IASC
NZSA
2017

Table of Contents

Welcomes

Welcome To The Conference From Ian Westbrooke, NZSA President

Welcome Address From Jung Jin Lee, IASC-ARS Chair-Person

Programme At A Glance

Monday

Tuesday

Wednesday

Thursday

Am I A Session Chair? Please Read!!!

Monday

Tuesday

Wednesday

Thursday

Index By First Name of Author

Index by Submission Number

Programme And Abstracts For Monday 11th Of December

R In Times Of Growing User Base And Data Sizes

Robust Principal Expectile Component Analysis

Effect Of Area Level Deprivation On Body Mass Index: Analysis Of NZ Health Surveys

Calendar-Based Graphics For Visualising People's Daily Schedules

Nonparametric Test For Volatility In Clustered Multiple Time Series

IGESS: A Statistical Approach To Integrating Individual Level Genotype Data And Summary Statistics In Genome Wide Association Studies

Author Name Identification For Evaluating Research Performance Of Institutes

A Computational Tool For Detecting Copy Number Variations From Whole Genome And Targeted Exome Sequencing

Clustering Using Nonparametric Mixtures And Mode Identification

Bayesian Curve Fitting For Discontinuous Function Using Overcomplete Representation With Multiple Kernels

Estimation Of A Semiparametric Spatiotemporal Models With Mixed Frequency

LSMM: A Statistical Approach To Integrating Functional Annotations With Genome-Wide Association Studies

A Study Of The Influence Of Articles In The Large-Scale Citation Network

Estimating Links Of A Network From Time To Event Data

Estimation Of A High-Dimensional Covariance Matrix

Innovative Bayesian Estimation In The von Mises Distribution

Evidence Of Climate Change From Nonparametric Change-Point Analysis

Joint Analysis Of Individual Level Genotype Data And Summary Statistics By Leveraging Pleiotropy

An Advanced Approach For Time Series Forecasting Using Deep Learning

Genetic Map Estimation Using Hidden Markov Models In The Presence Of Partially Observed Information

A Simple Method For Grouping Patients Based On Historical Doses

Semiparametric Mixed Analysis Of Covariance Model

Adaptive False Discovery Rate Regression With Application In Integrative Analysis Of Large-Scale Genomic Data

Structure Of Members In The Organization To Induce Innovation: Quantitatively Analyze The Capability Of The Organization

Vector Generalized Linear Time Series Models

Local Canonical Correlation Analysis For Multimodal Labeled Data

A Practitioners Guide To Deep Learning For Predictive Analytics On Structured Data

Clustering Of Research Subject Based On Stochastic Block Model

Zen And The aRt Of Workflow Maintenance

Canonical Covariance Analysis For Mixed Numerical And Categorical Three-Way Three-Mode Data

Variable Selection Algorithms

Estimating Causal Structures For Continuous And Discrete Variables

Incorporating Genetic Networks Into Case-Control Association Studies With High-Dimensional DNA Methylation Data

Adaptive Model Checking For Functional Single-Index Models

Mobile Learning In Teaching Bioinformatics For Medical Doctors

On Optimal Group Testing Designs: Prevalence Estimation, Cost Considerations, And Dilution Effects

The Use Of Bayesian Networks In Grape Yield Prediction

Pattern Prediction For Time Series Data With Change Points

Test For Genomic Imprinting Effects On The X Chromosome

Fluctuation Reduction Of Value-At-Risk Estimation And Its Applications

E-Learning Courses On Introductory Statistics Using Interactive Educational Tools

Estimation Of Animal Density From Acoustic Detections

Mixed Models For Complex Survey Data

Regression With Random Effects For Analysing Correlated Survival Data: Application To Disease Recurrences

Genetic Predictors Underlying Long-Term Cognitive Recovery Following Mild Traumatic Brain Injury

