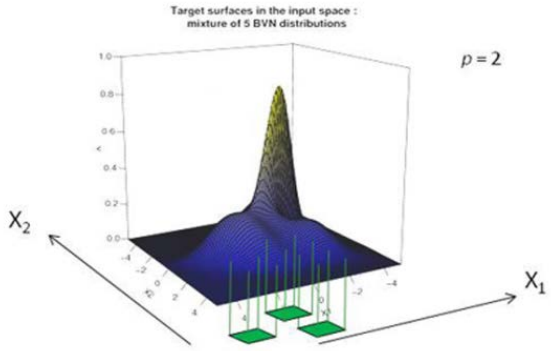
Research Interests - Developments

Bump hunting in high dimensional data:

There is often more structure and heterogeneity in the data than we initially thought of or assumed, and it is of great interest and a challenge to unveil these structures in high dimensional and noisy settings as is always the case with "omics" data. I have recently published a search algorithm of multiple modes in high dimensional data also known as "Local Sparse Bump Hunting" (Dazard & Rao, 2010).

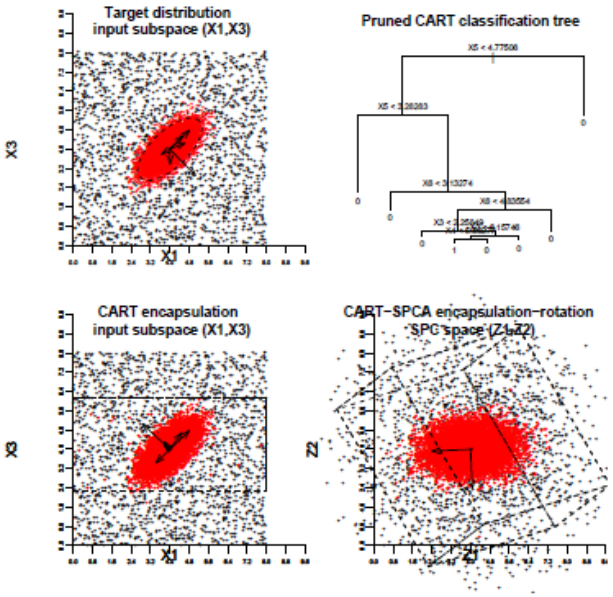
Example of bumps in a dimensional case (p = 2) simulated from a mixture of 5 bi-variate normal distributions. The goal is to identify the extrema (blue -> yellow) of the target function and their corresponding domain in the input space (green).



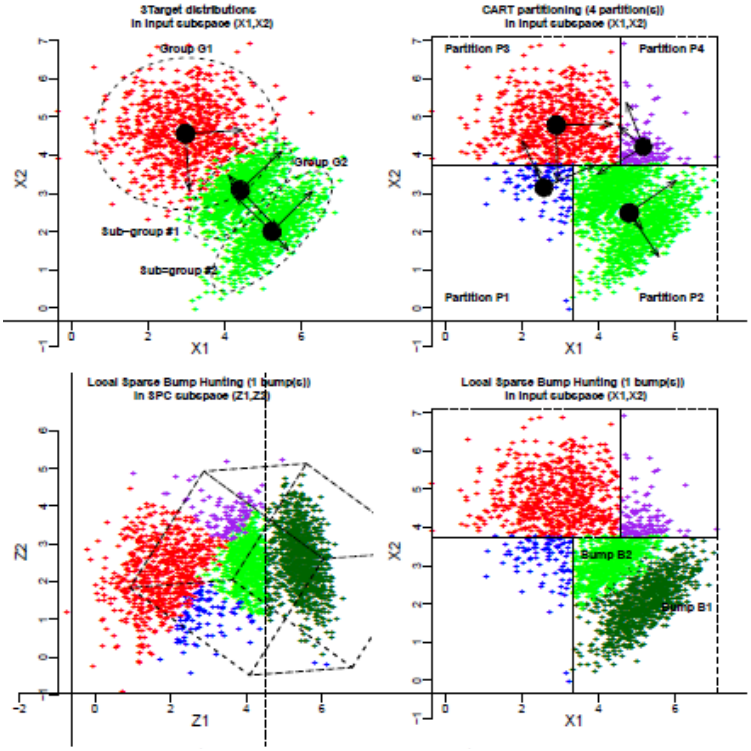
Here is how the Local Sparse Bump hunting works:

Algorithm 1 Local Sparse Bump Hunting.

