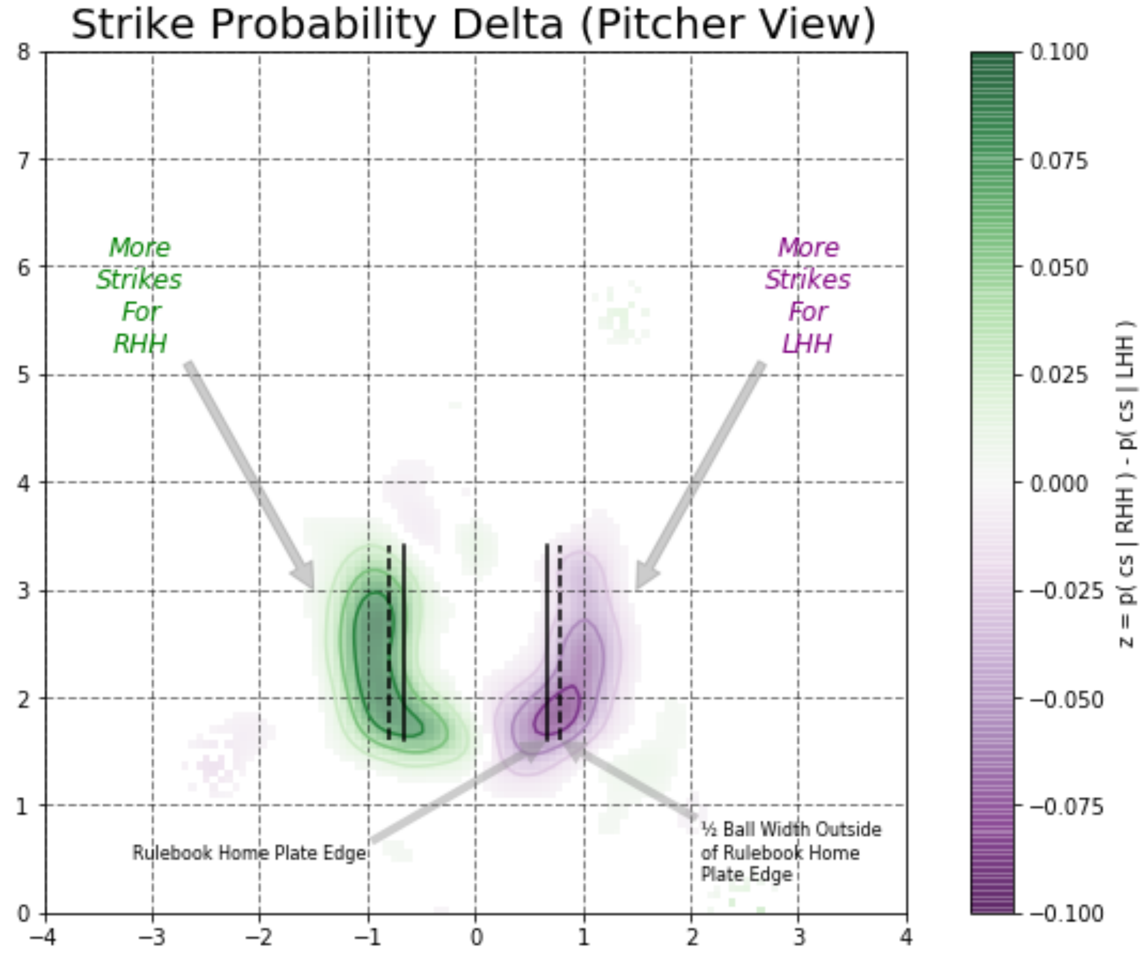
Toronto Blue Jays 2018 Analytics Exercise

Dan Goldberg - 2018-12-05

1. Produce a visualization illustrating how the strike zone differs for right-handed and left-handed batters.



2. **Which catcher was the most effective framer on a rate basis? Briefly describe (a paragraph or two) the process you used to come to this conclusion. Please include .csv with every catcherid along with some measurement of framing skill (how you choose to convey this is up to you).**

