Paired Comparisons with Ties: Modeling Game Outcomes in Chess

Daniel Shawul[1] and Rémi Coulom[2]

[1] University of Western Ontario, Canada
[2] Université Lille 3, INRIA, CNRS, France

Abstract. Bayesian rating of chess players requires a statistical model of the probabilities of a win, a draw, and a loss as a function of the rating difference between opponents. Some models are used in popular rating systems, but they were chosen rather arbitrarily, and it was not clear which fits the data best. In this paper, the goodness of fit of the Glenn-David (TrueSkill), Rao-Kupper (BayesElo), and Davidson models were measured for various databases of games between computers. Results demonstrate that the Davidson model fits the data best. The Davidson model features a draw distribution with longer tails, and, unlike the other models, makes two draws equivalent to one win and one loss. The Davidson model had not been used in any popular rating system, and the results presented in this paper will lead to a new improved version of BayesElo.

1 Introduction

[1], [2], [3], [4].

2 Models for Paired Comparisons with Ties

Linear models (not multidimensional, ...). Advantage of playing first may be handled the same way for all, not mentioned here.

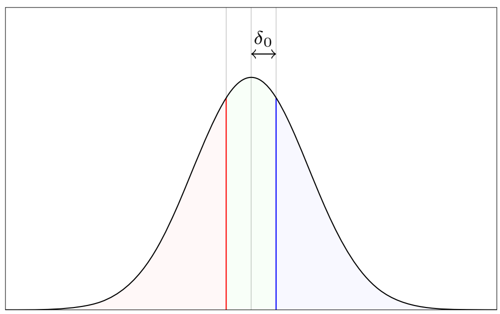
The Glenn-David model is used in TrueSkill [5], a rating system developed at Microsoft, and used in their Xbox game servers. The Rao-Kupper model is the basis of BayesElo [6], a freeware tool popular in the chess programming community. The Davidson model is used in Edo ratings [7].

2.1 The Glenn-David Model

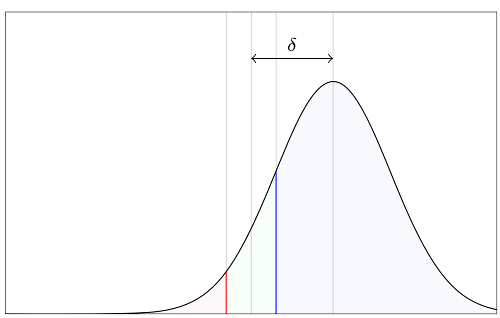
[8]

$$\begin{aligned} P(W|\delta) &= \Phi(+\delta - \delta_0) \ ; \\ P(L|\delta) &= \Phi(-\delta - \delta_0) \ ; \\ P(D|\delta) &= 1 - P(W|\delta) - P(L|\delta) \ . \end{aligned}$$

[9], Fig. 1, [10], [11]



(a) $\delta = 0$



(b) $\delta > 0$

Fig. 1. Principle of the Glenn-David model: the performance of a player in a game is assumed to be a random variable with a normal distribution. The difference between the performances of two opponents, plotted on these figures, is also normally distributed. A draw occurs when the performances of the opponents are within δ_0 of each other. The areas of the three regions represent the probabilities of a loss, a draw, and a win.

2.2 The Rao-Kupper Model

The Rao-Kupper model [12] is similar to the Glenn-David model, except that Φ is replaced by the logistic function:

$$f(x) = \frac{1}{1+e^{-x}} \ .$$

Outcome probabilities become

$$\begin{aligned} P(W|\delta) &= f(+\delta - \delta_0) \ ; \\ P(L|\delta) &= f(-\delta - \delta_0) \ ; \\ P(D|\delta) &= 1 - P(W|\delta) - P(L|\delta) \\ &= \big(e^{2\delta_0} - 1\big)P(W|\delta)P(L|\delta) \ . \end{aligned}$$

With the Rao-Kupper model, one win and one loss are equivalent to one draw. When $\delta_0 = 0$, the Rao-Kupper model becomes the Bradley-Terry model [13].