- Partition the input space into $\{P_r\}_{r=1}^{R}$ partitions, using, for example, CART.
- While $r \in \{1, \ldots, R\}$:
  - If $g_r > 1$
    - Run a local SPCA : Estimate the local Sparse Principal Components (SPCs), select a first few of them $j = 1, \ldots, q$, where $q \ll p$, and choose an optimal amount of shrinkage/sparsity for each of them, resulting in a total number of nonzero loadings s.
    - Rotate the local space according to SPCs main directions. Denote the corresponding transformation by $\mathcal{T}()$ and the transformed partition by $\mathcal{T}(P_r)$.
    - Test local multimodality $\hat{m}_0$ within transformed partition $\mathcal{T}(P_r)$.
    - If $\hat{m}_0 > 1$
      - Conditioning on $\hat{m}_0$, estimate PRIM meta-parameters α and β, in $\mathcal{T}(P_r)$.
      - Run a *local* and *tuned* PRIM-based bump hunting within $\mathcal{T}(P_r)$ and get descriptive rules of the bumps in the SPC space.
      - Rotate the local rules back into the input space to get rules in the form of "sparse linear combinations".
  - $r \leftarrow r + 1$
- Collect the local rules from all partitions to get a global rule giving a full description of the estimated bumps in the entire input space.



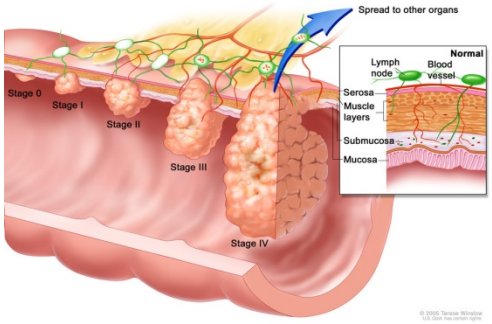
CART-SPCA encapsulation efficiency in supplemental synthetic dataset. In all plots (except otherwise stated), projection from (X_1 ,…, X_p) into subspace (X_1;X_3) only is shown (with perspective effect). Upper Left: single target group (D = 1) in the original input space (p = 6; n = 10; 000) with 95% Confidence Ellipse and Sparse Principal Components. A background noise of 5% was included, simulated from a p-variate uniform distribution. Upper Right: CART classification tree. Lower Left: target group with CART encapsulating partition in the original input space. Lower Right: target group with CART encapsulating partition in the rotated SPC space (projection in (Z_1;Z_2) is shown with perspective of CART encapsulating box and first three SPCs). Note how the SPCs are now parallel/orthogonal to the new coordinate axes. In each plot, the samples in red belong to the target distribution or encapsulating region respectively (Dazard & Rao, 2010).



All plots are projections from (X_1 ,…, X_p) into input subspace (X_1;X_2) or SPC subspace (Z_1;Z_2) (with superimposed perspective effect). Top Left: target groups from 3 overlapping target distributions (D = 3) with 2 class labels (G = 2) in the original input space (p = 100) with PC's and 95% Confidence Ellipse. Top Right: identification of the CART partition of interest with its SPCs. Bottom Left: two bumps found (dark green vs light green), shown in the SPC subspace with delineation by one side of the bump boundary (vertical solid line). Bottom Right: final 2 bumps shown back in the original input space (Dazard & Rao, 2010).

