April 15, 2021

Maths, Physics & Chem

Recovering data you have never seen

by Rik Voorhaar[1] | PhD Student

doi.org/10.25250/thescbr.brk513

[1]: Univeristy of Geneva, Geneva, Switzerland

This Break was edited by Ayala Sela, Scientific Editor - TheScienceBreaker

Imagine you spilled coffee over the spreadsheets containing your research data, and a large part of the data is completely unreadable. Before you throw away your data in despair, it might in fact be possible to recover it with the help of a bit of math.

Image credits: Pexels

What do you do when some of your data is missing? Imagine taking pictures with a camera and discovering some of the pixels are missing on all of the pictures. In this case, you can probably guess what the missing pixels should be using the surrounding pixels. In fact, if you replace the pixel with the average of the surrounding pixels, you probably won't be able to tell it was missing in the first place.

The problem of reconstructing missing values from a table (matrix) of data is called matrix completion. The reason we can solve this problem is that real data isn't random – it contains patterns. If we use the picture example again, it only works when the image has meaning. If it only contained random noise - like the snow pattern in an old television - then we would have no idea what to do with the missing pixels. Only if we can understand and model the patterns in data, can we recover missing values accurately.

Problems of recovering missing data are very common in science and industry. Services like Netflix use it to predict what kind of shows you like. They know which shows you and other people have watched (the known values), and using patterns from other people, they can recommend new shows to you (the unknown values). The authors of this study took the problem one step further: what if not only a large amount of data is missing, but a large part of the data we observe contains outliers - that is, random noise? With the right assumptions, we can still precisely recover the original data! But as with any mathematical problem, the assumptions are key.

In this case we need three assumptions. The first is that the missing values are spread throughout the data. If we have an image of a person, and random pixels are missing, we can probably recover the missing bits. However, if the entire face of the person in the image is missing, we cannot guess what they look like. The second assumption is that the original data should have a low rank. This assumption is related to the amount of information contained in the data. For a picture, this means that there should not be too many fast-changing or complex patterns; it is easier to recover an image of a clear blue sky than an image of a busy crowd. Finally, the third is that there is a limited number of missing values or outliers. Like the snow pattern of old television, if the data consists entirely of outliers or missing data, we are obviously at a loss.

Under these assumptions, we can recover missing bits of data by using an algorithm called the gradient descent algorithm. You can compare this algorithm to climbing a mountain. One strategy for reaching the summit of a mountain is to always step in the direction where the slope is the steepest. If you keep doing this, you either end up at the summit, or some smaller peak from where the only direction you can go is down. When recovering missing data, we start with a simple model, which takes in the known bits and is able to make predictions for the missing data. Our model of choice will use the three assumptions mentioned before in order to make good predictions for the missing data while automatically ignoring outliers.

At first the model's prediction of the missing data is not very accurate. To improve it, we evaluate the difference between the model's predictions and the known data and use it to find the best "direction" we need to move in. Then we take a small "step" in this direction and reevaluate the results. By repeating this multiple times, the predicted data will agree more with the known data. The "summit" we want to reach corresponds to the point where the predicted bits and the known bits are the most similar. Hopefully, at this "summit" the prediction will also closely match the original data – the data without outliers and missing data.

We then require two things from the algorithm. First, we want a guarantee that so long as we start close enough to the right solution, we will arrive at the right solution, that is, an exact recovery of all the original data without missing values and outliers. Secondly, we want to reach the solution quickly. Surprisingly, the conditions needed for these two requirements are identical, and are precisely the three assumptions mentioned earlier. Furthermore, when these three assumptions are met, up to half of the known data can consist of outliers, and still be recovered accurately.

After showing the theoretical effectiveness of this method for recovering missing data, the researchers used this method on an artificial test problem, and found that the algorithm exceeds the theoretical expectations. In fact, it seems that even if we don't start close to the right solution at all, we still end up reaching it.

There are many variations on the matrix completion problem, dealing with many different situations. What is nice about this study is that we get useful theoretical properties while making natural assumptions on the data. And so now you know: next time you spill coffee over your spreadsheets, you can give your local mathematician a call!