'Go-Variance': A Covariance Analysis of Final Positions in Go

Joshua Matt Turner

January 21, 2021

1 Introduction

1.1 Go

The ancient Chinese game of Go has been played continuously for thousands of years. In Go, two players alternate placing one of their pieces, called *stones*, on an empty intersection of a 19×19 grid. Players score points by surrounding empty spaces with their own stones (the empty spaces being called their *territory*) and by capturing enemy stones. A group of stones is captured when it has no open spaces, called *liberties*, around it. The game ends when both players consecutively pass their turns. Points are counted at the end of the game, and the player with the most points wins. These simple rules give rise to complex strategy and play styles.

Figure 1: A game of Go in progress. White has *territory* in the bottom left. The black stone on the bottom left has two *liberties*. If white fills these, it will be captured and removed from the board. Image by Goban1 - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=15223468.

1.2 Project Statement

In this project, I explore the following question: how do the final board states differ between low- and high-ranked amateur Go players? I investigate this question from both local and global perspectives. From the former perspective, we discover correlations between stone placements, and from the latter, we uncover the primary variations in overall board states within each skill level.

2 Methods

2.1 Data

To answer these questions, I analyze the `18k` (lowest rank) and `9d` (highest rank) subsets of featurecat's Go dataset, which consists of 21.1 million games played on the Fox Go server. Files which contained formatting irregularities, or were otherwise difficult to preprocess, were thrown out. In particular, preprocessing consisted of the following steps:

1. Within each folder, open each `.sgf` file.
2. For each `.sgf` file,
   (a) Extract only the sequence of moves, disregarding other game data;
   (b) Play out the game in the Go engine, which stores the board as a matrix B, where $B_{ij} = 1$ if a black stone is in row i, column j of the board, with the origin at the top left. The same rule applies with -1 indicating a white stone and 0 indicating no stone;
   (c) Export the final matrix to a `.csv` with the same name as the original `.sgf`.

The remaining portions of the `18k` and `9d` datasets which I used contained 298,615 and 146,487 games respectively.

Since the board size in these data is 19×19, we can encode a game as a 361-dimensional vector, and, more generally, view the datasets as draws of random vectors $\mathbf{X}_{\texttt{18k}}, \mathbf{X}_{\texttt{9d}} \in \mathbb{R}^{361}$. Framing the datasets in this way allows us to find the *covariance matrix* for each random vector, on which both the local and global analyses mentioned in the introduction rely. Since the covariance matrix of a random vector $\mathbf{X}$ is defined as

$$\Sigma_{\mathbf{X}} = \mathbb{E}[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^T],$$

we find the mean board-matrix for each dataset, then use this average to mean normalize the data.

2.2 Analyses

We now look at how we use this covariance matrix to extract meaningful information from the dataset.

2.2.1 Local

We first use this matrix to look at correlations between moves. In order to find these correlations, we first convert the covariance matrix of a dataset to a correlation matrix

$$P_{\mathbf{X}ij} = \frac{\Sigma_{\mathbf{X}ij}}{\sqrt{\Sigma_{\mathbf{X}ii} \cdot \Sigma_{\mathbf{X}jj}}}.$$

We then have $P_{\mathbf{X}ij} = \rho(\mathbf{X}_i, \mathbf{X}_j)$, which tells us the direction and strength of the linear relationship between stones at coordinates i and j. In other words, we will be able to see, given a stone placed at one coordinate, where other stones are likely to be on the board. Thus, this analysis will lend insight into the play styles of low- and high-rank players by revealing how particular moves relate to each other.

2.2.2 Global

We can also use this matrix to determine the directions (361-dimensional vectors) along which our datasets vary the most, or, more intuitively, the board shapes which encode the most information about our data. The process by which we determine these components is called *principal component analysis* (PCA). It consists of finding the eigenvectors of the covariance matrix. Those with larger eigenvalues are directions of larger variation in the data. PCA will help us see the primary structural components of games in each dataset, thus revealing how games differ between datasets from a broader perspective.

3 Results

3.1 Local: Correlation Matrix

By computing the `18k` and `9d` correlation matrices as described above, we obtain Figure 2.

