Tangled relationships unpicked

A statistical method discovers hidden correlations in complex data.

Philip Ball

15 December 2011

The US humorist Evan Esar once called statistics the science of producing unreliable facts from reliable figures. An innovative technique now promises to make those facts a whole lot more dependable.

Brothers David Reshef of the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, Yakir Reshef, now at the Weizmann Institute of Science in Rehovot, Israel, and their coworkers have devised a method to extract from complex sets of data relationships and trends that are invisible to other types of statistical analysis. They describe their approach in *Science* today[1].

"This appears to be an outstanding achievement," says Douglas Simpson, a statistician at the University of Illinois at Urbana–Champaign. "It opens up whole new avenues of inquiry."

Donald Miralle / Getty Images

Baseball produces a welter of data, from which correlations can be drawn – for example between the number of hits and a player's salary.

Dizzying complexity

Here is the basic problem. You have collected lots of data on some property of a system that could depend on many governing factors. To work out what depends on what, you plot them on a graph.

If you are lucky, you might find that one property changes in a simple way as a function of another factor: for example, people's health might steadily get better as their wealth increases. There are well known statistical methods for assessing how reliable such correlations are.

But what if there are many simultaneous dependencies in the data? Suppose that you are looking at how genes interact in an organism. The activity of one gene could be correlated with that of another, but there could be hundreds of such relationships all mixed together. To a cursory inspection, the data might look like random noise.

"If you have a data set with 22 million relationships, the 500 relationships in there that you care about are effectively invisible to a human," says Yakir Reshef.

And the relationships are all the harder to tease out if you don't know what you're looking for in the first place — if you have no reason to suspect that one thing depends on another.

The statistical method that Reshef and his colleagues have devised aims to crack those problems. It can spot many superimposed correlations between variables and measure exactly how tight each relationship is, on the basis of a quantity that the team calls the maximal information coefficient (MIC). The MIC is calculated by plotting data on a graph and looking for all ways of dividing up the graph into blocks or grids that capture the largest possible number of data points. MIC can then be deduced from the grids that do the best job.

To demonstrate the power of their technique, the researchers applied it to a diverse range of problems. In one case they looked at factors

that influence people's health globally, using data collected by the World Health Organization in Geneva, Switzerland. Here they were able to tease out superimposed trends — for example, female obesity increases with income in the Pacific Islands, where it is considered a sign of status, but there is no such link in the rest of the world.

In another example, the researchers identified genes that were expressed periodically, but with differing cycles, during the cell cycle of brewer's yeast (*Saccharomyces cerevisiae*). They also uncovered groups of human gut bacteria that proliferate or decline when diet is altered, finding that some bacteria are abundant precisely when others are not. Finally, the team identified performance factors for baseball players that are strongly correlated to their salaries.

Correlation and causation

Reshef cautions that finding statistical correlations is only the start of understanding relationships between variables. "At the end of the day you'll need an expert to tell you what your data mean," he says. "But filtering out the junk in a data set in order to allow someone to explore it is often a task that doesn't require much context or specialized knowledge."

He adds, "Our hope is that this tool will be useful in just about any field that is amassing large amounts of data." Reshef points to genomics, proteomics, epidemiology, particle physics, sociology, neuroscience and atmospheric science as just some of the fields that are "saturated with data". The method should also be valuable for 'data mining' in sports statistics, social media and economics.

One of the big questions remaining after a relationship has been uncovered is what causes what; the familiar mantra of statisticians is that correlation does not imply causality. "We see the issue of causality as a potential follow-up," says Reshef. "Inferring causality is an immensely complicated problem, but has been well studied previously."

Raya Khanin, a bioinformatician at the Memorial Sloan–Kettering Cancer Center in New York, acknowledges the need for a technique like the Reshefs', but reserves judgement about whether the MIC is the answer. "I'm not sure whether its performance is as good as and different from other measures," she says.

For example, she questions whether the findings about gut bacteria really needed this advanced statistical technique. "Having worked with this type of data, and judging from the figures, I'm quite certain that some basic correlation measures would have uncovered the same type of non-coexistence behaviour," she says.

Nature | doi:10.1038/nature.2011.9660

References

1. Reshef, D. N. *et al*. *Science* **334**, 1518–1524 (2011).

Nature ISSN 0028-0836 EISSN 1476-4687