2.3 The Davidson Model

[14] proposed another variation of the Bradley-Terry model. Unlike the Rao-Kupper model, the Davidson model assumes that one win and one loss are equivalent to two draws (instead of one):

$$\begin{aligned} d(\delta) &= \nu\sqrt{f(+\delta)f(-\delta)} \ ; \\ P(W|\delta) &= f(+\delta)/(1+d(\delta)) \ ; \\ P(L|\delta) &= f(-\delta)/(1+d(\delta)) \ ; \\ P(D|\delta) &= d(\delta)/(1+d(\delta)) \\ &= \nu\sqrt{P(W|\delta)P(L|\delta)} \ . \end{aligned}$$

ν is a parameter of the model that indicates the probability of draws. $\nu = 0$ is equivalent to the Bradley-Terry model.

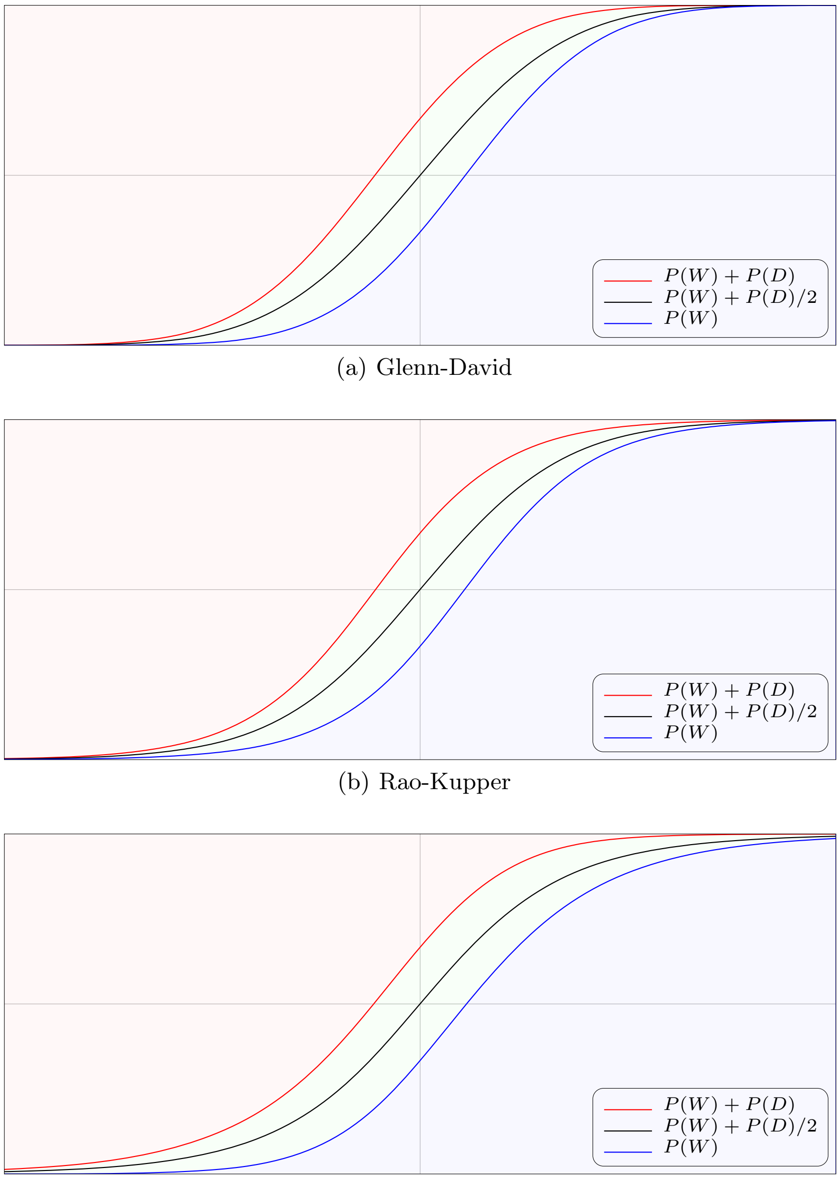
2.4 Individual Draw Percentages

[15], [16]

3 Model Selection

Use cross validation to determine a good model.

