PRIM-RPCA:
A NOVEL RECURSIVE BUMP HUNTING STRATEGY FOR HIGH DIMENSIONAL DATA

J-E. Dazard[1], J.S. Rao[1]

[1]*Case Western Reserve University, Cleveland, USA*

Email: *jxd101@case.edu*

The task in bump hunting is to locate regions of a multidimensional input space where the target function assumes local maxima. The search for these structures (bumps) in the data is important as these often reveal underlying phenomena. In addition, one has to deal today with high dimensional, noisy and correlated data, which often happens with genomic data. To that end, we introduce a recursive partitioning procedure for bump hunting, based on a tree-based method (CART) and a supervised approach known as the Patient Rule Induction Method (PRIM). Moreover, to integrate a dimension reduction feature to our procedure and to consider more complex relationships (correlation) between the predictors, we run PRIM in a sparse version of each PCA subspace of the predictors (SPCA). Potentially, this simplifies the descriptive rules of the bumps by imposing parsimony in the number of components and sparsity in their loadings. We call this method PRIM-RPCA. We illustrate how it behaves in a simple regression or classification setting of some synthetic data with respect to increasing noise and dimensionality. Results show that PRIM-RPCA outperforms a global approach in the above situations in terms of iterations and mean-support trade-offs. Results in a colon cancer microarray gene expression dataset are discussed. The novel method has practical potential applications in high dimensional, noisy and correlated datasets while minimal assumptions about the data generating mechanism are made.