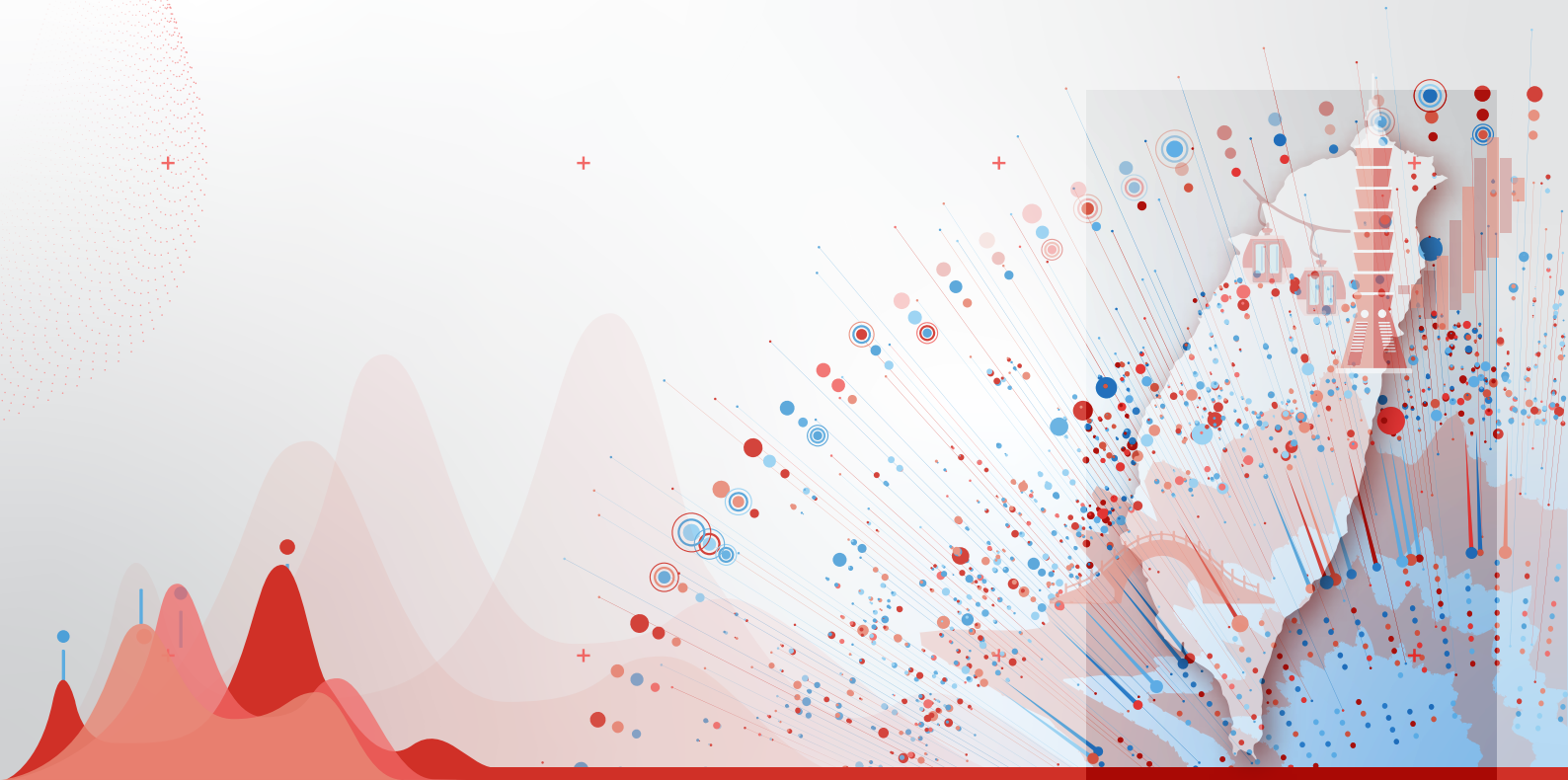
SCRI 2023

# Statistical Computing and Robust Inference for High Dimensional Data

December 11th (Mon.) - 14th (Thurs.)

# PROGRAM BOOK

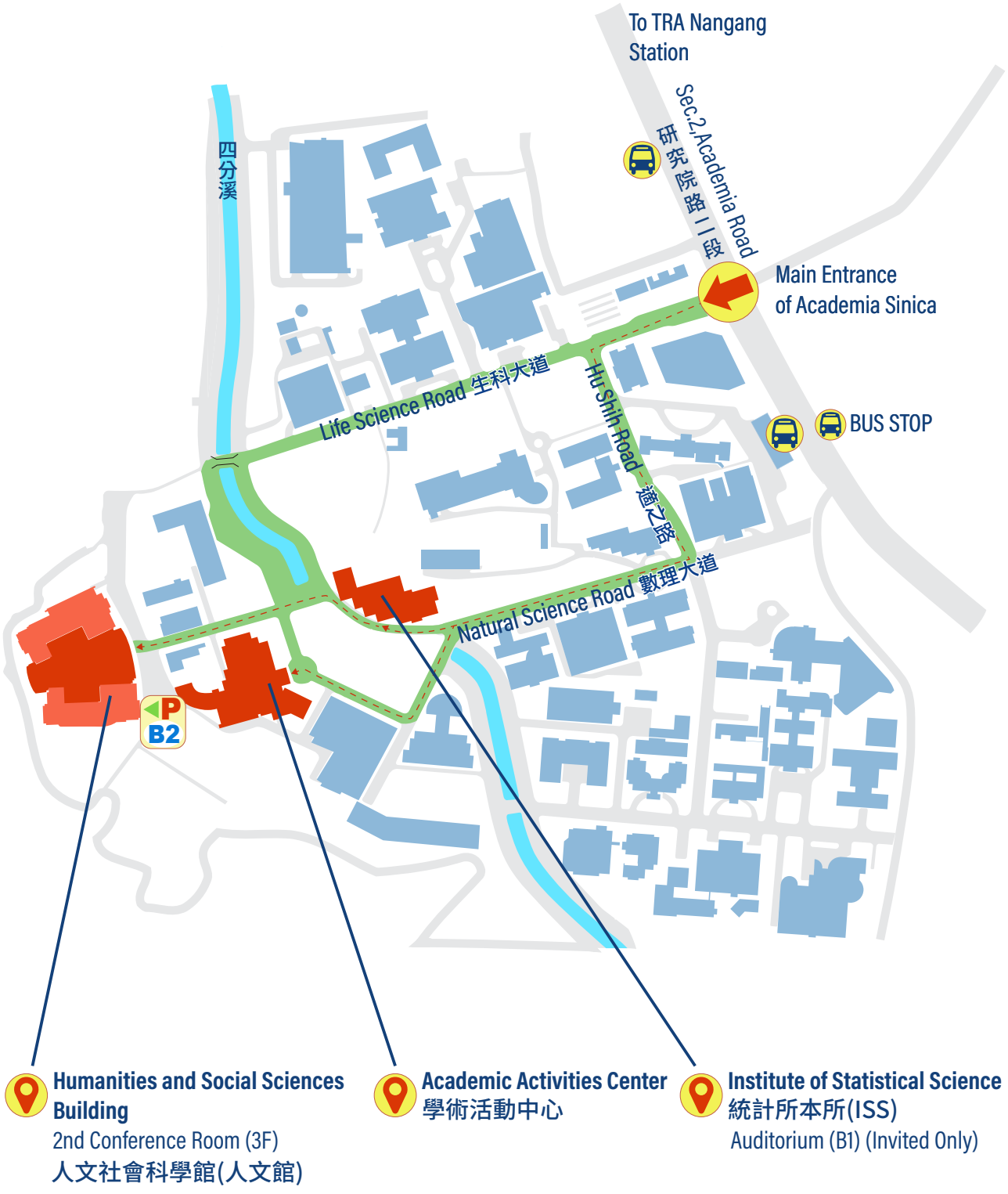


**2023SCRI Conference** Location: 2nd Conference Room, 3F of Humanities and Social Sciences Building (HSSB), Academia Sinica

(1) Additional Registration Time: Dec 10th (Sun.) 20:00-22:00 at the Lobby of Activity Center, Academia Sinica.