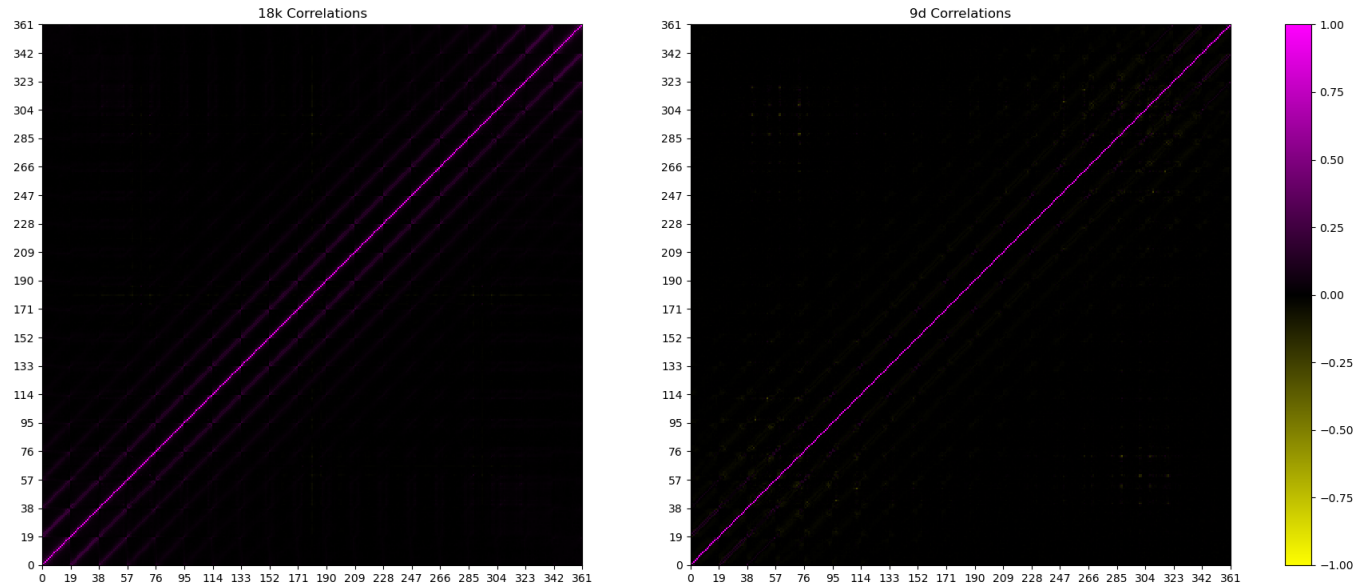

Figure 2: The correlation matrices for the `18k` and `9d` datasets

We immediately see that both plots contain stripes parallel to the diagonal. However, the stripes in the `9d` plot are much fainter, and contain mainly negative correlations. In fact, the `18k` plot hardly has any negative correlations, save for a faint cross shape about the center of the plot. Additionally, the `9d` matrix has a patch of anticorrelation in the top left. Although these are interesting and potentially meaningful features, they aren't really intepretable from these visualizations. We therefore leverage the fact that each row of each matrices represents the correlations between a stone at a particular coordinate and stones placed at every other coordinate on the board, in order to obtain a more natural and insightful visual representation of our data. Concretely, we reshape each row into its own 19×19 matrix, which allows us to visualize the row as a Go board.

Take, for instance, the below visualization of the 181st rows of the matrices, which corresponds to a stone placed at the center of the board.

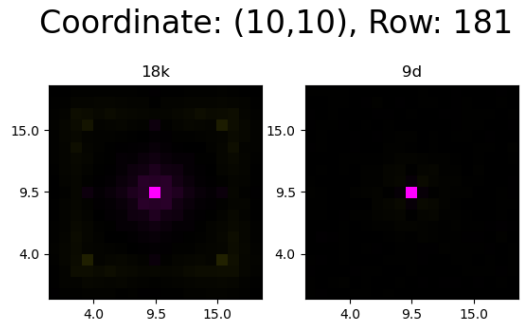

Figure 3: Correlations for a stone placed at the center of the board.

First, given that, in calculating the covariance matrix, we represented boards as 361-dimensional vectors, this reshaping shows us that this row does in fact correspond to the middle of the board. Second, we can now interpret its entries in terms of Go positions, where pinker squares indicate correlation, and yellower squares indicate anticorrelation. In particular, we see that, for 18k players, if one player has a stone at the center of the board, it is somewhat likely that there are friendly stones close by, and enemy stones near the corners. Meanwhile, for 9d players, there is no such pattern. Now that we know what we're looking at, we can note other interesting features of these plots.

Before getting to notable differences in specific positions, we first turn to the most prominent feature of the correlation matrix: the diagonal stripes. If we look at a few row plots, we notice that, for the 18k players, there is always a gaussian distribution of positive correlation centered about the stone.

