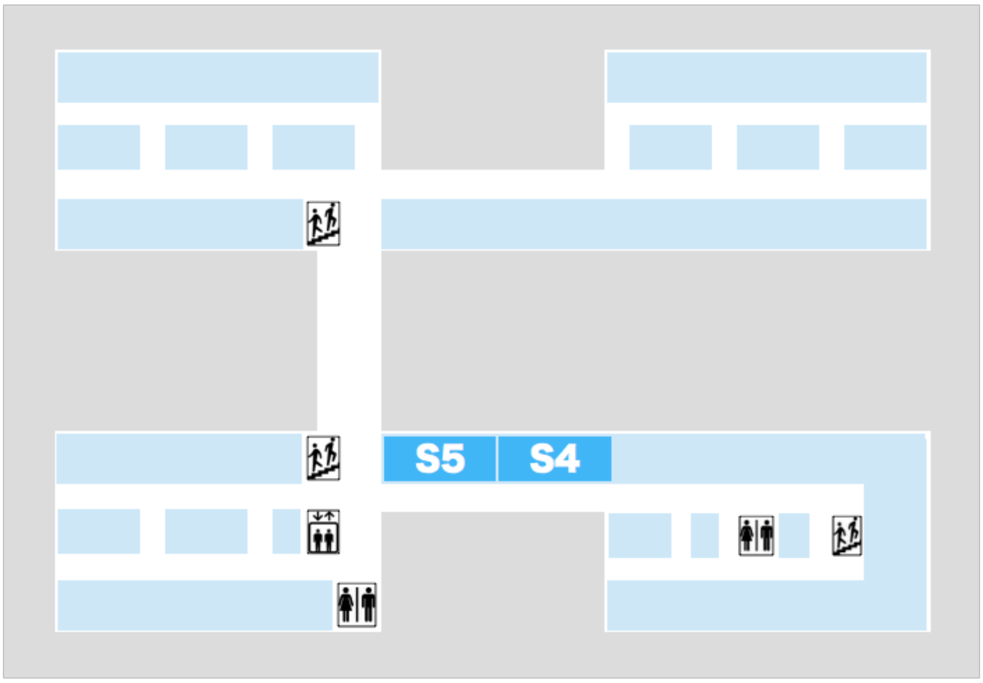
FIMI2018

19 February - 21 February 2018

The Institute of Statistical Mathematics

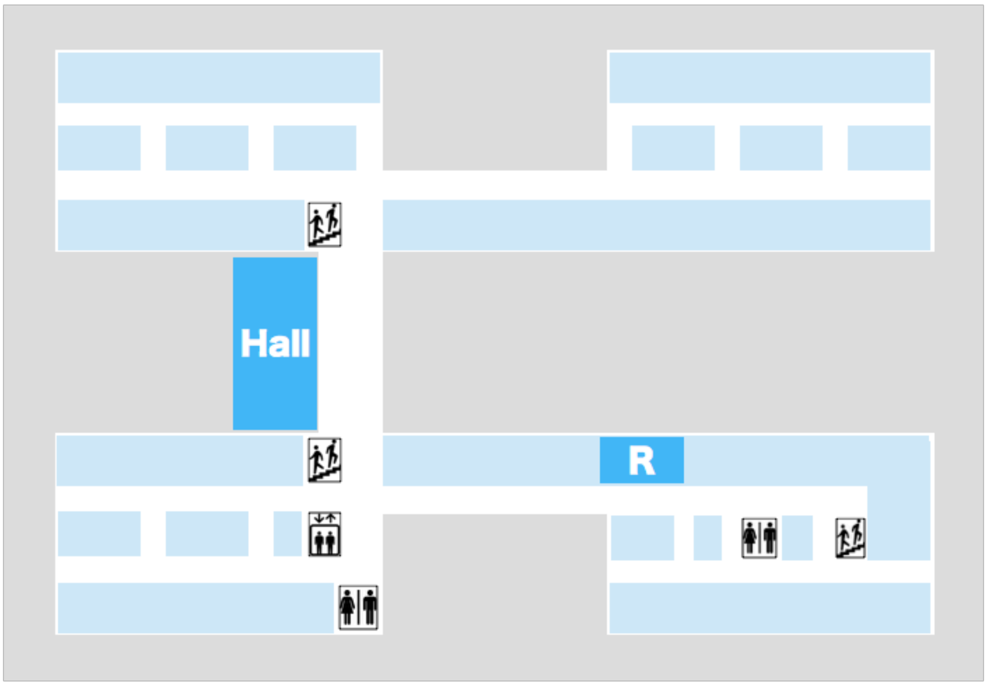
Floor Map

3rd Floor



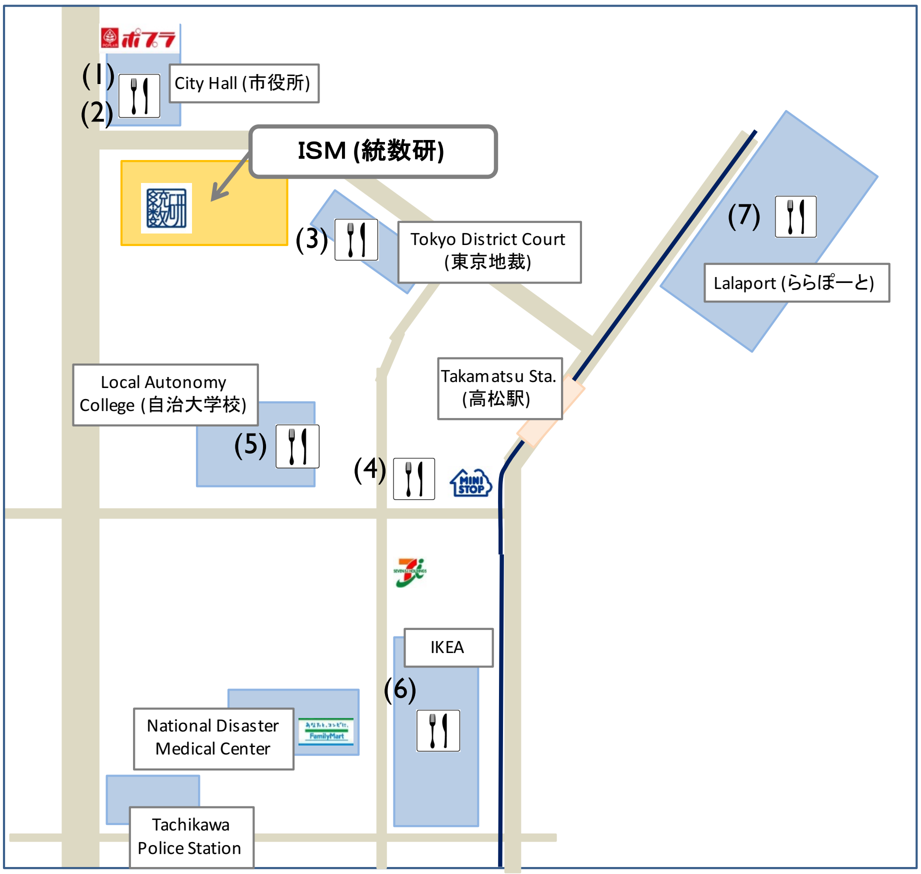
S5: Seminar Room 5 (Session)
S4: Seminar Room 4 (Coffee Room)

2nd Floor



Hall: Auditorium (Tuesday only)
R: Resting room D220 (Tuesday only)

Lunch Map



Pack lunches （at the lobby lounge） 11：00～13：00

Rin rin:	450~500yen each	
Haiji (curry)	500yen each	(a light dessert offered on Wednesdays)
Zuikyo (Chinese)	500yen each	
Hello Lunch	300~400yen each	(miso-soup included)