Bayesian Structure Selection For Vector Autoregression Model

Three-Dimensional Data Visualization Education With Virtual Reality

Talk Data To Me

Smooth Nonparametric Regression Under Shape Restrictions

Elastic-Band Transform: A New Approach To Multiscale Visualization

Meta-Analytic Principal Component Analysis In Integrative Omics Application

Flight To Relative Safety: Learning From A No-Arbitrage Network Of Yield Curves Model Of The Euro Area

Bayesian Analyses Of Non-Homogeneous Gaussian Hidden Markov Models

Robustness Of Temperature Reconstruction For Past 500 Years

Nonparametric Causal Inference By The Kernel Method

A Unified Regularized Group PLS Algorithm Scalable To Big Data

Evaluation Of Spatial Cluster Detection Method Based On All Geographical Linkage Patterns

Scoring Rules For Prediction And Classification Challenges

Meta-Analysis With Symbolic Data Analysis And Its Application For Clinical Data

Real-Time Transit Network Modelling For Improved Arrival Time Predictions

Visualization And Statistical Modeling Of Financial Big Data

Sparse Group-Subgroup Partial Least Squares With Application To Genomic Data

Genetic Approach And Statistical Approach For Association Study On DNA Data

Modeling Of Document Abstraction Using Association Rule Based Characterization

Bayesian Static Parameter Inference For Partially Observed Stochastic Systems

Bayesian Survival Analysis Of Batsmen In Test Cricket

Covariate Discretisation On Big Data

BIG-SIR A Sliced Inverse Regression Approach For Massive Data

Symbolic Data Analytical Approach To Unauthorized-Access Logs

My Knee Still Hurts; The Statistical Pathway To The Development Of A Clinical Decision Aid

Programme And Abstracts For Tuesday 12th Of December

Could Do Better … A Report Card For Statistical Computing

R&D Policy Regimes In France: New Evidence From A Spatio-Temporal Analysis

Analysing Scientific Collaborations Of New Zealand Institutions Using Scopus Bibliometric Data

Family Structure And Academic Achievements Of High School Students In Tonga

Analysis Of Multivariate Binary Longitudinal Data: Metabolic Syndrome During Menopausal Transition

Clustering Of Curves On A Spatial Domain Using A Bayesian Partitioning Model

The Uncomfortable Entrepreneurs: Bad Working Conditions And Entrepreneurial Commitment

Spatial Surveillance With Scan Statistics By Controlling The False Discovery Rate

Statistical Models For The Source Attribution Of Zoonotic Diseases: A Study Of Campylobacteriosis

Towards An Informal Test For Goodness-Of-Fit

Identifying Clusters Of Patients With Diabetes Using A Markov Birth-Death Process

Bayesian Temporal Density Estimation Using Autoregressive Species Sampling Models

How Does The Textile Set Describe Geometric Structures Of Data?

Intensity Estimation Of Spatial Point Processes Based On Area-Aggregated Data

Bayesian Inference For Population Attributable Measures

An Information Criterion For Prediction With Auxiliary Variables Under Covariate Shift

Analysis Of A Brief Telephone Intervention For Problem Gambling And Examining The Impact On Co-Existing Depression?

Prior-Based Bayesian Information Criterion

Geographically Weighted Principal Component Analysis For Spatio-Temporal Statistical Dataset

Dimensionality Reduction Of Multivariate Data For Bayesian Analysis

An EWMA Chart For Monitoring Covariance Matrix Based On Dissimilarity Index

Adjusting For Linkage Bias In The Analysis Of Record-Linked Data

Bayesian Semiparametric Hierarchical Models For Longitudinal Data Analysis With Application To Dose-Response Studies

Optimizing Junior Rugby Weight Limits

Spatial Scan Statistics For Matched Case-Control Data

Whitebait In All Its Varieties: One Fish, Two Fish, Three, Four, Five Fish.