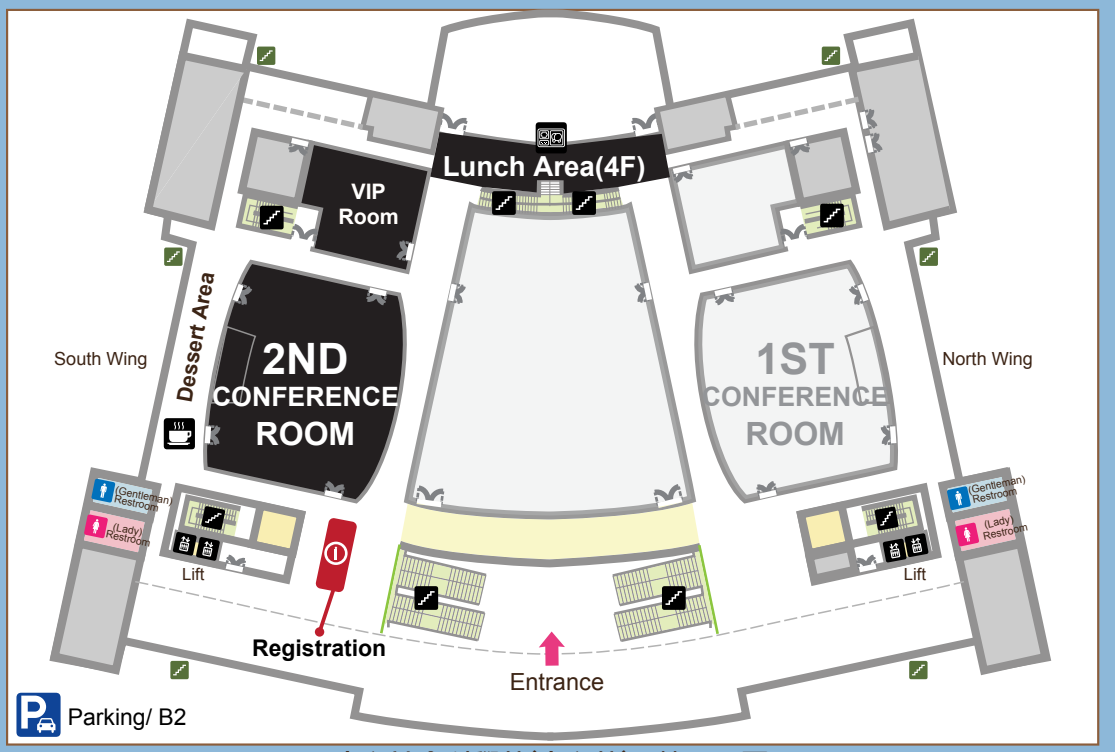
<table>
<tr><td>Time Table</td><td colspan="2">12/11 (Mon.)</td><td colspan="2">12/12 (Tue.)</td><td>12/13 (Wed.)</td><td>12/14 (Thurs.)</td><td>12/15 (Fri.)</td></tr>
<tr><td>Location</td><td colspan="2">2nd Conference Room,<br>3F of HSSB</td><td colspan="2">VIP Room, 3F of HSSB<br>Onsite and Online</td><td>2nd Conference Room,<br>3F of HSSB</td><td>2nd Conference Room,<br>3F of HSSB</td><td>Conference Rooms 3,<br>2F, Activity Center<br>Onsite and Online</td></tr>
<tr><td>8:00-9:00</td><td colspan="2">Registration</td><td colspan="2"></td><td rowspan="2">Registration<br>8:30-9:20</td><td rowspan="2">Registration<br>8:30-9:20</td><td></td></tr>
<tr><td>9:00~9:20</td><td colspan="2">Opening<br>(ISS-Hsin-Chou Yang、<br>ISI-Kerrie Mengersen、<br>UChicago- Ruey S. Tsay</td><td colspan="2" rowspan="2">IASC-EC Meeting<br>(By Invitation)</td><td rowspan="2">JDSSV Meeting<br>(By Invitation)</td></tr>
<tr><td>9:20~10:30</td><td colspan="2">Chair: Guan-Hua Huang<br>Karen Kafadar<br>Paulo Canas Rodrigues</td><td>Chair: Shih-Feng Huang<br>Yuichi Mori<br>Javier Trejos Zelaya</td><td>Chair: Nan-Cheng Su<br>Fabrizio Ruggeri<br>Ying Chen</td></tr>
<tr><td>10:30~10:50</td><td colspan="2">Tea Break</td><td colspan="2">Tea Break</td><td>Tea Break</td><td>Tea Break</td><td rowspan="3"></td></tr>
<tr><td>10:50~12:00</td><td colspan="2">Chair: Pei-Sheng Lin<br>Luis Firinguetti<br>Paula Brito</td><td colspan="2" rowspan="2">IASC-EC & JDSSV Joint Meeting<br>10:50 – 13:15<br>(By Invitation)</td><td>Chair: Chun-Shu Chen<br>Carlo Cavicchia<br>Paul Wu</td><td>Chair: Yi-Ting Hwang<br>Tomoyuki Higuchi</td></tr>
<tr><td>12:00~13:30</td><td>Lunch</td><td>IASC-EC& CIPS-EC<br>Joint Meeting<br>VIP Room, 3F of HSSB<br>(By Invitation)</td><td rowspan="2">Registration<br>and Lunch</td><td>Break</td><td>Lunch</td><td>Lunch</td><td>Lunch</td></tr>
<tr><td rowspan="2">13:30~14:40</td><td colspan="2" rowspan="2">Chair: Ching-Kang Ing<br>Zdeněk Hlávka<br>Elizabeth Ann Maharaj</td><td>2nd Conference Room,<br>3F of HSSB</td><td>VIP Room,<br>3F of HSSB<br>Onsite and Online</td><td rowspan="4">(Shuttle Departs @12:30)<br><br>Local Tour<br>(By Invitation)</td><td rowspan="2">Chair: Yufen Huang<br>Kwan-Liu Ma<br>Jürgen Symanzik</td><td rowspan="4">JDSSV Meeting<br>(By Invitation)</td></tr>
<tr><td>Chair: Nan-Jung Hsu<br>Faming Liang<br>Wei-Biao Wu</td><td rowspan="3">JDSSV Editorial<br>Board Meeting<br>(By Invitation)</td></tr>
<tr><td>14:40~15:00</td><td colspan="2">Tea Break</td><td>Tea Break</td><td>Tea Break</td></tr>
<tr><td>15:00~16:45</td><td colspan="2">Chair: Wen-Han Hwang<br>Stefan Van Aelst<br>Ray-Bing, Chen<br>Philip Leung-Ho Yu</td><td>Chair: Henry Horng-Shing Lu<br>Ming-Yueh Huang<br>Min Yang<br>Yoshiyuki Ninomiya</td><td>Chair: Ming-Chung Chang<br>Patrick J. F. Groenen<br>Kerrie Mengersen<br>15:00-16:10<br>Closing Remark<br>(ISS- Su-Yun Huang,<br>ISS- Chun-houh Chen)<br>16:10-16:30</td></tr>
<tr><td>16:45~20:00</td><td colspan="2">(Shuttle Departs @17:15)<br>Reception</td><td colspan="2">(Shuttle Departs @17:15)<br>Banquet<br>(By Invitation)</td><td></td><td></td></tr>
</table>

