# Table of Contents

Schedule webpage: http://ims.nus.edu.sg/events/2018/wstat/wk1.php

# Tutorial on High Dimensional Statistics with Rare and Weak Signals

Xin Tong and Wanjie Wang

*National University of Singapore*

ABSTRACT

Due to the developments of technology, high dimensional data sets are of great interest. Compared to the large amount of observations and features, the signals are usually rare and weak. It poses the question to current statistics. The tutorial mainly discusses the problems and solutions in this field, including the four areas:

1. Global detection that whether the features have any signals or not. Introduce Higher Criticism and innovated Higher Criticism for the global detection problem, and also the related phase diagram.

2. Classification. Discuss the Higher Criticism Classifier with its phase diagram.

3. Clustering. Discuss the Influential Features PCA algorithm, with gene-microarray data sets as examples.

4. Variable selection/signal recovery. Discuss the graphlet screening method.

# Principal Nested Submanifolds

J. S Marron

*The University of North Carolina at Chapel Hill, USA*

ABSTRACT

In the age of Big Data a less well understand, but perhaps greater challenge is the analysis of Complex Data. Many different data types, including various shape representations and other types of imaging based data, are most effectively analyzed in terms of points on a manifold. An important approach to data visualization is Principal Component Analysis (PCA). Various existing methods, including backwards approaches, and Barycentric Subspaces, are seen to be special cases of a general framework called Principal Nested Submanifolds. This framework is seen to present new optimization challenges and open problems.

# Robust Sparse Covariance Estimation by thresholding Tyler's M-estimator

Boaz Nadler[a], John Goes[b], and Gilad Lerman[b]

[a] *Weizmann Institute of Science, Israel*
[b] *University of Minnesota, USA*

ABSTRACT

Estimating a high-dimensional sparse covariance matrix from a limited number of samples is a fundamental problem in contemporary data analysis. Most proposals to date, however, are not robust to outliers or heavy tails. Towards bridging this gap, we consider estimating a sparse shape matrix from $n$ samples following a possibly heavy tailed elliptical distribution. We propose estimators based on thresholding either Tyler's M-estimator or its regularized variant. We derive bounds on the difference in spectral norm between our estimators and the shape matrix in the joint limit as the dimension $p$ and sample size $n$ tend to infinity with $p/n \to \gamma > 0$. These bounds are minimax rate-optimal. Results on simulated data support our theoretical analysis.

## References

[1] Goes J., Lerman G. and Nadler, B. *Robust Sparse Covariance Estimation by thresholding Tyler's M-estimator*, arXiv:1706.08020.

# On Some Statistical Challenges Coming with Non-Euclidean Data

Stephan F. Huckemann[a] and Benjamin Eltzner[a]

[a] *University of Göttingen, Germany*

ABSTRACT

A wide range of statistical applications rely on the central limit theorem (CLT) which assures for random vectors, upon existence of second moments, asymptotic $\sqrt{n}$-normality of the mean. For deviates taking values in a Riemannian manifold, the celebrated CLT of Bhattacharya and Patrangenaru [2] extends this result to Fréchet means of the intrinsic distance. This holds true only under a collection of additional assumptions, and their geometric meaning is not so obvious.

In the sequel the Battacharya-Patrangenaru CLT has been extended to Fréchet means of more general distances, e.g. Procrustes distances on shape spaces [6], to "distances" between data and a data-descriptor, a geodesic say, [7], and even to nested sequences of random descriptors, principal nested spheres [10]. In particular, the latter [9] is a generalization of the classical asymptotics of principal components by Anderson [1] from the 1960s. Still, for all of these general scenarios, analogs of the above mentioned additional assumptions are required to hold true.

Meanwhile, the geometric meaning of these assumptions has been exemplary studied: the distribution near the cut locus of the mean may be responsible for rates slower than $\sqrt{n}$ with non-normal limiting distributions [5], and we call this phenomenon *smeariness*. Asymptotic nonnormality may also be observed on nongeodesic manifolds, even without cut loci [8]. For nonmanifold spaces allowing only for a Riemannian stratification, for example shape spaces due to an isometric group action or phylogentic tree spaces [3], "infinitely fast" rates have been observed, called *stickiness* [4], making straightforward statistical testing impossible.

All of these phenomena may have serious statistical impact, calling for new methodology, and we conclude with some ideas into this direction.

