renormalization. All teams started out with a rating of 1/2, and only by exchange of rating points with other teams does one team's rating change. The first subsection below shows why the rating scheme explicitly conserves total ratings points. The subsection which follows establishes that the matrix in equation (17) is indeed positive definite, which allows for quick and stable solution of the matrix equation. The last subsection shows that the matrix C is singular if winning percentages are used in place of Laplace's formula, and thus, the method cannot be used with straight winning percentages.

7.1.   Conservation of Average Rating

Why does the matrix solution always preserve the average rating of 1/2? We can tackle this problem by examining the construction of the matrix C. From the definition of the matrix C in equation (19), it follows that the matrix can be represented as:

$$C = 2I + \sum_{k}^{\text{all games}} G^k, \tag{26}$$

where I is the identity matrix, and G^k is a matrix operator added for each game k. G^k always has the form that $G_{ii}^k = G_{jj}^k = 1$, and $G_{ij}^k = G_{ji}^k = -1$, with all other entries 0, like this:

$$
G^k = \begin{bmatrix}
 & \vdots & & \vdots & \\
\cdots & 1 & \cdots & -1 & \cdots \\
 & \vdots & & \vdots & \\
\cdots & -1 & \cdots & 1 & \cdots \\
 & \vdots & & \vdots &
\end{bmatrix}. \tag{27}
$$

Carrying out the multiplication $\vec{r}' = G^k\vec{r}$, we find that the ith and j^{th} entries in $\vec{r}'$ are $r_i - r_j$, and $r_j - r_i$, respectively, while all other entries in $\vec{r}'$ are obviously zeroes. The sum of all the $\vec{r}'$ values, is, therefore, zero, no matter what the values of $\vec{r}$. Hence,

$$
\sum_{i=1}^{N} (C\vec{r})_i = \sum_{i=1}^{N} (2I\vec{r})_i = 2\sum_{i=1}^{N} r_i. \tag{28}
$$

What about the other side of equation (17), $\vec{b}$? The definition of b_i from equation (18) is $b_i = 1 + (n_{w,i} - n_{\ell,i})/2$. It is easy to see that the total of the b_i's must be N, since each win by one team must be offset by a loss by that team's opponent, so that the total number of wins must equal the total number of losses, and therefore, $\sum b_i = N$.

So, summing both sides of equation (17), we have

$$
\begin{aligned}
\sum_i (C\vec{r})_i &= \sum_i b_i \\
2\sum_i r_i &= N;
\end{aligned} \tag{29}
$$

$$
\Rightarrow \frac{\sum r_i}{N} = \frac{1}{2}, \tag{30}
$$

i.e., the average value of r is exactly one-half.

7.2. The Colley Matrix is Positive Definite

In order to use Cholesky decomposition and back substitution to solve the matrix equation (17), the matrix C must be positive definite, such that, for any non-trivial vector $\vec{v}$, this inequality holds

$$
\vec{v}^T (C\vec{v}) > 0. \tag{31}
$$

Recalling our separation of C into $2I + \sum_k G^k$, and noting that matrix multiplication is distributive, $(A + B)\vec{v} = A\vec{v} + B\vec{v}$, we can examine the inequality piece-wise. Obviously the matrix $2I$ is

positive definite, which leaves the G^k's. In the subsection above, we discovered that the multiplication $G\vec{r}$ yielded zeroes in all entries, except the ith and j^{th} which contained $r_i - r_j$ and $r_j - r_i$, respectively. The product $\vec{r}^T(C\vec{r})$ is thus computed as

$$\vec{r}^T(G^k\vec{r}) = r_i(r_i - r_j) + r_j(r_j - r_i) = (r_i - r_j)^2 \geq 0. \tag{32}$$

Since $(r_i - r_j)^2 \geq 0$, and $\vec{r}^T(2I\vec{r}) > 0$, the matrix C must be positive definite.

7.3. Singularity of C for Straight Winning Percentages

We have seen that the iterative method would fail if straight winning percentages were used (i.e. if one removed the $+1$ and $+2$ from the numerator and denominator in equation [1]). In performing the same exercise with the matrix method, equation (19) would change C into a singular matrix!

For the one game case, the result is obvious.

$$C = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \tag{33}$$

obviously singular. In the general case, it's easy to see from equation (19) that if one removes the addend of 2 from c_{ii}, the total of the c_{ij}'s for any row or any column is zero. Therefore, in performing the legal row operation of adding all the rows, one has produced a row of all zeros; hence the matrix is singular. The matrix method, like the iterative method, cannot work with simple winning percentages.