# Map of Academia Sinica



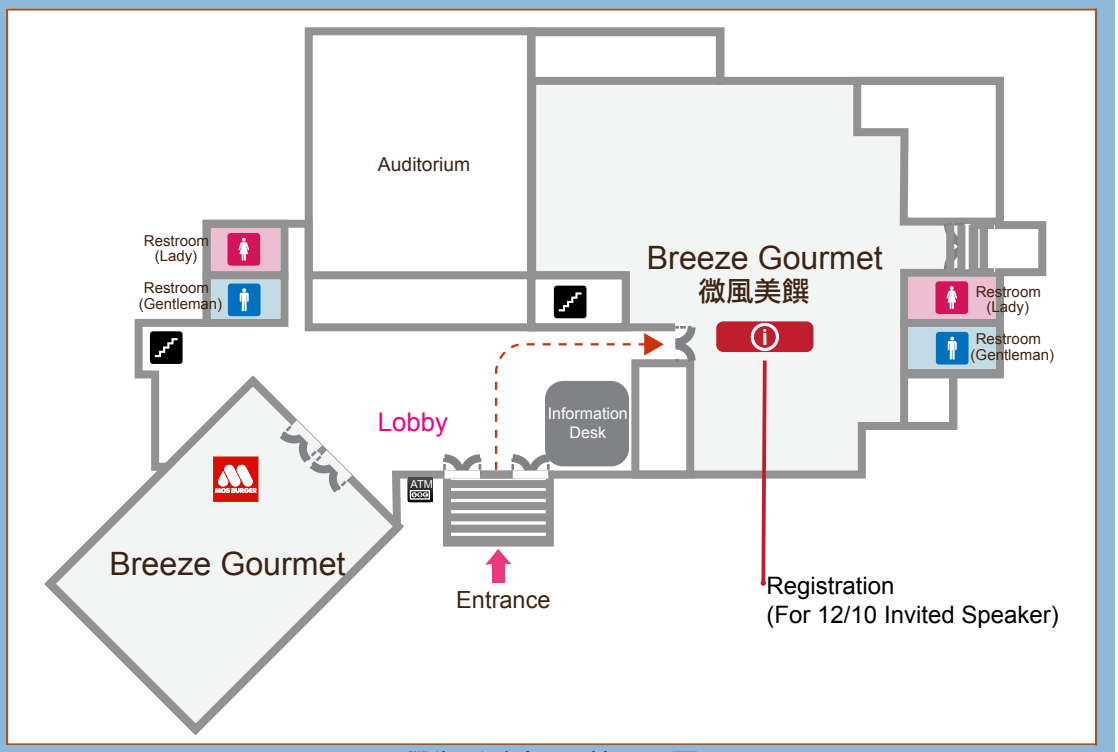
# Map of Academia Sinica

## Humanities and Social Sciences Building (Level 3)




人文社會科學館(人文館) 3樓平面圖

## Activity Center (Level 1)




學術活動中心1樓平面圖

# Abstracts

# Statistical Computing and Robust Statistical Methods for 'Big' Data

**Karen Kafadar**

**Commonwealth Professor**
**Department of Statistics, The University of Virginia**

Today's massive datasets require statistical methods for efficient computing and informative displays even more now than when the terms "Statistical Computing" and "Statistical Graphics" evolved as disciplines 50 years ago. Because the central goals of data analysis are insight and inference, and because rarely should all data be displayed, statistical methods continue to inform our algorithms, analyses, and displays. Further, `big data' invariably contain exotic values, outliers, or mixtures of distributions, and hence require robust techniques. Finally, more data may have more information, especially when they are not representative of their target populations. This talk will emphasize the role of statistics in uncovering sampling biases, classification, robust estimation of model parameters, and graphical displays, as well as offer some thoughts on when statistical models can be usefully replaced by `black-box' algorithms.

Topics:

1. Two motivating datasets (classification)
2. Are "big" data really informative? (Not always)
3. With so much (?) data, do we still need robust methods and statistical displays? (yes)
4. Why do we need efficient statistical computing?
5. Statistical thinking regarding when to use/avoid `black-box' algorithms

# Robust singular spectrum analysis: Methodology and application

**Paulo Canas Rodrigues**

**Federal University of Bahia, Salvador, Brazil**

paulocanas@gmail.com ; www.paulocanas.org

Singular Spectrum Analysis (SSA) is a powerful and widely used non-parametric method to analyze and forecast time series. Although SSA has proven to outperform traditional parametric methods for model fit and model forecasting, one of the steps of the SSA algorithm is the singular value decomposition (SVD) of the trajectory matrix, which is very sensitive to the presence of outliers because it uses the L2 norm optimization. Therefore, the presence of outlying observations has a significant impact on the SSA reconstruction and forecasts. The main aim of this talk is to introduce several robust alternatives to SSA, where the SVD is replaced by robust SVD and robust PCA alternatives. The SSA and the six robust SSA alternatives are compared in terms of model fit and model forecasting via Monte Carlo simulations based on synthetic and real data, considering several contamination scenarios. Joint work with Mohammad Kazemi, Rahim Mahmoudvand, and Vanda Lourenço.4. Why do we need efficient statistical computing? 5. Statistical thinking regarding when to use/avoid `black-box' algorithms

# An Extension of the Akash Distribution:

# Properties, Inference and Application

**L. Firinguetti, D. Gallardo, H. Gomez and Y. Gómez ´**

**Departamento de Estadística, Facultad de Ciencias Universidad del Bío-Bío**
**Avda. Collao 1202, Casilla 5C, Concepcion, Chile**

**lfiringu@ubiobio.cl**

A common problem encountered in regression modeling is that of multicollinear regressors. It is also the case that in Data Science high dimensional data matrices are commonplace, with explanatory variables which are highly collinear or even linearly dependent because the number of variables are often larger than the number of observations. This condition makes least squares methods infeasible for parameter estimation and special regression techniques are required to deal with estimation in the presence of multicollinearity.

**Keywords**
Density function; Estimation; Inference; Heavy Tail.

# Robust Sparse Models and Outlier Detection for Multivariate Distributional Data

**Pedro Duarte Silva[1], Peter Filzmoser[2], and Paula Brito[3]**

**[1] Católica Porto Business School & CEGE, Universidade Católica Portuguesa, Porto, Portugal**

psilva@ucp.pt

**[2] Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria**

peter.filzmoser@tuwien.ac.at

**[3] Faculdade de Economia, Universidade do Porto & LIAAD-INESC TEC, Porto, Portugal**

mpbrito@fep.up.pt

The classical data representation model, where for each statistical unit a single value is recorded for each variable, is too restrictive when the data to be analysed are not real numbers or single categories but comprise variability. In this work, we focus on numerical distributional data, i.e., data where units are described by histogram or interval-valued variables, representing the intrinsic variability of the corresponding observations. In our model, each distribution is represented by a location measure and interquantile ranges, for a given set of quantiles; typical cases consist in using the median, or else the midpoint, as central statistics, and quartiles, or other equally-spaced quantiles. The proposed model consists in assuming that the joint distribution of the central statistic and the logarithms of the ranges is Gaussian. Alternative sparse structures of the variance-covariance matrix are considered, which allow modelling the possible relations between the different indicators. A multivariate outlier detection method is then proposed that is based on a sparse robust estimator of the inverse of the variance-covariance matrix. The computations rely on an efficient adaptation of the graphical lasso algorithm. The proposed methodology is evaluated in a controlled simulation experiment, and illustrated with real distributional-valued data.

# Robust tests of mutual independence between functional time series

**Z. Hlavka, M. Huskova, S.G. Meintanis**

We propose a new test of mutual independence between two or more sequences of functional observations. Our method is based on empirical characteristic functions and, therefore, it requires less moment assumptions and also deals with true dependence and not just lack of covariance. After formulating the null hypothesis and suitable criteria, we investigate the asymptotic behavior of the test statistic and discuss computational issues. Simulated heavy-tailed functional observations are used to establish validity of block bootstrap and to investigate robustness of the proposed test in comparison to tests using empirical cross covariance operators or distance covariances. An application to real data investigates possible dependencies between time series of intraday returns of stock indices and crytocurrencies.

# Deployment of Renewable Energy Sources: Empirical Evidence in Identifying Clusters with Dynamic Time Warping

**Elizabeth Ann Maharaj[1], Livia De Giovanni[2], Pierpaolo D'Urso[3] and Mita Bhattacharya[4]**

**[1,4] Monash University, Australia.**
**[2]Luiss, Rome, 00197, Italy. [3]Sapienza, University of Rome, Italy.**

Deployment of renewable energy sources has caused a seismic shift in the world energy arena. Individual and coordinated efforts across countries and regions are shaping the world for the future, including business models which are supported globally to achieve net zero goals by 2050. This has resulted in changing cost structures, prices, and investment in energy uses, and approaching towards most sustainable environments for most of the regions. Our aim in this paper is to identify clusters of countries, where within a particular cluster, the levels of deployment of renewable energy sources are similar while across clusters, they are different. We propose a time series clustering method capturing the time-varying features of the renewable energy time series of 130 countries to enable the assessment of how similar or how different the usage is in relation to the Organisation for Economic Co-operation and Development (OECD) status of countries, their regional location, and their income grouping. Using Dynamic Time Warping (DTW) which is a method that calculates an optimal match between two given time series with certain restrictions, we the adopt the Partitioning Around Medoids (PAM) technique in a fuzzy framework to obtain cluster solutions. Our analysis shows that both 4-cluster and 5-cluster solutions best capture country separation based on OECD status, regional location, and income grouping.