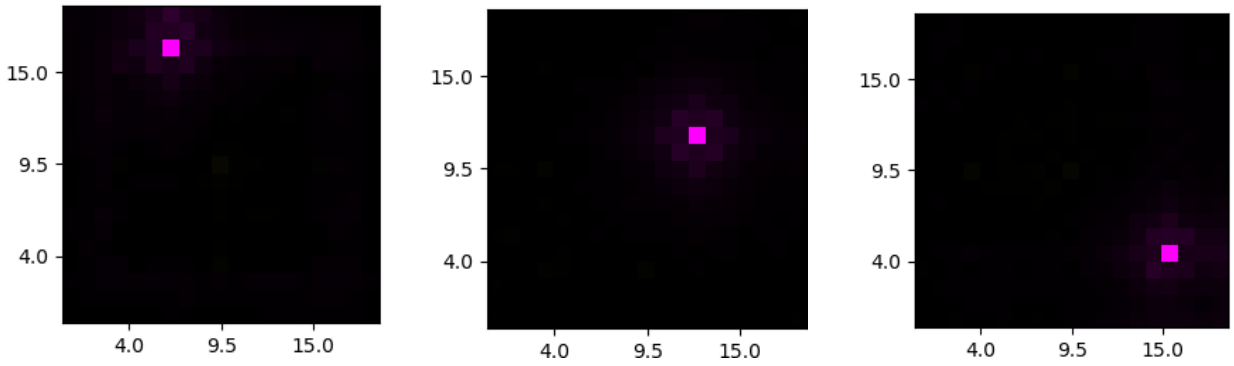

Figure 4: Across the board, 18k stones have friends nearby.

This indicates that 18k players regularly organize their stones in clumps. Of course, this is typical in Go, as connected groups of stones are much harder to capture than isolated stones. Given the fact that the stripes and gaussian are ever-present in their respective plots, combined with other subtle patterns in the covariance matrix (notice the 19×19 squares implied in the stripes), we see that the stripes correspond to this gaussian distribution, to the trend of placing stones close together.

Since strong position requires groups of stones, why do we not also see this trend for 9d players? In fact, why is there almost an *opposite* trend? If we look at some `9d` row plots, we notice that there is a similar gaussian distribution (albeit with less variance) near the edges.

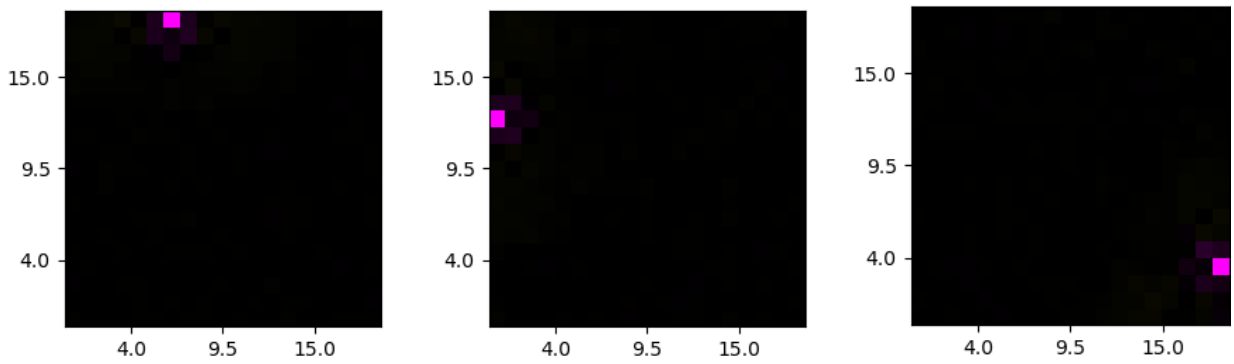

Figure 5: 9d stones also have friends near the edges.

However, once we get farther from the edges, we see that this pattern disappears, and is replaced with a faint pattern of diagonal enemy stones (Figure 6). These create the weaker diagonal stripes in the `9d` correlation matrix. Although I am not experienced enough a Go player to confidently answer, I suspect this difference may indicate that games between 9d players are much less uniform in terms of how parts

of the board look. In particular, play around the edges may look much different from play in the center, while 18k players might play similarly in all parts of the board. Alternatively, I would not be surprised if my interpretation of the pattern in Figure 4 were incorrect.