Latent Variable Models And Multivariate Binomial Data

Asking About Sex In General Health Surveys: Comparing The Methods And Findings Of The 2010 Health Survey For England With Those Of The Third National Survey Of Sexual Attitudes And Lifestyles

Bayesian Continuous Space-Time Model Of Burglaries

Tolerance Limits For The Reliability Of Semiconductor Devices Using Longitudinal Data

Model-Checking For Regressions: A Local Smoothing-Based Global Smoothing Test

Breeding Value Estimation In Partially-Genotyped Populations

BIVAS: A Scalable Bayesian Method For Bi-Level Variable Selection

Ranking Potential Shoplifters In Real Time

Two Stage Approach To Data-Driven Subgroup Identification In Clinical Trials

Inverse Regression For Multivariate Functional Data

Including Covariate Estimation Error When Predicting Species Distributions: A Simulation Exercise Using Template Model Builder

Adjusted Adaptive Index Model For Binary Response

Factors Influencing On Growth Of Garments Industry In Bangladesh

Comparison Of Exact And Approximate Testing Procedures In Clinical Trials With Multiple Binary Endpoints

Multiple Function-On-Function Linear Regression With Application To Weather Forecast Calibration

Modelling The Distribution Of Lifetime Using Compound Time-Homogenous Poisson Process

Detecting Change-Points In The Stress-Strength Reliability P(X<Y)

New Zealand Crime And Victims Survey: Filling The Knowledge Gap

Missing Data In Randomised Control Trials: Stepped Multiple Imputation

Programme And Abstracts For Wednesday 13th Of December

Promoting Your R Package

A Smoothing Filter Modelling Approach For Time Series

Online Learning For Bayesian Nonparametrics: Weakly Conjugate Approximation

Improving The Production Cycle At Stats NZ With RStudio

A Max-Type Multivariate Two-Sample Baumgartner Statistic

Random Search Global Optimization Using Random Forests

gridSVG: Then And Now

Probabilistic Outlier Detection And Visualization Of Smart Meter Data

The Joint Models For Nonlinear Longitudinal And Time-To-Event Data Using Penalized Splines: A Bayesian Approach

R – A Powerful Analysis Tool To Improve Official Statistics In Romania

Simultaneous Test For Mean Vectors And Covariance Matrices In High-Dimensional Settings

Dimension Reduction For Classification Of High-Dimensional Data By Stepwise SVM

Bringing Multimix From Fortran To R

Specification Of GARCH Model Under Asymmetric Error Innovations

Performance Of Bayesian Credible Interval For Binomial Proportion Using Logit Transformation

Statistical Disclosure Control With R: Traditional Methods And Synthetic Data

High Dimensional Asymptotics For The Naive Canonical Correlation Coefficient

Deep Learning High-Dimensional Covariance Matrices

R In Industry – Application On Pipe Renewal Planning

Empirical Comparison Of Some Algorithms For Automatic Univariate ARMA Modeling Using RcmdrPlugin.SPSS

Bayesian Optimum Warranty Length Under Type-II Unified Hybrid Censoring Scheme

Imputation Of The 2016 Economic Census For Business Activity In Japan

Applying Active Learning Procedure To Drug Consumption Data

R For Everything

R Package For New Two-Stage Methods In Forecasting Time Series With Multiple Seasonality

Analysis Of Official Microdata Using Secure Statistical Computation System

Presenting Flexi, A Statistical Program For Fitting Variance Models

Space And Circular Time Log Gaussian Cox Processes With Application To Crime Event Data

Cluster-Wise Regression Models Combined By A Quasi-Linear Function

Hierarchical Structural Component Analysis Of Gene-Environment Interactions

Wavelet-Based Power Transformation Of Non-Gaussian Long Memory Time Series

Cross Covariance Estimation For Integration Of Multi-Omics Data

Relationships Between Linguistic Characteristics And The Use Of Māori Loanwords In New Zealand English.

