Statement of Research

Haiyan Wang

My main research objective is to develop effective distribution-free methods for analysis and mining of complex data. In particular, I enjoy working on and producing powerful theoretical or computational methods for nonparametric hypothesis tests in big data, image analysis, high dimensional data modeling and data mining, and applying them to various fields. Microarray data, mass spectrum from lipidomics, proteinomics, etc, are some examples of high dimensional data. These data often contain thousands of variables with very small sample sizes. As high dimensional data have emerged more recently, statistical methodology development is facing urgent challenges while it also opened an exciting research area with plenty of opportunities. Due to cost concerns, the data typically are of small sample sizes rendering the large sample based classical methods infeasible. Distribution-free nonparametric methods typically make less assumptions and are often more meaningful for complex data. The classical nonparametric methods, however, are applicable at the expense of requiring large sample sizes. Scientific publications that apply traditional methods on high dimensional data can lead to misleading results due to violations of assumptions on distribution or a large sample size requirement relative to the number of variables. The current state of art of statistical analysis for high dimensional data contains some explorative analysis and theory on visual presentation, multiple testing, parametric model-based classification/regression, clustering of data, and some variable selection methods that often require the sample sizes to go to infinity. Direct inference for high dimensional data is very difficult to develop, especially if the inference is under no distributional assumptions.

My research during the past years at Kansas State University considered deriving the theory and inferences for hypothesis testing in high dimensional data and developing new algorithms that combine my nonparametric inferences with computational methods for complex data settings. As a result, I have published a series of articles jointly with my students and coauthors. These articles include a set of hypotheses testing theory for heteroscedastic high dimensional data [Wang and Akritas 2004, 2009, 2010a, 2010b, 2011; Wang, Tolos and Wang 2010; Wang, Higgins and Blasi 2010; Bathke et al. 2010; Zhang et al. 2011, Jin, Wang and Wang 2014], robust clustering methods to effectively group thousands of variables based on the observed patterns [Wang, Neill and Miller 2008; von Borries and Wang 2009], robust digital image similarity measures [Wang et al. 2011; Silwal et al. 2013], algorithms for image pixel classification and segmentation [Ghimire and Wang 2012], and variable selection and model estimation for high dimensional genomics, or medical/agricultural data [Qian, Wang and Yuan 2012; Zhou et al. 2012; Zhang et al. 2012; Li et al. 2012, Wang et al. 2013,

Xie et al. 2013, Dai et al. 2014, Chen et al. 2015]. Two of these articles earned a permanent 'highly accessed' designation from BMC Bioinformatics. The contributions and technical development in these articles spread across five different areas as described below.

(1). **Inference based on original observations for hypothesis testing in big data.** Wang and Akritas [2011] address the asymptotic theory for hypotheses testing in high dimensional analysis of variance when the distributions are completely unspecified. Wang and Akritas [2010a] and Wang, Higgins and Blasi [2010] provide the inference for testing several effects in nested heteroscedastic functional data that includes a large number of repeated measurements observed within a subject or stratum. We build our theory on novel models in which the random effects are assumed to be neither uncorrelated nor normal. The models leave the covariance structure unspecified and apply to both discrete and continuous data. The asymptotic theory of the test statistics is driven by a large number of factor levels or a large number of measurements per subject and the assumption of nonstationary α-mixing on the error term. Both weak and long range dependence are considered. Wang, Tolos and Wang [2010] present a test of independence between the response variable, which can be discrete or continuous, and a continuous covariate after adjusting for heteroscedastic treatment effects. This work was extended to the theory of lack-of-fit in heteroscedastic constant and nonlinear regression by my Ph.D. students [Gharaibeh, Sahtout and Wang 2015, Gharaibeh and Wang 2015]. Additionally, the results also made the first step toward a current research on nonlinear variable selection in additive models for high dimensional data that was studied in detail in the dissertation of one of my Ph.D. students. In Bathke et al. [2010], we derive asymptotic procedures as well as finite approximations for the analysis of data arising from series of randomized complete block designs with a large number of factor levels. The publication by Zhang et al. [2011] resulted from the dissertation work of my former student Ke Zhang provides a robust nonparametric approach to compare the expressions of longitudinally measured sets of genes under multiple treatments or experimental conditions.