Application of Local Sparse Bump Hunting (LSBH) Reveals Molecular Heterogeneity Of Colon Tumors.

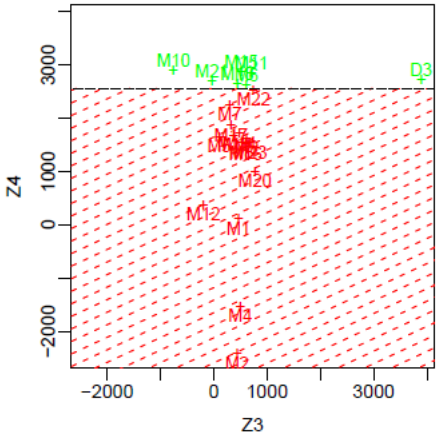
A major unresolved question in cancer is the molecular heterogeneity of the disease and of the tumoral phenotype. To understand the underlying molecular basis of this phenomenon in colon cancer, we analyzed genome-wide expression data of colon cancer metastasis samples, as these tumors are the most advanced and hence would be anticipated to be the most likely heterogeneous group of tumors, potentially exhibiting the maximum amount of genetic heterogeneity.



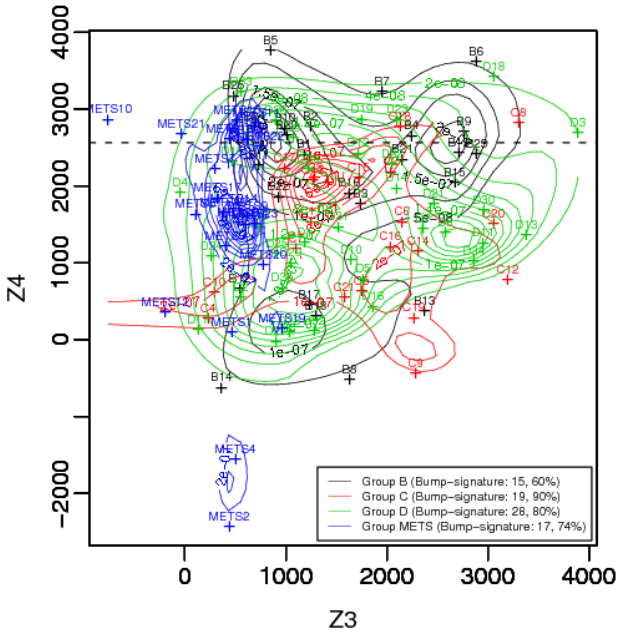
Colon Cancer Staging according to Duke's staging nomenclature. It shows groups of tumors at different stages of progression of the disease from primary tumor Stages 0, Stage I, Stage II ("B"), and Stage III ("C") to the metastatic Stage IV ("D & METS"). Inset shows serosa, muscle, submucosa and mucosa layers of the colon wall, and lymph nodes and blood vessels. As colon cancer progresses from Stage 0 to Stage IV, the cancer cells grow through the layers of the colon wall and spread to lymph nodes and other organs. This staging system is current at the time of this writing (with permission from author and rights holder: © 2005 Terese Winslow, U.S. Govt. has certain rights).

We successfully applied our LSBH algorithm on a real data collected from a large ongoing colon cancer study at CWRU. Thanks to the variable selection feature of the LSBH algorithm, a novel sparse gene expression signature was derived, which divides all colon cancer patients/tumors into two populations: a population whose expression pattern can be molecularly encompassed within the bump, and an outlier population that cannot be.

Graphical illustration of Local Sparse Bump Hunting (LSBH) on micro-array dataset for the latest stage of the disease "METS" (metastatic stage). The scatter plot shows metastatic samples in the partition of interest (and projected into SPC subspace (Z_3;Z_4)). The black dashed line represents the bump boundary (at threshold value z4 = 2562) of the bump rule, showing the separation of "bump-signature samples" (green) vs. "non-bump-signature samples" (red) respectively (Dazard et al., 2012).

