CSE702 Week 9+: General Methodological Issues Seen Via Intrinsic Performance Ratings

Let us first review the procedure for computing the Intrinsic Performance Rating (IPR) of a player P on a set of games G:

1. Do regression on the set of games G (taking the side of each game played by P) to fit the closest virtual player Y (which in the program is called a `TrialSpec` or just `Spec`).
2. Do `perfTest` ***on the reference set*** of 150 fixed games. Get the projected ASD figure a_Y.
3. The IPR is a function $r(a_Y)$, which is pre-determined by regression over the training sets.

That is to say, we do not actually use the actual ASD a_P by the player P on his/her own games G. That number would not be robust because the games G might have been unusually easy to play---or unusually hard. Instead, we fit Y to how P played in those games and do everything else with Y. The projected ASD comes with its own standard deviation σ_Y. Thus

$$[r(a_Y - 2\sigma_Y),\ r(a_Y + 2\sigma_Y)]$$

becomes the two-sigma confidence interval for the IPR measurement $IR_P = r(a_Y)$. Now we move on to address an important philosophical question:

What do these error bars represent?

What we want them to represent is the "95% confidence" interval for the actual strength R_P of the player P. That is, we want to interpret $IR_P^- = r(a_Y - 2\sigma_Y)$ and $IR_P^+ = r(a_Y + 2\sigma_Y)$ as estimates attributable to the player P. We want to say:

1. With "95%" confidence, the playing strength of P is between IR_P^- and IR_P^+.

Or at least we want to say:

2. With "95%" confidence, the playing strength that P showed on those games G is between IR_P^- and IR_P^+.

To see if these are justified, we need to say something about σ_Y that I've passed over before. It is not the projected standard deviation of ASD on the reference set. If it were, then the error bars would depend only on Y and have nothing else to do with the games G. Instead:

σ_Y is the projected sigma of ASD for Y on the original games G.

We have previously said why the IPR itself does not simply use the actual ASD of player P on the games G: the games may have involved less **hazard** (meaning: projected expectation loss or ASD)

than usual, or more hazard. The hazard in the games is, however, involved in the error bars of the IPR---that is, it factors into σ_Y.

To take in the big picture here, consider the following four ways a player P may have conducted the games G in a manner that produces *the same fit* $Y = (s_Y, c_Y, \ldots)$:

(a) P played smoothly at level Y in games that had relatively placid positions.
(b) P played smoothly at level Y in games that had choppy positions.
(c) P had games with placid positions but played in a choppy manner, alternating blunders with brilliance.
(d) P had games with choppy positions and played them in a choppy manner that still comes out to the same best-fit measurement Y.

All of this glomms over the issue of the error bars of the regression that produces Y itself---that is, the error bars on the fitted s and c (and etc.). But let's first make some intuitive observations about (a)--(d):

- Situation (b) will produce a higher σ_Y than (a). This says that the measurement Y itself is inherently less precise when the positions are choppier. Same with (d) versus (c), again keeping the player's manner of play the same. This can be summarized as: the more uncertain the "background", the less certain the *measurement*.
- In (c) versus (a), where the nature of the games G is the same, the greater uncertainty is squarely *about the inference of the player's skill*. A smooth player can be rated more precisely than a choppy one. This goes even more for (d) versus (b).

The σ_Y as defined above is based only on the positions---it is the *projected* ASD of Y on those positions---and so does not reflect the smooth-versus-choppy *variance* in the player's play. It does of course reflect the player's quality of moves through the fit of Y: each blunder lowered the measured quality; it took other series of good moves to raise it back to the level represented by Y; only the level goes into Y. To repeat: *the "intuitive variance" in how the player P got to level Y does not factor into the error bars σ_Y.*

For this reason, I regard the IPR error bars as confidence in the measurement only. They do not constrain judgment of the player's own skill---or the likelihood of achieving level $r(a_Y)$ on those games. Most in particular,

$$\frac{IR_P - Elo_P}{\sigma_Y}$$

shall not be treated as a z-score. There are two motivated reasons for keeping it this way, the second more important than the first:

- The cheating tests should be about more than "playing too well." They need elements of

specific concordance to machines.