See Table 1.

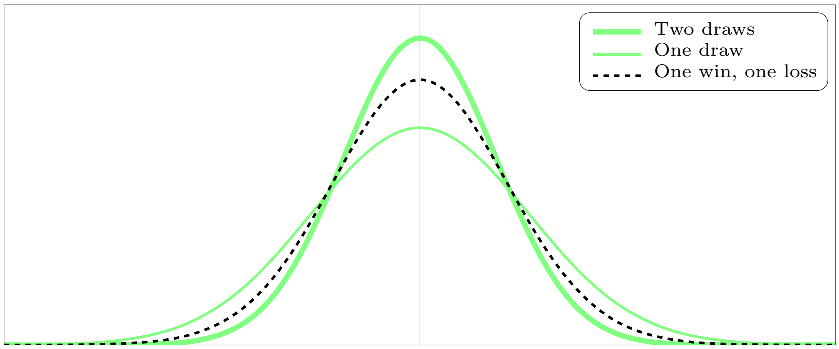


(a) Glenn-David

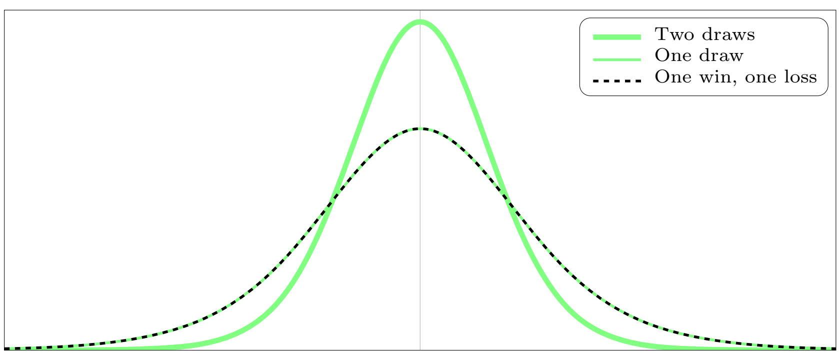
(b) Rao-Kupper

(c) Davidson

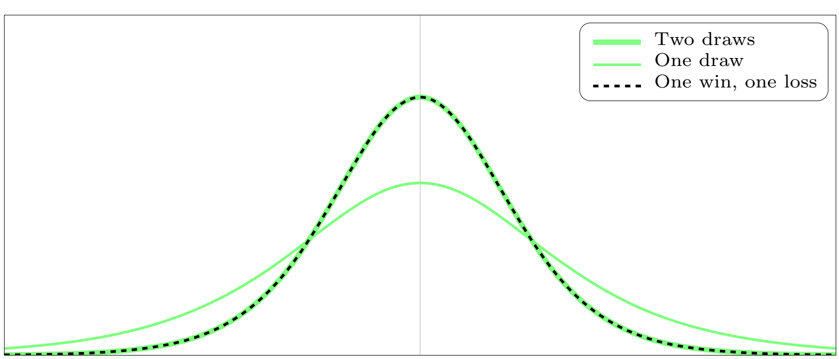
Fig. 2. Outcome probabilities as a function of rating difference δ. Parameters of the models were chosen so that $P(W|\delta = 0) = P(D|\delta = 0) = P(L|\delta = 0) = 1/3$. Horizontal axes were scaled so that $P(W)+P(D)/2$ has the same derivative at $\delta = 0$ for all models.



(a) Glenn-David



(b) Rao-Kupper



(c) Davidson

Fig. 3. Posterior rating probability densities with a uniform prior. Parameters and scales are like in Figure 2.

	RK	GD	DV	DV−RK	DV−GD
CCRL40/40	-846135	-845638	-845466	669	172
CEGT-blitz	-1840377	-1839400	-1838979	1398	421
CCRL-blitz	-1963747	-1962979	-1962675	1072	305

Table 1. Total log-likelihood of 10-fold cross-validation.