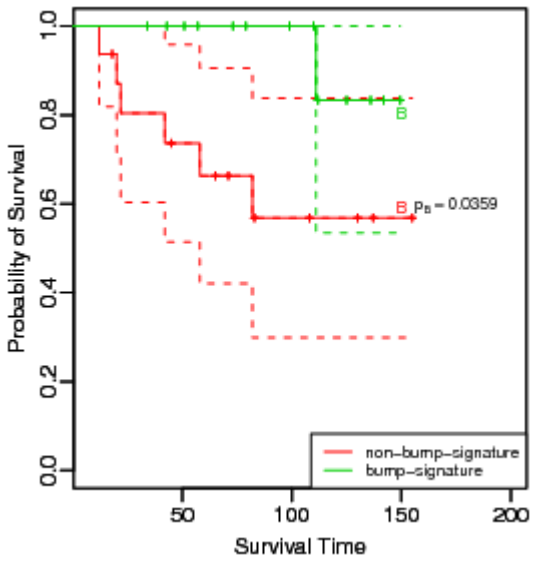


While all patients within any given stage of the disease, including the metastatic group, appear clinically homogeneous, our procedure revealed two subgroups in each stage with distinct genetic/molecular profiles:



Scatterplot smoothers with superimposed contours of the distributions of samples by Duke's stage group. Group membership: "Group B": black, "Group C": red, "Group D": green, and "Group METS: blue. The black dashed line represents the bump boundary or threshold value z4 = 2562 of the bump rule, showing the separation of "bump-signature samples" vs. "non-bump-signature samples". Counts of "bump-signature samples" and proportions w.r.t. groups are indicated (Dazard et al., 2012).

We identified new subtypes of colon tumors from the earliest stage of primary tumor formation to the latest metastatic stage with diagnostic and predictive values. This more refined classification allows clinicians to better classify patients and predict for them a clinical outcome of interest such as survival time:



Kaplan-Meier curves (solid lines) with 95% confidence intervals (dotted lines) for the earliest stage of the disease: "B". The specific censoring indicator is the death event: alive=0, DWD=1. Curves are marked at each censoring time. Survival time is in months (Dazard et al., 2012).

In conclusion, implications of such a finding are important in terms of early detection, diagnosis and prognosis (Dazard et al. 2012). In addition, this new patients sub-typing also bears some potential therapeutic value since patients can now be treated in a customized manner.

Variance estimation and stabilization of high dimensional data:

High-throughput technologies have given rise to datasets where thousands of individual variables are measured simultaneously with a relatively small number of replications. This raises challenging problems to traditional statistical methods, which literally breakdown in this situation (so called $p >> n$ paradigm). Among the problems posed by this type of dataset is that the variable-specific estimators of variances are not reliable and variable-wise tests statistics have low powers, both due to the small sample size. In addition, it has been observed in this type of data that the variance increases as a function of the mean, leading scientists to assume that the variance is a function of the mean. Several methods have been proposed to overcome either the lack of degrees of freedom or the variance stabilization problem, yet never realizing that both problem could be addressed at once. In addition, to remove sources of systematic variation due to experimental artifacts in the measured intensities and to ensure that the usual assumptions for statistical inferences are met, we need a variance stabilization and normalization procedure.