- The IPR measurement involves regression over the player's own *small data* of the games G. Whereas, the deployed z-tests are calibrated using only the *large data* of the main training sets. The error bars of that calibration are tiny, hence negligible. The tests involve only simple counting of the player's agreements and actual ASD over the games G.

The latter point is offset by the fact that whether a move counts as an agreement with the computer sporadically changes even between a high depth d and $d+1$. This luck-of-the-draw factor is in turn mitigated by requiring tests of concordance to multiple engines.

Nevertheless, one wants to make confidence judgments on inferences about players' skill. A particularly noteworthy instance has just arisen at this point of the collocated Open and Women's World Championship Candidates Tournaments after 10 of 14 rounds in Toronto. Despite the men averaging 2745 to the women's 2517, which is 228 Elo more, the women's playing level is within striking distance---well within 100 Elo. After showing this, we will talk about three different ways to generate "extensional" error bars for the IPR, as opposed to the "intensional" bars being for the measurements only:

1. Use the error bars of the IPR regression to calculate a "two-sigma Gaussian ball" around $Y = (s_Y, c_Y, \dots)$, then analytically maximize and minimize $Y' = (s'_Y, c'_Y, \dots)$ over this ball. Use that minimum and maximum as the error bounds.
2. Do sampling over said Gaussian ball, then use the empirical σ_{IR} from the sampling. Note that this is a standard deviation of a large number of IPR measurements directly.
3. Do *resampling* of the n tested moves in the games G, keeping the number n of samples the same but *with replacement*, so that in various trials, some moves will be skipped and others counted two or more times. Use the empirical σ_{IR} from the resulting bunch of IPR measurements.
4. Use σ_Y after all, on large enough data where you can presume that the projected ASD is accurately predicting the actual ASD and the variability "evens out"---especially when the sample includes different players.

Option 3 may seem weird at first blush---it did to me---but it is theoretically justified. It reflects choppy play insofar as blunders will be in some samples---even multiple times---and not in others. It is called the Efron Bootstrap, after a famous 1979 paper by Bradley Efron. (See also this introduction.) The consideration in 4 is why I've "rested on my laurels" and not given this point deeper attention in the past.

Measuring the Candidates

Through 10 of 14 rounds of the Open and Women's Candidates Tournaments in Toronto, here are the omnibus IPR measurements of all the players (8 in each section) taken together. [Note: "Open" designates a section that women may enter, even when it is "Closed" in the sense of being by qualification or invitation only. Judit Polgar played in the equivalent tournament in 2005 and just missed

in 2007, losing a play-in match.] Since this Open has all men, we'll call it "Men":

In the Unit Weights mode:
Stockfish 11:
- Men: 2735 ± 65
- Women: 2650 ± 70

Komodo 13.3:
- Men: 2710 ± 65
- Women: 2630 ± 75

Komodo 10:
- Men: 2690 ± 85
- Women: 2630 ± 85

Stockfish 7:
- Men: 2680 ± 60
- Women: 2665 ± 65.

In the EWN mode:
Stockfish 11:
- Men: 2745 ± 80
- Women: 2685 ± 85

Komodo 13.3:
- Men: 2650 ± 75
- Women: 2625 ± 80

Komodo 10:
- Men: 2685 ± 85
- Women: 2650 ± 90

Stockfish 7:
- Men: 2650 ± 75
- **Women: 2685 ± 75**.

Averages of eight measurements (given correlations, error bars are about ± 50):
- Men: 2693
- Women: 2653.

