Big Data: a new era for Statistics

Richard J. Samworth

Abstract

Richard Samworth (1996) is a Professor of Statistics in the University's Statistical Laboratory, and has been a Fellow of St John's since 2003. In 2012 he was awarded a five-year Engineering and Physical Sciences Research Council Early Career Fellowship – a grant worth £1.2million – to study 'New challenges in high-dimensional statistical inference'.

'Big Data' is all the rage in the media these days. Few people seem to be able to define exactly what they mean by it, but there is nevertheless consensus that in fields as diverse as genetics, medical imaging, astronomy, social networks and commerce, to name but a few, modern technology allows the collection and storage of data on scales unimaginable only a decade ago. Such a data deluge creates a huge range of challenges and opportunities for statisticians, and the subject is currently undergoing an exciting period of rapid growth and development. Hal Varian, Chief Economist at Google, famously said in 2009, 'I keep saying the sexy job in the next ten years will be statisticians'. This might raise eyebrows among those more familiar with Mark Twain's 'lies, damned lies and statistics', but there's no doubt that recent high-profile success stories such as Nate Silver's predictions of the 2012 US presidential election have given statisticians a welcome image makeover.

Let me begin by describing a simple, traditional statistical problem, in order to contrast it with today's situation. In the 1920s, an experiment was carried out to understand the relationship between a car's initial speed, x, and its stopping distance, y. The stopping distances of fifty cars, having a range of different initial speeds, were measured and are plotted in Figure 1(a). Our understanding of the physics of car braking suggests that y ought to depend on x in a quadratic way, though from the figure we see that we can't expect an exact fit to the data. We therefore model the relationship as

$$y = ax + bx^2 + \epsilon,$$

where ϵ represents the statistical error. Our aim is to estimate the unknown 'parameters' a and b, which reflect the strength of the dependence of y on each of x and x^2. Here we don't need to

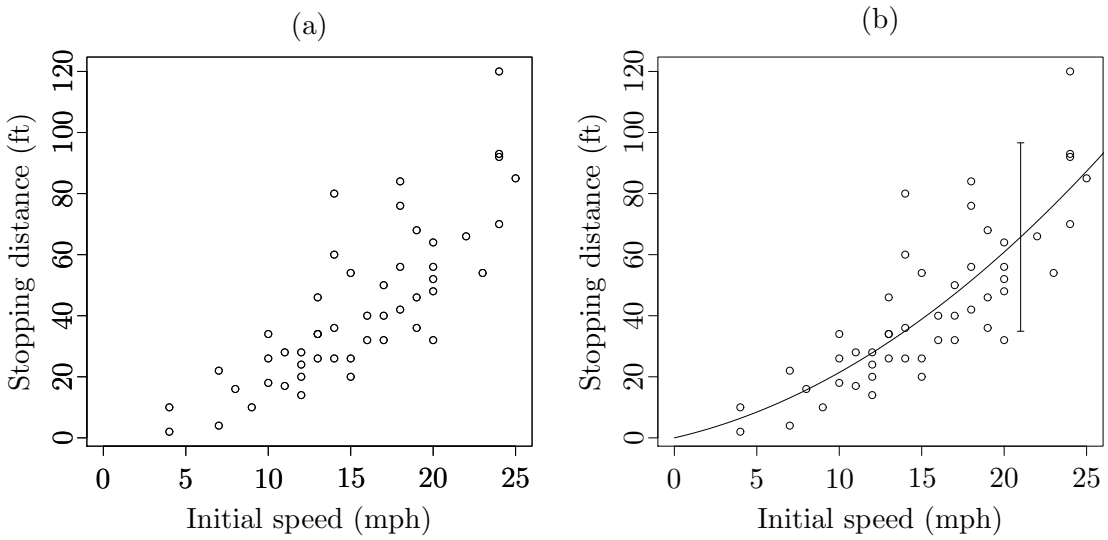

Figure 1: Panel (a) gives the stopping distances of 50 cars having a range of different initial speeds. Panel (b) also shows the fitted curve, as well as a 95 per cent prediction interval for the stopping distance of a car having an initial speed of 21mph.

include a constant term in the quadratic, because a car with zero initial speed doesn't take long to stop.

We would like to choose our estimates of a and b in such a way that the curve $ax + bx^2$ reflects the trend seen in the data. For any such curve, we can imagine drawing vertical lines from each data point to the curve, and a standard way to estimate a and b is to choose them to minimise the sum of the squares of the lengths of these lines. For a statistician, this is a straightforward problem to solve, yielding estimates $\hat{a} = 1.24$ and $\hat{b} = 0.09$ of a and b respectively, and the fitted curve displayed in Figure 1(b). From this, we can predict that a car with initial speed 21mph would take 65.8ft to stop. In fact, we can also quantify the uncertainty in this prediction: with 95 per cent probability, a car with this initial speed would take between 34.9ft and 96.6ft to stop. (Incidentally, a modern car would typically take around 43ft to stop at that initial speed.)

Of course, one can easily imagine that the stopping distance of a car would depend on many factors that weren't recorded in this experiment: the weather and tyre conditions, the state of the road, the make of car, and so on. Such additional information should allow us to refine our model, and make more accurate predictions, with less uncertainty.

