HDDA-XIII
Schedule and abstracts

Schedule (preliminary version)

- Day 1: August 27 (Tuesday), 2024.
  - 17:00–19:00 registration.
- Day 2: August 28 (Wednesday), 2024.
  - 09:20–10:30 coffee and registration.
  - 10:30–11:00 opening session.
  - 11:00–12:45 “High-dimensional matrices: theory and applications”.
  - 12:45–14:00 lunch.
  - 14:00–15:45 “Cutting-edge machine learning methods for complex biological problems”.
  - 15:45–16:15 coffee break.
  - 16:15–18:00 “High-dimensional models I: grouping and aggregating predictors”.
- Day 3: August 29 (Thursday), 2024.
  - 09:20–10:30 “Mean field and variational approximations”.
  - 10:30–11:00 coffee break.
  - 11:00–12:45 “Bayesian methods in statistics and machine learning”.
  - 12:45–14:00 lunch.
  - 14:00–15:45 “Disease modelling and epidemiology”.
  - 15:45–16:15 coffee break.
  - 16:15–18:00 “High-dimensional models II: robustness”.
- Day 4: August 30 (Friday), 2024.
  - 09:20–10:30 “Data science, AI, and high-dimensional spatiotemporal dynamics”.
  - 10:30–11:00 coffee break.
  - 11:00–12:45 “Recent advances in the theory of machine learning”.
  - 12:45–14:00 lunch.

High-dimensional matrices: theory and applications

Guangming Pan (Nanyang Technological University).

Asymptotic of eigenvectors of large sample covariance matrices. *The talk is about asymptotics of eigenvectors of large sample covariance matrices when the dimension and sample size both tend to infinity with their ratio being a positive constant. The eigenvectors are not limited to those corresponding to the spiked sample eigenvalues. We also explore its application in nonlinear shringage estimators for population covariance matrices.*

Wanjie Wang (National University of Singapore).

Recovery of Timestamps on Noisy High-Dimensional Data . *The analysis of proteins and biological macromolecules is of great interest today, with the development of single-particle cryo-electron microscopy (cryo-EM). The observation Y_i follows $X_{t_i} + Noise$ because of the motions of the molecule. Since the motions repeated a hidden pattern, the ordering of t_i does not have the same ordering with i. Hence, a proper ordering of Y_i will largely improve the recovery of the functional $X(t)$.*
In our work, we present a spectral method on the Laplacian matrix to order Y_i. We first reduce the noise in Y by taking the top eigenvectors of Y. Let Z be the matrix formed by these eigenvectors and we find the ordering of rows in Z. To do it, we first build the Gaussian kernel matrix on Z and then set L_Z to be the Laplacian of the kernel matrix. Ordering the second smallest eigenvector of Z will give the correct ordering of Y. We have set up the theoretical results to show consistency.

Antoine Ledent (Singapore Management University).

On approximate recovery in deep non-linear matrix completion with Schatten-p quasi-norm constraints. *Matrix Completion has a rich history of research from several angles. In the approximate recovery literature, it is known that regardless of the sampling distribution, $\widetilde{O}(n^{3/2}r^{1/2})$ entries are sufficient to approximate the ground truth matrix to a desired degree of accuracy by employing trace norm constraints, where n is the size of the matrix and r is a constraint of a similar scaling to the ground truth rank. In this talk, we extend those results to the case of Schatten p quasi norm constraints, which is known to be equivalent to the popular Deep Matrix Factorization framework. Our distribution-free generalization bound corresponds to a sample complexity of $\widetilde{O}(n^{1+\frac{p}{2}}r^{1-\frac{p}{2}})$. This demonstrates the power of increasing depth in terms of stable rank restriction, and to the best of our knowledge, this is the first non-vacuous result for this setting. Furthermore, we provide a weighted analogue of the Schatten p quasi norm which brings the rate down to $\widetilde{O}(nr)$ in a phenomenon similar to the effect of the weighting on the trace norm regularizer in existing literature. Next, we consider extensions of the model which include nonlinearities and show how to combine our results with neural network bounds in such scenarios. Lastly, we briefly touch upon earlier work on the question of incorporating side information in the form of linear constraints and discuss future avenues of research.*

Special session: "Cutting-edge machine learning methods for complex biological problems", organized by Yi Li

Yuedong Wang (University of California - Santa Barbara).