4 Conclusion

The experiments reported in this paper demonstrate that the Davidson model fits computer-chess data better than the Rao-Kupper and Glenn-David model. Of the three models tested, the Rao-Kupper fits the data worst. The Rao-Kupper model was used in the very popular BayesElo tool. The result presented in this paper will lead to a new improved version of BayesElo.

These experiments open some directions for future research. First, the models were tested with computer-chess data only. It would be interesting to test whether other games with draws (such as reversi, or Go with integer komi) produce similar results. It is also very clear that in games such as chess or reversi the draw rate increases with player strength. In order to deal with data sets that combine players of very different strengths, it would be good to have a statistical model where the draw rate depends on strength.

References

1. Elo, A.E.: The Rating of Chessplayers, Past and Present. Arco Publishing, New York (1978)
2. Glickman, M.E.: Parameter estimation in large dynamic paired comparison experiments. Applied Statistics **48** (1999) 377–394
3. Zermelo, E.: Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. Mathematische Zeitschrift **29** (1929) 436–460
4. Stern, H.: Are all linear paired comparison models empirically equivalent? Mathematical Social Sciences **23**(1) (February 1992) 103–117
5. Herbrich, R., Minka, T., Graepel, T.: TrueSkill$^{\mathrm{TM}}$: A Bayesian skill rating system. In Schölkopf, B., Platt, J., Hoffman, T., eds.: Advances in Neural Information Processing Systems 19, Vancouver, British Columbia, Canada, MIT Press (December 2006) 569–576
6. Coulom, R.: Bayeselo. `http://remi.coulom.free.fr/Bayesian-Elo/` (2005)
7. Edwards, R.: Edo historical chess ratings. `http://members.shaw.ca/edo1/` (2004)
8. Glenn, W.A., David, H.A.: Ties in paired-comparison experiments using a modified Thurstone-Mosteller model. Biometrics **16**(1) (March 1960) 86–109
9. Thurstone, L.L.: A law of comparative judgment. Psychological Review **34**(4) (July 1927) 273–286
10. Henery, R.J.: An extension to the Thurstone-Mosteller model for chess. The Statistician (41) (1992) 559–567
11. Batchelder, W.H., Bershad, N.J.: The statistical analysis of a Thurstonian model for rating chess players. Journal of Mathematical Psychology **19**(1) (1979) 39–60

12. Rao, P.V., Kupper, L.L.: Ties in paired-comparison experiments: a generalization of the Bradley-Terry model. Journal of the American Statistical Association **62** (1967) 194–204
13. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs. I. The method of paired comparisons. Biometrika **39**(3–4) (1952) 324–345
14. Davidson, R.R.: On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. Journal of the American Statistical Association **65**(329) (March 1970) 317–328
15. Joe, H.: Extended use of paired comparison models, with application to chess rankings. Journal of the Royal Statistical Society. Series C (Applied Statistics) **39**(1) (1990) 85–93
16. Kuk, A.Y.C.: Modelling paired comparison data with large numbers of draws and large variability of draw percentages among players. The Statistician **44**(4) (1995) 523–528
17. Hunter, D.R.: MM algorithms for generalized Bradley-Terry models. The Annals of Statistics **32**(1) (2004) 384–406

A MM Formula for the Rao-Kupper and Davidson Models

[17]

Data: w_{ij}, l_{ij}, d_{ij} are respectively wins, losses and draws of i against j, i playing as White.

Model parameters: γ_i is the strength of player i. θ_w is the advantage of playing as White. θ_d is the draw parameter.