I estimated Catcher 5385 as being the best performing framer on a rate basis with 73.37 Framing Runs Above Average per 7000 Called Pitches, though this point estimate has high variance according to the model (as that catcher only had around 50 chances) so the best low-variance catcher was Catcher 6428 with 52.6 Framing Runs Above Average per 7000 Called Pitches. Having trained this model a few times, there has been some significant variance in the scale of the estimates, but the stack-rank is relatively stable. To estimate this metric I implemented a simpler version of Baseball Prospectus' CSAA calculation as described in [1]. I tried using the same set of random effects but struggled to get the model to converge, so I removed the catcherid:cs_prob random effect to simplify the model. I ran this model in Python using the BinomialBayesMixedGLM class from the statsmodels package, using a probit link function, a fixed effect parameter variance prior of 0.5, a random effect parameter variance of 0.1, fitted using the built-in Variational Bayes fitting, with scipy.optimize.minimize fit method as BTGS, an L2 Norm on the gradient, and gradient tolerance of 500. The main fixed effect in the GLMM cs_prob (called strike probability) was generated by modeling the probability of a strike using what I thought would be informative features (location, release speed, break, pitch type, batside, pitch hand) using a RBF kernel SVM, trained using the Scikit-Learn svm.SVC class. After fitting the GLMM model, I calculated the baseline probability of a called strike for average pitch with average individuals (catcher, pitcher, etc.) and subtracted that from the probability of a called strike adjusted for the random effect of each catcher to get the CSAA (called strikes above average) for each catcher. Then, using 0.14 runs per called strike as a heuristic (as quoted in [1]) I calculated the runs added for each catcher's CSAA multiplied by 7000 average framing opportunities. The mixed effects model strategy in general is hard to argue with from the perspective of isolating each individual's impact on the average called pitch in average context, and though I'm sure it's possible it could be improved upon, I do think this is a pretty good strategy.

My specific implementation could almost certainly be better. It definitely could stand to have a better cs_prob model, and since this variable is such a vital fixed effect in the model, any shortcomings in the cs_prob model would surely and seriously hurt the accuracy of the downstream mixed effects estimation. I chose to train the RBF-SVM classifier on a small subset of the dataset (10%) and compare different settings of the C (regularization) parameter. I trained on a small subset of the data because I wanted to minimize using the training set cs_prob predictions in the GLMM (since the better SVM models as measured by test set accuracy had lower regularization and would overfit to the training set). An alternative could have been to train the cs_prob model on a small subset of the dataset and exclude that training set from the GLMM data altoghether, though I decided that more data for the GLMM was preferred if the SVM was still accurate and if the overfitting wasn't too bad. My implementation also varied from that of [1] since I excluded catcher_id:cs_prob from the random effects, which is supposed to account for the varying skill a catcher may have in framing depending on the pitch location (viewing cs_prob as a proxy for pitch location). Without this interaction effect the random intercept for catcherid (the parameter we're using to infer catcher framing skill) would probably contain some of this information, and so it's possible the CSAA Runs/7000 metric I've output here is more sensitive to the specific observed distribution of framing opportunities the catcher had, and how those opportunities fit with that catcher's location-dependent framing skill. Put another way, it's possible that a catcher getting unlucky in getting a higher proportion of their framing opportunities in regions of the zone that they are less skilled at framing would be penalized under this model, which is supposed to infer that catcher's skill in the counterfactual case of an average distribution of pitch locations, types, etc.

[1] Brooks, D., Pavlidis, H., Judge, J.; "Moving Beyond WOWY: A Mixed Approach To Measuring Catcher Framing"; Baseball Prospectus, 2015. Link

3. A coach asks you how the attributes of a pitch affect its likelihood of producing a called strike. In one sentence, how would you respond?

Based on a random forest model of strike probability I trained, the location of a pitch is by far the most predicative of a pitch being called strike probability, and based on my internal understanding of pitching and hitting along with the data I've seen it's my hypothesis that the other controllable attributes (pitch type, release speed, & break) interact in complex ways, both with each other and with pitch location, so that the impact of an increase or decrease of one attribute on called strike probability is possibly highly dependent on the other pitch attributes, leading me to hesitate in communicating any more heuristics.

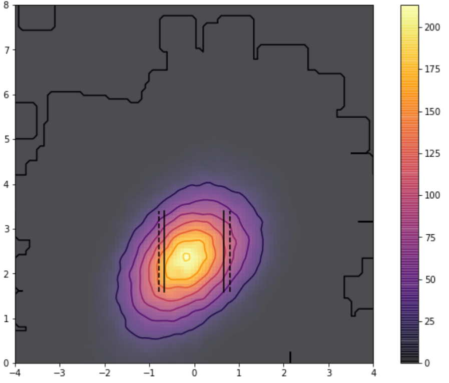
4. Tell us one other interesting conclusion you drew from the dataset.