8. Performance

With the mathematics of the ranking method on a firm foundation, the next question is how well does the method perform. Answering that question is difficult because we do not actually know the "truth." We simply do not know in any precise way how good Western Michigan is relative to Hawaii; therefore there is really nothing to check our results against.

The best we can do is compare our results to the venerable press polls, just as a mutual sanity check. Before I even begin, I should caution that the press polls are artificially correlated simply because the coaches are well aware of the AP poll, and the press writers are well aware of the Coaches' poll. Nonetheless we shall proceed.

Table 3 compares the final Colley rankings against the final AP rankings for 1998–2002 and against the Coaches' Poll rankings for 1999–2002. The first thing to notice is that the national

champion is agreed upon every year in all three ranking systems, which is ultimately the most important test for our purposes. Just scanning down the lists, the rankings are usually quite consistent within a few places, with an occasional outlier.

To quantify the agreement, I have found it more useful to think in terms of ranking ratios, or percentage differences, rather than simple arithmetic ranking differences. I have previously used the median absolute difference as a figure of merit, but have concluded this statistic poorly describes the behavior of the ranking comparisons. To show that, I have plotted in Fig. 2 the arithmetic differences *(top)* and ratios *(bottom)* of my rankings vs. the press rankings for 1999–2002 (all lumped together). In the plots, the histograms for the AP and Coaches' rankings are over-plotted (they're very similar).

Throwing all six groups together (three years, AP, Coaches), one can compute the direct average and variance of the distributions, which are listed as "avg" and "s" on the plots (yes, s is the square-root of the variance, and has a $\sqrt{n/(n-1)}$ factor). The corresponding normal curve is plotted as a dotted line. Another way of finding the mean and variance of a distribution is to fit the integral of the normal curve—that's $P(x)$ for you calculator statisticians, or the error function with $\sqrt{2}$'s in the right places for you math pedants—to the cumulative distribution. I have plotted those resulting normal curves as solid lines at top and bottom. If a distribution is normal, these two methods should be nearly identical. Obviously at top, the two curves are not so identical, but at bottom things seem much better.

Of course there are dozens of formal ways to check "Gaussianity" of a distribution, but I don't want to belabor the point. I just want to illustrate that the ratios are much better behaved than the arithmetic differences. What does this mean? It means that it's much more accurate to say my rankings disagree with the press rankings by a typical percentage, rather than by a typical number.

To compute that typical percentage, one may average the absolute differences of the logs of the rankings, as such,

$$\eta = \exp\left(\frac{1}{25} \sum_{i=1}^{25} |\log j_C(\text{team}_i) - \log i| \right), \tag{34}$$

where i is the press ranking (either AP or coaches), and $j_C(\text{team}_i)$ is the Colley ranking of that team. It's hard to come up with a good name for this statistic, η; perhaps "mean absolute ratio." Anyway, I list that statistic at the bottom of Table 3 for each of the 4 years of the BCS. The values are typically about 1.25, which means that my rankings agree with the press rankings within about 25% in either direction, so you might have an error of 1 ranking place at around #5, but about 5 places by #20. Inspecting the columns of Table 3 shows this to be quite a good description of the relative rankings.

Is that a *good* agreement?

Who knows, really? But to me (at least) the agreement is surprisingly good:

- The press polls started with a pre-season poll, with all the pre-conceived notions of history and tradition such an endeavor demands, then week by week allowed their opinions and judgments to migrate, being duly impressed or disappointed in the styles of winning and losing by certain teams, being more concerned about recent games than earlier ones, perhaps mentally weighting games seen on television as more important, perhaps having biases (good or bad) toward local schools one sees more often... *ad nauseam*.

- My computer rankings started with nothing, literally no information, but then, given only wins and losses, generated a ranking with pure algebra.

That two such processes produce even remotely consistent results is, frankly, remarkable to me.

I hope in this section we can agree to have learned, despite a lack of "truth" data, comparison of the press polls and my rankings shows both that the press and coaches must not be too loony, and that the Colley Matrix system yields common sense results.

9. Conclusions

Colley's Bias Free College Football Ranking Method, based on solution of the Colley Matrix, has been developed with several salient features, desirable in any computer poll that claims to be unbiased.

1. It has no bias toward conference, tradition, history, or prognostication.
2. It is reproducible; one can check the results.
3. It uses a minimum of assumptions.
4. It uses no *ad hoc* adjustments.
5. It nonetheless adjusts for strength of schedule.
6. It ignores runaway scores.
7. It produces common sense results that compare well to the press polls.

