# comparison of data sets

Alden Bradford

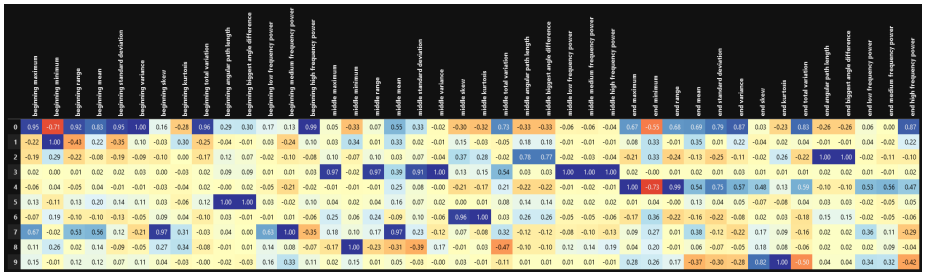
March 28, 2022

**Abstract**

In a collaboration between Weiyi, Matt, and myself, we identified three publicly available datasets which include accelerometer data from controlled falls. Our hope is to be able to transfer information from these controlled studies to our problem domain, or at least see how these other datasets were successfully classified. Here I describe some insights these datasets provide about our problem.

## 1 Factor analysis

Now that we have some other datasets available, we can begin to explore them in the same way we have explored the Makusafe data. One of the things we tried was factor analysis on the features – this allows us to see which features vary together. I scaled each feature to have mean zero and unit variance, then carried out factor analysis. Since the covariance matrix had ten eigenvalues greater than 1, I chose to look for ten factors. Here is a table comparing their factor loadings, scaled linearly so that each factor has a maximum loading of 1.




We can see that several features are redundant, in terms of what they encode. Looking at row 6, for example, we see that the skew and variance are highly correlated for the middle section of the incident. This makes sense, because both should be dominated by the central peak of the acceleration curve.

For each factor, I chose the feature with the greatest loading to include in a plot. Then, I generated a scatterplot for each pairing. The result is in Figure 1.

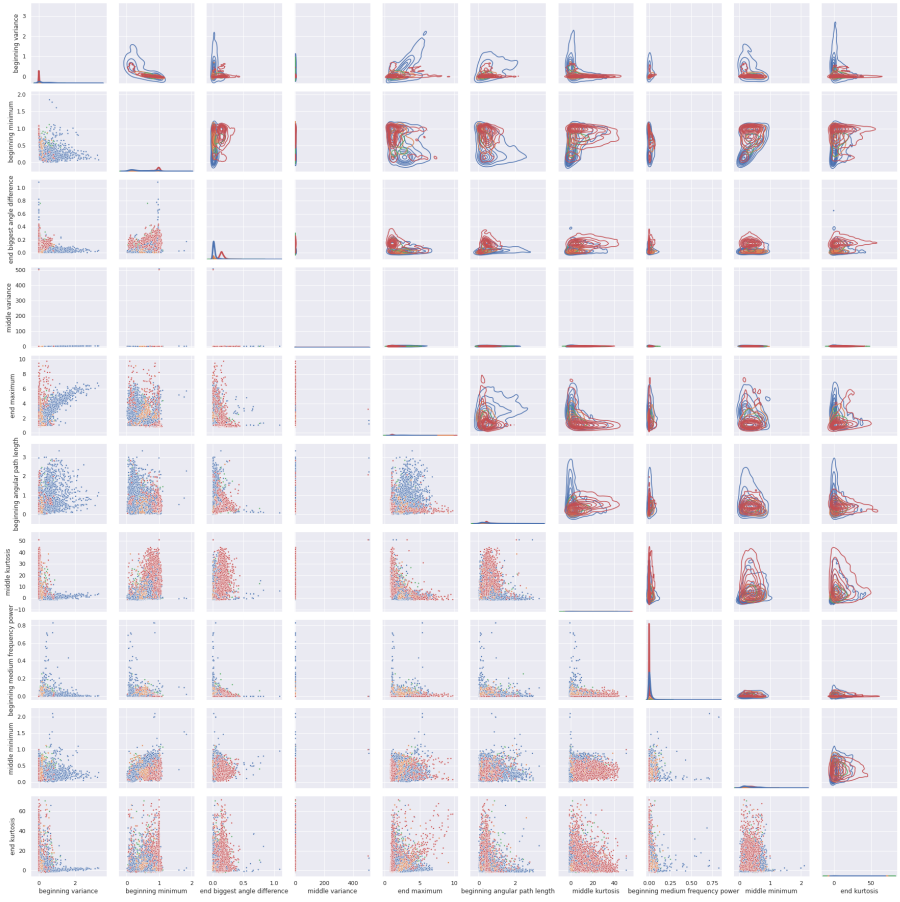



Figure 1: A selection of features. Each point is an incident, each color its motion.