In the course of creating a visualization for question #1 I was curious about the differences between fastballs (I grouped as FA, FC, FS, SI) and breaking balls/offspeed (CH, CU, SL, SC, KN) between each batside-pitcherhand combination (RHH vs. RHP, RHH vs. LHP, etc.). I compared the two pitchtype groups by visualizing each group's volume of usage by location, and probability of a strike or out (a positive result for a pitcher). Like in #1 I took the difference between the probabilities at each location to show where in and around the strike zone each pitchtype group leads to more positive results for the pitchers. I've included all plots at the end of this document for reference.

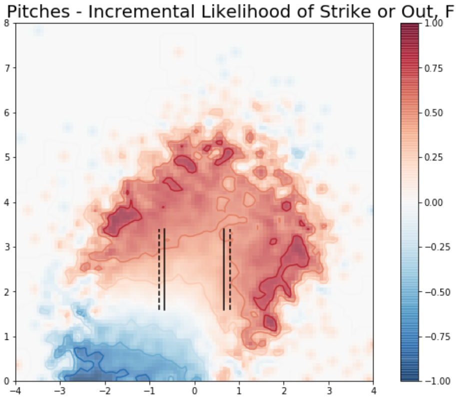
Most of what we can observe in these plots isn't all that surprising, though I found it interesting that opposite-side pitchers (LHP vs. RHH and RHP vs. LHH) find quite a bit of success low and inside with breaking balls/offspeed. The same is not true for same-side pitchers (RHP vs. RHH, LHP vs. LHH) as we can clearly observe a great advantage in fastballs in those locations. Perhaps the opportunity for opposite-side pitchers to create a severe horizontal break angle throwing across the plate allows them to find this success despite low and inside off the plate being more reachable for hitters than low and outside, the most extreme example possibly being Chris Sale and his slider against righties. This kind of pitch does have much in common with a RHP's breaking ball to RHH low and outside in that the pitcher has an ability to create a severe horizontal break angle and finish the pitch well off the plate. This also makes Tim Lincecum's changeup (one of my favourite pitches) all the more impressive, as he could use it so effectively low and inside on same-side batters, something is looks infrequent in the RHH-RHP breaking ball probability plot below.

RHH vs. RHP

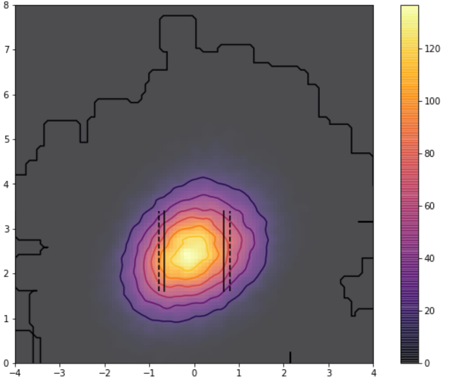
RHH vs RHP All Pitches - Observed Volume



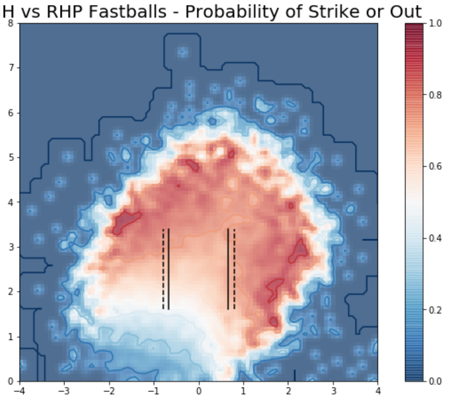
RHH vs RHP All Pitches - Incremental Likelihood of Strike or Out, FB vs. BB