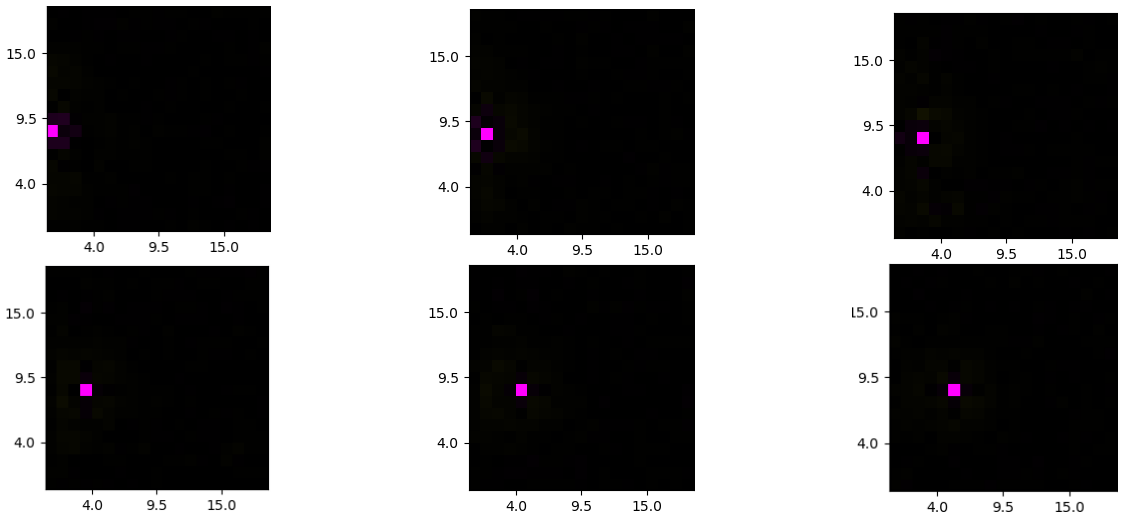

Figure 6: 9d stones lose their friends as they move away from the edge :(

We now turn to specific interesting plots. The most common opening move in Go is the 4-4 opening, so called because the player places a stone 4 rows and 4 columns away from the edge. By looking at the below plots, we can see how players of different skill levels structure the board around this move.

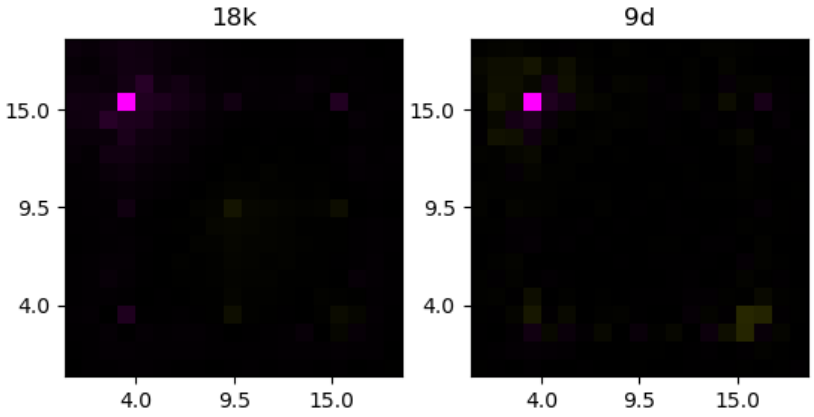

Figure 7: The classic 4-4 opening.

In the 18k plot, we see that there is still some gaussian-ness around the 4-4 stone. However, more notable is the grid-like arrangement of other stones. The points of the other significant correlations are called the *star points* of the board, which are marked on Go boards to help players visually orient themselves. Other `18k` plots where the stone is on a star point also exhibit this pattern, demonstrating the tendency of 18k players to structure their games around these points. In the `9d` plots, we see no such pattern. This makes sense, as 18k players are newer, and will structure their stones in more predictable ways.

Looking at the `9d` plot for the 4-4 point, we see light negative correlations in both the top left and bottom right corners. The anticorrelations in the top left might indicate an *invasion*: when one player tries to

establish territory, but their opponent jumps in to sabotage. Additionally, the bottom right indicates that, if a one player does the 4-4 opening, their opponent may try to claim the opposite corner. This pattern is part of the patch of anticorrelation in the top left of Figure 2. In general, this patch corresponds to common openings.

Finally, we investigate an offensive move called an *approach*, which serves to threaten an opponent's corner territory near the beginning of a game. We see one such move in row 112 of the matrix. In particular, we see that 9d players make this move much more consistently than 18k players.

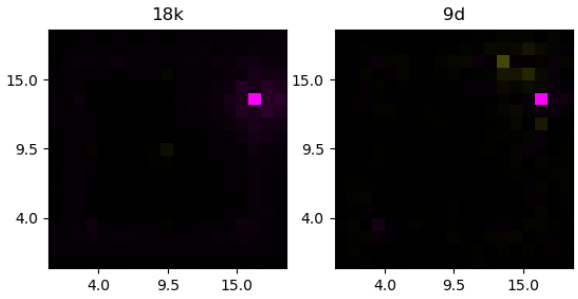

Figure 8: An approach.