(2) **Rank based inference for high dimensional data.** Such inference tests hypotheses specified in terms of distribution functions. Wang and Akritas [2004, 2010b] provide rank tests for the nonparametric main factor effects and interactions in two-way and multi-way high-dimensional analysis of variance when the cell distributions are completely unspecified, the sample size may be small and the number of factor levels may be large. Wang and Akritas [2009] consider rank based inferences for testing hypotheses in a fully nonparametric marginal model for heteroscedastic functional data. The asymptotic distribution of the rank statistics is obtained by showing their asymptotic equivalence to corresponding expressions based on the asymptotic rank transform. Compared with test procedures based on the original observations, the proposed rank procedures are free of moment conditions, converge to their limiting distribution faster, and have better power when

the underlying distributions are heavy tailed or skewed.

(3) **Rank-Test-based clustering for high dimensional data.** The idea of using rank tests in clustering for high dimensional data was developed in Wang et al. (2008) for agglomerative clustering of functional data and in von Borries and Wang (2009) for partition clustering of independent low sample size data. We define clusters through the unknown high dimensional multivariate distributions and provide test-based clustering algorithms that are invariant to monotone transformations of data. These test-based clustering methods can take all the information from thousands of variables to effectively detect unknown patterns or clusters in high dimensional independent or functional data.

(4) **Digital image quality assessment and image pixel classification and segmentation.** Current popular image similarity measures (mean squared error, signal to noise ratio, structure similarity measure and its variants) do not take into account of possible nonlinear dependence between the source image and the image being compared. In Wang, Maldonado and Silwal [2011] and Silwal, Wang and Maldonado [2013], we propose nonparametric rank-test-based similarity measures in frequency and wavelet domain, respectively. Applications of the methods on a variety of altered images showed superior performance. Ghimire and Wang [2011] introduce, implement and assess the idea of using combined evidence from the multiple hypothesis testing and minimum distance to carry out image pixel classification and image segmentation. Extensive experiments show that our test-based segmentation has excellent edge detection and texture preservation properties for both gray scale and color images.

(5) **Variable selection and data mining in big data**. Some of the research that I have been doing is on methods for variable selection and modeling in biochemistry, proteomics, and genomics data. Popular examples are cancer classification through molecular information from genomics data, prediction of gene annotation splice sites based on genomic sequence data, and quantitative structure-activity relationship modeling for drug design. The focus here is to produce methods that are effective for many datasets in external validation. Different from most literatures that focus on screening individual or pairs of variables without considering the possible interactions among variables, in Zhang et al. [2012] we introduce a new computational method for classification of cancer tissue samples based on gene expression data. The method takes potential variable interactions into account. During the variable selection process, the set of variables to be kept in the model was recursively refined and repeatedly updated according to the effect of a given variable on the contributions of other variables in reference to their usefulness in cancer classification. The variables selected from each data set leads to significantly improved leave-one-out classification accuracy across 10 data sets for multiple classifiers. In Qian et al. [2012], Zhou et al. [2012], Li et al. [2012],

Chen et al. [2015], we extract and summarize data-specific genomic sequence information to form new variables and achieve high accuracy in classification with algorithms using the support vector machine. In these articles, we dealt with both feature encoding and feature selection. Most statisticians understand the importance of feature selection. The feature encoding part is equally or more important in these studies since the input is the sequence. As an example, Chen et al. [2015] aim at identify protein glycosylation sites using only amino acid sequence information. For feature encoding, we extract features based on multi-scale composition of amino acids (the composition of *k* amino acids α_1 α_2 ...α_k is a scale *k* compostion, where α_i is one of the 21 amino acid residues) and then code each amino acid using 531 physiochemical and biochemical properties of amino acids. Such coding accounts for the relationship of amino acids at different locations in a sequence and turn the sequence into high dimensional feature variables. We found that such new feature coding followed with our feature selection method significantly improve the accuracy of glycosylation sites prediction. Wang et al. [2013] provide an algorithm that is a Chi-square-statistic-based Top Scoring Genes (TSG) classifier to perform informative gene selection in both binary and multi-class cancer classification. It overcomes the problem of only selecting gene pairs in top scoring pairs (TSP) family classifiers. Extensive comparison and validation with application to 9 binary and 10 multi-class gene expression datasets involving human cancers show that the our method has clear advantages. It outperforms TSP family classifiers by a big margin in most of the 19 datasets. In addition to improved accuracy, our classifier shares all the advantages of the TSP family classifiers including easy interpretation, invariant to monotone transformation, often selects a small number of informative genes allowing follow-up studies, and resistant to sampling variations due to within sample operations. In Xie et al. [2013] and Dai et al. [2014], we provide pipelines for prediction of multidimensional time series and quantitative structure-activity relationship analysis of peptides. The work introduces high dimensional semivariogram for near-neighbor sample selection for corresponding data settings, BMSF for feature selection, and weighted SVR regression for validation and prediction. Comparisons with published results suggest that our methods have much improved prediction accuracy.

