Predictive Modeling and Fraud Detection in Chess

Kenneth W. Regan
University at Buffalo (SUNY)

FIDE Ethics Commission, April 2017

Predictive Models

Given data and analysis on potential events $E_1, \ldots, E_L$ estimate probabilities $p_1, \ldots, p_L$ for them to occur.

Examples:

- Some of the events $E_1, \ldots, E_m$ are natural disasters.
- $E_1, \ldots, E_L$ are potential courses that a disease can take.
- The events are correct answers on an exam with L questions, and we want to estimate the distribution of results.
- The events are the legal moves in a chess position. They are *mutually exclusive* and (together with "draw" or "resign") *collectively exhaustive*: $\sum_i p_i = 1$.

Aggregates and Costs

- $\widehat{P} = \Pr[\text{some of events } E_1, \ldots, E_m \text{ occur}]$.
- $\widehat{P}_{k,j} = \Pr[\text{between } k-j \text{ and } k+j \text{ of them occur}]$.
- Suppose each E_i has a cost C_i. Then

$$\widehat{C} = \sum_i p_i C_i$$

  is the projected total cost.
- We may also wish to project

$$\widehat{P}_{C,j} = \Pr[\widehat{C} - j \leq C \leq \widehat{C} + j],$$

  which is the likelihood that our estimate $\widehat{C}$ will be within j of the actual cost C.
- In chess, $A_i = \delta(v_1, v_i)$ is the "cost" of an inferior move m_i. Then

$$\widehat{A} = \sum_i p_i A_i$$

  is the projected error on the move.

Multiple Trials

$$\text{Decision events } D = \begin{cases} E_{1,1}, \ldots, E_{L,1} \\ E_{1,2}, \ldots, E_{L,2} \\ \vdots \\ E_{1,T}, \ldots, E_{L,T} \end{cases}$$

- Estimate $\widehat{K}$ for $K =$ the number of times $E_{1,t}$ happens over turns t.
- Project J such that $Pr[\widehat{K} - J \leq K \leq \widehat{K} + J] = 0.95$, say.
- Predictive models need to project *how often they are wrong*.

Confidence Tests

- $E_{1,t}, \ldots, E_{L,t}$ = legal moves at turn t.
- $E_{1,t}$ = the analyzing engine's first line.
- $\widehat{K}$ = the expected agreement over T game turns, say $T = 250$ from 9 games.
- What if the actual K is outside $[\widehat{K} - J, \widehat{K} + J]$?
- The power to judge the model's rate of (in)accuracy *becomes* the power to judge the *unlikelihood* of $K > \widehat{K} + J$ in particular.
- To frame a **statistical confidence test** we need to quantfy the unlikelihood in general terms.

Multiple Multiple Trials

- Analyze multiple *sets* T_r of games.
- Get K_r, J_r, and $\widehat{K}_r$ for each r.
- Tally how often K_r is within $[\widehat{K}_r - J_r, \widehat{K}_r + J_r]$.
- For what J_r is $P_{K_r,J_r} = 95\%$, say?
- Provided $K_r \leq \widehat{K}_r + J_r$ at least 97.5% of the time, we are on sure ground to judge the unlikelihood of outcomes $K_r > q\widehat{K}_r + J_r$.
- Similar *confidence intervals* apply to the $\widehat{A}$ error estimation.

Multiple Multiple Multiple Trials

- In fact, I have run suites of 10,000 9-game trials at each Elo rating level 1600 through 2700.
- For 2400, for instance, I use games with both players rated between 2390 and 2410.
- My intervals are "snug" for the 2300–2500 range, *conservative* outside it.
- But there aren't 90,000 standard games for each rating level—only several hundred or thousand such games in ChessBase Big discs.
- Hence I *resample* 9-game subsets S_R of the set for each level taking a random side of each game.
- For each "virtual player" S_R, generate K_R, J_R, and $\widehat{K}_R$, and similarly with the error measure $\widehat{A}$.