## References

[1] Anderson, T. (1963). Asymptotic theory for principal component analysis. *Ann. Math. Statist. 34*(1), 122–148.

[2] Bhattacharya, R. N., Patrangenaru, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds II. *The Annals of Statistics.* 33 (3), 1225–1259.

[3] Billera, L., S. Holmes, and K. Vogtmann (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics 27*(4), 733–767.

[4] Hotz, T., S. Huckemann, H. Le, J. S. Marron, J. Mattingly, E. Miller, J. Nolen, M. Owen, V. Patrangenaru, and S. Skwerer (2013). Sticky central limit theorems on open books. *Annals of Applied Probability 23*(6), 2238–2258.

[5] Hotz, T. and S. Huckemann (2015). Intrinsic means on the circle: Uniqueness, locus and asymptotics. *Annals of the Institute of Statistical Mathematics 67*(1), 177–193.

[6] Huckemann, S. (2011). Inference on 3D Procrustes means: Tree boles growth, rank-deficient diffusion tensors and perturbation models.*Scandinavian Journal of Statistics* 38 (3), 424–446.

[7] Huckemann, S. (2011). Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth. *The Annals of Statistics 39*(2), 1098–1124.

[8] Huckemann, S., J. C. Mattingly, E. Miller, and J. Nolen (2015). Sticky central limit theorems at isolated hyperbolic planar singularities. *Electronic Journal of Probability*. 20 (78), 1–34.

[9] Huckemann, S. F. and B. Eltzner (2016). Backward nested descriptors asymptotics with inference on stem cell differentiation. *The Annals of Statistics* (accepted) *arXiv preprint arXiv:1609.00814*.

[10] Jung, S., I. L. Dryden, and J. S. Marron (2012). Analysis of principal nested spheres. *Biometrika 99*(3), 551–568.

# Minimax Estimation of Large Precision Matrices with Bandable Cholesky Factor

Zhao Ren

*University of Pittsburgh, USA*

## ABSTRACT

Last decade witnesses significant methodological and theoretical advances in estimating large precision matrices. In particular, there are scientific applications such as longitudinal data, meteorology and spectroscopy in which the ordering of the variables can be interpreted through a bandable structure on the Cholesky factor of the precision matrix. However, the minimax theory has still been largely unknown, as opposed to the well established minimax results over the corresponding bandable covariance matrices. In this paper, we focus on two commonly used types of parameter spaces, and develop the optimal rates of convergence under both the operator norm and the Frobenius norm. A striking phenomenon is found: two types of parameter spaces are fundamentally different under the operator norm but enjoy the same rate optimality under the Frobenius norm, which is in sharp contrast to the equivalence of corresponding two types of bandable covariance matrices under both norms. This fundamental difference is established by carefully constructing the corresponding minimax lower bounds. Two new estimation procedures are developed: for the operator norm, our optimal procedure is based on a novel local cropping estimator targeting on all principle submatrices of the precision matrix while for the Frobenius norm, our optimal procedure relies on a delicate regression-based block-thresholding rule. We further establish rate optimality in the nonparanormal model. Numerical studies are carried out to confirm our theoretical findings. This talk is based on a joint work with my student Yu Liu.

# Inference for Complex Extreme Events

Anthony Davison

*Ecole Polytechnique Fd́érale de Lausanne, Switzerland*

ABSTRACT

The complexity of modern society makes it increasingly vulnerable to rare events, for example financial events such as stock market meltdowns or environmental events such as major storms or wildfires. Fortunately such events are rare, and data on them is limited; indeed, one is often interested in estimating the probabilities of events that have never before occurred. Such estimation involves strong structural assumptions on the tails of the distributions of the underlying data, and these lead naturally to the use of max-stable and related extremal processes for extreme events. Efficient estimation for such models is, however, vexed by a variety of issues. In this talk I shall sketch the area, outline the major problems, and suggest some approaches to dealing with them, using examples from environmental science.

The talk is based on joint work with numerous others, most notably Raphaël de Fondeville.

## References

Asadi, P., Davison, A. C. and Engelke, S. (2015) Extremes on river networks. *Annals of Applied Statistics* **9**, 2023—2050.

Davison, A. C., Padoan, S. A. and Ribatet, M. (2012) Statistical modelling of spatial extremes (with Discussion). *Statistical Science* **27**, 161–186.

de Fondeville, R. and Davison, A. C. (2018) High-dimensional peaks-over-threshold inference. *Biometrika* **105**(to appear).