A Nonparametric Mixed-Effects Mixture Model for Patterns of Clinical Measurements Associated with COVID-19. *Some patients with COVID-19 show changes in signs and symptoms, such as temperature and oxygen saturation, days before being positively tested for SARS-CoV-2, while others remain asymptomatic. It is important to identify these subgroups and to understand what biological and clinical predictors are related to these subgroups. This information will provide insights into how the immune system may respond differently to infection and can further be used to identify infected individuals. We propose a flexible nonparametric mixed-effects mixture model that identifies risk factors and classifies patients with biological changes. We model the latent probability of biological changes using a logistic regression model and trajectories in the latent groups using smoothing splines. We developed an EM algorithm to maximize the penalized likelihood for estimating all parameters and mean functions. We evaluate our methods by simulations and apply the proposed model to investigate changes in temperature in a cohort of COVID-19-infected hemodialysis patients.*

Kin Yau Wong (The Hong Kong Polytechnic University).

Robust score tests with incomplete covariates and high-dimensional auxiliary variables for genomic studies. *Analyses of modern genomic data are often complicated by missing data. When variables of interest are missing for some subjects, it is desirable to use observed auxiliary variables, which are sometimes high dimensional, to impute or predict the missing values in order to improve statistical efficiency. Although many methods have been developed for prediction using high-dimensional variables, it is challenging to perform valid inference based on such predicted values. In this study, we develop a general association test for an outcome variable, which may be continuous, discrete, or censored, and a potentially missing covariate. The missing covariate values can be predicted using variables selected from a set of high-dimensional auxiliary variables. The test is flexible and robust against model misspecification. We establish the validity of the test under data-driven model-selection procedures. We also demonstrate the validity of the proposed method and its advantages over existing methods using extensive simulation studies and applications to major cancer genomic studies.*

Yi Li (University of Michigan).

Penalized Deep Partially Linear Cox Models: Error Rate and Selection Consistency. *Partially linear Cox models have gained popularity for survival analysis by dissecting the hazard function into parametric and nonparametric components, allowing for the effective incorporation of both well-established risk factors (such as age and clinical variables) and emerging risk factors (e.g., image features) within a unified framework. However, when the dimension of parametric components exceeds the sample size, the task of model fitting becomes formidable, while nonparametric modeling grapples with the curse of dimensionality. We propose a novel Penalized Deep Partially Linear Cox Model (Penalized DPLC), which incorporates the SCAD penalty to select important texture features and employs a deep neural network to estimate the nonparametric component of the model. We prove the convergence and asymptotic properties of the estimator and compare it to other methods through extensive simulation studies, evaluating its performance in risk prediction and feature selection.*

High-dimensional models I: grouping and aggregating predictors

Anand N. Vidyashankar (George Mason University).

Grouping Predictors using network-wide metrics and divergences. *When multitudes of features can plausibly be associated with a response, privacy considerations and model parsimony suggest grouping them to increase the predictive power of a regression model. If the set of predictors possesses common characteristics, methods such as group LASSO integrated with tools that account for model selection uncertainty can be used for downstream data analysis. However, if little is known about the features, ad hoc grouping can lead to erroneous inference. In these situations, grouping the predictors using alternate metrics is convenient. We describe a new weighted implicit network approach to group variables where the weights represent the relative association of the variables to the response. Specifically, we describe a new algorithm (supervised learning algorithm) that utilizes network-wide metrics and a sequential testing procedure to find groups of variables that have significant association with the response. For this reason, we describe asymptotic distributional results concerning the network-wide metrics and a novel bootstrap approach for uncertainty assessment. We illustrate the performance of the results using numerical experiments and data from sports analytics.*

Ha (Van) Hoang (University of Science, Vietnam National University).

