Researchers discover methods to find 'needles in haystack' in data

December 5 2005

A Case Western Reserve University research team from physics and statistics has recently created innovative statistical techniques that improve the chances of detecting a signal in large data sets. The new techniques can not only search for the "needle in the haystack" in particle physics, but also have applications in discovering a new galaxy, monitoring transactions for fraud and security risk, identifying the carrier of a virulent disease among millions of people or detecting cancerous tissues in a mammogram.

Case faculty members Ramani Pilla and Catherine Loader from statistics and Cyrus Taylor from physics report their findings in the article, "A New Technique for Finding Needles in Haystacks: A Geometric Approach to Distinguishing between a New Source and Random Fluctuations," December 2, in the journal, *Physical Review Letters*.

"As haystacks of information grow ever larger--and the needles ever smaller--the search for a signal becomes increasingly difficult to find using traditional approaches. There is a need for sophisticated new statistical methods," the researchers report.

Researchers working with large amounts of data encounter the fundamental problem of determining a real signal from random variation in the data. In many practical problems, a suspected signal may only be a small blip in a noisy experimental background.

The Case team discovered a technique that is built on the principle of

comparing a set of summary characteristics for any sub region of the observations with the background variation. From these characteristics, attempts are made to find small regions that appear significantly different from the background--a difference that cannot simply be attributed to random chance.

"Methods used in high-energy particle physics problems traditionally have searched for any departure from a background model; that is, anything that is not a haystack," said Pilla, the project leader. "Our method efficiently incorporates information about the type of disorder expected, thereby enabling us to find the signal of interest more accurately."

At the core of the breakthrough is the idea of posing the problem in terms of a "hypothesis-based testing" paradigm to detect statistical disorder in the data. The method further exploits the flexibility behind a long-established geometric formula in creating a technique that significantly enhances the ability to distinguish a signal.

The researchers said the challenge is two-fold: defining efficient test statistics, and determining the critical cut-off. That is, to help the scientist find what is random variation as opposed to what is the signal. The detection problem involves a large number of comparisons, and the researchers caution that experimentalists should not be fooled into false discoveries by random variation.

"The experimenter wants to control the experiment-wise error rate: if there is nothing in the data, then there must be minimal probability of falsely discovering a signal. On the other hand, we want to maximize our chance of discovering any real signal that may be present in the massive data set," said Loader.

"The probabilistic problem associated with this scenario is reduced to

one of finding the areas of certain regions on the surface of high-dimensional spheres," explains Pilla.

The Case researchers then exploit the geometric methods pioneered in 1939 by Harold Hotelling and Hermann Weyl. They tested the statistical techniques by using computer simulated particle physics experiments that mimic the real experiments conducted in colliders to demonstrate that the new technique significantly increased detection probabilities.

"In high-energy particle physics and astrophysics problems, chi-square goodness-of-fit tests are widely employed, although they have relatively low power to detect the signal," notes Taylor. "Through my collaborative work with Professors Pilla and Loader, we will be able to develop powerful statistical tests for detecting a signal from noisy data with high probability, a fundamental problem encountered in many scientific disciplines."

Taylor added that "conducting experiments in a particle collider may cost tens of millions of dollars. Improving efficiency in the analysis of experimental results can lead to enormous cost savings. Furthermore, we can obtain the same results with much smaller experiments, or effectively find much smaller departures from the background model."

"Detecting a real signal (the needle) present in random and chaotic data (the haystack) will lead to scientific success," conclude the researchers.

Source: Case Western Reserve University