

# Research Statement on Statistics

Jun Zhang (junzhang@uchicago.edu)

My general interests on statistics include statistical learning theory, generalized linear models, nonparametric statistics and nonlinear times series. My recent work on statistics is to detect and understand the patterns of usually high dimensional correlated data with complex structure, particularly when applied to statistical genetics and medical images processing. Specific interests include:

- (1) developing statistical tools to detect population structure from population genotype data, and correct it for association studies of complex diseases;
- (2) developing methods for pedigree based gene mapping;
- (3) developing diagnostic tools for medical images analysis to reduce false discoveries.

## Population Structure Detection and Correction

Correctly detecting population structure (PS) is critical for inference of human migration history and understanding the evolution. The confounding errors due to population structure in the rapidly planned disease association studies, i.e., false discoveries due to the systematic allele frequency differences among subpopulations, makes the issue urgent. The prevailing method to analyze PS is to use the top principal component (PC) of covariance matrix of subjects to summarize the global genetic variations across space.

Does PCA always work? Under what conditions does it fail and how to adjust it? To answer these questions, in [1] we fitted a model of population structure  $Beta(p_l \frac{1-\tau_i}{\tau_i}, (1-p_l) \frac{1-\tau_i}{\tau_i})$ ,  $i = 1, \dots, k$ , for the allele frequencies,  $\tau_i$  are analogous to  $F_{st}$  parameters which specifies how far the subpopulation's allele frequencies tend to be from typical values. Under the model, we were able to derive the top theoretical principal component and express association test statistics explicitly in terms of the disease prevalence and numbers of case/controls in each subpopulation, which gives the conditions for correct type one errors. We also studied association testing for admixed populations under simplified admixture processes which were motivated from African Americans.

While in [2, 3] from the point view of manifold learning, I propose using the Laplacian eigenfunctions to infer PS, instead of PCs. The idea is to construct an adjacent graph where each node represents one subject and it is connected only to its close neighbors, since subjects who are less correlated are usually meaningless. Then one can study the geometry of the intrinsic dependence graph. In particular, Laplacian eigenfunctions associated to the graph are the generalized harmonic functions which

contain useful geometric information of the graph, and they are more meaningful than the PCs. Compared with PCA, our method is less noise and can successfully detect the ancestral populations from admixed individuals such as European Americans while PCA fails. When applied to disease genome-wide association studies in a regression setting, our method has slightly improved power than PCA in demonstrated simulations. Our method, LAPSTRUCT, is expected to become a promising tool for population structure detection and correction in disease association studies.

### **Association Testing for Related Individuals from Structured Population**

In [4], I propose a novel method to combine family based individuals together with unrelated case-controls for association testing. The method allows the presence of population structure in the samples. Given the kinship coefficients of related individuals, the method first give unbiased estimates of the allele frequencies using quasi-likelihood method, and then infer the population structure from the suitably correlation adjusted genotype data using Laplacian eigenfunctions. Finally I give two association testing statistics, namely, a modified Armitage statistic and a quasi-score statistic  $U^T i^{rr} U$ , where  $U$  is the derivative of the log-quasi-likelihood and  $i^{rr}$  is the (r,r)th entry of Fisher information matrix. A similar but easier situation for nuclear families is illustrated in [5], using a variance component adjustment method. The method generalizes straightly to quantitative trait locus and haplotype association testing, and can handle missing data.

### **Diagnosis of Medical Images**

In computerized diagnosis of medical images such as CT colonography and digital mammography for breast cancer it is critical to have an efficient statistical learning algorithm to distinguish malignant lesions from benign ones. One approach is to select a massive number of subregions as features from golden standard samples and train them according to their approximated likelihood of being malignant. In [6] I took the principal components of the features and trained by a multi-layer Artificial Neural Network (ANN), which reduced the training time to a quarter while the statistical significance is maintained at the same sensitivity level. In [7] I apply Support Vector Machine (SVM) instead of ANN in the framework of massive training, but with an additional penalty term in the cost function which accounts for the correlation of features, noting that the features actually have some intrinsic geometric structures since they are overlapped. The software incorporating this algorithm, LAP-SVM, has achieved significant reduction of false positives on real data sets. The LAP-SVM can also be used as dimension reduction tool for general feature selection

purposes. Future research on this direction is to select better features instead of just the pixel values.

### Statistical Manifold Learning

Manifold methods have become increasingly important and popular in machine learning and have seen numerous recent applications in data analysis including dimensionality reduction, visualization, clustering and classification. The central modeling assumption in all of these methods is that the data resides on or near a low-dimensional submanifold in a higher-dimensional space. However, one does not have access to the underlying manifold but instead approximates it from a point cloud usually by constructing an adjacency graph. The underlying intuition has always been that since the graph is a proxy for the manifold, inference based on the structure of the graph corresponds to the desired inference based on the geometric structure of the manifold. In [8], we give some theoretical results to justify this intuition.