For the rest of the document, we will focus on just one of these figures: the beginning variance versus the middle kurtosis. This one appears to show a good separation between the hazardous and non-harzardous incidents.

## 2 Beginning variance, middle kurtosis

It is important to remember that each chosen feature is just one representative of a class of related features. For example, the beginning variance aligns very well with several other features also computed on the "beginning" portion of the signal (the portion spanning from 4 seconds pre-incident to one second pre-incident). It aligns well with the total variation, as well as the range. We can consider this an indication of how much the acceleration was going up-and-down during whatever activity preceded the incident. Recall that this is strictly looking at the magnitude of the acceleration, so we do not expect it to be affected by rotation or positioning. It makes sense that this would be a negative predictor of hazards; if the person was moving vigorously, their base levels of acceleration would be higher and so they would necessarily have a higher chance of falsely triggering the acceleration threshold.

The kurtosis in the middle is highly correlated with the skew in the middle. The "middle" here means the period extending from one second pre-incident to one second post-incident. A high skew or kurtosis would be caused by a small number of samples which are very different from the majority of samples. We can consider this an index for how high the peak acceleration was, compared to the usual fluctuations in acceleration during this interval.

See the results of this in Figure 2. It is clear that a high variance in the beginning interval is a predictor of a non-hazardous incident. In particular, there were no hazards with a beginning variance greater than 1. Compare this with the kurtosis during the middle interval, and we see that a kurtosis greater than about 15 is a very strong indicator that the incident was hazardous. This pattern shows up independently in each data set we considered; it is not an artifact of combining data from different sources. A similar plot showing only the points from each of our datasets on its own is included in Appendix A.

## 3 The Makusafe data

It is natural to ask: how does the makusafe data compare to this? Let's inspect a plot showing the same features, this time only looking at data from Makusafe. This is in Figure 3. What we see looks familiar, in the sense that is is a similar shape – it is just missing the major non-hazardous branch. We have mostly these high-kurtosis incidents, which in the other three datasets would almost certainly be classified as hazards. We have very few of the high-variance incidents which would be classified as false alarms. What we see instead is that the "true hazards" are all mixed in with the false alarms. There would be no way, based on these two

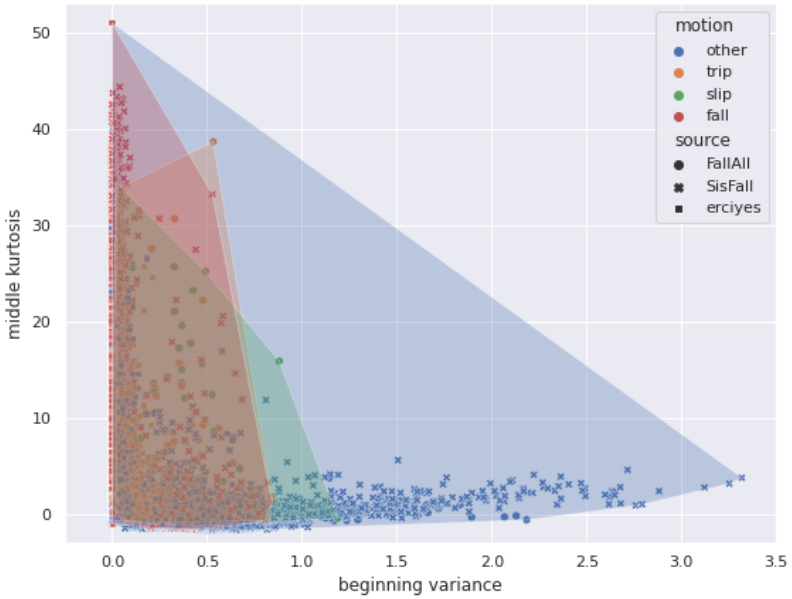



Figure 2: The plot of begiinning-variance versus middle-kurtosis. Notice there are mostly blue dots in the lower-right, and mostly red dots in the upper-left.