Transfer Regression And Predictive Distributions

An Overview Of The Correspondence Analysis Family

Testing For Genetic Associations In Arbitrarily Structured Populations

Threshold Determination For The Meteorological Data Quality Control In Korea

Regularized Noise-Reduction Methodology For High-Dimensional Data

Computation Of Influence Functions For Robust Statistics

Adaptive Model Averaging In High-Dimensional Linear Regression

Model-Based Clustering For Multivariate Categorical Data With Dimension Reduction

Phylogenetic Tree-Based Microbiome Association Test

Fitting Additive Hazards Model Using Calibrated Weights For Case-Cohort Data

Selecting The Number Of Principal Components

Rolling Survival Extrapolation Algorithm For Estimating Life Years Lost Among Subjects Exposed To Long-Term Air Pollution

Enhancing The Flexibility Of Regression Modeling By Liquid Association

Clusterwise Low-Rank Correlation Analysis Based On Majorization

Bayesian Analysis For Fitting Zero-Inflated Count Data With Data Augmentation

Towards A Sparse, Scalable, And Stably Positive Definite (Inverse) Covariance Estimator

Tick-By-Tick Effect On The Inference Of Ultra-High Frequency Data

High Mortality Predictions With Lines Or Curves Fitted To Over-Dispersed Exposure-Mortality Data

Lattice Polytope Samplers

Statistical Modelling And Analysis Of Cosmic Microwave Background Data

Measure Of Departure From Marginal Average Point-Symmetry For Two-Way Contingency Tables With Ordered Categories

Sparse Estimates From Dense Precision Matrix Posteriors

Dimension Reduction Strategies For Modeling Bi-Clustered High Dimensional Data

Investigating Methods To Produce Price Indexes From Big Data

Computing Entropies With Nested Sampling

Spline-Based Drift Models For High Temperature Operating Life Tests

A New Approach To Distribution Free Tests In Contingency Tables

A Bayesian Inference For Time Series Via Copula-Based Markov Chain Models

Modified Gene Shaving Algorithm - A Dimension REduction And Clustering Method

The Potential Of Web Scraping

A Simple Method To Construct Confidence Bands In Functional Linear Regression

Separation Of Symmetry For Square Contingency Tables With Ordinal Categories

Testing For Presence Of Clustering Effect In Multilevel Model With High-Dimensional Predictors

Towards A Big Data CPI For New Zealand

Programme And Abstracts For Thursday 14th Of December

ALTREP: Alternate Representations Of Basic R Objects

Penalized Vector Generalized Additive Models

A Package For Multiple Precision Floating-Point Computation On R

Dissimilarities Between Groups Of Data

Comparison Of Tests Of Mean Difference In Longitudinal Data Based On Block Resampling Methods

SSREM: A Summary-Statistics-Based Random Effect Model To Estimating Heritability, Co-Heritability And Effect Sizes In GWAS Data Analysis

Consistency Of Linear Mixed-Effects Model Selection With Inconsistent Covariance Parameter Estimators

An Incomplete-Data Fisher Scoring With An Acceleration Method

Interactive Visualization Of Aggregated Symbolic Data

Analysis Of Spatial Data With A Gaussian Mixture Markov Random Field Model

Forward Selection In Regression Models Based On Robust Estimation

Selecting Generalised Linear Models Under Inequality Constraints

Improvement Of Computation For Nonlinear Multivariate Methods

Feature Selection In High-Dimensional Models With Complex Block Structures

Statistical Generalized Derivative Applied To The Profile Likelihood Estimation In A Mixture Of Semiparametric Models

New Zealand Statistical Association and the International Association of Statistical Computing (Asian Regional Section) Joint Conference 2017

Auckland
New Zealand

10-14th December 2017

New Zealand Statistical Association and the International Association of Statistical Computing (Asian Regional Section) Joint Conference 2017