Model (i is White):

A.1 Rao Kupper Model

Outcome probabilities:

$$
\begin{aligned}
P(i \text{ beats } j) &= \frac{\theta_w \gamma_i}{\theta_w \gamma_i + \theta_d \gamma_j} \\
P(j \text{ beats } i) &= \frac{\gamma_j}{\theta_w \theta_d \gamma_i + \gamma_j} \\
P(i \text{ ties } j) &= (\theta_d^2 - 1) P(i \text{ beats } j) P(j \text{ beats } i)
\end{aligned}
$$

Update rules:

$$\gamma_i \leftarrow \frac{\sum_j w_{ij} + d_{ij} + l_{ji} + d_{ji}}{\sum_j \frac{(d_{ij} + w_{ij})\theta_w}{\theta_w\gamma_i + \theta_d\gamma_j} + \frac{(d_{ij} + l_{ij})\theta_d\theta_w}{\theta_d\theta_w\gamma_i + \gamma_j} + \frac{(d_{ji} + w_{ji})\theta_d}{\theta_w\gamma_j + \theta_d\gamma_i} + \frac{d_{ji} + l_{ji}}{\theta_d\theta_w\gamma_j + \gamma_i}}$$

$$\theta_w \leftarrow \frac{\sum_{ij} w_{ij} + d_{ij}}{\sum_{ij} \frac{(w_{ij} + d_{ij})\gamma_i}{\theta_w\gamma_i + \theta_d\gamma_j} + \frac{(l_{ij} + d_{ij})\theta_d\gamma_i}{\theta_d\theta_w\gamma_i + \gamma_j}}$$

$$\theta_d \leftarrow \alpha + \sqrt{\alpha^2 + 1}, \text{ with } \alpha = \frac{\sum_{ij} d_{ij}}{\sum_{ij} \frac{(w_{ij} + d_{ij})\gamma_j}{\theta_w\gamma_i + \theta_d\gamma_j} + \frac{(l_{ij} + d_{ij})\theta_w\gamma_i}{\theta_d\theta_w\gamma_i + \gamma_j}}$$

A.2 Davidson Model

Outcome probabilities:

$$P(i \text{ beats } j) = \frac{\theta_w\gamma_i}{\theta_w\gamma_i + \gamma_j + \theta_d\sqrt{\theta_w\gamma_i\gamma_j}}$$

$$P(j \text{ beats } i) = \frac{\gamma_j}{\theta_w\gamma_i + \gamma_j + \theta_d\sqrt{\theta_w\gamma_i\gamma_j}}$$

$$P(i \text{ ties } j) = \theta_d\sqrt{P(i \text{ beats } j)P(j \text{ beats } i)}$$

Update rules:

$$\gamma_i \leftarrow \frac{\sum_j w_{ij} + \frac{d_{ij}}{2} + l_{ji} + \frac{d_{ji}}{2}}{\sum_j \left(\theta_w + \theta_d\sqrt{\frac{\theta_w\gamma_j}{\gamma_i}}\right)\frac{w_{ij}+d_{ij}+l_{ij}}{\theta_w\gamma_i+\gamma_j+\theta_d\sqrt{\theta_w\gamma_i\gamma_j}} + \left(1+\theta_d\sqrt{\frac{\theta_w\gamma_j}{\gamma_i}}\right)\frac{w_{ji}+d_{ji}+l_{ji}}{\theta_w\gamma_j+\gamma_i+\theta_d\sqrt{\theta_w\gamma_i\gamma_j}}}$$

$$\theta_w \leftarrow \left(\frac{-b+\sqrt{b^2+16ac}}{4a}\right)^2, \text{ with}$$

$$a = \sum_{ij}\frac{(w_{ij}+d_{ij}+l_{ij})\gamma_i}{\theta_w\gamma_i+\gamma_j+\theta_d\sqrt{\theta_w\gamma_i\gamma_j}},$$

$$b = \sum_{ij}\frac{(w_{ij}+d_{ij}+l_{ij})\theta_d\sqrt{\gamma_i\gamma_j}}{\theta_w\gamma_i+\gamma_j+\theta_d\sqrt{\theta_w\gamma_i\gamma_j}}, \text{ and}$$

$$c = \sum_{ij} w_{ij} + \frac{d_{ij}}{2}$$

$$\theta_d \leftarrow \frac{\sum_{ij} d_{ij}}{\sum_{ij}\frac{(w_{ij}+d_{ij}+l_{ij})\sqrt{\theta_w\gamma_i\gamma_j}}{\theta_w\gamma_i+\gamma_j+\theta_d\sqrt{\theta_w\gamma_i\gamma_j}}}$$