Restaurant

(1) City Hall dining room, 3rd floor. (立川市役所食堂(3F))	300~500yen
(2) Café Harmony at the City Hall, 1st floor. (カフェハーモニー, 立川市役所 1F)	600yen

(3) Pole Light, Tokyo District Court dining room, 1st basement floor. (食堂ポールライト, 東京地裁 B1F)

(4) Chinese Restaurant Zuikyo, near MINISTOP. (中華料理店 瑞京),

(5) Local Autonomy College dining room. (自治大学校食堂)

(6) IKEA restaurant and café, 2nd floor. (IKEA レストラン 2F)

(7) Restaurants in Lalaport (shopping mall). (ららぽーと レストラン街)

Convenience Stores

Popula (City Hall 1F) / MINISTOP / 7-Eleven / Family Mart

Program

19 February (Mon)

10:00-10:10	Opening
10:10-11:00	Arthur Gretton (University College London) Conditional Densities and Efficient Models in Infinite Exponential Families
11:15-12:05	Bharath K Sriperumbudur (Pennsylvania State University) On Approximate Kernel PCA Using Random Features
12:05-13:45	Lunch Break
13:45-14:35	Motonobu Kanagawa (Max Planck Institute for Intelligent Systems) Convergence Analysis of Deterministic Kernel-Based Quadrature Rules in Misspecified Settings
14:50-15:40	Krikamol Muandet (Mahidol University) Eigendecompositions of Transfer Operators in Reproducing Kernel Hilbert Spaces

20 February (Tue)

09:35-10:25	Song Liu (Bristol University) Density Ratio Estimation using Stein Method and Its Applications
10:40-11:30	Kenji Fukumizu (The Institute of Statistical Mathematics) Machine Learning Approach to Topological Data Analysis
11:30–13:10	Lunch Break
13:10–14:00	Alexandre Tsybakov (CREST, ENSAE) Optimal Variable Selection and Noisy Adaptive Compressed Sensing
14:15-15:05	Mladen Kolar (University of Chicago) Estimation and Inference for Differential Networks
15:20-16:10	Wittawat Jitkrittum (Max Planck Institute for Intelligent Systems) A Linear-Time Kernel Goodness-of-Fit Test
18:00-	Workshop Banquet

Program

21 February (Wed)

09:35-10:25	Taiji Suzuki (The University of Tokyo) Connecting Model Compression and Generalization Analysis for Deep Neural Network
10:40-11:30	Yarin Gal (University of Oxford) Bayesian Deep Learning
11:30–13:10	Lunch Break
13:10–14:00	Johannes Schmidt-Hieber (Leiden University) Statistical Theory for Deep Neural Networks with ReLU Activation Function
14:05-14:55	Masaaki Imaizumi (The Institute of Statistical Mathematics) Statistical Estimation for Non-Smooth Functions by Deep Neural Networks
14:55-15:00	Closing

Abstract (Monday)

Arthur Gretton (University College London)

Conditional Densities and Efficient Models in Infinite Exponential Families

The exponential family is one of the most powerful and widely used classes of models in statistics. A method was recently developed to fit this model when the natural parameter and sufficient statistic are infinite dimensional, using a score matching approach. The infinite exponential family is a natural generalisation of the finite case, much like the Gaussian and Dirichlet processes generalise their respective finite models.

In this talk, I'll describe two recent results which make this model more applicable in practice, by reducing the computational burden and improving performance for high-dimensional data. The first is a Nytsrom-like approximation to the full solution. We prove that this approximate solution has the same consistency and convergence rates as the full-rank solution (exactly in Fisher distance, and nearly in other distances), with guarantees on the degree of cost and storage reduction. The second result is a generalisation of the model family to the conditional case, again with consistency guarantees. In experiments, the conditional model generally outperforms a competing approach with consistency guarantees, and is competitive with a deep conditional density model on datasets that exhibit abrupt transitions and heteroscedasticity.

