Workshop program
Statistical Modelling and Applications

Venue: University of Economics Ho Chi Minh City, Campus B,
279 Nguyen Tri Phuong Street, Dien Hong Ward, Ho Chi Minh City

---

25 July 2024

7:30-8:00	**Arriving/Registration**		
8:00-8:10	**Opening Speech by UEH and VIASM delegates**		**Chair**
8:10 - 8:50	**Edwin Fong** University of Hong Kong, Hong Kong	*Martingale Posteriors for Structured Data*	**Tran Minh Ngoc** University of Sydney, Australia
8:50 – 9:30	**Hien Nguyen** La Trobe University, Australia	*A generalized functional delta method with applications to Bayesian model selection*	
9:30 – 10:10	**Minwoo Chae** Pohang University of Science and Technology, South Korea	*Online Bernstein-von Mises theorem*	
10:10 - 10:30	**Coffee Break**		
10:30 – 11:10	**Trinh Thi Huong** Thuongmai University, Vietnam	*Testing the equality of mean densities with an application to climate change in Vietnam*	**Nguyen Xuan Long** University of Michigan, US
11:10 - 11:50	**Hoang Van Ha** University of Science – VNUHCM, Vietnam	*Goodness-of-fit testing for stationary density of a size-structured PDE*	
11:50 - 14:00	**Lunch**		
14:00- 14:40	**Le Hong Van** Czech Academy of Sciences, Czech Republic	*A universal model for Bayesian supervised learning*	**Ngo Hoang Long** Hanoi National University of Education, Vietnam
14:40 – 15:10	**Antonio Lijoi** Bocconi University, Italy	*Dependent random measures for modelling exchangeable and non-exchangeable data*	
15:10 – 15:30	**Coffee Break**		
15:30 – 16:10	**Takuo Matsubara** University of Edinburgh, UK	*Wasserstein gradient boosting with application for individual-level Bayesian inference and prediction*	**Tran Minh Ngoc** University of Sydney, Australia

16:10 – 16:50	**Duong Duc Lam** LUT University, Finland	*Scalable diffusion posterior sampling in nonparametric Bayesian inverse problems*	

26 July 2024

8:00 - 8:30	**Arriving/Registration**		
08:30 - 9:10	**Ha Quang Minh** RIKEN, Japan	*Statistical Distances and Divergences between Stochastic Processes*	**Ngo Hoang Long** Hanoi National University of Education, Vietnam
9:10 - 9:50	**Jun Zhang** University of Michigan, US	*Box-Cox Transformation and Lambda-Deformation to Hessian Information Geometry*	
9:50 - 10:30	**Dau Hai Dang** National University of Singapore	*Recent advances in Parallel Tempering algorithms*	
10:30-11:00	**Coffee Break**		
11:00-11:40	**Weining Shen** University of California, Irvine, US	*Large language models in sports analytics*	**Nguyen Xuan Long** University of Michigan, US
11:40-12:30	**Yanxun Xu** John Hopkins University, US	*Robust Bayesian learning for individualized treatment rules under unmeasured confounding*	
12:15-14:00	**Lunch**		
Free Afternoon			

Title and abstract details

Antonio Lijoi

Title: Dependent random measures for modeling exchangeable and non-exchangeable data

Dependent discrete nonparametric priors are powerful tools in Bayesian nonparametrics for modeling non-exchangeable data that are affected by some source of heterogeneity. Dependence among samples can be introduced in various ways. A widely used strategy involves compositions of random measures, whose infinite-dimensional structures can capture latent features, account for data heterogeneity, and enable information sharing. An alternative approach employs kernel-weighted discrete random measures, which introduce covariate dependence through their jumps. It will be shown how dependent priors can be effective also when modeling exchangeable data in a survival analysis setting, with competing risks. This talk aims to explore both constructions, emphasizing their key distributional properties, particularly the dependence and clustering structures they induce.

Hien Nguyen

Title: A generalized functional delta method with applications to Bayesian model selection

Delta methods are useful devices for obtaining limiting distributions of transformations of weakly converging statistics using the derivative of the transformation. We extend upon the usual delta method constructions by proposing a variant for general normed vector spaces where the transformation is allowed to depend on the sample size. Our construction allows for the derivation of limiting distributions of various objects defined via posterior averaging or generalisations of such averages. In particular, our derivation allows for the statement of a model selection theorem for models that are characterised via generalized posterior measures.

Duong Duc Lam

Title: Unconditional representation of the conditional score in nonparametric Bayesian inverse problems

Abstract: Score-based diffusion models (SDMs) have emerged as a powerful tool for sampling from the posterior distribution in Bayesian inverse problems. However, existing methods often require multiple evaluations of the forward mapping to generate a single sample, resulting in significant computational costs for large-scale inverse problems. To address this limitation, we propose an unconditional representation of the conditional score-function (UCoS) tailored to linear nonparametric inverse problems, which avoids forward model evaluations during sampling by shifting computational effort to an offline training phase. In this phase, a task-dependent score function is learned based on the linear forward operator. Crucially, the conditional posterior score is derived exactly from the trained score using affine transformations, eliminating the need for conditional score approximations. Our approach is shown to work in infinite-dimensional diffusion models and is supported by rigorous convergence analysis. We validate UCoS through high-dimensional computed tomography (CT) and image deblurring experiments, demonstrating both scalability and accuracy. This is a joint work with F. Schneider, M. Lassas, M. de Hoop and T. Helin.