Functional mixtures-of-experts. *We consider the statistical analysis of heterogeneous data for prediction in situations where the observations include functions, typically time series. We extend the modeling with Mixtures-of-Experts (ME), as a framework of choice in modeling heterogeneity in data for prediction with vectorial observations, to this functional data analysis context. We first present a new family of ME models, named functional ME (FME) in which the predictors are potentially noisy observations, from entire functions. Furthermore, the data generating process of the predictor and the real response, is governed by a hidden discrete variable representing an unknown partition. Second, by imposing sparsity on derivatives of the underlying functional parameters via Lasso-like regularizations, we provide sparse and interpretable functional representations of the FME models called iFME. We develop dedicated expectation–maximization algorithms for Lasso-like (EM-Lasso) regularized maximum-likelihood parameter estimation strategies to fit the models. The proposed models and algorithms are studied in simulated scenarios and in applications to two real data sets, and the obtained results demonstrate their performance in accurately capturing complex nonlinear relationships and in clustering the heterogeneous regression data.*
Joint work with F. Chamroukhi, N. T. Pham and G. J. McLachlan.

Chiung-Yu Huang (University of California, San Francisco).

Synthesizing external aggregated information in the presence of population heterogeneity: A penalized empirical likelihood approach. *With the increasing availability of data in the public domain, there has been a growing interest in exploiting information from external sources to improve the analysis of smaller-scale studies. An emerging challenge in the era of big data is that the subject-level data are high dimensional, but the external information is at an aggregate level and of a lower dimension. Moreover, heterogeneity and uncertainty in the auxiliary information are often not accounted for in information synthesis. We propose a unified framework to summarize various forms of aggregated information via estimating equations and develop a penalized empirical likelihood approach to incorporate such information in logistic regression. When the homogeneity assumption is violated, we extend the method to account for population*

heterogeneity among different sources of information. When the uncertainty in the external information is not negligible, we propose a variance estimator adjusting for the uncertainty. The proposed estimators are asymptotically more efficient than the conventional penalized maximum likelihood estimator and enjoy the oracle property even with a diverging number of predictors. Simulation studies show that the proposed approaches yield higher accuracy in variable selection compared with competitors. We illustrate the proposed methodologies with a pediatric kidney transplant study.

Mean-field and variational approximations

Kamélia Daudel (ESSEC Business School, Paris campus).

Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics. *Variational Inference methods are optimization-based methods that have generated a lot of attention in Bayesian Statistics due to their applicability to high-dimensional machine learning problems. In particular, several algorithms involving the Variational Rényi (VR) bound have been proposed to optimize an alpha-divergence between a target posterior distribution and a variational distribution. Despite promising empirical results, those algorithms resort to biased stochastic gradient descent procedures and thus lack theoretical guarantees. In this paper, we formalize and study the VR-IWAE bound, a generalization of the Importance Weighted Auto-Encoder (IWAE) bound. We show that the VR-IWAE bound enjoys several desirable properties and notably leads to the same stochastic gradient descent procedure as the VR bound in the reparameterized case, but this time by relying on unbiased gradient estimators. We then provide two complementary theoretical analyses of the VR-IWAE bound and thus of the standard IWAE bound. Those analyses shed light on the benefits or lack thereof of these bounds. Lastly, we illustrate our theoretical claims over toy and real-data examples.*
Reference: K. Daudel, J. Benton, Y. Shi and A. Doucet (2023). Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics. Journal of Machine Learning Research, 24(243):1–83.

Atsushi Nitanda (A*STAR CFAR, Singapore).

Convex Analysis of the Mean Field Optimization. *Optimization of mean-field models has recently attracted attention due to its connection to training two-layer neural networks under the mean-field regime. We have established the theory of convex analysis for these models to analyze their optimization dynamics and have derived various optimization methods with convergence guarantees. In this talk, I present recent advances in mean-field optimization methods, including mean-field Langevin dynamics.*

Linda S. L. Tan (National University of Singapore).

Variational inference based on a subclass of closed skew normals. *Gaussian distributions are widely used in Bayesian variational inference to approximate intractable posterior densities, but the ability to accommodate skewness can improve approximation accuracy significantly, especially when data or prior information is scarce. We study the properties of a subclass of closed skew normals constructed using affine transformation of independent standardized univariate skew normals as the variational density, and illustrate how this subclass provides increased flexibility and accuracy in approximating the joint posterior density in a variety of applications by overcoming limitations in existing skew normal variational approximations. The evidence lower bound is optimized using stochastic gradient ascent, where analytic natural gradient updates are derived. We also demonstrate how problems in maximum likelihood estimation of skew normal parameters occur similarly in stochastic variational inference and can be resolved using the centered parametrization.*

Bayesian methods in statistics and machine learning

David Tyler Frazier (Monash University).

