Resistant Multiple Sparse Canonical Correlation with High-Throughput Data

Canonical Correlation Analysis (CCA) is a multivariate technique that takes two datasets and forms the most highly correlated possible pairs of linear combinations between them. Each subsequent pair of linear combinations is orthogonal to the preceding pair, meaning that new information is gleaned from each pair. By looking at the magnitude of coefficient values, we can find which variables can be grouped together, thus better understanding multiple interactions that are otherwise difficult to compute or grasp intuitively.

CCA appears to have quite powerful applications to high throughput data, as we can use it to discover, for example, relationships between gene expression and gene copy number variation. One of the biggest problems of CCA is that the number of variables (often upwards of 10,000) makes biological interpretation of linear combinations nearly impossible. To limit variable output, we have employed a method known as Sparse Canonical Correlation Analysis (SCCA), while adding estimation which is resistant to extreme observations or other types of deviant data. We demonstrate the success of resistant estimation in variable selection using SCCA. Additionally, we have used SCCA to find multiple canonical pairs for extended knowledge about the datasets at hand. Again, using resistant estimators provided more accurate estimates than standard estimators in the multiple canonical correlation setting.

For more information, visit the Statistics Seminar Series website or contact Lisa Stieler at lstieler@uci.edu or (949) 824-3806.