Standard Deviations

- When the decision events (i.e., game turns) $D_1, \ldots, D_T$ are *independent*, then the outcomes K_R of the projected agreements from each subset become *normally distributed* for high enough T.
- The **Central Limit Theorem** establishes this for simple linear aggregates from any distribtution.
- The standard deviation of the outcomes is called σ (*sigma*).
- Values of J are expressed as multiples of σ.
- Two "tests of confidence" in my system are (i) that it centers $\widehat{K}$ correctly (for each Elo level) and (ii) that its J accurately eflects the Gaussian normal confidence intervals.
- (In fact, my theoretical σs for $\widehat{K}$ and $\widehat{A}$ need to be *adjusted* wider because the independence and overall modeling are not perfect.)

Some Sigma Levels

- $J = 1.00\sigma$: abouit 68% confidence of an outcome within the interval, 14% above it.
- $J = 2.00\sigma$: about 95% within, 2.5% above.
- $J = 3.00\sigma$: the *natural frequency* of outcomes above it is about 1-in-740.
- $J = 4.00\sigma$: about 1-in-32,000.
- $J = 5.00\sigma$: about 1-in-3 million. Level used by phyiscists to declare the Higgs Boson and gravitational waves discovered.

Readings of $K - \widehat{K}$ expressed in multiples of σ are called **z-scores**.

Odds for general distributions are called *p-values*.

The Focal Question

When do we suspect a high-outlier z-score for K or A to be outside natural frequency?

Two main factors weigh against taking z at "face value":

1. The number of players at any given time.
2. The *prior likelihood* of abnormality—meaning a player whose K_r is not governed by the Gaussian normal distribution.

The sense of "governed" allows that from time to time people play very well. It is important to examine how these two factors relate.

Littlewood's Law

- Suppose 1,000 people play in FIDE tournaments every week.
- Maybe 5,000 in summer.
- *The Week in Chess* (TWIC) gives about 35,000 player-performances per year.
- The "law" by the British mathematician John Littlewood says that for any 1,000 people and any confidence test you can expect to find one 1-in-1,000 outlier for that test ($z \geq 3.09$).
- In chess, let's say the player who is playing so well "has a beaming head."

"Beamers" and "Cherry-Picking"

- With 1,000 players in the hall, "people can easily see the beaming head"—especially suspicous people with engines...
- You 'catch' the "beamer" and get $z = 3.10$.
- Is he guilty? Of course not.
- If he is a former FIDE President, strangely, guilt is more likely because we have few living former FIDE presidents.
- But also we have FIDE VPs, IAs who play when not working,..., even a jazz violinist I met in Hilton Head SC three weeks ago took 4 years of lessons from GM Julio Bolbochán. "We all have distinctions."
- Which kinds of distinctions matter?

Most Mis-Applied Equation in Science

$$\Pr[Hyp \mid Evi] = \Pr[Evi \mid Hyp] \cdot \frac{\Pr[Hyp]}{\Pr[Evi]}.$$

- Called the "Transposed Conditional" in *Bayes's Theorem*.
- Here "*Evi*" means the statistical result—we will contrast it with human *evidence*.

"Flat View" of Bayes's Theorem

- Suppose you take a cancer test that is correct 99.9% of the time on both its 'yes' and 'no' answers.
- The natural frequency of the cancer is 1 in 10,000 people.
- You test positive. What are the odds you have the cancer?
- Let's take 10,000 typical people.
- One has the cancer.
- The test probably says 'yes' on the one.
- It also expects to say 'yes' on 10 other people who are factually negative.
- So you are one of 11 people; your odds of having the cancer are 1-in-11.