# Local linear regression on manifolds and its geometric implications

Ming-Yen Cheng[a], Liang Peng[b], Lu-Hung Chen[c], and Kee-Hoon Kang[d]

[a] *Hong Kong Baptist University, Hong Kong*
[a] *National Taiwan University, Taiwan*
[b] *Georgia State University, USA*
[c] *National Chung Hsing University, Taiwan*
[d] *Hankuk University of Foreign Studies*

## ABSTRACT

Multiparameter local likelihood models and generalizations have been widely applied to analyze relationship between responses and covariates in data arising from climatology, econometrics, environmetrics, finance, etc. Both point and interval estimation for the unknown parameter functions have been extensively investigated. However, it remains an unsolved problem how to formally test specification of the conditional density in the model. On one hand, the estimated parameter functions become meaningless or even misleading if specification of the conditional density is inadequate. On the other hand, it is challenging to construct proper test statistics. We can construct Kolmogorov-Smirnov and Cramér-von Mises test statistics and show that formal tests can be constructed if undersmoothing is employed. The asymptotic null distributions of the proposed test statistics depend on the unknown parameter functions, so bootstrap tests are suggested. A simulation study shows the proposed tests have accurate level and nontrivial power. Application to an environmental dataset is also given.

# Inferring gene interactions and functional modules beyond standard statistical models

Haiyan Huang[a], Peter Bickel[a], Rachel Wang[a], Ke Liu[a], Marisa Medina[b], and others[a,b]

[a] *University of California, Berkeley, USA*
[b] *Children's Hospital Oakland Research Institute, USA*

ABSTRACT

Identifying gene interactions has been one of the major tasks in understanding biological processes. However, due to the difficulty in characterizing/inferring different types of biological gene relationships, as well as several computational issues arising from dealing with high-dimensional biological data, finding groups of interacting genes remains challenging. In this talk, I will introduce our recent effort on identifying higher-level gene-gene interactions (i.e., gene group interactions) by evaluating conditional dependencies between genes, i.e., the relationships between genes after removing the influences of a set of other functionally related genes. The detailed technique involves performing sparse canonical correlation analysis with repeated subsampling and random partition. This technique is especially unique and powerful in evaluating conditional dependencies when the correct dependent gene sets are unknown or only partially known. When used effectively, this is a promising technique to recover gene relationships that would have otherwise been missed by standard methods. Comparisons with other methods using simulated and real data show this method achieves considerably lower false positive rates. In addition, I will discuss an ongoing work on using a bagged semi-supervised clustering approach to study changes in the membership of functional gene pathways in response to genetic or phenotypic variation.

## References

[1] YX Rachel Wang, Keni Jiang, Lewis J Feldman, Peter J Bickel, Haiyan Huang (2015). *Inferring gene–gene interactions and functional modules using sparse canonical correlation analysis*, The Annals of Applied Statistics, 9(1), pp. 300–323.

# Differential Connectivity Analysis

Ali Shojaie

*University of Washington, USA*

ABSTRACT

Recent evidence suggests that changes in biological networks, e.g., rewiring or disruption of key interactions, may be associated with development of complex diseases. These findings have motivated new research initiatives in computational and experimental biology that aim to obtain condition-specific estimates of biological networks, e.g. for normal and tumor samples, and identify differential patterns of connectivity in such networks, known as *differential network analysis.*

In this talk, we focus on testing whether two Gaussian graphical models are the same. Existing methods try to accomplish this goal by either directly comparing their estimated structures, or testing the null hypothesis that the partial correlation matrices are equal. Unfortunately, existing inference procedures for this task may lead to misleading results. To address this shortcoming, we propose a two-step inference framework, for testing the null hypothesis that the edge sets in two networks are the same. The proposed framework is especially appropriate if the goal is to identify nodes or edges that show differential connectivity. We investigate theoretical and numerical properties of the proposed framework and illustrate its utility in a study of changes in brain connectivity network associated with mild trauma.