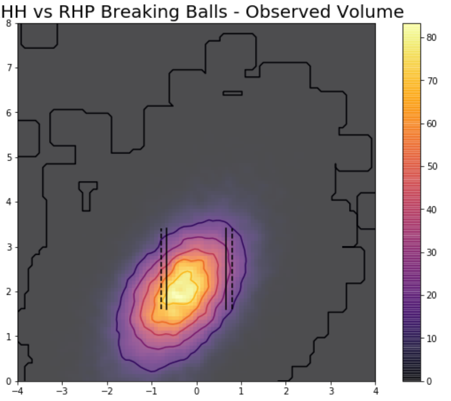
RHH vs RHP Fastballs - Observed Volume



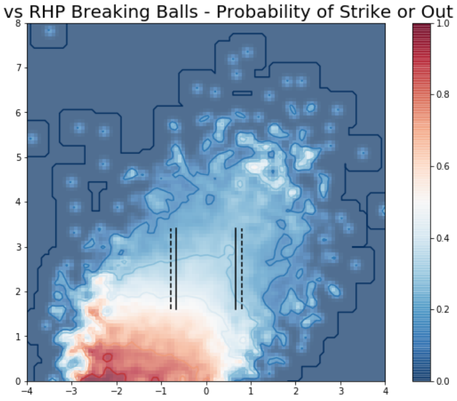
RHH vs RHP Fastballs - Probability of Strike or Out



RHH vs RHP Breaking Balls - Observed Volume

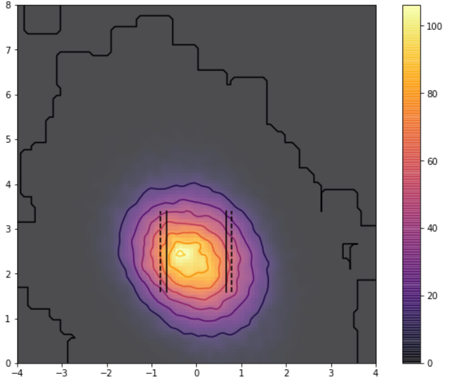


RHH vs RHP Breaking Balls - Probability of Strike or Out

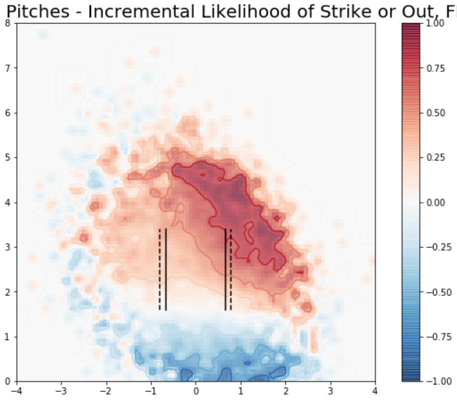


RHH vs. LHP

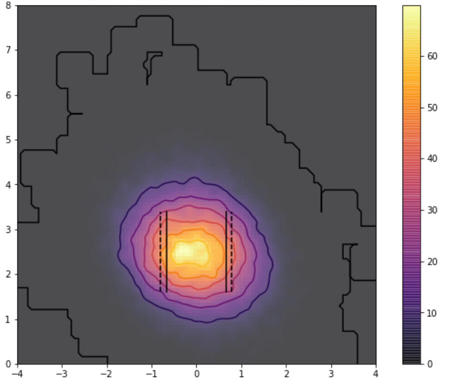
RHH vs LHP All Pitches - Observed Volume



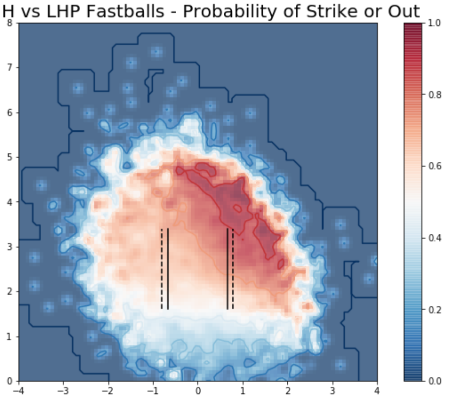
RHH vs LHP All Pitches - Incremental Likelihood of Strike or Out, FB vs. BB