My current research in progress is targeted at methods that reduce false positive rate for variable selection in high dimensional data. The work with my Ph.D. student Girly Ramirz considers a conditional screening approach that extends the least angle regression to additive models. The method selects each variable by comparing the relative contribution of the variable with those already selected into the model. The effect of a covariate is defined partly through a nonlinear dependence evaluation via the test in Wang, Tolos and Wang 2010 mentioned earlier. Comparison with the methods by Fan et al. (2011) and Hall and Miller (2009) showed that our method not only has significantly less false positive rate than both methods, but also has accuracy and true positive rate comparable to greedy-INIS of Fan et al. that is computationally extensive for large dataset. The work

with my Ph.D. student Mohammad Sahtout extends the widely used Nearest Shrunken Centroids classifier by Tibshirani et al. (2002) to heteroscedastic settings in addition to employing two other thresholding methods. We obtained much improved results on 10 multi-class microarray human cancer datasets than the original Nearest Shrunken Centroids algorithm. Additional work in this line of research is the inference of the model parameters. This is demanded by applied researchers who often want some evidence of the significance or confidence intervals. Such inference barely exists for high dimensional data modeling and thus provides rich opportunities for future research.

In summary, the multiple lines of my research together with my students and collaborators not only enrich the pool of statistical and computational methods on high dimensional data, but also provide a set of effective tools that can be utilized by scientists in other disciplines to accelerate research findings. The most direct application fields of these methods are medical, agricultural, and other sciences. As a result, I have been actively engaged in collaborative projects with researchers both on campus and off campus. Majority of the collaborative publications are about biological and biomedical research. The inter-disciplinary nature of these collaborations not only offers me insights to envision challenges that require the development of new and innovative statistical methodologies, but also provides sound analysis support for my collaborating scientists.

References

Bathke, A.C., Harrar S.W., **Wang, H.**, Zhang, K., Piepho, H. (2010). Series of randomized complete block experiments with non-normal data. *Computational Statistics and Data Analysis.* 54(7), 1840-1857.

Chen, Y., Zhou, W., **Wang, H.** Li, L., Wang, L., Yuan, Z. (2015) Prediction of O-glycosylation sites based on Multi-Scale Composition of Amino Acids and Feature selection. Medical & Biological Engineering & Computing, Jun;53(6):535-44. doi: 10.1007/s11517-015-1268-9.

Dai, Z., Wang, L., Chen, Y., **Wang**, **H.,** Bai, L., and Yuan Z. (2014) A pipeline for improved QSAR Analysis of Peptides: physiochemical property parameter selection via BMSF, near-neighbor sample selection via semivariogram, and weighted SVR regression and prediction. *Amino acid*. 46(4):1105-19. doi: 10.1007/s00726-014-1667-5.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557.

Hall, P. and Miller, H. (2009). Using generalised correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18:533–550.

Jin, L. Wang, S. and **Wang, H.** (2014). A new nonparametric stationarity test of time series in time domain. Journal of the Royal Statistical Society: Series B (Statistical Methodology). doi: 10.1111/rssb.12091

Li, J., Wang, L., **Wang, H.**, Bai, L., Yuan, Z. (2012) High-accuracy splice sites prediction based on sequence component and position features. *Genetics and Molecular Research*. **11** (3): 3432-3451. http://dx.doi.org/10.4238/2012.September.25.12.