# Gradient Boosting: Overview, Theory and Applications to Big Data Analytics

Tze Leung Lai

*Stanford University, USA*

## ABSTRACT

We begin with a review of the history of gradient boosting, dating back to the LMS algorithm of Widrow and Hoff in 1960 and culminating in Freund and Schapire's AdaBoost and Friedman's gradient boosting and stochastic gradient boosting algorithms in the period 1999-2002 that heralded the big data era. The role played by gradient boosting in big data analytics, particularly with respect to deep learning, is then discussed. We also present some recent work on the mathematical theory of gradient boosting, which has led to some refinements that greatly improves the convergence properties and prediction performance of the methodology.

# Low Rank Matrices with Random Perturbation and Applications

Van Ha, Vu

*National University of Singapore*

ABSTRACT

Many real-life problems (such as clustering, data recovery, sparse sampling etc) have the following common mathematical model: Given a data set of the form A+E, where A is a low rank deterministic matrix and E is random noise, we would like to recover A, with as little error as possible. For instance, the Netflix problem can be Put into this setting.

We are going to present new mathematical tools, which compare the most important spectral parameters of A+E and those of A. These tools enable us to develop a Simple solution to the above problem, and we are going discuss a few applications.

# A Level Set Based Variational Principal Flow Method for Nonparametric Dimension Reduction on Riemannian Manifolds

HAO LIU[a], ZHIGANG YAO[b], SHINGYU LEUNG[a], AND TONY F. CHAN[a]

[a] *The Hong Kong University of Science and Technology, Hong Kong, China*
[b] *National University of Singapore, Singapore*

ABSTRACT

We present a variational formulation for dimension reduction on Riemannian manifolds. The algorithm is developed based on the level set method together with a recently developed principal flow algorithm. The original principal flow algorithm is a Lagrangian technique which extends the principal component analysis (PCA) to dimension reduction on Riemannian manifolds. We propose to incorporate the level set method to obtain a fully implicit formulation so that the overall algorithm can naturally handle various topological changes in the curve evolution. The variational energy consists of two terms which try to balance the contributions from both the dataset itself and the principal direction by the PCA. We will demonstrate that the method is insensitive to the initial guess and is robust enough for noisy data. This is a joint work with H. Liu, Z. Yao and T.F. Chan.

## References

[1] Hao Liu, Zhigang Yao, Shingyu Leung and Tony F. Chan. *A Level Set Based Variational Principal Flow Method for Nonparametric Dimension Reduction on Riemannian Manifolds.* SIAM J. Sci. Comput., 39(4), A1616-A1646, 2017.

# Interaction screening for high dimensional data

Yongdai Kim

*Seoul National University, Korea*

ABSTRACT

LARS [3] is an algorithm to solve the $l_1$ penalized residual sum of squares. When the number of covariates is large and we are interested in interactions, the number of regression coefficients to be considered would be too large to apply the LARS. A typical remedy is to consider a some hierarchical structure between the main effects and interactions such as the weak or strong heredity. In this talk, we propose a modified LARS algorithm which works under such an hierarchical structure. Advantages of our proposed algorithm compared to other penalized methods such as the heirNet by [1] and SHIM of [2] are that our algorithm is much faster and easy to be parallelized so that it can be applied to huge data sets. Moreover, the modified LARS is superior to the forward selection approach proposed by [4] for interaction screening. We also discuss a method which can select interactions without heredity condition, which is a very challenge problem.

## References

[1] Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of statistics*, 41(3), 1111-1141.

[2] Choi, N. H., Li, W., and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489), 354-364.

[3] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of statistics*, 32(2), 407-499.

[4] Hao, N., and Zhang, H. H. (2014). Interaction screening for ultrahigh-dimensional data. Journal of the American Statistical Association, 109(507), 1285-1301.

# Model-free Variable Transformation in High-dimensional Data Analysis

Hui Zou

*University of Minnesota, USA*

## ABSTRACT

Variable transformation is a classical concept in regression. For example, the famous Box-Cox model uses a power or log transformation of the response variable in order to better satisfy the classical normal-linear regression model assumptions. In the Box-Cox model, the transformation is estimated together with other regression parameters. We call this model-based variable transformation. In this talk, we will present a systematic study of model-free variable transformation in some popular high-dimensional learning problems. The message is that variable transformation can greatly facilitate the theoretical and methodological developments of high-dimensional learning algorithms.