Bharath K Sriperumbudur (Pennsylvania State University)

On Approximate Kernel PCA Using Random Features

Kernel methods are powerful learning methodologies that provide a simple way to construct nonlinear algorithms from linear ones. Despite their popularity, they suffer from poor scalability in big data scenarios. Various approximation methods, including random feature approximation, have been proposed to alleviate the problem. However, the statistical consistency of most of these approximate kernel methods is not well understood except for kernel ridge regression wherein it has been shown that the random feature approximation is not only computationally efficient but also statistically consistent with a minimax optimal rate of convergence. In this work, we investigate the efficacy of random feature approximation in the context of kernel principal component analysis (KPCA) by studying the statistical behavior of approximate KPCA in terms of the convergence of eigenspaces and the reconstruction error.

Abstract (Monday)

Motonobu Kanagawa (Max Planck Institute for Intelligent Systems)

Convergence Analysis of Deterministic Kernel-Based Quadrature Rules in Misspecified Settings

In this talk, we present convergence analysis of kernel-based quadrature rules in misspecified settings, focusing on deterministic quadrature in Sobolev spaces. In particular, we deal with misspecified settings where a test integrand is less smooth than a Sobolev RKHS based on which a quadrature rule is constructed. We provide convergence guarantees based on two different assumptions on a quadrature rule: one on quadrature weights, and the other on design points. More precisely, we show that convergence rates can be derived (i) if the sum of absolute weights remains constant (or does not increase quickly), or (ii) if the minimum distance between distance design points does not decrease very quickly. As a consequence of the latter result, we derive a rate of convergence for Bayesian quadrature in misspecified settings. We reveal a condition on design points to make Bayesian quadrature robust to misspecification, and show that, under this condition, it may adaptively achieve the optimal rate of convergence in the Sobolev space of a lesser order (i.e., of the unknown smoothness of a test integrand), under a slightly stronger regularity condition on the integrand.
(Joint work with Bharath K. Sriperumbudur and Kenji Fukumizu)

Krikamol Muandet (Mahidol University)

Eigendecompositions of Transfer Operators in Reproducing Kernel Hilbert Spaces

Transfer operators such as the Perron-Frobenius or Koopman operator play an important role in the global analysis of complex dynamical systems. The eigenfunctions of these operators can be used to detect metastable sets, to project the dynamics onto the dominant slow processes, or to separate superimposed signals. We extend transfer operator theory to reproducing kernel Hilbert spaces and show that these operators are related to Hilbert space representations of conditional distributions, known as conditional mean embeddings in the machine learning community. Moreover, numerical methods to compute empirical estimates of these embeddings are akin to data-driven methods for the approximation of transfer operators such as extended dynamic mode decomposition and its variants. In fact, most of the existing methods can be derived from our framework, providing a unifying view on the approximation of transfer operators. One main benefit of the presented kernel-based approaches is that these methods can be applied to any domain where a similarity measure given by a kernel is available. We illustrate the results with the aid of guiding examples and highlight potential applications in molecular dynamics as well as video and text data analysis.

Abstract (Tuesday)

Song Liu (Bristol University)

Density Ratio Estimation using Stein Method and Its Applicaitons

In this research, we estimate the ratio between the data generating probability density function and a model density function with the help of Stein operator. The estimated density ratio allows us to approximate the Kullback-Leibler divergence from a model to the data efficiently. We explore applications using such a goodness of fit measure including parameter fitting, Bayesian learning and change point detection.
This is a joint work with Wittawat Jitkrittum and Carl Henrik Ek.

Lizhen Lin (The University of Notre Dame)

Geometry and Statistics: Nonparametric Statistical Inference of Non-Euclidean Data

This talk presents some recent advances in nonparametric inference on manifolds and other non-Euclidean spaces. The focus is on nonparametric inference base on Frechet means.
In particular, we present omnibus central limit theorems for Frechet means for inference, which can be applied to general metric spaces including stratified spaces, greatly expanding the current scope of inference. Applications are also provided to the space of symmetric positive definite matrices arising in diffusion tensor imaging. A robust framework based on the classical idea of median-of-means is also proposed which yields estimates with provable robustness and improved concentration. In addition to inferring i.i.d data, we also consider nonparametric regression problems where predictors or responses lying on manifolds. Various simulated or real data examples are considered.

