Speaker: Mark Berman, Commonwealth Scientific and Industrial Research Organisation (CSIRO)

Title: Group Subset Selection and Non-Negative Group Subset Selection With Applications in Spectroscopy

Abstract:

Every material has a distinctive spectrum. The spectrum of a material tells us about its chemistry. A typical reflectance spectrum consists of several hundred numbers representing the amount of light reflected at several hundred wavelengths. At each wavelength, the measurement is actually a (weighted) average over a pixel. Commonly, the material covered by the pixel actually consists of several materials, so that the spectrum measured from a single pixel actually represents the spectrum of a mixture of materials.

Over many years, we have been developing fast algorithms and software for “unmixing” spectra into their constituent parts, especially for tens of thousands of mineral spectra measured in individual drill holes. The algorithms are based on relatively simple linear mixture models. The weights in the model tell us (in a semi-quantitative way) about how much of each material is present in the mixture. Therefore it is natural to constrain the weights to be non-negative. Typically we build libraries of pure materials and unmix against them. For instance, in our main application, we have built a library of shortwave infrared (SWIR) spectroscopy, consisting of 60 materials (mostly minerals), with between 2 and 12 spectra per material class, in order to give us an estimate of typical within-class variation. After certain background corrections, we use the mean of the 60 materials in our model. However, in SWIR mineral spectra, more than 3 materials are seldom apparent, even to experts. So, for each spectrum in our data sets, until recently we have used Alan Miller’s variable subset selection (VSS) code (the basis of the leaps package in R) to choose best subsets of sizes 1, 2 and 3 and then certain ad hoc rules to estimate the correct subset size. The software implementing this algorithm, called The Spectral Assistant, has been incorporated in a commercial CSIRO package called The Spectral Geologist.
The variation in some of the classes is insufficiently modelled by a single spectrum. Therefore, more recently, we have investigated fitting a mixture of two materials in some classes in our model. This leads naturally to the concept of group subset selection (GSS), where a group consists of two spectra per class. We have refined and extended Miller’s code to deal with groups of variables, including (optionally) constraining the weights to be no less than specified lower bounds (e.g. zero). GSS is seldom much slower than VSS, and in some cases it is much faster.

Even more recently, we have been working with thermal infrared spectra, which are far more variable than SWIR spectra. So we need to model many classes as mixtures of up to five materials. GSS often fails here, because it will omit groups where any of the fitted weights is negative. So we have developed non-negative GSS (NNGSS), where the group weights are constrained to be non-negative, by embedding non-negative least squares within the GSS framework. Not surprisingly, NNGSS is much slower than either VSS or GSS. So we have been investigating screening techniques to speed up the unmixing.

In this presentation, I will introduce VSS, GSS and NNGSS and compare them with versions of the Lasso and Group Lasso with non-negativity constraints imposed. I will also discuss possible screening techniques for NNGSS.

Joint work with Yi Guo, Junbin Gao, and Andy Green.