Dau Hai Dang

Title: Recent advances in Parallel Tempering algorithms

State-of-the-art Monte Carlo techniques such as Sequential Monte Carlo (SMC) and Parallel Tempering (PT) overcome the poor mixing of MCMC by traversing a sequence of intermediate distributions. While the theory and practice of SMC have been well investigated, Parallel Tempering has received less attention. In this talk we explore the basics of Parallel Tempering algorithms and recent advances in this research area which incorporate neural maps and more efficiently exploit parallel hardware by using multiple particles per temperature.

Edwin Fong

Title: Martingale Posteriors for Structured Data

The martingale posterior is a framework for Bayesian inference where posterior uncertainty is generated through predictive imputation. As a result, the Bayesian model can be directly specified using a sequence of predictive distributions, bypassing the need for explicit likelihood and prior specifications.

Extending this framework beyond the i.i.d. setting however presents challenges, particularly in eliciting predictive distributions that account for the inherent structure of the data. This talk outlines recent advancements in constructing nonparametric martingale posteriors for more complex settings, including quantile regression, dependent density estimation, and hierarchical models.

Hoang Van Ha

Title: Goodness-of-fit testing for stationary density of a size-structured PDE

We consider two division models for structured cell populations, where cells can grow, age and divide. These models have been introduced in the literature under the denomination of 'mitosis' and 'adder' models. In the recent years, there has been an increasing interest in Biology to understand whether the cells divide equally or not, as this can be related to important mechanisms in cellular aging or recovery. We are therefore interested in testing the null hypothesis H0 where the division of a mother cell results into two daughters of equal size or age, against the alternative hypothesis H1 where the division is asymmetric and ruled by a kernel that is absolutely continuous with respect to the Lebesgue measure. The sample consists of i.i.d. observations of cell sizes and ages drawn from the population, and the division is not directly observed. The hypotheses of the test are reformulated as hypotheses on the stationary size and age distributions of the models, which we assume are also the distributions of the observations. We propose a goodness-of-fit test that we study numerically on simulated data before applying it on real data.

Le Hong Van

Title: A universal model for Bayesian supervised learning

In my talk I shall propose a universal model for Bayesian supervised learning encompassing Bayesian density and regression estimations, using functorial properties of Bayesian inversions. I shall also discuss the choice of priors for this general model when a label space is a Polish space.

Takuo Matsubara

Title: Wasserstein Gradient Boosting with Application for Individual-Level Bayesian Inference and Prediction

Gradient boosting is a sequential ensemble method that fits a new weaker learner to pseudo residuals at each iteration. We propose Wasserstein gradient boosting, a novel extension of gradient boosting that fits a new weak learner to alternative pseudo residuals that are Wasserstein gradients of loss functionals of probability distributions assigned at each input. It solves vector-to-distribution regression, in which the output value associated with each input value is probability distribution. The main application of Wasserstein gradient boosting is individual-level Bayesian inference, where the model and posterior are given for every single datum pointwise similarly to mixed-effects modelling. Given unobserved new datum as input, Wasserstein gradient boosting provides a nonparametric particle-based prediction of the corresponding posterior. We empirically demonstrate the superior performance of the probabilistic prediction by Wasserstein gradient boosting in comparison with existing uncertainty quantification methods.

Minwoo Chae

Title: Online Bernstein-von Mises theorem

Online learning is an inferential paradigm in which parameters are updated incrementally from sequentially available data, in contrast to batch learning, where the entire dataset is processed at once. In this talk, we assume that mini-batches from the full dataset become available sequentially. The Bayesian framework, which updates beliefs about unknown parameters after observing each mini-batch, is naturally suited for online learning. At each step, we update the posterior distribution using the current prior and new observations, with the updated posterior serving as the prior for the next step. However, this recursive Bayesian updating is rarely computationally tractable unless the model and prior are conjugate. When the model is regular, the updated posterior can be approximated by a normal distribution, as justified by the Bernstein-von Mises theorem. We adopt a variational approximation at each step and investigate the frequentist properties of the final posterior obtained through this sequential procedure. Under mild assumptions, we show that the accumulated approximation error becomes negligible once the mini-batch size exceeds a threshold depending on the parameter dimension. As a result, the sequentially updated posterior is asymptotically indistinguishable from the full posterior.