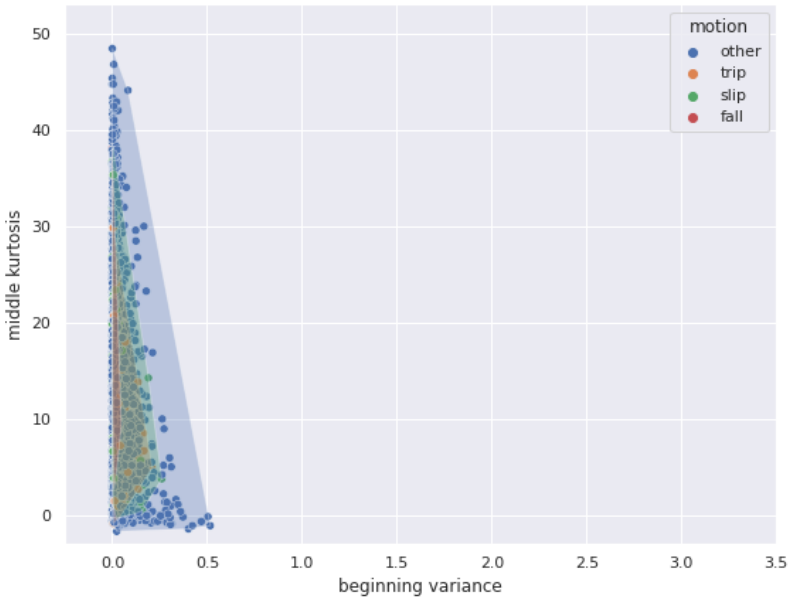



Figure 3: The data from Makusafe

features, to distinguish them.

# 4 Conclusions

This was an exercise in data exploration. There are not many things we can conclude definitively. What follows are a few speculations based on this observation.

## 4.1 The other data sets are fundamentally different – what works for these may not work at all for the Makusafe data.

The other studies we considered relied on computing simple statistics such as variance and skew. We can see from this simple demonstration that this makes sense for the data they have. When comparing these variables, a clear separation emerged.

We can see from this analysis that it would be unreasonable to suspect that same strategy to work for the Makusafe data. The data are simply not as well separated, at least not along the same features. This also suggests that, while support vector machines were appropriate for the

other data sets, they may be inappropriate in our case.

## 4.2 Potential causes

Why do these sets behave so differently? I think it is safe to rule out differences due to position of sensor, or sampling frequency. The Erciyes data set, for example, had the same sampling frequency as the Makusafe data. Each of the other data sets had sensors mounted at a different point on the body – the wrist, the waist, the chest. It would be very strange if the upper-arm was so very different.

I see three potential sources of the difference:

### 4.2.1 controlled vs uncontrolled

The other data sets used simulated falls, while the Makusafe data uses real observed falls. It could be that a simulated fall has very different characteristics from a real fall.

### 4.2.2 wider range of non-hazardous activities

The other data sets had a limited set of activities being performed – walking, running, jumping, standing up, lying down, etc. The activites that a person does in a factory for example may not be well-represented by these so-called "activities of dailiy living".

### 4.2.3 unreliable classification

It could be that many of the incidents in the Makusafe set which are classified as "other" would be better classified as "a fall, but not the kind of fall we are worried about." That is, it may be that a high kurtosis is in fact a good predictor of a fall, and the incidents in the Makusafe set were misclassified because of a misalignment about what we think should constitute an incident which is worth reporting.

## 4.3 Future directions

Based on this exploration, it would be unreasonable to think a suport vector machine, for example, could be trained on these other data sets and give a good model which predicts the ways that the Makusafe data are labeled. We should even reconsider many of our choices for features – we chose them because they worked well for these other data sets, which we now have reason to believe will not lead to a good performance on our data. Alternatively, we should reconsider what our goal should be. Perhaps instead of trying to predict what the labels are, we should try to produce some easily-computed index of risk which could be used in aggregate to identify hazardous locations or activities, for example.