This information, the weekly poll updates, as well as useful college football links may be found on the Internet home for Colley's Rankings:

`http://www.colleyrankings.com/.`

discussions on the subject of rankings. All programming for this method was done by WNC in IDL, FORTRAN, C, C++, Perl and shell script in the Solaris Unix and Linux environments.

REFERENCES

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P., 1992, *Numerical Recipes in FORTRAN*, Cambridge University Press, Cambridge, UK, pp. 89-91

Convergence of Ratings via Iteration

Iteration	Winning Team's Rating	Losing Team's Rating
0	0.666667	0.333333
1	0.611111	0.388889
2	0.629630	0.370370
3	0.623457	0.376543
4	0.625514	0.374486
5	0.624829	0.375171
6	0.625057	0.374943
7	0.624981	0.375019
8	0.625006	0.374994
9	0.624998	0.375002
10	0.625001	0.374999
11	0.625000	0.375000

Table 1: Convergence of ratings to final, stable values, after 11 iterations, for the simple two team, one game case. The initial ratings are 2/3 for the winner and 1/3 for the loser, before any adjustment for schedule strength. Moving down the table are successive adjustments for strength of schedule. Because the winning team beat a below average (0-1) team, while the losing team lost to an above average (1-0) team, the final ratings are lower for the winning team, and greater for the losing team than were the initial ratings.

Iterative Convergence of Ratings in a Round-Robin

iter.	team 1	other teams
1.	$r \to \rho + \delta_1$ $n_w^{eff} = \nu_w^{eff}$	$r = \rho$ $n_w^{eff} \to \nu_w^{eff} + \delta_1$
2.	$r = \rho$ $n_w^{eff} = \nu_w^{eff} + n\delta_1/(2+n)$	$r = \rho + \delta_1/(2+n)$ $n_w^{eff} = \nu_w^{eff} + (n-1)\delta_1/(2+n)$
3.	$r = \rho + n\delta_1/(2+n)^2$ $n_w^{eff} = \nu_w^{eff} + n(n-1)\delta_1/(2+n)^2$	$r = \rho + (n-1)\delta_1/(2+n)^2$ $n_w^{eff} = \nu_w^{eff} + [(n-1)^2 + n]\delta_1/(2+n)^2$
4.	$r = \rho + n(n-1)\delta_1/(2+n)^3$ $\vdots$	$r = \rho + [(n-1)^2 + n]\delta_1/(2+n)^3$ $\vdots$

Table 2: Iterative convergence in a round-robin. Ratings r and effective wins n_w^{eff} are computed from equations (1) and (7) in each iteration. Starting with the correct (final) values, ρ and ν_w^{eff}, an error δ_1 is added to team 1's rating. Column 1 gives the propagation of that error in team 1 through 4 iterations; Column 2 does the same for all other teams (whose errors will be equivalent). As one moves down the table, the errors shrink (slowly).

Comparison of Final Rankings with Press Polls

	Colley Ranking for Teams with Given Press Rank								
press	**1998**	**1999**		**2000**		**2001**		**2002**	
rank	**AP**	**AP**	**Coaches**	**AP**	**Coaches**	**AP**	**Coaches**	**AP**	**Coaches**
1	1	1	1	1	1	1	1	1	1
2	2	5	2	3	3	3	3	3	3
3	3	2	5	2	2	4	4	2	2
4	6	11	11	5	6	2	2	4	4
5	8	3	3	6	5	6	6	5	5
6	4	7	7	4	4	10	10	6	11
7	10	4	4	7	8	8	5	11	6
8	5	6	6	8	11	5	8	8	8
9	11	10	10	11	7	7	7	7	7
10	9	8	8	9	13	11	13	12	12
11	7	9	9	13	9	13	11	10	14
12	13	13	12	22	22	9	9	14	17
13	12	12	15	19	19	15	15	13	13
14	17	15	13	17	12	12	12	19	20
15	16	14	14	10	17	16	16	17	18
16	22	24	24	12	10	14	17	18	19
17	14	28	23	15	30	17	14	9	9
18	19	23	17	20	23	31	31	20	24
19	20	17	28	29	15	18	18	24	32
20	23	18	22	30	20	20	20	21	23
21	24	21	18	23	29	30	21	15	21
22	15	27	27	28	27	24	28	25	28
23	27	22	21	14	18	28	30	28	15
24	21	16	29	27	14	33	19	32	26
25	26	26	16	18	32	19	24	23	25
η									
Colley vs. poll	1.224	1.309	1.281	1.287	1.331	1.262	1.232	1.200	1.253
AP vs. Coaches	n/a	1.071		1.098		1.037		1.082	