To be precise, earlier Nigoyi introduced a framework based on Laplacian Beltrami operator on a manifold to motivate using the graph Laplacian associated to point-cloud data, namely, *Laplacian Eigenmap*. Assuming  $\mathcal{M}$  is a compact Riemannian submanifold of  $\mathbb{R}^n$ , the operator  $\Delta_{\mathcal{M}}$  is defined as  $\Delta_{\mathcal{M}}f = -\text{div}(\nabla f)$ , where  $f \in \mathcal{C}^2(\mathcal{M})$ . The eigenfunctions of Laplacian form a basis for  $L^2(\mathcal{M})$ , and play a central role in a variety of algorithms for data analysis. If the manifold is taken with a measure  $\nu$  (given by  $d\nu(x) = P(x)d\mu(x)$ ) for some density function  $P(x)$  and with  $d\mu$  being the canonical measure to the volume form, then the *weighted Laplacian* is defined as  $\Delta_{\mathcal{M},\nu}f(x) = \frac{1}{P(x)}\text{div}(P(x)\nabla_{\mathcal{M}}f)$ . Given data points  $\{x_1, \dots, x_n\}$  sampled i.i.d from an arbitrary distribution  $\mathcal{P}$  on  $\mathcal{M}$ , we construct a weighted graph associated to the point cloud using Gaussian kernel. We define the *point cloud Laplace operator* by

$$\mathbf{L}_n^t f(x) = f(x) \frac{1}{n} \sum_j e^{-\frac{\|x-x_j\|^2}{4t}} - \frac{1}{n} \sum_j f(x_j) e^{-\frac{\|x-x_j\|^2}{4t}}$$

We justify the following: Let  $t_n = n^{-\frac{1}{k+2+\alpha}}$ , where  $\alpha > 0$  and let  $f \in \mathcal{C}^\infty \mathcal{M}$ , then the following equality holds:

$$\lim_{n \rightarrow \infty} \frac{1}{t_n (4\pi t_n)^{\frac{k}{2}}} \mathbf{L}_n^t f(x) = \frac{1}{\text{vol}(\mathcal{M})} \Delta_{\mathcal{M},\nu} f(x).$$

## Planned Research in Near Future

### Gene Regulatory Network and Graphical Models

Identifying variations in DNA that increase susceptibility to disease is one of the primary aims of genetic studies using a forward genetics

approach such as linkage and association testings. However, such studies provide limited functional information on how genes lead to diseases. An alternative is to identify gene networks that are perturbed by susceptibility loci and that in turn lead to diseases. Bayesian network has been recently employed as a tool to infer the interactions between genes. It is a graphical model of joint multivariate probability distributions that captures properties of conditional independence between variables. Given genotyping and expression profiles, I am interested in developing certain graphical models which can better *learn* transcriptional regulatory networks and infer causal relations from the noisy data. The gene regulatory network is actually also a *dynamic* network. With the additional time course gene expression data, it will be very valuable to combine tools from time series into the network framework. Another closely related direction is to develop certain graphical models based tools to incorporate the known biological pathway knowledge into association studies.

### **Medical Image Reconstruction and Segmentation**

Recently Xiaochuan Pan revealed the relation between the backprojection of a derivative of projection data and the Hilbert transform of an image along certain segments of lines covering region of interest (ROI). This naturally leads to a data sufficiency condition for 2D or 3D ROI reconstruction from a limited family of line integrals. This sufficiency condition can be further generalized by showing that unique and stable reconstruction can be achieved from an even more restricted family of data sets. The condition is derived by analyzing the inversion of the *truncated* Hilbert transform, defined as the problem of recovering a function of one real variable from the knowledge of its Hilbert transform along a segment which only partially covers the support of the function but has at least one end point outside that support. However, only iterative numerical algorithm is available and the reconstruction computation usually takes several weeks for slow convergent issues. Together with Xiaochun Pan, I try to derive a closed form analytic inversion formula for the truncated Hilbert transform, with the hope that this would reduce the reconstruction time to minutes.

Lesion segmentation on mammograms is a challenging task since mass lesions are usually embedded and hidden in varying densities of parenchymal tissue structures. Methods for automatic delineation of lesion boundaries on digital mammograms are highly expected. I propose to utilize a geometric active contour model that minimizes an energy function based on the homogeneities inside and outside of the evolving contour. Prior to the application of the active contour model, a radial gradient index based segmentation method will be applied to yield an initial contour closer to the lesion boundary location. The performance in segmentation needs further investigation.

### **Random Matrix Theory for Dependent Data and Dependence Measure for Nonlinear Time Series**

The classical random matrix theory (RMT) gives a distribution of suitably normalized eigenvalues of covariance matrix for i.i.d Gaussian samples, namely, the Tracy-Widom law. However, the real data are always dependent. A natural question comes up: is there a similar asymptotic distribution for dependent samples? It seems hopeless for the general dependent situation. Instead, I am investigating the question in a weak sense of dependence such as *m-dependence*, and the samples could be some observations of certain simple stationary time series. I have obtained some encouraging preliminary results which are very close to Tracy-Widom law. This is a joint work with Wei-Biao Wu.

## References

- [1] Mary Sara Mcpeek, Jun Zhang, Mark Abney **Association testing based on principal components correction for stratification: When and how does it work?** reviewing at American Journal of Human Genetics
- [2] Jun Zhang, Partha Niyogi, **Laplacian eigenfunctions learn population structure**, reviewing at Nature Genetics
- [3] Jun Zhang, Chunhua Weng, Partha Niyogi, **Graphical analysis of population structure on Rheumatoid arthritis data**, reviewing at BMC Genetics
- [4] Jun Zhang, **Association testing for related individuals in structured population**, to be submitted
- [5] Jun Zhang, Xiaofeng Zhu, Richard Cooper, **An integrated genome-wide association analysis on Rheumatoid arthritis data**, BMC Proceedings 2007, 1(Suppl 1):S35
- [6] Jun Zhang, Kenji Suzuki, **Improved massive training ANN using principal components for computer aided detection of polyps in CT colonography**, reviewing at Medical Physics
- [7] Jun Zhang, Kenji Suzuki, **Geometrically regularized SVM for massive training in computerized diagnosis**, to be submitted
- [8] Jun Zhang, Partha Niyogi, **Convergence of Laplacian Eigenmaps on weighted manifold**, in preparation