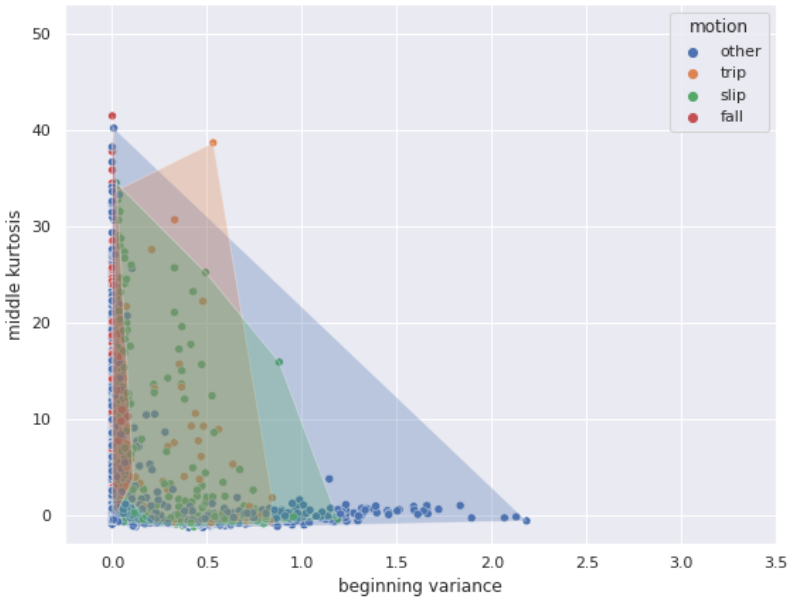



Figure 4: The data from FallAllD

# A Separate plots of other datasets

Since each of these data sets came from a different source, using different accelerometers under different testing cicumstances, it would be reasonable to suspect that the bahavior documented above – the split of the data into two well-separated clumps – could be due to only differences in the data collection methods. However, this does not seem to be the case. To one degree or another, we see this pattern in each data set independently. You can see this for yourself in Figures 4-6.

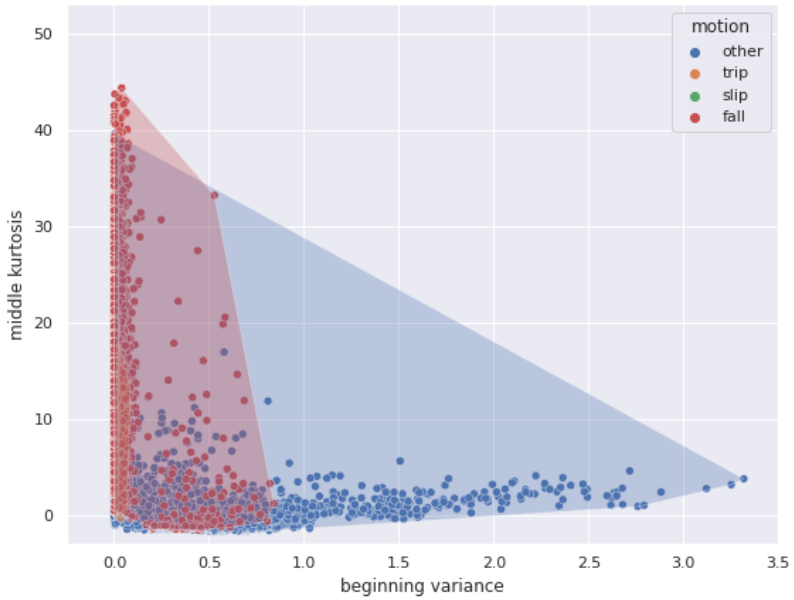



Figure 5: The data from SisFall

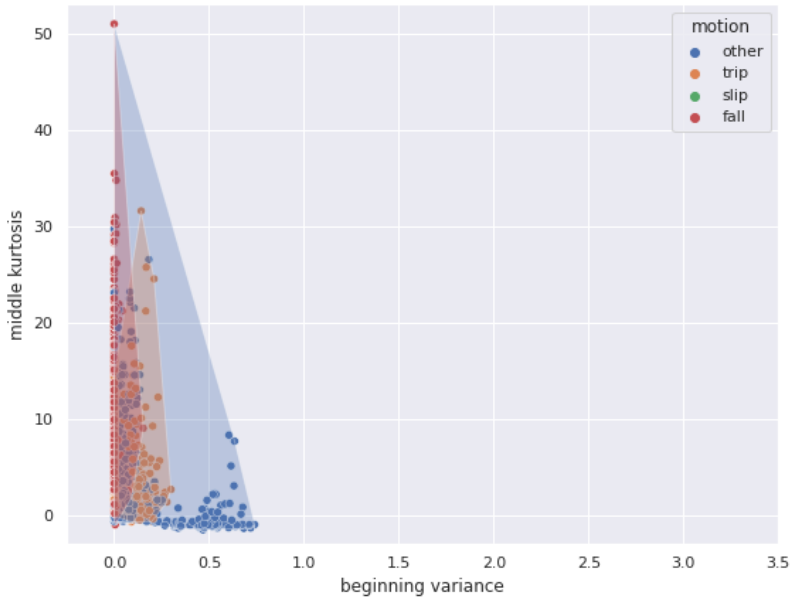



Figure 6: The data from Erciyes