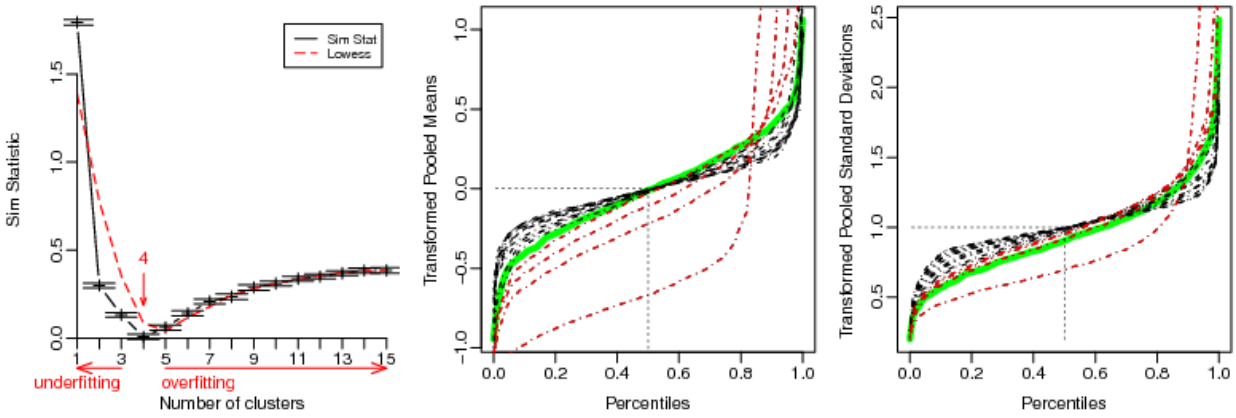
I recently developed a non-parametric adaptive regularization procedure, called Joint Adaptive Mean-Variance Regularization (R Package 'MVR') that generates joint and local shrinkage estimators of the mean and the variance to better estimate the population marginals (Dazard & Rao, 2010, 2011, 2012).

Here is how Joint Adaptive Mean-Variance Regularization works:

Algorithm 1 Joint Adaptive Mean-Variance Regularization

1. for $l = 1$ to C_{max} do
   - Select a variable cluster configuration $\mathcal{C}$ with l clusters.
   - Standardize each variable individually using corresponding estimates $\{\hat{\mu}(l_j), \hat{\sigma}^2(l_j)\}$ where $l_j \in \{C_l\}_{l=1}^{l}$.
   - Compute the corresponding *Similarity Statistic* estimates $\{\widehat{Sim}_p(l)\}_{l=1}^{l}$ as in 8.
2. Find the optimal cluster configuration $\mathcal{C}$ with $\hat{C}$ clusters, where $\hat{C}$ is determined as in 9.
3. Standardize all variables individually using this optimal cluster configuration $\mathcal{C}$. After which, all means $\hat{\mu}_j^*$ and variances $\hat{\sigma}_j^{*2}$ of the transformed data are assumed to follow sampling distributions with target first moments, i.e. $(0, 1)$ respectively.

We showed that regularized t-like statistics derived from these joint shrinkage estimators offer significant more statistical power in hypothesis testing than their standard sample counterparts, or regular common value-shrinkage estimators, or when the information contained in the sample mean is simply ignored. This latter result is also a direct consequence of the strong mean-variance dependence inherent to this type of data. Finally, we show that these estimators feature interesting properties of variance stabilization and normalization that can be used as a rational for preprocessing high-dimensional multivariate data.



Typical similarity statistic profile in a single group design (left) showing the estimated number of clusters for the optimal clustering configuration. The vertical red arrow indicates the result of the stopping rule: i.e. the largest value of l for which Simp(l) is minimal up to one standard deviation. Directions of over/under-fitting are indicated. Red dashed line depicts the LOESS scatterplot smoother. Empirical quantile profiles of means (middle) and standard deviations (right) for each clustering configuration (dashed red and black lines) are shown to check how the distributions of first and second moments of the transformed data fit their respective theoretical null distributions under a given cluster configuration. The single cluster configuration, corresponding to no transformation, is the most vertical curve, while the largest cluster number configuration reaches horizontality. Notice how empirical quantiles of transformed pooled means and standard deviations converge (from red to black) to the theoretical null distributions (solid green lines) for the optimal configuration (Dazard & Rao, 2010, 2011, 2012).

