Radioisotope Studies with Principal Component Analysis

Boris Mesits

21 May 2018

Procedure

Here we are trying to determine the relative abundance of radioactive isotopes in a sample, only by measuring decay events with a pair gamma-ray spectrometer. Suppose we have spectrometer data for a sample containing known radioactive isotopes A and B, but we don't know the relative abundance of A and B in the sample. Our goal is an algorithm that tells us the relative abundance if we give it the spectrometer data. Our first step is to gather spectrometer data from samples of known abundances, so that the algorithm can compare these to the unknown sample. In our case, we simulate the spectrometer data using GEANT. We can pick different relative abundances for A and B and generate theoretical data using the simulation. This data takes the form of a 2D histogram, which displays the number of radioactive decay events that fall into a certain energy range for each of the two detectors in the pair spectrometer.

Sometimes a rough qualitative estimation of the relative abundance of A and B can be made just by comparing a few histograms. For example, in Figure 2, the unknown mixture histogram looks to be some combination of the pure ^{238}U histogram and pure ^{60}Co histogram. However, to get precise result, we use a technique inspired by computer vision methods, involving principal component analysis (PCA).

The basic idea of PCA is illustrated in Figure 1. Principal components (PCs) are eigenvectors for a covariance matrix, which effectively define the primary features of a data distribution. A data set has "observables", or data points, and variables that define each data point's location in the variable space. In our case, the "observables" are a few histograms (corresponding to different isotope ratios) and the variables are the values of every bin in the histogram - 40,804 bins in total. Therefore, we perform PCA on a few data points scattered in 40,804-dimensional space. Generally the number of PCs equals the number of dimensions, but this is only the case when there are more data points than dimensions. Otherwise, as in our case, for N data points, there are N-1 PCs. After applying PCA, we can entirely describe all the histogram data with a few PCs, described by 5×1 vectors. We can simplify our analysis by throwing out all but the first, largest PC, which is responsible for the vast majority of variance among histograms. We then find the position of each data point along this first PC, done automatically by MATLAB in an output called "scores", also illustrated in Figure 1. With only one PC, we now can encode the information of an entire histogram in a single number, to a very good approximation (as long as we are dealing with only two isotopes A and B, see the Next Steps section for more isotopes). The practical steps of this procedure are shown in Figure 2 and described in the code documentation.

Once we have the PC score for each histogram, we can plot the PC score versus the isotope ratio in each histogram. As it turns out, this creates a linear relationship. Using the relationship, given a histogram's PC score, we can accurately estimate its isotope ratio. We can pass several histograms, some with known isotope ratios and some without, into the PCA function. Creating a linear relationship with the known-ratio histograms can tell us the isotope ratios of the other.

Results

I used three radioactive isotopes - ^{238}U, ^{60}C, and ^{22}Na - to develop and test our method. I used four GEANT-generated histograms as test cases. The actual abundances, unknown to the algorithm but specified by the

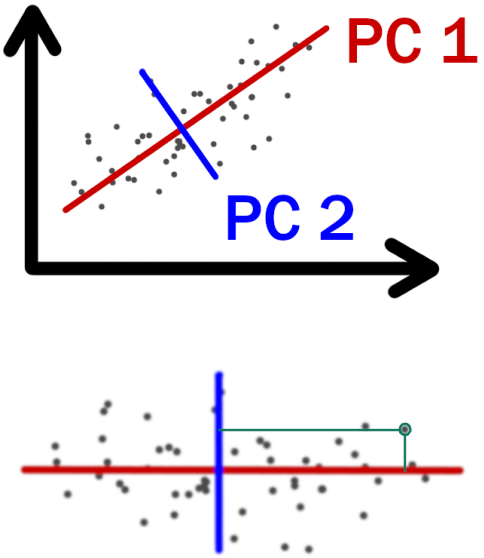

Figure 1: This shows the meaning of the "scores" output of MATLAB's PCA function. For a 2D distribution of data, two principal components can be generated. The position of any data point relative to the coordinates defined by the principal components is the score of that data point. In our case, we are not plotting dozens of data points in two dimensional space, but rather a few data points in 40804-dimensional space.

macro file passed into the GEANT simulation, are compared to the abundances predicted by the algorithm.

Actual Ratio (specified in GEANT)	Actual Abundances	Predicted Abundances
Co 2:1 Na	66.67% Co, 33.33% Na	66.88% Co, 33.12% Na
Co 1:5 Na	16.67% Co, 83.33% Na	16.94% Co, 83.06% Na
Co 2:1 U	66.67% Co, 33.33% U	64.95% Co, 35.05% U
Co 1:10 U	9.09% Co, 90.91% U	9.12% Co, 90.88% U

Table 1: The largest percent error in this table is 5.2%, which demonstrates the ability of the algorithm to obtain meaningful results.

Next Steps

I chose ^{238}U, ^{60}C, and ^{22}Na for testing because their characteristic histograms are visually distinctive, which made for easy troubleshooting - I could confirm a simulation worked as intended just by inspecting the histogram. However, the fact that the three isotopes looked so distinctive may have given the algorithm an easier task. The algorithm will be more thoroughly tested when A and B have very similar histograms.