1. Welcomes
    1. Welcome To The Conference From Ian Westbrooke, NZSA President
    2. Welcome Address From Jung Jin Lee, IASC-ARS Chair-Person
2. Programme At A Glance
    1. Monday
    2. Tuesday
    3. Wednesday
    4. Thursday
3. Am I A Session Chair? Please Read!!!
    1. Monday
    2. Tuesday
    3. Wednesday
    4. Thursday
4. Index By First Name of Author
5. Index by Submission Number
6. Programme And Abstracts For Monday 11th Of December
    1. R In Times Of Growing User Base And Data Sizes
    2. Robust Principal Expectile Component Analysis

3. Effect Of Area Level Deprivation On Body Mass Index: Analysis Of NZ Health Surveys
4. Calendar-Based Graphics For Visualising People's Daily Schedules
5. Nonparametric Test For Volatility In Clustered Multiple Time Series
6. IGESS: A Statistical Approach To Integrating Individual Level Genotype Data And Summary Statistics In Genome Wide Association Studies
7. Author Name Identification For Evaluating Research Performance Of Institutes
8. A Computational Tool For Detecting Copy Number Variations From Whole Genome And Targeted Exome Sequencing
9. Clustering Using Nonparametric Mixtures And Mode Identification
10. Bayesian Curve Fitting For Discontinuous Function Using Overcomplete Representation With Multiple Kernels
11. Estimation Of A Semiparametric Spatiotemporal Models With Mixed Frequency
12. LSMM: A Statistical Approach To Integrating Functional Annotations With Genome-Wide Association Studies
13. A Study Of The Influence Of Articles In The Large-Scale Citation Network
14. Estimating Links Of A Network From Time To Event Data
15. Estimation Of A High-Dimensional Covariance Matrix
16. Innovative Bayesian Estimation In The von Mises Distribution
17. Evidence Of Climate Change From Nonparametric Change-Point Analysis
18. Joint Analysis Of Individual Level Genotype Data And Summary Statistics By Leveraging Pleiotropy
19. An Advanced Approach For Time Series Forecasting Using Deep Learning
20. Genetic Map Estimation Using Hidden Markov Models In The Presence Of Partially Observed Information

21. A Simple Method For Grouping Patients Based On Historical Doses

22. Semiparametric Mixed Analysis Of Covariance Model

23. Adaptive False Discovery Rate Regression With Application In Integrative Analysis Of Large-Scale Genomic Data

24. Structure Of Members In The Organization To Induce Innovation: Quantitatively Analyze The Capability Of The Organization