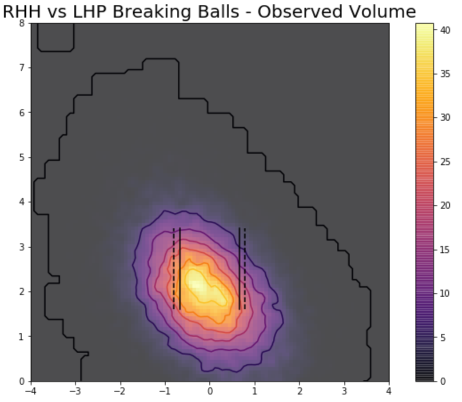
RHH vs LHP Fastballs - Observed Volume



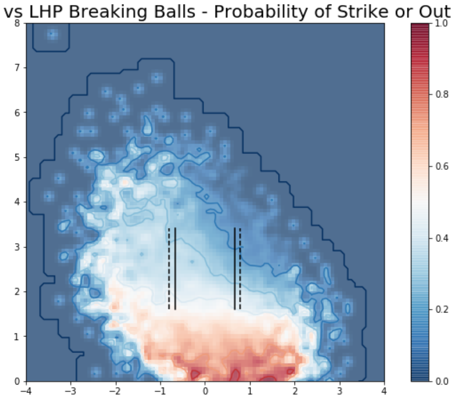
RHH vs LHP Fastballs - Probability of Strike or Out



RHH vs LHP Breaking Balls - Observed Volume

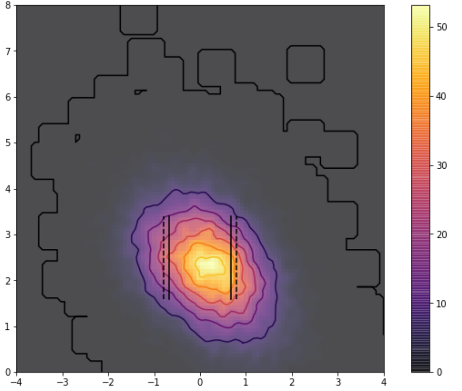


RHH vs LHP Breaking Balls - Probability of Strike or Out

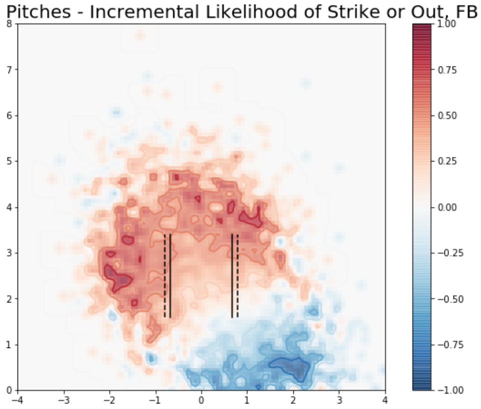


LHH vs LHP

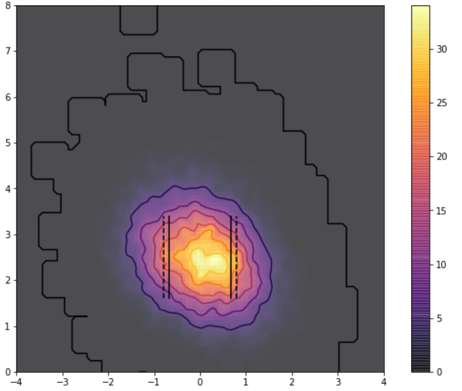
LHH vs LHP All Pitches - Observed Volume



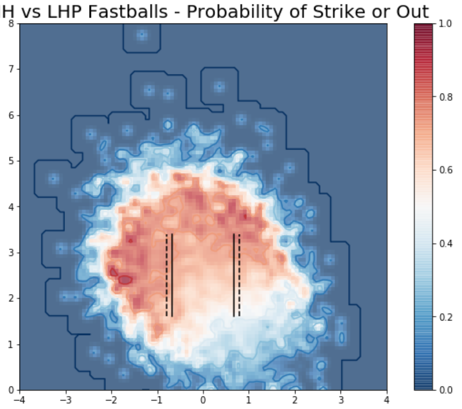
LHH vs LHP All Pitches - Incremental Likelihood of Strike or Out, FB vs. BB



