How many different spectra are there? Can we learn anything from consideration of this question?

Anthony M.C. Davies

Norwich Near Infrared Consultancy, 75 Intwood Road, Cringleford, Norwich NR4 6AA, UK
td@nnirc.co.uk

Introduction

Is there a limit to the possible number of different near infrared (NIR) spectra? At first this may seem a very silly question but if we want to define the limits of NIR spectroscopy it may be a question worth considering.

The number of different spectra would be a measure of the information content of the NIR region. The information content of the NIR region must be limited so it follows that there is a limit to the number of different spectra. Can we make an estimate of that number?

Let us begin with a single datapoint measured at some wavelength and displayed, in Figure 1, on a log 1/R scale.

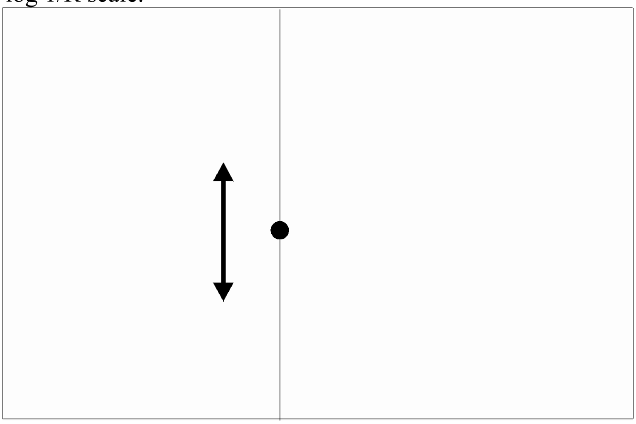

Figure 1. A single datapoint measure on a log 1/R scale.

The measurement of this datapoint is subject to error which is estimated from the noise value of the instrument. A typical value for noise in a modern spectrometer is 30 micro Au (mAu). This is a standard deviation so if we want to be very conservative we can say that the point could be +/- 120 mAu from its measured position. Let's call it a span of 250 mAu. If we limit our absorption scale to a value of 1.0 Au [=1,000,000 mAu], then we could specify 4,000 different levels which a datapoint could occupy. This means that if we only measured at one wavelength we could have 4,000 different spectra. If we made measurements at two wavelengths then we would have a potential of 4,000 x 4,000 = 4000^2 = 16x106 different spectra. But we have typically 700 datapoints in a spectrum that spans 1100 to 2500 nm. Does this mean that we have the potential for 4000^{700} different spectra? The answer must be NO! We have to think about the wavelength scale. We are making measurements at 2 nm intervals and the precision of these measurements is probably better than +/- 0.1 nm. So there is no possibility of confusion between consecutive datapoints on the wavelength scale. But there is a problem and it is called collinearity. Collinearity means that the value of a datapoint is very similar to the proceeding datapoint. It is measured by the correlation coefficient and in order to see its importance we need to look at some real spectra.

Computations and Results

The spectra shown in Figure 2 come from Karl Norris' collection of spectra1 which he kindly made available. This particular selection of twelve spectra was used in a recent study[2] of data compression by Fourier transform or wavelets. It was chosen to be a diverse set of spectra and thus it represents an extreme set of samples. Any random set of samples would be much more similar. The spectra have been limited to the 1100 to 2500 nm range. The datapoint interval is 1.6 nm.

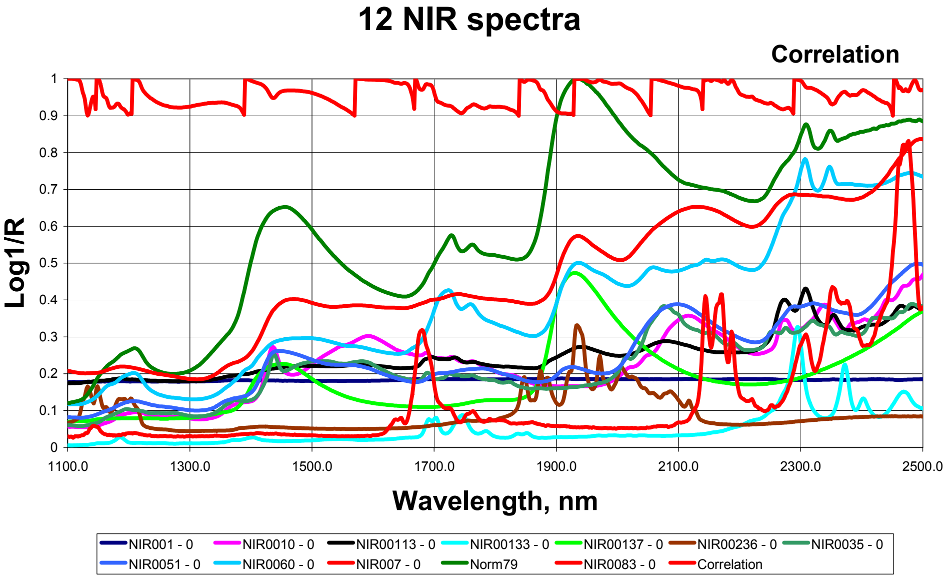

Figure 2. Twelve spectra from Karl Norris' collection.

In addition to the spectra, Figure 2 also contains a line of correlation value. We start by calculating the correlation of the twelve datapoints at 1100 nm with the 12 at the next wavelength (1101.6) and so on until the correlation drops below at determined value. For this experiment a correlation value of 0.9 was chosen as the limit. When the limit is reached we then start again comparing the values at this new wavelength with successive measurements until the correlation is again below 0.9. With this data set, this limit was reached at eleven values: 1147, 1206, 1390, 1571, 1669, 1840, 1930, 2056, 2141, 2290 and 2453. These eleven wavelengths are not free of collinearity (there is no way of deciding what the limit should be) but it does indicate that rather than 700 wavelength measurements, we have less than eleven independent measurements.

If we did have 11 then we would estimate that there could be $4{,}000^{11} = 4.19 \times 10^{39}$ different spectra!

This is a very large number. Can we reduce it?

Well the absorption scale, for practical NIR measurements, does not extend from 0 to 1.0 over the whole wavelength range. It is normally low at 1100 nm and much higher at 2500 nm. Perhaps the average practical range is 0.75 Au, which gives us 3,000 different levels. If we lower the correlation requirement we can easily get to a value of eight for the number of different wavelength measurements so a very conservative estimate would be:

$$3{,}000^8 = 6.56 \times 10^{27}$$

This is still a very large number!

Conclusions

What can we learn from this?

- That the NIR region has incredible potential for discrimination.
- That the very low noise on the absorption measurement is very important.
- Why the Technicon InfraAlyzer with 19 filters was such a successful instrument!

References

1. P. Williams and K.H. Norris (eds), Near-Infrared Technology in the Agricultural and Food Industries, , 2nd edition, American Association of Cereal Chemists, St. Paul, USA (2001).
2. T. Fearn and A.M.C. Davies, J. Near Infrared Spectrosc., 11, 3 (2003).