Trinh Thi Huong

Title: Testing the equality of mean densities with an application to climate change in Vietnam

Given samples of density functions on an interval $(a,b),$ categorized according to a factor variable, we aim to test the equality of their mean functions both overall and across the groups defined by the factor. While the Functional Analysis of Variance (FANOVA) methodology is well-established for functional data, its adaptation to density functions (DANOVA) is necessary due to their inherent constraints of positivity and unit integral. To accommodate these constraints, we naturally use Bayes spaces methodology by mapping the densities using the centered log-ratio transformation into the $L^2_0(a,b)$ space where we can use FANOVA techniques. Traditional contrasts in FANOVA rely on squared differences and can be reinterpreted as squared distances between Bayes perturbations within the densities space. We illustrate our

methodology on a dataset comprising daily maximum temperatures across Vietnamese provinces between 1987 and 2016. Within the context of climate change, we first investigate the existence of a non zero temporal trend of the densities of daily maximum temperature over Vietnam and then examine whether there is any regional effect on these trends. Finally, we explore odds ratio based interpretations allowing us to describe the trends more locally.
This is joint work with Camille Mondon (Toulouse School of Economics, France), Christine Thomas-Agnan (Toulouse School of Economics, France) and Josep Antoni Martín-Fernández (Universitat de Girona, Spain).

Weining Shen

Title: Large language models in sports analytics

In this talk, I will discuss recent work on evaluating the sports understanding of LLMs, using newly introduced benchmark datasets. Our evaluation covers a range of tasks, from basic queries about rules and historical facts to complex, context-specific reasoning, as well as assessing the sports reasoning capabilities of video language models. Experiments show that models fall short on hard tasks that require deep reasoning and rule-based understanding. We hope the published benchmarks will serve as a critical step toward improving models' capabilities in sports understanding and reasoning.

Yanxun Xu

Title: Robust Bayesian Learning for Individualized Treatment Rules Under Unmeasured Confounding

Data-driven personalized decision-making has become increasingly important in many scientific fields. Most existing methods rely on the assumption of no unmeasured confounding to establish causal inferences before proceeding with decision-making for identifying the optimal individualized treatment rule (ITR). However, this assumption is often violated in practice, especially in observational studies. While techniques like instrumental variables or proxy variables can help address unmeasured confounding, such additional data sources are not always available. Moreover, robustly learning the optimal ITR from observational data is challenging when data are unbalanced, where certain combinations of treatments and patient characteristics are underrepresented. In this paper, we develop a novel Bayesian approach to robustly learn the optimal ITR for continuous treatments under unmeasured confounding. For causal identification, we propose a Bayesian causal model that achieves unique identification under certain mild distributional assumptions, without requiring additional data sources. For policy optimization, we develop a practical algorithm that robustly learns the optimal ITR by identifying a conservative policy. Through simulations and an application to a large-scale kidney transplantation dataset, we demonstrate the proposed method's identifiability, utility, and robustness, highlighting its value in advancing precision medicine.

Ha Quang Minh

Title: Statistical Distances and Divergences between Stochastic Processes

Distances and divergences between probability distributions play a crucial role in many areas of probability theory, statistics, machine learning, and their applications. While a large part of the literature is focused on divergences between finite-dimensional distributions, there is a growing body of work on infinite-dimensional distances/divergences, which are motivated by applications in functional data analysis, Bayesian inverse problems, and functional Bayesian neural networks, among others. In this talk, we present an overview of recent results on some of the most important distances/divergences being studied, including the Wasserstein and Fisher-Rao distances, Kullback-Leibler, Renyi, and Geometric Jensen-Shannon divergences. We discuss the many challenges that arise in the infinite-dimensional setting, e.g. the lack of a natural reference measure such as the Lebesgue measure and the fact that many functions such as determinants and logarithm are only well-defined in specific settings, along with their resolutions. For the case of Gaussian measures on Hilbert space, we present closed form formulas for all of the above distances/divergences. Using the methodology of reproducing kernel Hilbert spaces (RKHS), we furthermore obtain consistent finite-dimensional approximations of the above divergences in the Gaussian process setting, with dimensional-independent sample complexities. The resulting numerical algorithms can be readily employed in practical applications.

Jun Zhang

Title: Box-Cox Transformation and Lambda-Deformation to Hessian Information Geometry

Box-Cox transformation (using a parameter lambda) is a familiar tool in statistics for converting a random variable into one with more normal shape. This transformation also connects Renyi entropy to Tsallis entropy, two popular versions of deformation to Shannon entropy. Here, an information-geometric analysis of this transformation/deformation is performed using the method of abstract convex duality (notably c-duality related to the optimal transport theory). After a review classical information geometry, especially the tight interlocking between exponential (and mixture family) family, maximum entropy inference and Hessian geometry, we investigate the lambda-deformation to the Legendre duality to derive the lambda-exponential family and mixture families, along with deforming Kullback-Leibler divergence to Renyi and Tsallis divergence and deforming Bregman divergence to logarithmic divergence. This suite of lambda-deformation turns out, geometrically, to reveal a dually projectively-flat (and hence conformal Hessian) generalization to the dually flat Hessian information geometry. (Joint work with T.K. Leonard Wong)