There are many more interesting images which we do not have the space to cover/analyze here. The curious reader will find all 361 visualizations in the GitHub repository.

3.2 Global: PCA

Having looked at which moves tend to co-occur, we now shift to how games differ overall. Performing PCA as described in 2.2.2 yields the following top 36 principal components for each dataset.

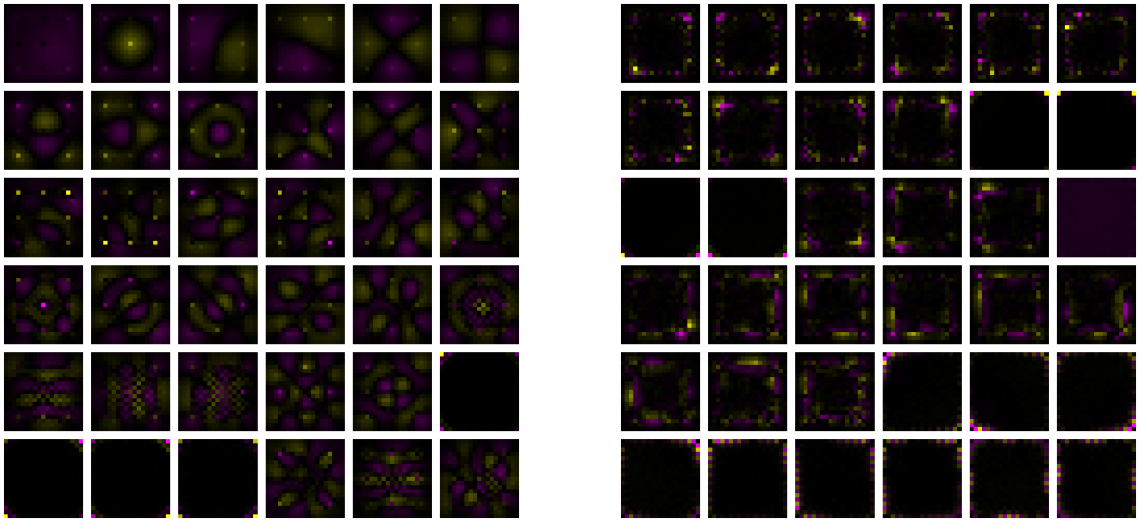

Figure 9: The top 36 PCs of the `18k` (left) and `9d` (right) datasets

We immediately see a stark difference between 18k and 9d games. In particular, 18k games differ primarily in terms of large, well-defined groups of stones. For instance, the second principal component indicates that one player occupies the center and the other controls the edges. We'll call this type of principal

component *territorial.* On the other hand, the 9d PCs focus more on control of the edges. In addition, they do not indicate clearly defined regions. We'll call this type of principal component *noisy.*

The implication of this difference is that low-rank players tend to divide territory very cleanly in simple, large sections, since the presence or absence of these divisions provides the most information about the final board state. In fact, the first 29 PCs of the `18k` dataset are territorial! Meanwhile, it seems that high-rank games are more determined by play around the edges, with territory in the center playing a less substantial role. Additionally, the noisiness of the PCs suggests that territory for 9d players does not come in large chunks, as it does for 18k players. The first territorial PC for the 9d dataset is the 46th, further supporting this claim.

Another significant type of principal component observed for both datasets is a frame-like component: one which includes only pieces on the edges. These are ranked much higher for the `9d` dataset (32 in `9d` vs 71 for `18k`), which again indicates that play around the edges is more important in 9d games.

For the 18k dataset, we observe that many of the top PCs highlight the star points. This supports what we said in the previous section: that 18k players structure their play around the star points.

Finally, there is a type of principal component present in each dataset which looks very strange. As we see in Figure 9, both datasets have principal components consisting of only the corners. Not only are these components limited in how much of the board they include, but they are also fairly highly ranked! I wouldn't have expected corner occupation to be so important in describing final positions.

As with the correlation matrix visualizations, there are many more interesting principal components left out (PC 116 for `9d` has a cool yin-yang pattern), but fret not! All 361 of the principal components for each dataset are again available on the GitHub repository.

4 Summary

By investigating the covariance matrices for amateur games of Go, we have noticed significant differences between the final board positions of low- and high-ranked players. In particular, we have learned that 18k players play in larger clumps and love structuring around the star points, while 9d games are largely determined by play on the edges and do not contain large chunks, particularly in the center of the board.

5 Acknowledgement

Many thanks to Alexander Turner for help with brainstorming and to Andrew Ng for his great machine learning course on Coursera.