Abstract (Tuesday)

Alexandre Tsybakov (CREST, ENSAE)

Optimal Variable Selection and Noisy Adaptive Compressed Sensing

We consider variable selection based on n observations from a high-dimensional linear regression model. The unknown parameter of the model is assumed to belong to the class S of all s-sparse vectors in $\mathbb{R}^p$ whose non-zero components are greater than $a > 0$. Variable selection in this context is an extensively studied problem and various methods of recovering sparsity pattern have been suggested. However, in the theory not much is known beyond the consistency of selection. For Gaussian design, which is of major importance in the context of compressed sensing, necessary and sufficient conditions of consistency for some configurations of n, p, s, a are available. They are known to be achieved by the exhaustive search decoder, which is not realizable in polynomial time and requires the knowledge of s. This talk will focus on the issue of optimality in variable selection based on the Hamming risk criterion. The benchmark behavior is characterized by the minimax risk on the class S. We propose an adaptive algorithm independent of \$s,a\$, and of the noise level that nearly attains the value of the minimax risk. This algorithm is the first method, which is both realizable in polynomial time and is consistent under the same (minimal) sufficient conditions as the exhaustive search decoder.

Mladen Kolar (University of Chicago)

Estimation and Inference for Differential Networks

We present a recent line of work on estimating differential networks and conducting statistical inference about parameters in a high-dimensional setting. First, we consider a Gaussian setting and show how to directly learn the difference between the graph structures. A debiasing procedure will be presented for construction of an asymptotically normal estimator of the difference. Next, building on the first part, we show how to learn the difference between two graphical models with latent variables. Linear convergence rate is established for an alternating gradient descent procedure with correct initialization. Simulation studies illustrate performance of the procedure. We also illustrate the procedure on an application in neuroscience. Finally, we will discuss how to do statistical inference on the differential networks when data are not Gaussian.

Abstract (Tuesday)

Wittawat Jitkrittum (Max Planck Institute for Intelligent Systems)

A Linear-Time Kernel Goodness-of-Fit Test

We propose a novel adaptive test of goodness of fit, with computational cost linear in the number of samples. We learn the test features that best indicate the differences between observed samples and a reference model, by minimizing the false negative rate. These features are interpretable, indicating where the model does not fit the samples well.The features are constructed via Stein's method, meaning that it is not necessary to compute the normalising constant of the model.We analyse the asymptotic Bahadur efficiency of the new test, and prove that under a mean-shift alternative, our test always has greater relative efficiency than a previous linear-time kernel test, regardless of the choice of parameters for that test. In experiments, the performance of our method exceeds that of the earlier linear-time test, and matches or exceeds the power of a quadratic-time kernel test. In high dimensions and where model structure may be exploited, our goodness of fit test performs far better than a quadratic-time two-sample test based on the Maximum Mean Discrepancy, with samples drawn from the model.

Kenji Fukumizu (The Institute of Statistical Mathematics)

Machine Learning Approach to Topological Data Analysis

Topological data analysis (TDA) is a recent methodology for extracting topological and geometrical features from complex geometric data structures. Persistent homology, a new mathematical notion proposed by Edelsbrunner (2002), provides a multiscale descriptor for the topology of data, and has been recently applied to a variety of data analysis. In this talk I will introduce a machine learning framework of TDA by combining persistence homology and kernel methods. As an expression of persistent homology, persistence diagrams are widely used to represent the lifetimes of generators of homology groups.While they serve as a compact representation of data, it is not straightforward to apply standard methodology of data analysis since they consist of a set of points in 2D space expressing the lifetimes.We introduce a method of kernel embedding of the persistence diagrams to obtain their vector representation, which enables one to apply any kernel methods in topological data analysis, and propose a persistence weighted Gaussian kernel as a suitable kernel for vectorization of persistence diagrams. Some theoretical properties including Lipschitz continuity of the embedding are discussed. I will also present applications to change point detection and time series analysis in the field of material sciences and biochemistry.