**Keywords**
Renewable energy; Dynamic time warping; Partitioning around medoids; Fuzzy clustering.

# Subset Selection Ensembles

**Anthony-Alexander Christidis, Stefan Van Aelst, Ruben Zamar**

**Department of Mathematics, University of Leuven**

Two key approaches for high-dimensional regression are sparse methods such as best subset selection and ensemble methods such as random forests. Sparse methods have the advantage that they yield interpretable models. However, they are often outperformed in terms of prediction accuracy by “blackbox” multi-model ensemble methods. We propose an algorithm to optimize an ensemble of penalized regression models by extending recent developments in optimization for sparse methods to multi-model regression ensembles. The algorithm learns sparse and diverse models in the ensemble simultaneously from the data. Each of these models provides an explanation for the relationship between a subset of predictors and the response variable. To initialize our algorithm forward stepwise regression is generalized to multi-model regression ensembles. The resulting ensembles achieve excellent prediction accuracy by exploiting the accuracy-diversity tradeoff of ensembles. The ensembles can outperform state-of-the-art competitors on both simulated and real data.

# Bayesian Selection Approach for Multinomial Probit Models

**Ray-Bing Chen**

**Department of Statistics & Insitute of Data Science**
**National Cheng Kung University, Taiwan**

In this study, a multinomial probit model is proposed to analyze a categorical response variable. Our primary goal is to identifying the influential variables in the model. To achieve this goal, we introduce a Bayesian selection method that incorporates two nested indicators. The first indicator signifies the activity of a variable in relation to the categorical response, while the second indicator is associated with both a specific variable and a categorical level, helping determine its significance within that level. The selection process relies on posterior indicator samples generated through an MCMC algorithm. We illustrate the effectiveness of our proposed Bayesian selection approach through simulations and a real-world example.

# Variable Selection for High-Dimensional Incomplete Data

**Philip Leung-Ho Yu**

**Department of Mathematics and Information Technology**
**The Education University of Hong Kong**

Regression analysis is often impacted by several challenges such as high dimensionality, severe multicollinearity, and missing data. These problems can obscure important relationships and result in biased conclusions. In this talk, we propose a new method that effectively addresses these challenges by combining data imputation and variable selection. Specifically, our approach incorporates a novel multiple imputation algorithm based on matrix completion (Multiple Accelerated Inexact Soft-Impute), a more stable and accurate randomized lasso method (Hybrid Random Lasso), and a consistent method to integrate a variable selection method with multiple imputation. The proposed method is applied to analyze the Asian American minority subgroup in the 2017 National Youth Risk Behavior Survey. We focus on studying key risk factors related to the intention for suicide among Asian Americans. Through simulations and real data analyses on various regression and classification settings, our method demonstrates improved accuracy, consistency, and efficiency in both variable selection and prediction. This is a joint work with Lixing Liang and Yipeng Zhuang.

# An Introduction to Stochastic Deep Learning

**Faming Liang**

**Department of Statistics, Purdue University, USA.**

We have developed a new type of stochastic neural network (StoNet), which is formulated as a composition of many simple linear/logistic regression models, and designed an adaptive stochastic gradient MCMC algorithm for its training. The StoNet fits into the framework of statistical modeling, allowing us not only to address fundamental issues in deep learning, such as structural interpretability and uncertainty quantification, but also to provide a platform for transferring the theory and methods developed for linear models to deep learning. We showcase the integration of reproducing kernel methods into deep neural networks to enhance their training and prediction performance. Furthermore, we demonstrate how to use the StoNet to perform nonlinear sufficient dimension reduction and causal inference on high-dimensional data. Lastly, we illustrate how to leverage the StoNet to handle special types of data, such as those with missing values or measurement errors, and how to use it to perform statistical inference for conventional deep neural networks. This talk is based on joint works with Yan Sun, Siqi Liang, and Yaxin Fang.

# Asymptotics for Constant Step Size Stochastic Gradient Descent

**Wei Biao Wu**

**Department of Statistics, University of Chicago**

I will discuss a novel approach to understanding the behavior of Stochastic Gradient Descent (SGD) with constant step size by interpreting its evolution as a Markov chain. Unlike previous studies that rely on the Wasserstein distance, our approach leverages the functional dependence measure and explore the Geometric-Moment Contraction (GMC) property to capture the general asymptotic behavior of SGD in a more refined way. In particular, our approach allow SGD iterates to be non-stationary but asymptotically stationary over time, providing quenched versions of the central limit theorem and invariance principle valid for averaged SGD with any given starting point. These asymptotic results allow for the initialization of SGD with multiple distinct step sizes, which is a widespread practice in the discipline. We subsequently show a Richardson-Romberg extrapolation with an improved bias representation to bring the estimates closer to the global optimum. We establish the existence of a stationary solution for the derivative SGD process under mild conditions, enhancing our understanding of the entire SGD procedure across varied step sizes. Lastly, we propose an efficient online method for estimating the long-run variance of SGD solutions. This aligns with the recursive nature of SGD, thereby facilitating fast and efficient computations. The work is joint with Jiaqi Li, Zhipeng Lou and Stefan Richter.

# Efficient Data Integration Under Prior Probability Shift

**Ming-Yueh Huang**

**Institute of Statistical Science, Academia Sinica**

Conventional supervised learning usually operates under the premise that data are collected from a homogeneous underlying population. However, challenges may arise when integrating new data from different populations, resulting in a phenomenon known as dataset shift. This paper focuses on prior probability shift, a specific form of dataset shift, where the distribution of the outcome varies across different datasets but the conditional distribution of features given the outcome remains the same. To tackle the challenges posed by this shift, we propose a maximum likelihood estimation method that efficiently amalgamates information from multiple sources under prior probability shift. Unlike existing methods that are restricted to discrete outcomes, the proposed approach accommodates both discrete and continuous outcomes. It also handles high-dimensional covariate vectors through variable selection using an adaptive LASSO penalty, producing efficient estimates that possess the oracle property. Moreover, a novel semiparametric likelihood ratio test is proposed to check the validity of prior probability shift assumptions by embedding the null conditional density function into Neyman's smooth alternatives and testing study-specific parameters. We demonstrate the effectiveness of our proposed method through extensive simulations and two real data examples. The proposed methods serve as a useful addition to the repertoire of tools for addressing challenges that arise from dataset shifts in machine learning. This is a joint work with Dr. Jing Qin in National Institute of Allergy and Infectious Diseases, National Institutes of Health and Prof. Chiung-Yu Huang in University of California, San Francisco.

# Optimal Subdata Selection for Large-scale Multi-class Logistic Regression

**Min Yang**

**Department of Mathematics, Statistics, and Computer Science**
**University of Illinois at Chicago**

Big data presents the unprecedented challenge of analysis due to its immense size. One common solution is to select a subset of the data that can be managed with existing computational resources. Although various subset selection methods exist, the optimal approach, in theory, would be to choose a subset that minimizes the variance-covariance matrix from all possible data subsets. However, this is a classic NP-hard problem. In this paper, we target multi-class logistic regression models and introduce an optimal subset selection algorithm for large-scale datasets. This algorithm aims to derive near-optimal subsets under various setups. Empirical studies show that our proposed algorithm significantly surpasses existing subsampling approaches in statistical efficiency while also reducing computational time

# Selective Inference in Propensity Score Analysis

**Yoshiyuki Ninomiya, Yuta Umezu, Ichiro Takeuchi**

**Department of Mathematical Analysis and Statistical Inference, The Institute of Statistical Mathematics**

Selective inference (post-selection inference) is a methodology that has attracted much attention in recent years in the fields of statistics and machine learning. Naive inference based on data that are also used for model selection tends to show an overestimation, and so the selective inference conditions the event that the model was selected. In this presentation, we develop selective inference in propensity score analysis with a semiparametric approach, which has become a standard tool in causal inference. Specifically, for the most basic causal inference model in which the causal effect can be written as a linear sum of confounding variables, we conduct Lasso-type variable selection by adding an L1 penalty term to the loss function that gives a semiparametric estimator. Confidence intervals are then given for the coefficients of the selected confounding variables, conditional on the event of variable selection, with asymptotic guarantees. An important property of this method is that it does not require modeling of nonparametric regression functions for the outcome variables, as is usually the case with semiparametric propensity score analysis.