Generalized Bayesian Inference with Intractable Discrepancies. *Generalized or Gibbs posteriors are currently limited to situations where the loss function used in the analysis is analytically tractable. This limitation ensures that many interesting classes of loss functions, such as those based on robust measures of distance, e.g., maximum mean discrepancy, cannot be used within Generalized Bayesian inference. We show that so long as one can obtain a reasonable estimate of the analytically intractable loss function, then it is feasible to construct a novel type of Gibbs posterior that is nearly as accurate as if one had access to the intractable loss function. We demonstrate that the price to pay for replacing the intractable loss with an estimated version is a (possible) reduction in the rate of posterior concentration, with the speed of concentration ultimately depending on the quality of the estimated loss. We demonstrate these issues through several examples including generalized Bayesian inference based on maximum mean discrepancy, and likelihoods estimated using kernel densities.*

Cheng Li (National University of Singapore).

Bayesian fixed-domain asymptotics for covariance parameters in spatial Gaussian process models. *Gaussian process models typically contain finite dimensional parameters in the covariance function that need to be estimated from the data. We study the Bayesian fixed-domain asymptotics for the covariance parameters in spatial Gaussian process regression models with an isotropic Matern covariance function, which has many applications in spatial statistics. For the model without nugget, we show that when the dimension of the domain is less than or equal to three, the microergodic parameter and the range parameter are asymptotically independent in the posterior. While the posterior of the microergodic parameter is asymptotically close in total variation distance to a normal distribution with shrinking variance, the posterior distribution of the range parameter does not converge to any point mass distribution in general. For the model with nugget, we derive new evidence lower bound and consistent higher-order quadratic variation estimators, which lead to explicit posterior contraction rates for both the microergodic parameter and the nugget parameter. We further study the asymptotic efficiency and convergence rates of Bayesian kriging prediction. All the new theoretical results are verified in numerical experiments and real data analysis.*

Disease modelling and epidemiology

Jeremy Heng (ESSEC Business School, Singapore campus).

Statistical inference for individual-based models of transmission. *Individual-based models of transmission involve stochastic rules that specify how a number of individuals would infect one another, recover or be removed from the population. To facilitate statistical inference, common yet stringent assumptions stipulate interchangeability of individuals and that all pairwise contacts are equally likely. In this talk, I will discuss two computationally tractable inference strategies when such modeling assumptions are relaxed.*

Shuangge Ma (Yale University).

Heterogeneous Network Analysis of Disease Clinical Treatment Measures via Mining Electronic Medical Record Data. *The analysis of clinical treatment measures has been extensively conducted and can facilitate more effective resource management and planning and also assist better understanding diseases. Most of the existing analyses have been focused on a single disease or a large number of diseases combined. Partly motivated by the successes of gene-centric and phenotypic human disease network (HDN) research, there has been growing interest in the network analysis of clinical treatment measures. However, the existing studies have been limited by a lack of attention to heterogeneity and relevant covariates, ineffectiveness of methods, and low data quality. In this study, our goal is to mine the Taiwan National Health Insurance Research Database (NHIRD), a large population-level electronic medical record (EMR) database, and construct HDNs for number of outpatient visits and medical cost. Significantly advancing from the existing literature, the proposed analysis accommodates heterogeneity and effects of covariates (for example, demographics). Additionally, the proposed method effectively accommodates the zero-inflation nature of data, Poisson distribution, high-dimensionality, and network sparsity. Computational and theoretical properties are carefully examined. Simulation demonstrates competitive performance of the proposed approach. In the analysis of NHIRD data, two and five subject groups are identified for outpatient visit and medical cost, respectively. The identified interconnections, hubs, and network modules are found to have sound implications.*

Yaqing Xu (Shanghai Jiao Tong University School of Medicine).