Abstract (Wednesday)

Taiji Suzuki (The University of Tokyo)

Connecting Model Compression and Generalization Analysis for Deep Neural Network

In this talk, we consider a model compression problem for deep neural network models and show its connection to generalization error analysis. The generalization analysis is based on the eigenvalue distribution of the kernel functions defined in the internal layers. It gives a fast learning rate and the obtained convergence rate is almost independent on the network size unlike the previous analysis. Based on the analysis, we develop a simple compression algorithm for the neural network which is applicable to wide range of network models.

Yarin Gal (University of Oxford)

Bayesian Deep Learning

Bayesian models are rooted in Bayesian statistics and easily benefit from the vast literature in the field. In contrast, deep learning lacks a solid mathematical grounding. Instead, empirical developments in deep learning are often justified by metaphors, evading the unexplained principles at play. These two fields are perceived as fairly antipodal to each other in their respective communities. It is perhaps astonishing then that most modern deep learning models can be cast as performing approximate inference in a Bayesian setting. The implications of this are profound: we can use the rich Bayesian statistics literature with deep learning models, explain away many of the curiosities with this technique, combine results from deep learning into Bayesian modeling, and much more.

In this talk I will review a new theory linking Bayesian modeling and deep learning and demonstrate the practical impact of the framework with a range of real-world applications. I will also explore open problems for future research—problems that stand at the forefront of this new and exciting field.

Abstract (Wednesday)

Johannes Schmidt-Hieber (Leiden University)

Statistical Theory for Deep Neural Networks with ReLU Activation Function

The universal approximation theorem states that neural networks are capable of approximating any continuous function up to a small error that depends on the size of the network. The expressive power of a network does, however, not guarantee that deep networks perform well on data. For that, control of the statistical estimation risk is needed. In the talk, we derive statistical theory for fitting deep neural networks to data generated from the multivariate nonparametric regression model. It is shown that estimators based on sparsely connected deep neural networks with ReLU activation function and properly chosen network architecture achieve the minimax rates of convergence (up to logarithmic factors) under a general composition assumption on the regression function. The framework includes many well-studied structural constraints such as (generalized) additive models. While there is a lot of flexibility in the network architecture, the tuning parameter is the sparsity of the network. Specifically, we consider large networks with number of potential parameters being much bigger than the sample size. Interestingly, the depth (number of layers) of the neural network architectures plays an important role and our theory suggests that scaling the network depth with the logarithm of the sample size is natural.

Masaaki Imaizumi (Institute of Statistical Mathematics)

Statistical Estimation for Non-Smooth Functions by Deep Neural Networks

We theoretically discuss why deep neural networks (DNNs) performs better than other models in some cases by investigating statistical properties of DNNs for non-smooth functions. While DNNs have empirically shown higher performance than other standard methods, understanding its mechanism is still a challenging problem. From an aspect of the statistical theory, it is known many standard methods attain optimal convergence rates, and thus it has been difficult to find theoretical advantages of DNNs. This paper fills this gap by considering learning of a certain class of non-smooth functions, which was not covered by the previous theory. We derive convergence rates of estimators by DNNs with a ReLU activation, and show that the estimators by DNNs are almost optimal to estimate the non-smooth functions, while some of the popular models do not attain the optimal rate. In addition, our theoretical result provides guidelines for selecting an appropriate number of layers and edges of DNNs. We provide numerical experiments to support the theoretical results.

Organizers

Masaaki Imaizumi (The Institute of Statistical Mathematics)
Arthur Gretton (University College London)
Kenji Fukumizu (The Institute of Statistical Mathematics)

This workshop is supported by
Research Center for Statistical Machine Learning, ISM, and Grant-in-Aid for Scientific Research on Innovative Areas, "Initiative for High-Dimensional Data-driven Science through Deepening of Sparse Modeling".