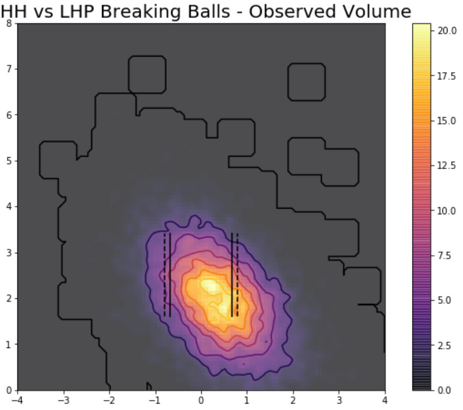
LHH vs LHP Fastballs - Observed Volume



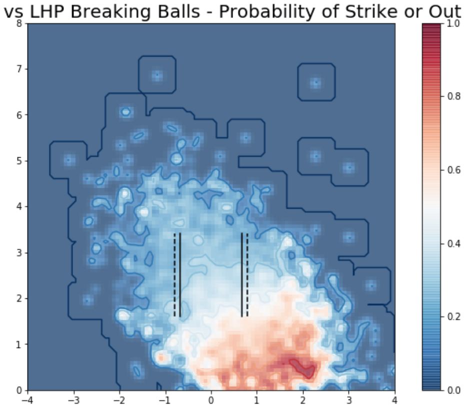
LHH vs LHP Fastballs - Probability of Strike or Out



LHH vs LHP Breaking Balls - Observed Volume

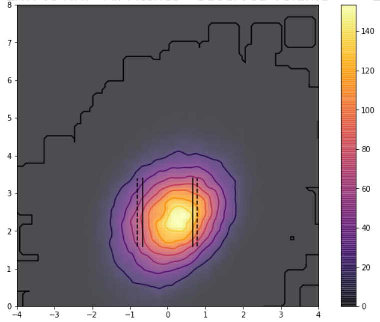


LHH vs LHP Breaking Balls - Probability of Strike or Out

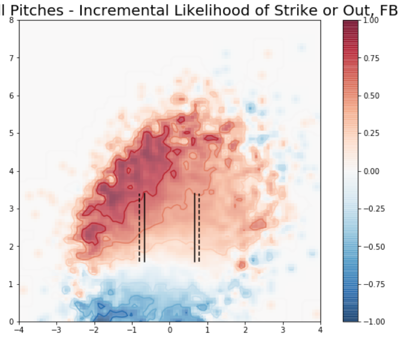


LHH vs. RHP

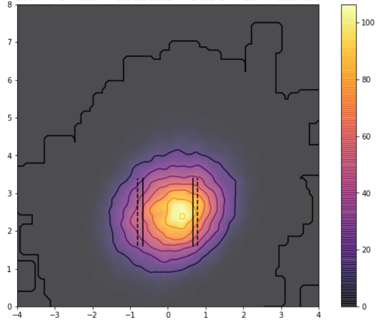
LHH vs RHP All Pitches - Observed Volume



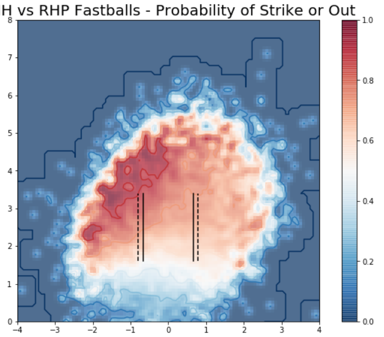
LHH vs RHP All Pitches - Incremental Likelihood of Strike or Out, FB vs. BB



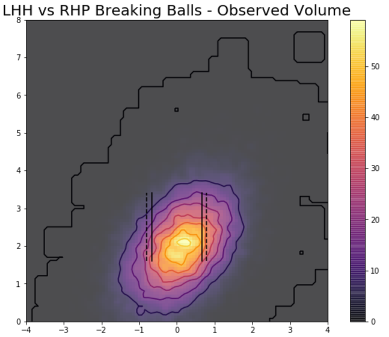
LHH vs RHP Fastballs - Observed Volume



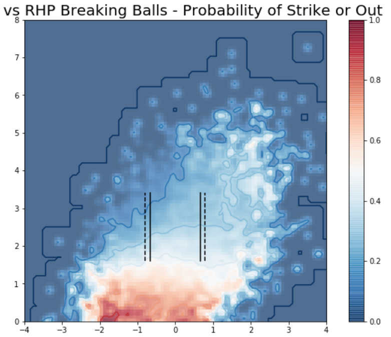
LHH vs RHP Fastballs - Probability of Strike or Out



LHH vs RHP Breaking Balls - Observed Volume



LHH vs RHP Breaking Balls - Probability of Strike or Out


8
7
6
5
4
3
2
1
0
−4
−3
−2
−1
0
1
2
3
4
1.0
0.8
0.6
0.4
0.2
0.0