Judging and Applying Prior Likelihoods

- If *nobody* ever gets the cancer then you can say definitively that your odds will be 0-in-10.
- Suppose the prior likelihood of cheating in chess is not 0 but rather 1-in-10,000.
- Now our sample of 10,000 people (over 2-3 months) expects to have 1 factual positive and 10 other test posiitives—maybe 10 "beaming" players.
- What are the odds one of the 11 positives is cheating? 1-in-11?
- But this depends completely on belief about the *prior* $\Pr[hyp]$, not the test results—if it is zero then it means you just had a slight deviation from 10 to 11 in the number of "beaming players."

Distinguishing Marks

- In general, Bayes's Theorem is the gateway to deep and murky areas of statistical science.
- However, the issues can be resolved when the "*Evi*" is supplemented by "what humans call evidence."
- Story of "Thirteen Sigma" and a Rostov-on-Don tournament.
- ACC Regulations: Physical and/or behavioral evidence apart from engine "evi."

Two Main Effects

1. No "Littlewood's Law"—now a player is distinguished by reasons other than "beaming" (or being FIDE President, etc.)
2. The prior likelihood is no longer "general."

- We could ask the conditional question: what is the chance of cheating given the reported objects/mannerisms?
- But with the "flat-view" approach we should start by asking, *how many people* (in a week or month or ...) have those objects or mannerisms?
- In the case of people being just caught with a cellphone, or going out every move to smoke, perhaps quite a few...

Statistical Effects

- Such evidence makes the face-value odds from the z-score close to the true odds.
- Perhaps it leaves the prior Pr$[Hyp]$ still about 0.5 or 0.3.
- The $z = 2.75$ threshold in the ACC Guidelines is roughly 3-in-1,000, 99.7% confidence.
- The prior of 0.3 still leaves 99.0% confidence.
- In civil cases, $z = 2.00 =$ 1-in-44 (face-value) odds might be allowed—just as it is the (controversial but common) norm in academic publishing.
- Still short of CAS "reasonable comfort" standards at various levels.
- But what about further effects of *evidence* besides changing the priors on the "evi"?

Other Indicators and Cross-Checks

- (Besides re-runs of the same test with some difference in sample or engine...)
- In medical real-life, if you're not reasonably comfortable with a diagnosis you get a second test.
- If your first test was complex and new, you might fall back on a test with older, well-worn diagnostic indicators.
- The "Basic Sanity Check" of statistical odds: gathering large samples around an outlier's score and tallying the negatives.

The Screening Test

- Less acute than the main test.
- Does not *predict* at all: no $\widehat{K}_r$ let alone J_r projections, just tallying the actual $K_r, A_r, \ldots$ results.
- Takes only 5-10 minutes per core per game as opposed to 4-8 hours for ful test. Over a million games maybe now.
- Simple "Raw Outlier Index" (ROI), which is a function *only* of the player's Elo rating E and $K_r, A_r, \ldots$ results.
- If K_r, A_r hit the average for players rated E then the ROI is 50 on a 0–100 scale.
- The σ nominally "should be' 5, but in fact it's set to basically 7.
- Like issue of whether the standard deviation of Elo rating performance intends to be 200 or 280+.
- Real purpose is to "identify a reasonable subset of the lambs so that if there is a wolf then the wolf is likely in that set." Some "famous names" hit just 70.

The Screening Data

- Since fall 2014, every event in TWIC.
- About 35,000 player-performances from about 240,000 games.
- Opens with only games on top boards preserved generate considerable upward bias in ROI (of course, rapid and Blitz dampen is strongly).
- Until 2012 I did all with Rybka 3. E.g. had over 5,000 player-perfs. in Opens.
- Only player to top 70% matching to Rybka 3 was 71% by Sebastian Feller at the 2010 Paris International Ch. (not the Olympiad).
- Made my 3.25 z-score = 1-in-1,733 odds more concrete: 1,732 other players did not match Rybka as much as Feller did—indeed, three times that many.