# Estimating a covariance function from fragments of functional data

Aurore Delaigle[a], Peter Hall[a], Wei Huang[a], and Alois Kneip[b]

[a] *The University of Melbourne, Australia*
[b] *University of Bonn, Germany*

ABSTRACT

Functional data are often observed only partially, in the form of fragments. In that case, the standard approaches for estimating the covariance function do not work because entire parts of the domain are completely unobserved. In previous work, Delaigle and Hall (2013, 2016) have suggested ways of estimating the covariance function, based for example on Markov assumptions. In this work we take a completely different approach which does not rely on such assumptions. We show that, using a tensor product approach, it is possible to reconstruct the covariance function using observations located only on the diagonal of its domain.

# Detecting Rare and Faint Signals via Thresholding Maximum Likelihood Estimators

Yumou Qiu[a], Song Xi Chen[b], and Dan Nettleton[c]

[a] *University of Nebraska, USA*
[b] *Peking University, China*
[c] *Iowa State Univeristy, USA*

ABSTRACT

Motivated by the analysis of RNA sequencing (RNA-seq) data for genes differentially expressed across multiple conditions, we consider detecting rare and faint signals in high-dimensional response variables. We address the signal detection problem under a general framework, which includes generalized linear models for count-valued responses as special cases. We propose a test statistic that carries out a multi-level thresholding on maximum likelihood estimators (MLEs) of the signals, based on a new Cramér type moderate deviation result for multi-dimensional MLEs. Based on the multi-level thresholding test, a multiple testing procedure is proposed for signal identification. Numerical simulations and a case study on maize RNA-seq data are conducted to demonstrate the effectiveness of the proposed approaches on signal detection and identification.

*Keywords*: Detection boundary, False discovery proportion, Generalized linear model, Moderate deviation, Multiple testing procedure, RNA-seq data

## References

[1] Fujikoshi, Y. (1985). Selection of variables in two-group discriminant analysis by error rate and Akaike's information criteria. *J. Multivariate Anal.*, **17**, 27–37.

[2] Fujikoshi, Y., Sakurai, T. and Yanagihara, H. (2013). Consistency of high-dimensional $AIC$-type and $C_p$-type criteria in multivariate linear regression. *J. Multivariate Anal.*, **149**, 199–212.

[3] Nishii, R. , Bai, Z. D. and Krishnaia, P. R. (1988). Strong consistency of the information criterion for model selection in multivariate analysis. *Hiroshima Math. J.*, **18**, 451–462.

[4] Yanagihara, H., Wakaki, H. and Fujikoshi, Y. (2015). A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large. *Electron. J. Stat.*, **9**, 869–897.

# High-dimensional consistency of efficient screening methods based on information criterion and misclassification rate criterion in discriminant analysis

Yasunori Fujikoshi[a] and Tetsuro Sakurai[b]

[a] *Hiroshima University, Japan*
[b] *Tokyo University of Science, Suwa, Japan*

ABSTRACT

This paper is concerned with asymptotic properties of variable selection methods in discriminant analysis. Under a large sample framework it is known (Fujikoshi [1]) that the selection methods by the misclassification rate criterion (MRC) and the AIC are asymptotically equivalent. Nishii et al. [3] studied consistency properties of a general information criterion (GIC) in multiple discriminant analysis. In this paper we are interested in consistency properties in a high-dimensional framework when the dimension and the sample size are large. Such high-dimensional properties have been studied in multivariate regression model by Fujikoshi et al. [2], Yanagihara et.al. [4], etc. The number of subsets of variables in discriminant analysis becomes extremely large as the dimension is large, and so the methods based on AIC and GIC will be not feasible when the dimension is moderately large. In this paper we consider efficient screening methods, which are feasible, based on AIC, GIC and MRC in two-groups discriminant analysis. Some consistency results are given in high-dimensional framework.

## References

[1] Fujikoshi, Y. (1985). Selection of variables in two-group discriminant analysis by error rate and Akaike's information criteria. *J. Multivariate Anal.*, **17**, 27–37.

[2] Fujikoshi, Y., Sakurai, T. and Yanagihara, H. (2013). Consistency of high-dimensional AIC-type and $C_p$-type criteria in multivariate linear regression. *J. Multivariate Anal.*, **149**, 199–212.

[3] Nishii, R. , Bai, Z. D. and Krishnaia, P. R. (1988). Strong consistency of the information criterion for model selection in multivariate analysis. *Hiroshima Math. J.*, **18**, 451–462.

[4] Yanagihara, H., Wakaki, H. and Fujikoshi, Y. (2015). A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large. *Electron. J. Stat.*, **9**, 869–897.