In none of the nine comparisons are the men and women statistically distinguished at two-sigma confidence. I don't regard "not distinguished" or "within the margin of error" as justification for saying "statistically tied" the way pollsters often do. However, the fact that the women come out ahead on one measurement, and that all but two of the comparisons have the values within the lower individual error bar of one, make it IMPHO allowable in this case. (IMPHO = "In my professional humble opinion.")

After 12 Rounds:

In the Unit Weights mode:
Stockfish 11:

- Men: 2730 +- 55
- Women: 2645 +- 65

Komodo 13.3:

- Men: 2700 +- 60
- Women: 2620 +- 70

Komodo 10:

- Men: 2700 +- 75
- Women: 2620 +- 75

Stockfish 7:

- Men: 2690 +- 55
- Women: 2640 +- 60.

In the EWN mode:
Stockfish 11:

- Men: 2740 +- 70
- Women: 2680 +- 80

Komodo 13.3:

- Men: 2650 +- 65
- Women: 2600 +- 75

Komodo 10:

- Men: 2700 +- 75
- Women: 2630 +- 80

Stockfish 7:

- Men: 2670 +- 65
- Women: 2645 +- 70.

Averages of eight measurements (given correlations, error bars are about ± 45):

- Men: 2698
- Women: 2635.

After 13 Rounds:

In the Unit Weights mode:
Stockfish 11:

- Men: 2725 +- 55
- Women: 2640 +- 60

Komodo 13.3:

- Men: 2695 +- 55

- Women: 2620 +- 65

Komodo 10:

- Men: 2700 +- 70
- Women: 2610 +- 75

Stockfish 7:

- Men: 2685 +- 50
- Women: 2640 +- 55.

In the EWN mode:

Stockfish 11:

- Men: 2735 +- 65
- Women: 2670 +- 75

Komodo 13.3:

- Men: 2645 +- 65
- Women: 2605 +- 70

Komodo 10:

- Men: 2695 +- 70
- Women: 2620 +- 75

Stockfish 7:

- Men: 2665 +- 60
- Women: 2640 +- 65.

Averages of eight measurements (given correlations, error bars are about ± 40-45):

- Men: 2693
- Women: 2631.

Final: After 14 Rounds:

In the Unit Weights mode:

Stockfish 11:

- Men: 2730 +- 55
- Women: 2630 +- 60

Komodo 13.3:

- Men: 2700 +- 55
- Women: 2620 +- 65

Komodo 10:

- Men: 2695 +- 70
- Women: 2610 +- 65

Stockfish 7:

- Men: 2690 +- 50
- Women: 2630 +- 55.

In the EWN mode:
Stockfish 11:

- Men: 2745 +- 65
- Women: 2650 +- 75

Komodo 13.3:

- Men: 2660 +- 60
- Women: 2595 +- 70

Komodo 10:

- Men: 2700 +- 70
- Women: 2610 +- 75

Stockfish 7:

- Men: 2670 +- 60
- Women: 2635 +- 65.

Averages of eight measurements (given correlations, error bars are about ± 40-45):

- Men: 2699
- Women: 2623.

Second-Half Figures:

In the Unit Weights mode:
Stockfish 11:

- Men: 2715 +- 70
- Women: 2625 +- 85

Komodo 13.3:

- Men: 2675 +- 75
- Women: 2635 +- 85

Komodo 10:

- Men: 2700 +- 90
- Women: 2605 +- 100

Stockfish 7:

- Men: 2695 +- 65
- Women: 2610 +- 75.

In the EWN mode:
Stockfish 11:

- Men: 2730 +- 85
- Women: 2650 +- 105

Komodo 13.3:

- Men: 2645 +- 80
- Women: 2610 +- 100

Komodo 10:

- Men: 2695 +- 85
- Women: 2620 +- 105

Stockfish 7:

- Men: 2675 +- 75
- Women: 2595 +- 90.

Averages of eight measurements (given correlations, error bars are about ± 60-65):

- Men: 2691
- Women: 2619.