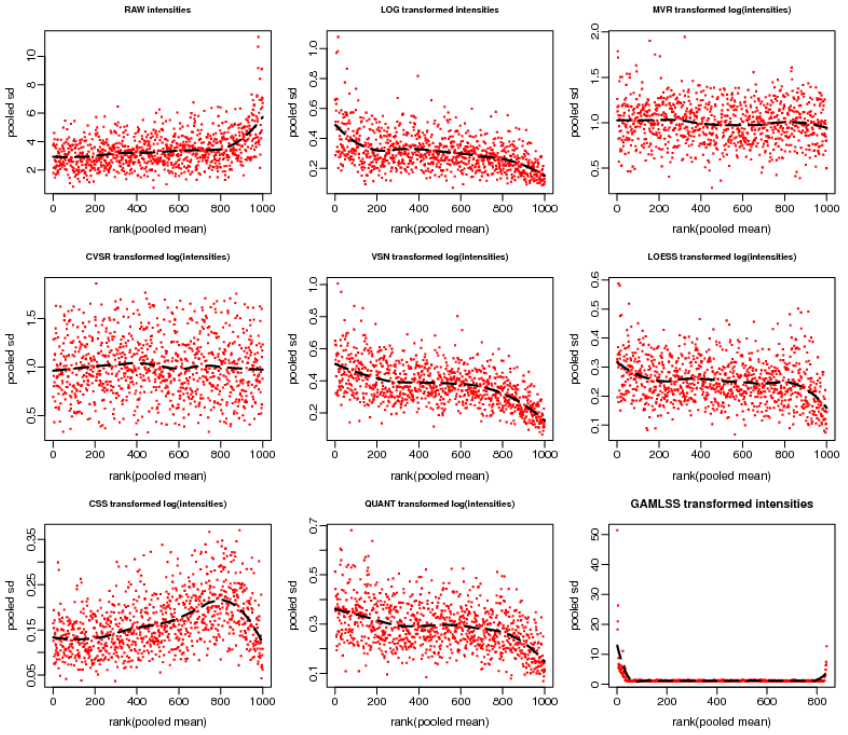
For simulation purposes, we used a realistic model mimicking the real life situations of mean-variance dependency and unequal sample group variances. In this model, the individual response (signal, intensity) can be written as an additive and multiplicative multi group error model: problem, where for each variable $j = 1,\ldots,p$ (gene, peptide, protein, …) and each group $k = 1,\ldots,G$.

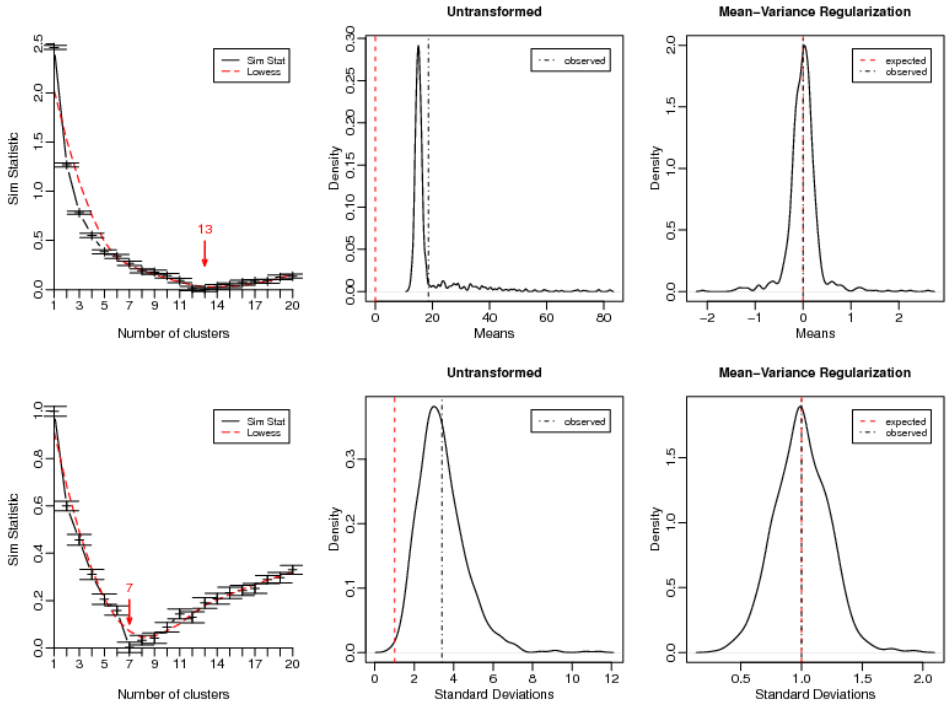
$$Y_{i,j} = \mu_{k,j} + (\mu_{k,j} + \beta_k)e^{\rho_k \eta_{i,j}} + \nu_k \varepsilon_{i,j}$$