Table 3: Comparison of final rankings to AP Poll for 1998–2002, and to the Coaches' Poll for 1999–2002. At bottom is a statistic η, described in the text. Essentially, it is the typical ratio of the Colley ranking to the poll ranking, or vice versa, so that the larger of the two always in the numerator, (specifically η is the exponent of the mean of the absolute values of the logs of the ratios), so $\eta = 1.25$ means the rankings would differ by typically one place at #4, and 5 places at #20.
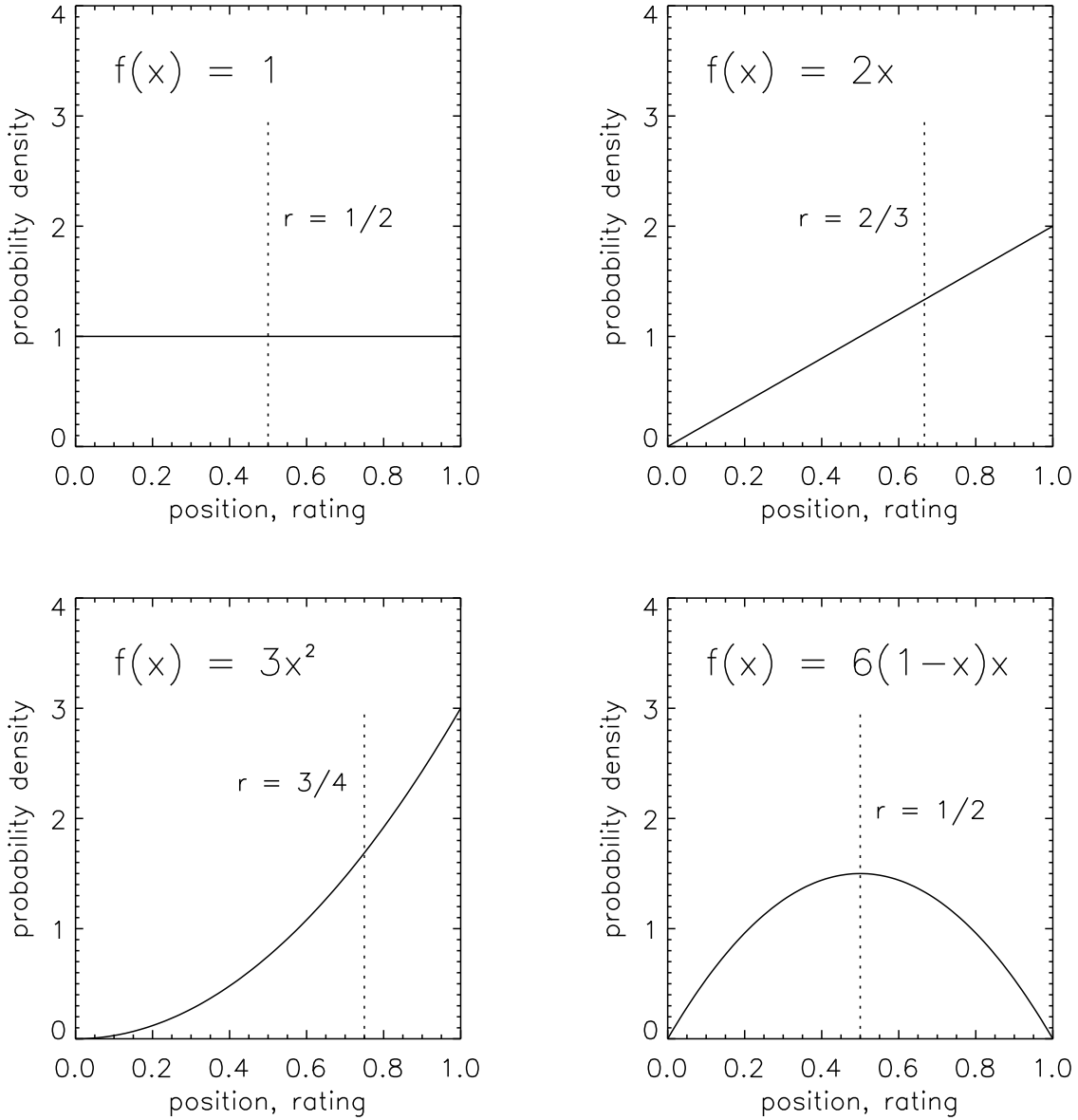
Fig. 1.— Probability distribution functions for the Laplace dice problem, analogous to rating by wins and losses. At top left is the initial condition {no dies thrown; no games played}. At top right is {one die left; one game won}: the probability density must be zero at the left. At bottom left is {two dies left; two