# Acceleration and quantification with dimension reduction in fuzzy clustering

**Tatsuki Akaki*, Yuichi mori**, Masahiro Kuroda**, Masaya Iizuka*****

*** Graduate School of Management, Okayama University of Science, Japan**
**** Okayama University of Science, Japan**
***** Okayama University, Japan**

Fuzzy clustering, e.g., fuzzy c-means clustering (FCM), is a useful tool in the variety of fields, but we sometimes meet the following two problems: computational time due to iterative convergence and difficulty of finding the latent clusters due to high dimensionality. For the former case, we have developed an acceleration algorithm for FCM using Wynn's vector epsilon algorithm (ve-FCM), which generates a new accelerated convergent sequence based on the original linearly convergent sequence in estimating two parameters, the membership matrix and the cluster centroid matrix, alternatingly. Some numerical experiments demonstrate that the ve-FCM accelerates the computation twice or more as faster as the original one. For the latter case, although the reduced k-means method is a well-known method to find clusters with the k-means clustering and dimension reduction simultaneously, this method is hard clustering for numerical data, but not soft clustering, not for categorical data. So we propose a modified fuzzy clustering method with dimension reduction for categorical data (catRFCM). The catRFCM quantifies the original categorical data and estimates low-dimensional cluster centers by implementing fuzzy c-means (FCM) with quantification and FCM with dimension reduction simultaneously. Numerical experiments are conducted to evaluate the performance of the catRFCM by comparing with, e.g., a ordinary FCM with the original categorical data as a numerical data, and a tandem method using quantification, dimension reduction, and fuzzy clustering in sequence.

# Multidimensional scaling by particle swarm optimization

**Javier Trejos-Zelaya**

**Research Center for Pure and Applied Mathematics (CIMPA) & School of Mathematics, University of Costa Rica, San José, Costa Rica.**

javier.trejos@ucr.ac.cr

**Víctor Bazán**

**Graduate Program in Mathematics, University of Costa Rica, San José, Costa Rica.**

victor.bazan@ucr.ac.cr

Multidimensional scaling (MDS) is a dimension reduction method for representing n points in a low dimension space, given a matrix of nxn distances or dissimilarities [1]. Several methods have been proposed for solving this problem by the minimization of a minimum sum of squares criterion, called stress; however, these methods usually stack in local minima of stress. We use the particle swarm optimization [3] paradigm for minimizing stress in metric multidimensional scaling. For this, we use particles in a very high dimensional space identifying particles to solutions of the MDS problem and tune the approximation of the optimum by weighting the terms of the iterative velocity equation. The method performs well, and we illustrate the results on several data sets. Comparisons are made with some well-known methods for MDS, as well as with other metaheuristics applied to MDS, such as simulated annealing, tabu search and genetic algorithms [2], [4].

## References

[1] Borg, I., Groenen, P. J. F., & Mair, P. (2013). Applied multidimensional scaling. Springer Science + Business Media. Doi: https://doi.org/10.1007/978-3-642-31848-1

[2] Groenen, P.J.F., Mathar, R. & Trejos, J. (2000). "Global optimization methods for multidimensional scaling applied to mobile communications", in: W. Gaul, O. Opitz & M. Schader (Eds.) Data Analysis. Scientific Modeling and Practical Application, Springer, Berlin: 459-469. Doi: https://doi.org/10.1007/978-3-642-58250-9_37

[3] Kennedy, J. & Eberhart, R. (1995). "Particle swarm optimization", Proceedings of ICNN'95 - International Conference on Neural Networks vol.4, pp. 1942-1948, doi: https://doi.org/10.1109/ICNN.1995.488968

[4] Trejos, J., Castillo, W., González, J., & Villalobos, M. (2000). "Application of simulated annealing in some multidimensional scaling problems", in: H.A.L. Kiers, J.-P. Rasson, P.J.F. Groenen, & M. Schader (Eds.): Data Analysis, Classification and Related Methods. Springer, Berlin: 297-302. Doi: https://doi.org/10.1007/978-3-642-59789-3_47

# Convex Clustering of Mixed Numerical and Categorical Data


## Carlo Cavicchia

**Erasmus School of Economics, Erasmus University of Rotterdam**



Clustering analysis is an unsupervised learning technique widely used for information extraction. Current clustering algorithms often face instabilities due to the nonconvex nature of their objective function. The class of convex clustering methods does not suffer from such instabilities and finds a global optimum for the clustering objective. Whereas convex clustering has previously been established for single-type data, real-life data sets usually comprise both numerical and categorical, or mixed, data. Therefore, we introduce the mixed-data convex clustering (MIDACC) framework. MIDACC combines likelihood-based loss functions for numerical and categorical data, weighted by parameters controlling the importance of both data types on the clustering outcome. The penalty term fuses centroids, and thus allows for clustering of observations. The presence of few parameters characterizes MIDACC and allows the user to tailor the analysis for the problem at hand. For instance, in contrast to, who perform subgradient descent (in the case of numerical data only) for a path for values of the regularization term in order to retrieve an entire clusterpath, we run our algorithm for fixed values of it. We therefore present our framework as a partitional clustering methods and we do not provide the solution in a hierarchical fashion. Another crucial parameter is the one controlling the contribution of the data type; this one is fine tuned to obtain the optimal solution. We present two different implementations for this framework. The first consists of a dedicated subgradient descent algorithm. However, our current implementation follows a majorization-minimization approach (similar to what propose for only numerical data) that results remarkably faster and more efficient. Through numerical experiments, we show that, in contrast to baseline methods, MIDACC achieves near-perfect recovery of both spherical and non-spherical clusters, is able to capture information from mixed data while distinguishing signal from noise, and has the ability to recover the true number of clusters present in the data. Furthermore, MIDACC outperforms all baseline methods on a real-life data set.

# Optimal Partitioning of Directed Acyclic Graphs with Dependency Between Clusters

**Paul Pao-Yen Wu**

**School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia**

Directed Acyclic Graphs (DAGs) such as Bayesian Networks (BNs) can be used to infer parameter values or predict behaviour in complex systems with high dimensional data. Oftentimes, this inference process could be made more computationally efficient by partitioning (i.e. mapping) the DAG into clusters. In BNs, inference could refer to computation of posterior marginal probabilities given observations (evidence), or to update conditional probabilities of one or more nodes given data. Inference could also involve finding clusters such as homogeneous portions of a non-homogeneous system. The cost function to optimise for is arbitrary, and could include computational cost, AIC, likelihood or variance given a dataset. Computational cost is important as BN and Dynamic BN (DBN) inference, for instance, is NP-hard or worse. In addition, optimal partitioning is NP-hard, and the challenge is exacerbated by statistical inference as the cost to be optimised is dependent on both nodes within a cluster, and the mapping of clusters connected via parent and/or child nodes, which we call dependent clusters.

We discuss a novel algorithm called DCMAP which can, given an arbitrarily defined, positive cost function, iteratively and rapidly find near-optimal, then optimal cluster mappings. Shown analytically to converge to optimal solutions using dynamic programming, we use a simple example and a complex systems seagrass DBN to demonstrate the algorithm. For this 25 (one time-slice) and 50-node (two time-slices) DBN, the search space size was $9.91\times10^{9}$ and $1.51\times10^{21}$ possible cluster mappings, respectively, but near-optimal solutions with 88% and 72% similarity to the optimal solution were found at iterations 170 and 865, respectively. The first optimal solution was found at iteration 934 (95% CI 926,971), and 2256 (2150,2271) with a cost that was 4% and 0.2% of the naive heuristic cost, respectively. DCMAP opens up new research opportunities for combining optimisation with inference to support prediction and learning of DAGs with high dimensional data.

# Unsupervised Statistical Tools for Anomaly Detection: The Case of Healthcare Frauds

**Fabrizio Ruggeri**

**Institute for Applied Mathematics and Information Technologies, Milano, Italy**

The research is motivated by the increased interest in detecting possible frauds in healthcare systems. We propose some unsupervised statistical tools (Lorenz curve, concentration function, sum of ranks, Gini and Pietra indices) to provide efficient and easy-to-use methods aimed to signal possible anomalous behaviours. A more sophisticated method, based on Bayesian co-clustering, is presented as well.