where

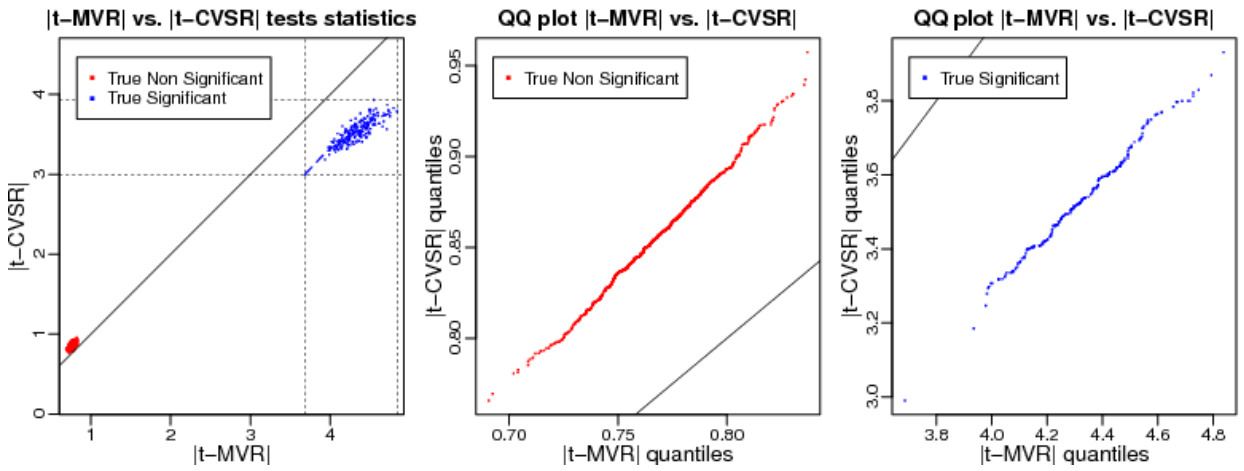
$$\begin{cases} \varepsilon_{i,j}, \eta_{i,j} \overset{iid}{\sim} N(0,1) \\ i : k_i = k \end{cases}$$

The Mean-SD scatterplot allows to visually verify whether there is a dependence of the variance on the mean. The black dotted curve depicts the running median estimator (equal window-span=0.5 for all procedures). If there is no variance-mean dependence, then this curve should be approximately linear horizontal (Dazard & Rao, 2010, 2011, 2012)





First column: similarity statistic profiles giving the estimated number of variable clusters. Red arrows indicate results of the stopping rule. Middle and right columns: density distributions of pooled means and pooled standard deviations before (middle) and after (right) multi-group MVR-transformation. Notice how the sampling distributions of transformed means and standard deviations are centered about their target first moments (0; 1) respectively (Dazard & Rao, 2012)



Comparison of performance of our regularized test statistics t-MVR with its best competitor: t-CVSR on simulated dataset. Shown are Monte Carlo estimates of regularized test statistics |t-MVR| and |t-CVSR| (in absolute value), based on B = 100 replicated synthetic datasets. Black solid line : identity line. Left: scatter-plot of |t-MVR| vs. |t-CVSR| tests statistics. Middle: Quantile-Quantile plot of |t-MVR| vs.|t-CVSR| for non-significant variables (red dots). Right: Quantile-Quantile plot of |t-MVR| vs. |t-CVSR| for significant variables (blue dots) (Dazard & Rao, 2010, 2011, 2012).