Hierarchical Multi-Label Classification with Gene-Environment Interactions in Disease Modeling. *In biomedical studies, gene-environment (G–E) interactions have been demonstrated to have important implications for disease prognosis beyond the main G and main E effects. Many approaches have been developed for G-E interaction analysis, yielding important findings. However, hierarchical multi-label classification, which provides insightful information on disease outcomes, remains unexplored in G-E analysis literature. Moreover, unlabeled data is commonly observed in practical settings but omitted by many existing methods of hierarchical multi-label classification. In this study, we consider a semi-supervised scenario and develop a novel approach for the two-layer hierarchical response with G-E interactions. A two-step penalized estimation is then proposed using an efficient expectation-maximization (EM) algorithm. Simulation shows that it has superior accuracy in classification and feature selection performance. The analysis of The Cancer Genome Atlas (TCGA) data on lung cancer demonstrates the practical utility of the proposed approach. Overall, this study can fill the important knowledge gap in G-E interaction analysis by providing a widely applicable framework for hierarchical multi-label classification for complex disease outcomes.*

High-dimensional models II: robustness

David Kepplinger (George Mason University).

Robust Variable Selection Under Adversarial Contamination in High-Dimensional Data. *Outliers and adversarial contamination are common issues when dealing with high-dimensional data. Including a large number of variables in an analysis opens the door for adversarial contamination to render variable selection and statistical inference unreliable. Most often such adversarial values are present in only a few entries of the data table, but we demonstrate that these values can completely distort variable selection and estimation if not handled appropriately. We further highlight that common pre-processing strategies to deal with these outliers poses the risk of removing valuable information, thereby lowering the efficiency of classical variable selection methods, and simultaneously instilling unwarranted confidence in the results. In this talk we present a robust and reliable alternative to classical penalized regression estimators. We show the superior stability of variable selection in the presence of a wide range of adversarial contamination settings and discuss in detail how these properties are achieved.*

Badr-Eddine Chérief-Abdellatif (CNRS, Paris).

Label Shift Quantification via Distribution Feature Matching. *Quantification learning deals with the task of estimating the target label distribution under label shift. In this talk, we present a unifying framework, distribution feature matching (DFM), that recovers as particular instances various estimators introduced in previous literature. We derive a general performance bound for DFM procedures and extend this analysis to study robustness of DFM procedures in the misspecified setting under departure from the exact label shift hypothesis, in particular in the case of contamination of the target by an unknown distribution.*

Mohamed Amezziane (Central Michigan University).

Properties of Coefficients Estimators in Random Design Linear Regression. *We derive the standard errors of the regression coefficients in the general case when the columns of the design matrix are correlated, and we use them to obtain confidence intervals for the coefficients. We extend the results to the high-dimensional setting and investigate through simulation the effects the effects of the curse of dimensionality on the results' accuracy.*

Special session: "Data science, artificial intelligence, and high-dimensional spatiotemporal dynamics", organized by Ivo D. Dinov

Ivo D. Dinov (University of Michigan).

AI and Spacekime Analytics in Health Research and Biomedical Inference. *This talk will present a direct connection between quantum mechanical principles, data science foundations, AI, and statistical inference on repeated longitudinal data. By extending the concepts of time, events, particles, and wavefunctions to complex-time (kime), complex-events, data, and inference-functions, spacekime analytics provides a new foundation for representation, modeling, analyzing, and interpreting dynamic high-dimensional data. We will show the effects of kime-magnitude (longitudinal time order) and kime-phase (related to repeated random sampling) on the induced predictive AI analytics, forecasting, regression, and classification.*
The mathematical foundation of spacekime analytics also provides mechanisms to introduce spacekime calculus, expand Heisenberg's uncertainty principle to reveal statistical implications of inferential uncertainty, and a develop a Bayesian formulation of spacekime inference. Lifting the dimension of time opens a number of challenging theoretical, experimental, and computational data science problems. It leads to a new representation of commonly observed processes from the classical 4D Minkowski spacetime to a 5D spacekime manifold. Using simulated data and clinical observations (e.g., structural and functional MRI), we will demonstrate alternative strategies to transform time-varying processes (time-series) to kime-surfaces and show examples of spacekime analytics.

Eric (Tatt Wei) Ho (Universiti Teknologi PETRONAS).

