31 March 2021. Brief description of the methodology of a paper to be presented at SORMA-2021.
Keywords: Expectation maximization, Bayesian learning, unsupervised learning, clustering, probability mixture models.

Detecting Hidden Signals in a Data Sets Using Probability Density Mixture Models

James B. Cole, james.b.cole.gamma@gmail.com

Many real-world data sets are produced by a superposition of stochastic processes with overlapping probability distributions. The signature of one or more of the processes which generated the data can thus be hidden in the overlap with stronger signals. Using a form of unsupervised learning within the framework of a probability mixture model, these hidden signals can be extracted.

As an example, consider the data set shown in Fig.1. Let *x* be the result of a measurement (for example, energy deposited in a detector by three different particles emitted in a radioactive decay. The histogram tells us how many events fall in each energy bin, but it does not tell us how many particles of each type were detected. It is impossible to know the exact answer, but with a few plausible assumptions we can make a reasonable probabilistic estimate.

Assuming, for example, that the data due a superposition of three probability density functions (pdfs) which we take to be gaussians. We seek the means (μ_i), the standard deviations (σ_i) and the relative weights (w_i), which best describe the data; where $i=1,2,3$. The data are thus modeled by $\rho(x)=\sum_{i=1}^{3} w_i\rho_i(\mu_i,\sigma_i,x)$, where the weights, sum to 1. We need to determine the values of w_i, μ_i, and σ_i which best fit the data.

This can be done using an algorithm called expectation maximization [1] based on the Bayesian methodology Iteratively updating the initial guesses $\left(w_i^0,\mu_i^0,\sigma_i^0\right)$, $\left(w_i^j,\mu_i^j,\sigma_i^j\right)\to\left(w_i^{j+1},\mu_i^{j+1},\sigma_i^{j+1}\right)$, we terminate the process using a χ^2 stop criterion.

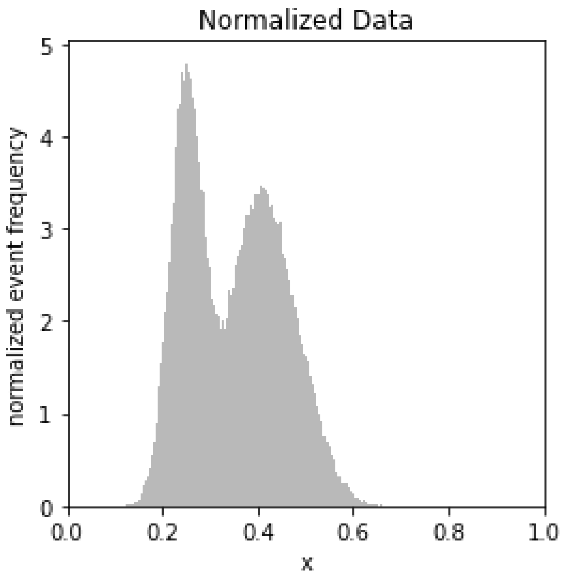

Fig. 1. Normalized data histogram

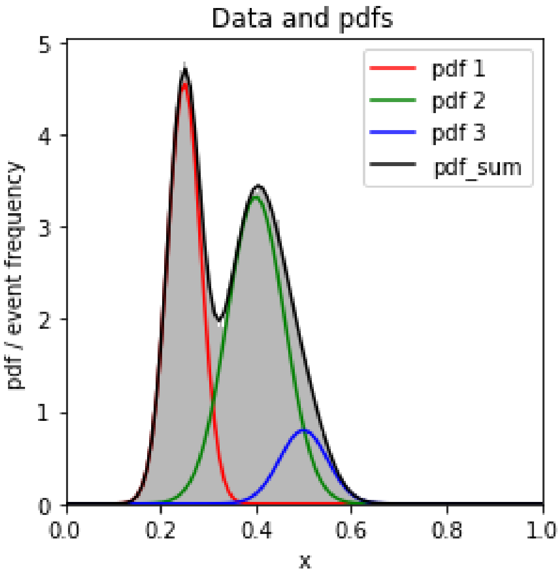

Fig. 2. Detection of hidden signal (blue).

Details of are given in [2].

In Fig. 2, pdf 3 (blue curve) depicts the probability density function that best describes the distribution particle type 3 in the data. Here we have used the *a priori* knowledge that there are only three particle types. If we do not know the number of particle types we could use an information criterion such as the Akaike Information Criterion to produce a best estimate.

Having found the parameters of the ρ_i we can estimate the numbers of each particle type in the data set and hence the decay spectrum. This kind of analysis can be applied to any kind of data. For example, *x* could be the concentration of some blood protein in patients' with three different possible diseases.Then pdf-*i* gives the relative probability of disease *i* for a given value of *x*.

The probability density functions need not be gaussian. We have extended the expectation maximization algorithm to truncated gaussian and lognormal pdfs.

This methodology is a form of cluster analysis and unsupervised machine learning.

References

[1] P. M. Lee, "Bayesian Statistics," Wiley (2012), pp. 283-290.

[2] J. B. Cole, et al. SORMA-2021.