Model selection:
One can cast a variety of bioinformatics question into that of a model selection problem. Bayesian variable selection algorithm like shrunken Bayesian ANOVA and model selection algorithm like the "Fence" (recently developed by Dr. J.S. Rao and H. Ishwaran) have proven extremely efficient in reducing the complexity and dimensionality of the data with often dramatic improvement in model fitting and prediction error as compared to existing methods. My interest is in developing and applying these model selection algorithms and variations to specific settings in bioinformatics such as Differential Expression, Data Integration, Pathway Analysis/Gene Set Enrichment Analysis, Association Studies for simple and complex traits and Quantitative Trait Loci (QTL) mapping, and Protein interaction network.

Parallel computing:
The Bioinformatics Division completed the installation of a pilot high performance computing PC cluster in order to meet the future instructional and research needs in the areas of parallel computing and/or analyses of large datasets ("omics" data) as generated by microarray, proteomics and sequencing high-throughput technologies. The pilot cluster is fully operational (http://bioinfo.meds.cwru.edu/ganglia/). It consists of 4 compute nodes, each equipped with two Intel Xeon E5430 Quad-Core processors (32 CPUs total), running at 2.66 GHz. The master node is equipped with 64GB of main memory and each slave node with 32GB (160 GB of total main memory). Each node has a 1TB Hitachi SCSI hard drive. The system is easily scalable. It is anticipated to upgrade the cluster to at least 8 compute nodes with twice as much main memory per node and a data backup system with up to 24TB of disk storage.

All nodes run under the RedHat Centos 5.2 Linux Operating System, interconnected with Gigabit Ethernet for Parallel Virtual Machine (PVM) or Message Passing Interface (MPI). The master node of the cluster also serves as a server to host main services for the Bioinformatics Division. At this moment, the server hosts user accounts (for faculty, staff, or student); a standard batch job scheduler for multi users (SGE), the Parallel Virtual machine (PVM) and/or Message Passing Interface (MPI) for enabling parallel computing, software server services including compilers, libraries and supported software installations; dedicated storage service and

database storage facility; laser color printing service; and a future dedicated website-hosting service to replace the former one currently at http://bioinfo.case.edu/bioinfo/bioinfo.html. Current software installations include commercially available statistical computing packages (SAS, S+), Ingenuity Pathway Analysis (IPA) for bioinformatics analyses, and OPEnTEXT (formerly Hummingbird) for enabling X-Server and allowing secured SSH and SFTP connections, as well as open source packages such as R for statistical computing. It is anticipated to buy new licenses of Minitab, Matlab, Mathematica as well.

Acknowledgement:
I received invaluable help from Alberto Santana (http://starnix.case.edu/staff/asantana.html), System Administrator of the department of Epidemilology and Biostatistics. He provided advices during the steps of design, implementation and current support.

A few links:
- My cluster in action: http://bioinfo.meds.cwru.edu/ganglia/
- The R Project for Statistical Computing: http://www.r-project.org/
- An interesting paper on the R language that recently appeared in the The New York Times:
http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html
http://bits.blogs.nytimes.com/2009/01/08/r-you-ready-for-r/

- My collaborators:
J.S. Rao http://dev.case.edu/rao/software.html
H. Ishwaran http://www.bio.ri.ccf.org/Resume/Pages/Ishwaran/ishwaran.html