A survey of opportunities through case studies of Generative AI, adaptation of Large Foundation Models and Physics Informed Neural Networks for high dimensional data analysis. *Deep neural networks have demonstrable ability to learn effective feature representations for state-of-art classifiers and regressors from data. From the perspective of data analytics, the task of extracting salient features from data shares similarities with the signal processing task of learning parsimonious and descriptive feature representations. When paired with sensitivity analysis methods from the explainable AI community, these form a powerful new toolbox to explore the complex associations embedded in high dimensional data. However, training deep neural networks on high dimensional data remains challenging with no guarantee of convergence to a good solution,presumably stymied by the curse of dimensionality. Deep neural network models tend to be overparameterized to facilitate convergence towards a good solution via gradient descent optimization. Often, it is also challenging practically to acquire sufficient high-quality labeled training data. This results in a sparse sampling of the high dimensional data space which introduces challenges to generalization. Training deep neural networks via supervised learning can be conceived as solving an under-determined system of nonlinear equations so model overparameterization and paucity of constraints from training data can be understood as limitations to converging the training of an accurate neural network model. In linear algebra, under-determined systems are solved by imposing additional constraints via regularization. Drawing inspiration from this, I discuss how recent advances in generative AI, adaptation of large foundation models and physics informed neural networks can be conceptualized as imposing additional constraints to ameliorate the challenge of sparse sampling in high dimensional data space. While generative AI attempts to learn additional constraints directly from the training data, transfer learning from large foundation models such as Low Rank Adaptation of Large Models attempt to borrow generalizable constraints from a different data domain whereas physics-informed neural networks impose constraints expressed as differential equations directly in the gradient descent training. Through case studies, I propose some practical approaches to apply* *these concepts and conclude with brief sharing on a method to reduce overparameterization of deep neural networks.*

Recent advances in the theory of machine learning

Makoto Yamada (Okinawa Institute of Science and Technology).

Approximating 1-Wasserstein distance with Trees and its application to KNN and self-supervised learning. *The Wasserstein distance, which measures the discrepancy between distributions, shows efficacy in various types of natural language processing and computer vision applications. One of the challenges in estimating the Wasserstein distance is that it is computationally expensive and does not scale well for many distribution-comparison tasks. In this study, we propose a regression-based approach for approximating the* 1*-Wasserstein distance by the tree-Wasserstein distance (TWD), where the TWD is a* 1*-Wasserstein distance with tree-based embedding that can be computed in linear time with respect to the number of nodes on a tree. We first apply the proposed method for nearest neighbor search problems in NLP tasks and then introduce to use TWD for self-supervised learning.*

Jonathan Scarlett (National University of Singapore).

Recent Developments in High-Dimensional Estimation with Generative Priors. *The problem of estimating an unknown vector (or image) from linear or non-linear measurements has a long history in statistics, machine learning, and signal processing. Classical studies focus on the "$n \gg p$" regime (#measurements $\gg$ #parameters), and more recent studies handle the "$n \ll p$" regime by exploiting low-dimensional structure such as sparsity or low-rankness. Such variants are commonly known as compressive sensing. In this talk, I will overview recent methods that move beyond these explicit notions of structure, and instead assume that the underlying vector is well-modeled by a data-driven generative model (e.g., produced by deep learning methods). I will focus primarily on theoretical developments, including upper and lower bounds on the sample complexity in terms of various properties of the generative model, such as its number of latent (input) parameters, its Lipschitz constant, and its width and depth in the special case of neural network models.*

Masaaki Imaizumi (The University of Tokyo / RIKEN AIP).

Statistical Analysis on Overparameterized Models and In-Context Learning. *Deep learning and artificial intelligence technologies, one of the modern data science technologies, have made great progress, and their mathematical understanding is required to efficiently control and develop these technologies. In this talk, we present two types of research related to this topic. (I) The first is high-dimensional statistics for excess parameter models as typified by large-scale neural networks. Traditional high-dimensional statistics havs developed a methodology to reduce excess dimension. However, since recent large-degree-of-freedom models do not have explicit excess dimension, another theoretical approach has been developed in recent years. We present several results on the application of this approach to more practical statistical models. (II) The second is a statistical analysis of a scheme called in-context learning, which explains foundation models for artificial intelligence such as ChatGPT. We argue that in-context learning can achieve efficient learning under certain conditions, owing to the property of the transformer, which can handle the entire property of empirical distributions.*