# Optimal Market Making in a Multi-agent Market under Model Uncertainty: A Reinforcement Learning Approach

**Ying Chen**

**Department of Mathematics, National University of Singapore, Singapore**

We delve into the optimal market-making challenge in order-driven electronic markets, emphasizing model uncertainty. The study factors in ambiguity concerning order arrival intensities, aiming to derive a strategy robust enough for diverse market conditions. By incorporating a tractable model for the limit order book using Markov Decision Processes, we plan on leveraging Reinforcement Learning to tackle the intricate optimization problem. This methodology allows for a precise depiction of order book dynamics with tick structures, diverging from typical stochastic approaches that focus on price dynamics. We also factor in the scenario of several market makers vying in the same marketplace and analyze the impact of confidential information on the optimal strategy. Both simulations and numerical evaluations are employed to gauge the efficiency of varying methods within our market context. This research is a collaborative effort with Yijiong Zhang, Hoang Hai Tran, and Julian Sester.

# Domain Adaptation for Application to Smart Meters of Electricity Usage in Each Household

**Hisashi Oshima*, Tomoyuki Higuchi*, and Tsuyoshi Ishizone[+]**

***Graduate School of Science and Engineering, Chuo University, Japan**

**[+]Graduate School of Advanced Mathematical Sciences, Meiji University, Japan**

The measurement of electricity consumption by each household is becoming increasingly automated with the installation of smart meters, and the use of this data is being considered in a variety of ways. In Japan and the U.K., data is collected and stored every 30 minutes from the standpoint of data storage efficiency. One promising use of the data is to analogize the presence or absence of residents in a home. If this can be achieved, the cost of re-delivery to absent residences in home delivery, which has increased dramatically in recent years due to the Corona disaster, can be significantly reduced. Therefore, there has been considerable research in recent years on determining whether a homeowner is present or absent from his or her electricity smart meter.

From a machine learning perspective, the study of algorithms for determining presence/absence is relatively easy to formulate in the framework of supervised learning. The problem is that supervised data are rarely available. Experimentally, it is easy to obtain supervised data by installing carbon dioxide monitoring device or sensors that detect people's movements. However, because people's behavior is so varied, machine learning models based on a small number of examples of experimental data are almost useless. As for the possibility of obtaining more data, it is quite difficult to obtain cooperation for the installation of devices in homes to monitor presence and absence due to privacy and security concerns. Also, while it would be possible to solve this problem by having security companies and power companies share data, it is not feasible from a business standpoint to begin with.

Therefore, to summarize the problem from a machine learning perspective, a presence/absence discriminator must be constructed based on a small number of supervised data sets and a large amount of unsupervised data sets. A more troublesome problem arises here. The problem is that the distribution of the explanatory variable vectors differs between the supervised and unsupervised data sets, a so-called covariate shift. This cannot be avoided even in the analysis of electricity usage data.

Techniques for dealing with this type of problem are referred to in recent machine learning as domain adaptation, and domain adaptation can also be considered a form of

transfer learning. Although the definition of domain adaptation is very broad, in this study we define it simply. The domain with supervised data is usually called the source domain and the domain with unsupervised data is usually called the target domain. In domain adaptation, a domain discriminator that discriminates between source and target from the explanatory variable vector and a discriminator that discriminates between presence and absence from the source's explanatory variable vector are learned simultaneously. However, feature representation learning, which consists of explanatory variable vectors, is learned in an adversarial manner so that the differences between domains are small. In other words, the feature extractor is learned in such a way that it tricks the domain discriminator to reduce the differences between domains.

Although the application of this methodology is very broad, in this presentation we focus on the problem of presence/absence discrimination from electricity smart meters. This methodology allows us to assign pseudo-labels to unlabeled, unsupervised data from a vector of explanatory variables. Even a small increase in the accuracy of the labels can have a tremendous economic impact. In this presentation, we consider a methodology that can handle cases where the distribution differs not only between the source and target, but also between other factors for both. Due to the limited presentation time, this presentation will introduce our aims and outline our methodology. We also present the results of our application to publicly available data. Please note that our methodology is based on the real big data of a company.

# Visual Analytics for Emerging Data-Intensive Problems

**Kwan-Liu Ma**

**Department of Computer Science, University of California-Davis, USA**

In many domains of study and practice, tremendous amounts of data are being collected but not fully utilized, due to a lack of capable data-driven discovery and decision-support tools to extract value from the data. Visual analytics, which integrates statistical analysis and machine learning methods with interactive visualization, can uncover essential patterns and relations in the data to guide the data analysis process, leading to findings crucial to decision making and problem solving. I will present visual analytics designs for emerging data-intensive problems found in healthcare, supercomputing, and document analysis.

# Lessons Learned from an Eye-Tracking Study on Human Postures

**Juergen Symanzik**

**Department of Mathematics and Statistics, Utah State University, USA**

juergen.symanzik@usu.edu

Eye-tracking has been used in many scientific fields, such as behavioral sciences, education, marketing, and sports. Visualization usually plays an important role in the analysis of eye tracking data. After a brief glance at the underlying R software, we will focus in this presentation on a study on human postures where participants from a treatment group (those with yoga experience) and a control group (those with no or only little yoga experience) were asked to look at images of 22 different human postures and assess the stability of each posture while wearing eye-tracking equipment. The core question for the part of the analysis presented here was whether viewing patterns of the postures for participants from the treatment group differ from those from the control group. We answered this question at the group level (treatment vs. control), for individual participants within each group, and with respect to individual body parts viewed by the participants from both groups. This is joint work with Joanna Coltrin, Chunyang Li, Eric McKinney, Sarah Schwartz, and Breanna Studenka.

# Interpretable Kernals

**Patrick Groenen[1] and Michael Greenacre[2]**

**[1]Erasmus University Rotterdam, Econometric Institute, Rotterdam**
**[2]Universitat Pompeu Fabra, Barcelona, Spain**

The use of kernels for nonlinear prediction is widespread in machine learning. They have been popularized in support vector machines and used in kernel ridge regression, amongst others. These methods share three aspects. First, instead of the original n × p matrix of predictor variables, each row is mapped into a high dimensional feature space. Second, a ridge penalty term is used to shrink the weights (coefficients) on the predictors in the high-dimensional feature space. Third, the solution is not obtained in this feature space, but through solving a dual problem. A major drawback in the use of kernels is that the interpretation in terms of the original predictor variables is lost. In this paper, we argue that in the case of a wide n × p matrix of predictor variables (with p > n), the kernel solution can be re-expressed in terms of a linear combination of the original matrix of predictor variables and a ridge penalty that involves a special metric. Consequently, the exact same predicted values can be obtained as a weighted linear combination of the predictor variables in the usual manner and thus can be interpreted. In the case p × n, we discuss a least-squares approximation of the kernel matrix that still allows the interpretation in terms of a linear combination. It is shown that these results hold for any function of a linear combination that minimizes the weights and has a ridge penalty of these weights such as in kernel logistic regression and kernel Poisson regression. When the objective function is minus the log likelihood, standard likelihood theory can be used to estimate the standard deviations of the weights. As an extension, it is possible to apply an approximation in a k-dimensional space with k < p thereby enforcing dimension reduction in the predictor space.

# Dealing with Sensitive Data


## Kerrie Mengersen

**School of Mathematical Sciences, Queensland University of Technology (QUT)**
**ARC Centre of Excellence for Mathematical & Statistical Frontiers**
**Brisbane Australia**
**k.mengersen@qut.edu.au**



Many datasets of interest to statisticians are subject to privacy conditions. This can constrain access, analysis, sharing and release of results. In this presentation, we will consider two ways in which this issue might be addressed. The first is through federated learning, in which the analysis is undertaken in such a way that the data remain in situ and private. The second is synthetic generation of the data, such that the simulated data retains salient characteristics but retains the required privacy. We provide some extensions to the class of models that can be considered in federated learning, and an overview of synthetic generation of tabular data. The exposition of these ideas will be motivated by the creation of an Australian Cancer Atlas.



This research is in collaboration with QUT colleagues Conor Hassan and Dr Robert Salomone, and is funded by the Australian Research Council and Cancer Council Queensland.