Ghimire S. and **Wang H.** (2012). Classification of image pixels based on minimum distance and hypothesis testing. *Computational Statistics and Data Analysis*. **56**, 2273-2287.

Gharaibeh, M. and **Wang, H.** (2015). A nonparametric lack-of-fit test of nonlinear regression in presence of heteroscedastic variances. In revision.

Gharaibeh, M., Sahtout, M., and **Wang, H.** (2015). A nonparametric lack-of-fit test of constant regression in presence of heteroscedastic variances. In revision.

Qian, G., **Wang, H.**, and Yuan, Z. (2012) Using homology information from PDB to improve the accuracy of protein β-turn prediction by NetTurnP. *Progress in Biochemistry and Biophysics*. **39**(5): 472-482.

Silwal, S. and **Wang, H.**, Maldonado, D. (2013). Image similarity assessment via nonparametric hypothesis testing on wavelet coefficients. *Statistics and Its Interface*. **6**:117135.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572.

von Borries, G. and **Wang, H.** (2009). Partition clustering of high dimensional low sample size data based on p-values. *Computational Statistics and Data Analysis*. 53, 3987-3998.

Wang, H. and Akritas, M.G. (2004). Rank Tests for ANOVA with Large Number of Factor Levels, *Journal of Nonparametric Statistics*, **16**(3-4): 563-589.

Wang, H. and Akritas, M.G. (2009). Rank tests in heteroscedastic multi-way HANOVA. *Journal of Nonparametric Statistics*. 21(6), 663-681.

Wang, H. and Akritas, M. (2011). Asymptotically distribution free tests in heteroscedastic unbalanced high dimensional ANOVA. *Statistica Sinica*. 21(3), 1341-1377.

Wang, H. and Akritas, M.G. (2010a). Inference from heteroscedastic functional data. *Journal of Nonparametric Statistics*. 22(2), 149-168.

Wang, H. and Akritas, M. (2010b). Rank test for heteroscedastic functional data. *Journal of Multivariate Analysis.* 101, 1791-1805.

Wang, H., Higgins, J., and Blasi, D., (2010). Distribution-free tests for no effect of treatment in heteroscedastic functional data under both weak and long range dependence. *Statistics and Probability Letters*. 80, 390-402.

Wang, H., Maldonado, D. and Silwal S. (2011). A nonparametric-test-based structural similarity measure for digital images. *Computational Statistics and Data Analysis*. 55, 2925-2936.

Wang, H., Neill, J.W. and Miller, F.R. (2008). Nonparametric clustering of functional data. *Statistics and Its Interface*. 1, 47-62.

Wang, H., Tolos, S. and Wang, S. (2010). A distribution free nonparametric test to detect dependence between a response variable and covariate in presence of heteroscedastic treatment effects. *The Canadian Journal of Statistics*. 38(3), 408–433.

Wang, H., Zhang, H., Dai, Z., Chen, M., and Yuan, Z. (2013) TSG: A new algorithm for binary and multi-class cancer classification and informative genes selection . *BMC Medical Genomics*. **6**(Suppl 1):S3. doi:10.1186/1755-8794-6-S1-S3

Xie, Y., Zhang, H., **Wang, H.**, Wang, L. and Yuan, Z. (2013) Prediction of multidimensional time series based on GS-RSR-SVR and its application in agricultural economy. *Bulgarian Journal of Agricultural Science*. **19**: 1327-1336.

Zhang, H., **Wang, H.**, Dai, Z., Chen, M., and Yuan, Z. (2012) Improving accuracy for cancer classification with a new algorithm for genes selection. *BMC Bioinformatics*, **13**:298. doi:10.1186/1471-2105-13-298.

Zhang, K., **Wang, H.**, Bathke, A.C., Harrar, S.W., Piepho Hans-Peter, and Deng, Y. (2011).

Gene set analysis for longitudinal gene expression data. *BMC Bioinformatics*. 12:273. http://www.biomedcentral.com/1471-2105/12/273

Zhou, W., Dai, Z., Chen, Y., **Wang, H.**, Chen, M., Yuan, Z. (2012) High-dimensional descriptor selection and computational QSAR modeling for antitumor activity of ARC-111 analogues based on SVR*. International Journal of Molecular Sciences*, **13**, 1161-1172.