My point is that, by contrast with the 1920s, in today's world we can often record a whole raft of variables whose values may influence the 'response' of interest. In genetics, for instance,

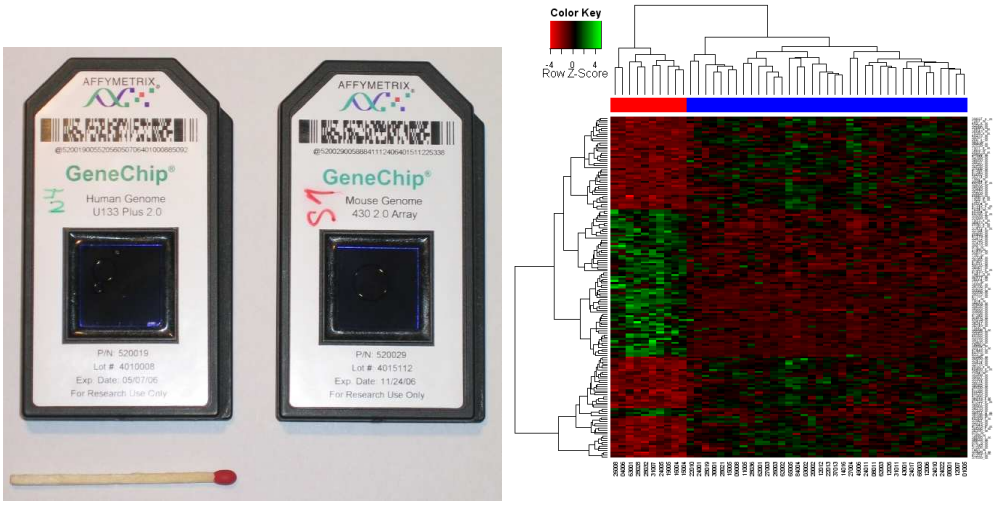

Figure 2: The left panel shows a photograph of a typical microarray. The right panel shows the complexity of the output from a typical microarray experiment.

microarrays (see Figure 2) are used in laboratories to measure simultaneously the expression levels of many thousands of genes in order to study the effects of a treatment or disease. An initial statistical model analogous to our car model, then, would require at least one variable for each gene. Interestingly, for microarray data, there may still be only around fifty replications of the experiment, though many other modern applications have vast numbers of replications too. Suddenly, estimating the unknown parameters in the model is not so easy. The method of least squares we used with our car data, for example, can't be used when we have more variables than replications. What saves us here is a belief in what is often called *sparsity*: most of the genes should be irrelevant for the particular treatment or disease under study. The statistical challenge, then, is that of 'variable selection' – which variables do I need in my model, and which can I safely discard?

Many methods for finding these few important needles among the huge haystack of variables have been proposed over the last two decades. One could simply look for variables that are highly correlated with the response, or use the exotically-named 'Lasso' (Tibshirani, 1996), which can be regarded as a modification of the least squares estimate. In Shah and Samworth (2013), we gave a very general method for improving the performance of any existing variable selection method: instead of applying it once to the whole dataset, we showed that there are advantages to applying it to several random subsamples of the data, each of half the original sample size, eventually choosing the variables that are chosen on a high proportion of the subsamples. We were able to prove bounds that allow the practitioner to choose the threshold for this proportion to control the trade-off between 'false negatives' and 'false positives'.

Another problem I've worked on recently is 'classification'. Imagine that a doctor wants to diagnose diabetes. On a sample of diabetics, she makes measurements that she thinks are relevant for determining whether or not someone has the disease. She also makes the same measurements on a sample of non-diabetics. So when a new patient arrives for diagnosis, she again takes the same measurements. On what basis should she classify (diagnose) the new individual as coming from the diabetic or non-diabetic population? From a statistical point of view, the problem is the same as that encountered by banks that have to decide whether or not to give someone a loan, or an email filter that has to decide whether a message is genuine or spam.

One can imagine that an experienced doctor might have a notion of distance between any two individuals' sets of measurements. So one very simple method of classification would be to assign the new patient to the group of his nearest neighbour (i.e. the person closest according to the doctor's distance) among all n people, say, in our clinical trial. But intuitively, we might feel there was too much chance about whether or not the nearest neighbour happened to be diabetic, so a slightly more sophisticated procedure would look at the patient's k nearest neighbours, and would assign him to the population having at least half of those k nearest neighbours. In Hall, Park and Samworth (2008), we derived the optimal choice of k, in the sense of minimising the probability of misclassifying the new individual. For those interested, it should be chosen proportional to $n^{4/(d+4)}$, where d is the number of measurements made on each individual.

An obvious drawback of the k-nearest neighbour classifier is that it gives equal importance to the group associated with the nearest neighbour as it does the kth nearest neighbour. This observation prompts us to consider weighted nearest neighbours, with weights that decay as one moves further from the individual to be classified. In Samworth (2012), I derived the optimal weighting scheme, as well as a formula for the improvement attainable over the unweighted k nearest neighbour classifier. It's between a 5 per cent and 10 per cent improvement when $d \leq 15$, which might not seem like much, until it's you that requires the diagnosis!

Five years ago, I set up the Statistics Clinic, where once a fortnight any member of the University can come and receive advice on their statistical problems from one of a team of helpers (mainly my PhD students and post-docs). The sheer range of subjects covered and the diversity of problems they present to us provide convincing evidence that statistics is finally being recognised for its importance in making rational decisions in an uncertain world. The twenty-first century is undoubtedly the information age – even Mark Twain would agree!

Professor Richard J. Samworth

References

P. Hall, B. U. Park and R. J. Samworth 'Choice of neighbor order in nearest-neighbor classification', *Annals of Statistics*, **36** (2008), 2135–2152.

R. J. Samworth 'Optimal weighted nearest neighbour classifiers', *Annals of Statistics*, **40** (2012), 2733–2763.

R. D. Shah and R. J. Samworth 'Variable selection with error control: Another look at Stability Selection', *Journal of the Royal Statistical Society, Ser. B*, **75** (2013), 55–80.

R. Tibshirani 'Regression shrinkage and selection via the Lasso', *Journal of the Royal Statistical Society, Ser. B*, **58** (1996), 267–288.