### Key Reading


C Hassan, R Salomone, K Mengersen (2023) Federated variational inference methods for structured latent variable models. arXiv preprint arXiv:2302.03314

C Hassan, R Salomone, K Mengersen (2023) Deep generative models, synthetic tabular data and differential privacy: an overview and synthesis. arXiv preprint arXiv:2307.15424

# Local Tour

| Dec 13. | Agenda | Note |
|---|---|---|
| **12:30** | Departure from Academic Activities Center | |
| 12:30-13:15 | Traffic Time | Duration 40mins |
| **13:30-15:30** | **National Palace Museum** | **Guided tour 1 hr.** |
| 15:30-15:50 | Traffic Time | Duration 20 mins, Snack |
| **16:00-17:00** | **Taipei Confucius Temple** | **Guided tour 1hr.** |
| 17:00-17:30 | Picture & Free Time (Dalongdong Baoan Temple) | |
| 17:30-18:00 | Traffic Time | Duration 30 mins |
| **18:00-19:00** | **Lungshan Temple** | **Guided tour 1 hr.** |
| **19:00-20:00** | **Huaxi Street Tourist Night Market** | **Dinner on one's own** |
| -20:00 | Return to hotels: Green World Hotel Nangang, Academic Sinica Activities Center<br>or<br>Free Tour | |

**Important Notice**

1. We've got a packed schedule, so please make sure to gather and get on/off the bus on time according to the schedule to make things run smoothly.
2. If you're not feeling well or need any assistance, don't hesitate to let one of the staff members know right away.
3. Respect the historical sites, artifacts, exhibits, and facilities. Keep the noise down during the visit, and be mindful of keeping the surroundings clean.
4. Inside the National Palace Museum exhibitions: (1) No eating or smoking, (2) Avoid using flash photography, tripods, and selfie sticks, (3) If your backpack is larger than A3 size, kindly store it, and (4) Since the exhibition hall is maintained at a temperature between 20 to 24 degrees Celsius, consider bringing warm clothing based on your needs.

**Traffic from Huaxi Street Tourist Night Market to Activity Center of Academia Sinica**

1. Taxi or Uber. Approximate time: 30 minutes, Price: 500-650 NTD (20 USD)
2. MRT. Blue Line (to Nangang Exhibition Center)
   → Longshan Temple (BL10) to Nangang Exhibition Center (BL23) Exit 5
   → Take any bus at "Nangang Exhibition Center" Stop in the opposite street
   → Get off at Academia Sinica (中研院) station

Welcome to Taipei, a city rich in cultural heritage and vibrant traditions. We are excited to introduce you to some of the most iconic landmarks in the heart of this bustling metropolis.

| | Introduction |
|---|---|
|  <br>Source: https://www.travelking.com.tw/tourguide/scenery151.html<br>  <br>Source: https://www.npm.gov.tw/Articles.aspx?sno=04012814&l=1 | National Palace Museum (國立故宮博物院)<br><br>Imagine a treasure trove of 700,000 Chinese artifacts spanning 8,000 years, all under one roof. These treasures, once housed in the Forbidden City, found their new home in Taipei during the ROC retreat. The National Palace Museum, established in 1965, showcases exquisite items such as delicate porcelain, intricate jade carvings, and masterpieces of calligraphy. Its expansive building, constructed between 1964 and 1965, underwent subsequent expansions, making it one of the world's largest museums. Linked to the prestigious Palace Museum of Beijing, the artifacts highlight the craftsmanship and creativity of China's past, offering invaluable insights into the country's cultural heritage.<br><br>For more information, visit National Palace Museum.<br>https://www.npm.gov.tw/index.aspx?l=2 |
|  <br>Source: https://taipeiwalker.walkerl | Dalongdong Baoan Temple (大龍峒保安宮)<br><br>Next, we invite you to marvel at the magnificence of Dalongdong Baoan Temple that has graced Taipei for over 200 years! Dalongdong Baoan Temple was designated as a national historic site in 2019. This temple stands as a unique fusion of religion, culture, education, and art. It holds the distinction of being the only temple in Taiwan to receive the UNESCO Cultural Heritage Conservation Award. Renowned for its decorative artistry, the temple boasts an impressive collection of woodwork, paintings, and ceramic pieces crafted by contemporary Taiwanese master artists. The temple has played |

| | |
|---|---|
| and.com.tw/articles/view/2371 | a significant role in revitalizing local culture in Taipei City, leaving a profound impact on the community.<br><br>Learn about its fascinating history and cultural significance at Dalongdong Baoan Temple.<br>https://www.baoan.org.tw/history.php?lang=en |
|  <br>Source: https://shotrip.com/attractions/552/%E3%80%8C%E5%8F%B0%E5%8C%97%E5%AD%94%E5%BB%9F%E3%80%8D%E4%BA%BA%E6%96%87%E5%8F%A4%E8%B9%9F%E4%B9%8B%E6%97%85 | Taipei Confucius Temple (孔廟)<br><br>Our journey continues to the Taipei Confucius Temple, a sanctuary of wisdom and learning. The Taipei Confucius Temple is the only existing community-donated Confucius temple in Taiwan. It venerates Confucius, the greatest teacher in Chinese history, as well as the sages and scholars of the past. Confucius, the founder of Confucianism, emphasized the importance of morality, family, government, and social order, establishing a crucial philosophical foundation for East Asian culture. His teachings have profoundly influenced Chinese history and culture. Through its unique architecture, decorations, and rituals, the temple provides an opportunity to experience the deep reverence Chinese people have for traditional values, knowledge, wisdom, and their embodiment. Immerse yourself in the serene ambiance of this ancient temple dedicated to the great philosopher Confucius!<br><br>Discover more at Taipei Confucius Temple.<br>https://www.tctcc.taipei/en-us/index.htm |
|  <br>Source: https://www.expedia.com.tw/Lungshan-Temple-Wanhua.d6080203.Place-To-Visit | 龍山寺<br>Lungshan Temple<br><br>No visit to Taipei is complete without a stop at Lungshan Temple, a spiritual oasis nestled in the heart of the city.<br><br>Wanhua (Bangka) holds significant historical and cultural importance in Taipei, Taiwan. As the city's oldest neighborhood, Wanhua has been a central hub since the 1820s, serving as the economic, political, and military center of northern Taiwan. |

| | |
|---|---|
| | Wanhua Longshan Temple has been a focal point of religious and cultural activities for centuries.<br>It is not only a center of religious devotion for residents but also a place where various matters, such as discussions and legal disputes, were resolved through prayers and seeking divine guidance. Over the years, it has become a place where generations of locals have visited to seek blessings for good fortune, health, and even advice on matters like marriage. The temple's rich history and cultural significance continue to attract both locals and tourists, making it a must-visit destination in Taipei.<br><br>Experience the tranquility and spirituality of this revered temple, a testament to Taiwan's rich religious heritage.<br><br>Find out more at Lungshan Temple.<br>https://www.lungshan.org.tw/tw/index.php |
|  <br>Source:<br>https://supertaste.tvbs.com.tw/infocard/20754 | Huaxi Street Tourist Night Market (華西街觀光夜市)<br><br>Lastly, indulge your senses at the Huaxi Street Tourist Night Market, where the vibrant atmosphere, delectable street food, and unique souvenirs create an unforgettable experience.<br><br>Hua Xi Street Night Market, Taiwan's pioneering tourist hotspot, tells Taipei's story through its diverse culinary landscape. From sizzling shrimp to famous soups, the market tantalizes taste buds with local delights. Beyond food, it offers foot massages, fitness centers, and intriguing snake and turtle shops, adding a touch of adventure. Dive into the market's cultural richness with traditional herbal stores, Lingzhi tea sellers, and medicinal plaster shops, offering a glimpse into Taiwan's heritage. More than just a food haven, Hua Xi Street Night Market is a vibrant tapestry of traditions, making every visit a flavorful adventure and a delightful cultural exploration.<br><br>Explore the bustling night market scene at Huaxi Street Tourist Night Market. |

| | |
|---|---|
| | https://www.travel.taipei/en/attraction/details/1687 |

In the 21st century, Taipei is a bustling and modern metropolis, but with clear and distinct roots in the past. Travelling through these areas will allow the you to gain knowledge of the old city and also experience firsthand what life may have been like when Taipei was young. Prepare yourselves for a captivating journey through Taipei's cultural tapestry, where ancient traditions blend seamlessly with modern allure. Enjoy your exploration!