The algorithm has so far only been tested on simulated data. Since GEANT is not a perfect virtual analog of real detectors, the algorithm should be tested with data from a real spectrometer to see if the results are reasonable. The downside, of course, for real data is that the exact abundances of contained isotopes is not known (it is, after all, the desired unknown), but certain sanity checks ought to be performed.

Furthermore, it is important to expand the use of the algorithm to more than two isotopes at a time. The current algorithm only works when the unknown sample is a combination of two isotopes, but hopefully

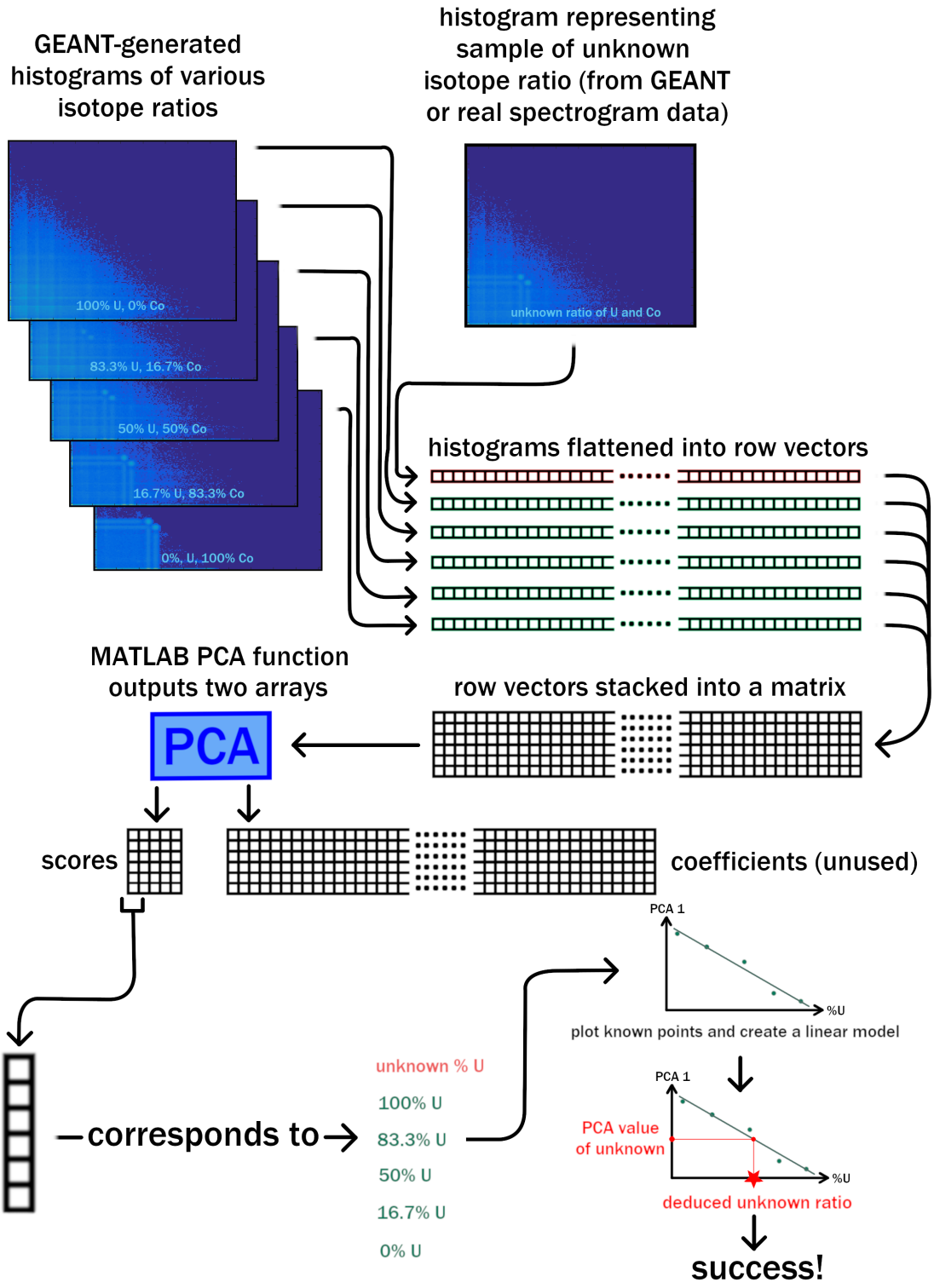

Figure 2: A visualization of how results were obtained. Note that before the matrix is passed into the PCA function, it is standardized (or normalized).

mixtures of several isotopes can be analyzed. This, however, would probably require using more than just the first principal component. For example, one principal component can only contain the information needed to find the ratio between isotopes A and B. A second PC is needed for the ratio between A and C, a third PC for A and D, and so on for the full suite of possible isotopes in a real sample.

References

[1] Spruyt, Vincent. "Feature extraction using PCA". Computer Vision for Dummies. Accessed on May 21, 2018 at `http://www.visiondummy.com/2014/05/feature-extraction-using-pca¿`.

[2] Starmer, Josh. "StatQuest: Principal Component Analysis (PCA) clearly explained." YouTube video. Accessed on May 21, 2018 at `https://www.youtube.com/watch?v=_UVHneBUBW0\&t=1028s`.