The "Floating Z" Caveat

- The ROI list (adjusting for biases) behaves like a giant bell curve—as it should since it falls under the Central Limit Theorem.
- Hence you *can* interpret rank in the curve as a z-score.
- Fallacious for outlier inference, however: "Someone has to be #1."
- "Curving" exams after-the-fact basically does this too.
- *Rasch modeling*, *Item-response theory*, and other psychometric models try to make personnel scoring more scientific via a so-called ability parameter θ, for instance.
- (A Jan. 2015 paper on my work by Barnes and Hernandez-Castro seems to misunderstand this as the procedure.)

Full Theory Behind Z

- Two principal ability parameters:
  - 's' for "sensitivity": lower is better.
  - 'c' for "consistency": higher is better.
  - (New work introduces a third trying to capture "depth of thinking.")
- Define a formal "Virtual Player" $P(s, c, \dots)$.

Main Principle

The probability of $P(s, c, \dots)$ playing move i depends on its value v_i in relation to the overall position value v_1 and the values of ther moves.

- A move with a clear standout value will be most likely for humans as well as computers.
- If moves have nearly equal values *throughout the search* then they should have nearly equal probabilities.

Main Equation and Training

Subject to $p_1 + p_2 + \cdots + p_L = 1$, given s and c we solve:

$$\frac{\log(p_1)}{\log(p_i)} = \exp(-\left(\frac{\delta_i}{s}\right)^c).$$

Then fit s and c to minimize the least-squares difference between the projections of our main quantities and their actuals in the training data. This makes $\widehat{K}_r$ and $\widehat{A}_r$ *unbiased estimators*

Across the Elo Spectrum

- Games between players both within 10 of the same Elo century point (15 in some places).
- Given an Elo level E, training yields s_E, c_E.
- The individual s_E sequence and the individual c_E sequence give a "decent" linear fit to E. (Not as sharp as the K_E and A_E sequences from the data, but workable.)
- Combining them yields the "central fit" in my 2011 paper with Haworth—updated somewhat in summer 2014 with renewed resampling trials. Call it now $\tilde{s}_E$ and $\tilde{c}_E$.
- To give slack, given an Elo rating E, use $\tilde{s}_{E+25}$ and $\tilde{c}_{E+25}$. (The line tails down to a 10 or so pt. dofference below 1500.)

Procedure

1. Eliminate opening turns 1–8 and all moves played by 2300+ players.
2. Elim. turns with one side ahead more than 3.00 and turns in repeating sequences.
3. (It doesn't matter to elim. positions with just 1 legal move; the system adjusts both $\widehat{K}_r$ and the actual K_r by $+1$,)
4. Enter $s_E, c_E, \ldots$ from the "central fit" for the *post*-tournament rating E.
5. Obtain probabilities $p_{1,t}, \ldots, p_{L,t}$ for each turn t.
6. Report the resulting z-scores for $K_r - \widehat{K}_r$ and $A_r - \widehat{A}_r$ (plus a third test for equal-top value recommended by Barnes-HC paper).
7. And report a combined test value.

Sample Size Note and IPRs

Important that the procedure does not involve regression on the player's games ("small data"), only applying parameters from fits on the larger training data.

- Regression on the player's games produces an "Intrinsic Performance Rating" (IPR).
- Error bars on IPRs are not a confidence test but only errors of measurement.
- If IPR is in the Elo 3000+ "computer range" this is another indicator.
- Whereas if a 1800 player has an IPR "only' 2500 it is "within the realm of human possibility" and might convey some doubt.

Some Reasonable Defenses

- Argue that the defending player is "more tactical" or "more positional" than the average of like-rated players.
- Argue that certain moves beyond the book-by-2300+ limit were home prep.
- Argue that some moves were more/less critical than others.
- Argued that some sequences of moves followed a *plan*.
- Argue that engines differ on certain critical moves.
- The test is *deterministic* and *reproducible* (*Rybka3 caveat), but it is also *aleatory*.

The two last points can lead into murky matters but are largely covered by the provision of multiple tests with different engines, and in a different mode ("Multi-PV" vs. "Single-PV") from the screening data.