**Accommodation Service Notification**

Phone: +886-2-2785-2717, Email：ac@gate.sinica.edu.tw

**Check-in and check-out**

1. Standard check-in time: after 2:00 pm, check-out time: before 12:00 noon
2. Please be noted that the fee is in included in the per diem we provide, and you are required to complete the payment process during check-out.
3. If your expected check-in and check-out time falls outside the designated time, you are subject to additional charges.
   (1) Early check-in: incurs a half-day fee (Single Room: 550 NTD; Double Room: 750 NTD)
   (2) Late check-out: 100 NTD per hour

**Amenities**

The accommodation at our center provides body wash and shampoo. In response to environmental conservation, disposable amenities such as toothbrushes, combs, razors, shower caps, soap, and paper slippers are **not** provided. You are kindly requested to bring your own or purchase them at the following locations,

**Px Mart** (8:00am-10:00pm)

**Hi Life convenience store** (24hrs)

**Co-op grocery store** (8:30 am- 5:30pm)

**Breakfast**

Breakfast will be provided from 8:30 to 9:00 a.m.

Location: Conference venue on Dec. 11, 13, 14

Lounge, 2F of the Institute of Statistics on Dec. 12

Other options:

**Mos burger** (7:00am-8:30pm) on the 1st floor of this activity center: Japanese fast-food chain known for its fresh ingredients and innovative Japanese-style burgers, curry rice, and fries.

**Trine & Zen** (8:00am-7:00pm) at the Tsai Yuan-pei Memorial Hall: Blending the essence of Chinese Zen cuisine with authentic American flavors, including the proudest feature – freshly ground coffee.

**Self-service Laundry (Service Hours: 24 hours)**

Located on B1 of the Academic Activity Center, we offer self-service laundry and clothes drying facilities. The cost is 20 New Taiwan Dollars per wash and 20 New Taiwan Dollars per drying cycle.

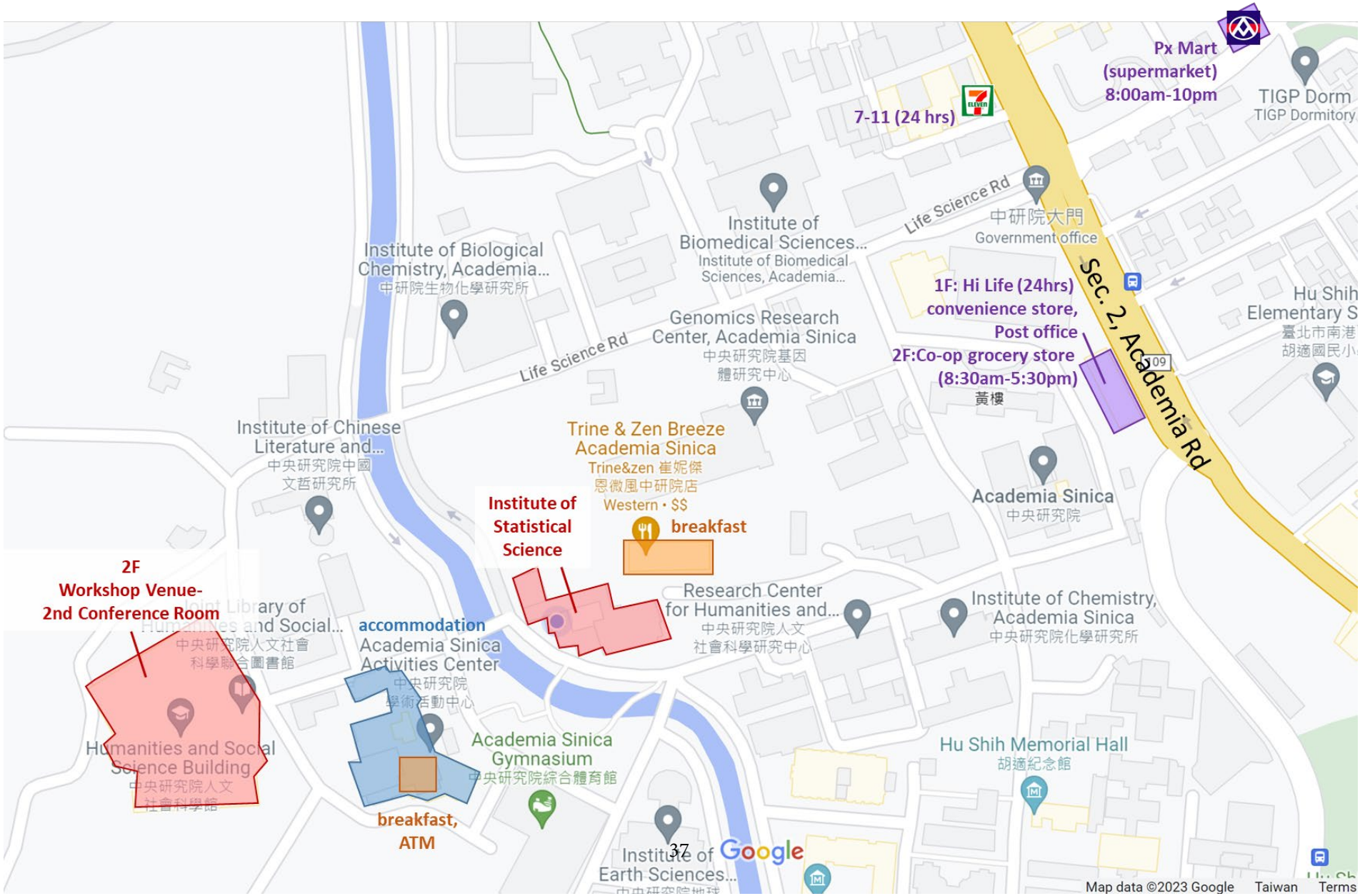
**Vending Machines (Service Hours: 24 hours)**

Located on B1 of the Academic Activity Center, our vending machines provide a variety of items such as instant noodles, biscuits, snacks, and beverages.

Service hours: 24 hours

**ATM (Service Hours: 24 hours)**

Located on the 1st floor of the Academic Activity Center.


Px Mart
(supermarket)
8:00am-10pm
TIGP Dorm
TIGP Dormitory
7-11 (24 hrs)
Life Science Rd
中研院大門
Government office
Institute of
Biomedical Sciences...
Institute of Biomedical
Sciences, Academia...
Institute of Biological
Chemistry, Academia...
中研院生物化學研究所
Genomics Research
Center, Academia Sinica
中央研究院基因
體研究中心
Life Science Rd
1F: Hi Life (24hrs)
convenience store,
Post office
2F:Co-op grocery store
(8:30am-5:30pm)
黃樓
Sec. 2, Academia Rd
209
Hu Shih
Elementary S
臺北市南港
胡適國民小
Institute of Chinese
Literature and...
中央研究院中國
文哲研究所
Trine & Zen Breeze
Academia Sinica
Trine&zen 崔妮傑
恩微風中研院店
Western · $$
breakfast
Institute of
Statistical
Science
Academia Sinica
中央研究院
2F
Workshop Venue-
2nd Conference Room
Joint Library of
Humanities and Social...
中央研究院人文社會
科學聯合圖書館
accommodation
Academia Sinica
Activities Center
中央研究院
學術活動中心
Research Center
for Humanities and...
中央研究院人文
社會科學研究中心
Institute of Chemistry,
Academia Sinica
中央研究院化學研究所
Humanities and Social
Science Building
中央研究院人文
社會科學館
Academia Sinica
Gymnasium
中央研究院綜合體育館
Hu Shih Memorial Hall
胡適紀念館
breakfast,
ATM
Institute of
Earth Sciences...
Google
Map data ©2023 Google
Taiwan
Terms

# COMMITTEE

## SCIENTIFIC PROGRAM COMMITTEE

**RUEY S. TSAY**
(Booth School of Business, University of Chicago)

**CHUN-HOUH CHEN**
(Academia Sinica, Taiwan R.O.C.)

**SU-YUN HUANG**
(Academia Sinica, Taiwan R.O.C.)

**HSIN-CHOU YANG**
(Academia Sinica, Taiwan R.O.C.)

**HSIN-CHENG HUANG**
(Academia Sinica, Taiwan R.O.C.)

**MING-CHUNG CHANG**
(Academia Sinica, Taiwan R.O.C.)

## LOCAL ORGANIZING COMMITTEE

**MING-YUEH HUANG**
(Academia Sinica, Taiwan R.O.C.)

**YI-JU LEE**
(Academia Sinica, Taiwan R.O.C.)

**ZI-XIU LAI**
(Academia Sinica, Taiwan R.O.C.)

**YU-TING CHEN**
(Academia Sinica, Taiwan R.O.C.)

Institute of Statistical Science, Academia Sinica

The Chinese Institute of Probability and Statistics

iasc isi International Association for Statistical Computing