games won}: the probability densities multiply. At bottom right is {one die left, one die right; one game won, one game lost}: the probability density must be zero at the left and at the right. The functions here have been normalized to have an integral of one, which is irrelevant in section 3 of the text, because the normalizations explicitly divide out.
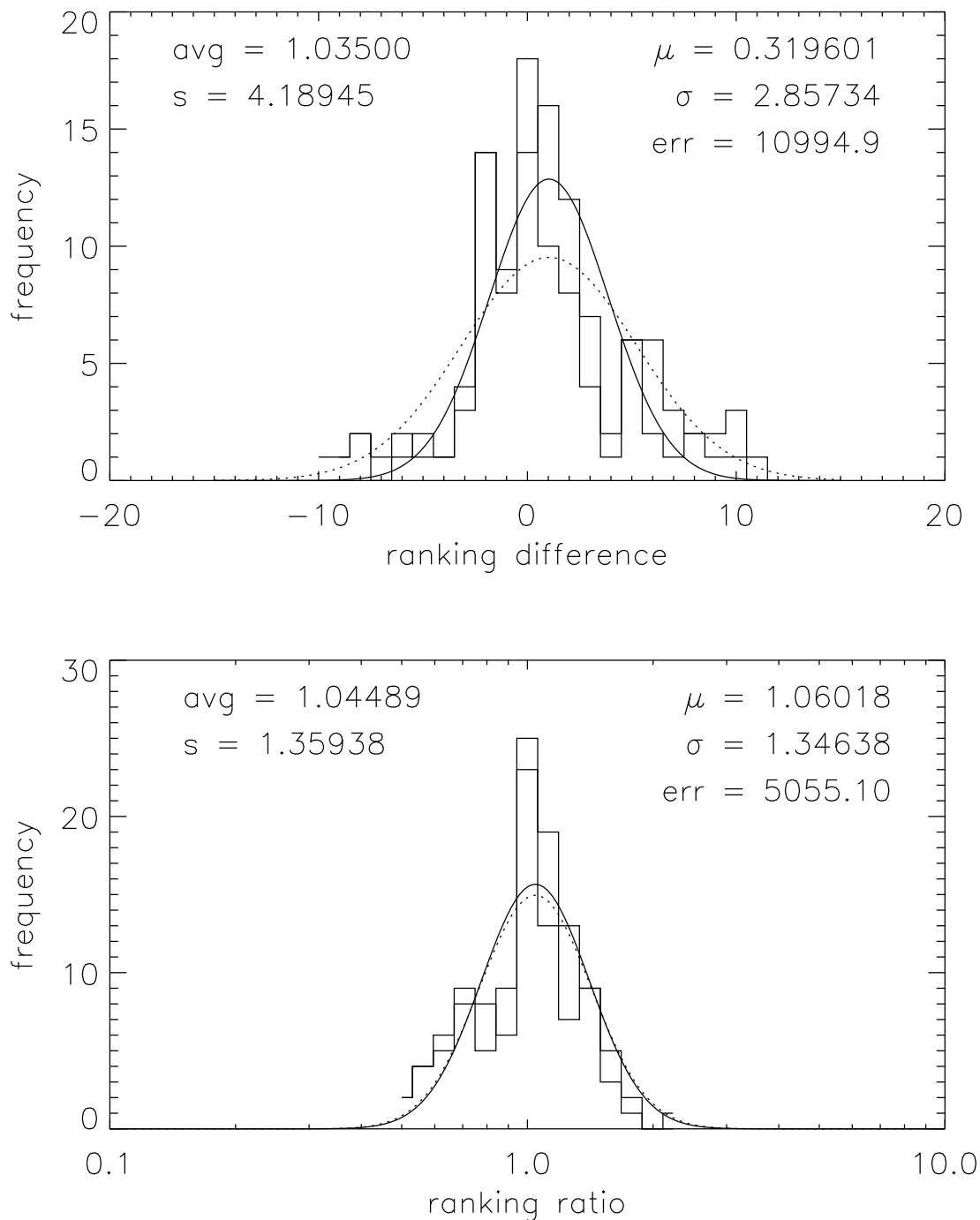
Fig. 2.— Two different ways for comparing the Colley Rankings with the Press Polls. *(at top)* Arithmetic differences, (Colley − press), between the final rankings for 1999–2002 for both the AP and Coaches' polls. Over-plotted are the normal curves from direct measurement of mean (= *avg*) and standard deviation (= *s*), and from fitting for the mean (= μ) and standard deviation (= σ). *(at bottom)* Same plots, but for ratios (Colley ÷ press) in logarithmic space.