25. Vector Generalized Linear Time Series Models

26. Local Canonical Correlation Analysis For Multimodal Labeled Data

27. A Practitioners Guide To Deep Learning For Predictive Analytics On Structured Data

28. Clustering Of Research Subject Based On Stochastic Block Model

29. Zen And The aRt Of Workflow Maintenance

30. Canonical Covariance Analysis For Mixed Numerical And Categorical Three-Way Three-Mode Data

31. Variable Selection Algorithms

32. Estimating Causal Structures For Continuous And Discrete Variables

33. Incorporating Genetic Networks Into Case-Control Association Studies With High-Dimensional DNA Methylation Data

34. Adaptive Model Checking For Functional Single-Index Models

35. Mobile Learning In Teaching Bioinformatics For Medical Doctors

36. On Optimal Group Testing Designs: Prevalence Estimation, Cost Considerations, And Dilution Effects

37. The Use Of Bayesian Networks In Grape Yield Prediction

38. Pattern Prediction For Time Series Data With Change Points

39. Test For Genomic Imprinting Effects On The X Chromosome

40. Fluctuation Reduction Of Value-At-Risk Estimation And Its Applications

41. E-Learning Courses On Introductory Statistics Using Interactive Educational Tools

42. Estimation Of Animal Density From Acoustic Detections

43. Mixed Models For Complex Survey Data

44. Regression With Random Effects For Analysing Correlated Survival Data: Application To Disease Recurrences

45. Genetic Predictors Underlying Long-Term Cognitive Recovery Following Mild Traumatic Brain Injury

46. Bayesian Structure Selection For Vector Autoregression Model

47. Three-Dimensional Data Visualization Education With Virtual Reality

48. Talk Data To Me

49. Smooth Nonparametric Regression Under Shape Restrictions

50. Elastic-Band Transform: A New Approach To Multiscale Visualization

51. Meta-Analytic Principal Component Analysis In Integrative Omics Application

52. Flight To Relative Safety: Learning From A No-Arbitrage Network Of Yield Curves Model Of The Euro Area

53. Bayesian Analyses Of Non-Homogeneous Gaussian Hidden Markov Models

54. Robustness Of Temperature Reconstruction For Past 500 Years

55. Nonparametric Causal Inference By The Kernel Method

56. A Unified Regularized Group PLS Algorithm Scalable To Big Data

57. Evaluation Of Spatial Cluster Detection Method Based On All Geographical Linkage Patterns

58. Scoring Rules For Prediction And Classification Challenges

59. Meta-Analysis With Symbolic Data Analysis And Its Application For Clinical Data

60. Real-Time Transit Network Modelling For Improved Arrival Time Predictions

61. Visualization And Statistical Modeling Of Financial Big Data

62. Sparse Group-Subgroup Partial Least Squares With Application To Genomic Data

63. Genetic Approach And Statistical Approach For Association Study On DNA Data

64. Modeling Of Document Abstraction Using Association Rule Based Characterization

65. Bayesian Static Parameter Inference For Partially Observed Stochastic Systems

66. Bayesian Survival Analysis Of Batsmen In Test Cricket

67. Covariate Discretisation On Big Data

68. BIG-SIR A Sliced Inverse Regression Approach For Massive Data

69. Symbolic Data Analytical Approach To Unauthorized-Access Logs

70. My Knee Still Hurts; The Statistical Pathway To The Development Of A Clinical Decision Aid

7. Programme And Abstracts For Tuesday 12th Of December

1. Could Do Better ... A Report Card For Statistical Computing

2. R&D Policy Regimes In France: New Evidence From A Spatio-Temporal Analysis

3. Analysing Scientific Collaborations Of New Zealand Institutions Using Scopus Bibliometric Data

4. Family Structure And Academic Achievements Of High School Students In Tonga

5. Analysis Of Multivariate Binary Longitudinal Data: Metabolic Syndrome During Menopausal Transition

6. Clustering Of Curves On A Spatial Domain Using A Bayesian Partitioning Model

7. The Uncomfortable Entrepreneurs: Bad Working Conditions And Entrepreneurial Commitment

8. Spatial Surveillance With Scan Statistics By Controlling The False Discovery Rate

9. Statistical Models For The Source Attribution Of Zoonotic Diseases: A Study Of Campylobacteriosis

10. Towards An Informal Test For Goodness-Of-Fit

11. Identifying Clusters Of Patients With Diabetes Using A Markov Birth-Death Process

12. Bayesian Temporal Density Estimation Using Autoregressive Species Sampling Models

13. How Does The Textile Set Describe Geometric Structures Of Data?

14. Intensity Estimation Of Spatial Point Processes Based On Area-Aggregated Data

15. Bayesian Inference For Population Attributable Measures

16. An Information Criterion For Prediction With Auxiliary Variables Under Covariate Shift

17. Analysis Of A Brief Telephone Intervention For Problem Gambling And Examining The Impact On Co-Existing Depression?

18. Prior-Based Bayesian Information Criterion

19. Geographically Weighted Principal Component Analysis For Spatio-Temporal Statistical Dataset

20. Dimensionality Reduction Of Multivariate Data For Bayesian Analysis

21. An EWMA Chart For Monitoring Covariance Matrix Based On Dissimilarity Index

22. Adjusting For Linkage Bias In The Analysis Of Record-Linked Data

23. Bayesian Semiparametric Hierarchical Models For Longitudinal Data Analysis With Application To Dose-Response Studies

24. Optimizing Junior Rugby Weight Limits

25. Spatial Scan Statistics For Matched Case-Control Data

26. Whitebait In All Its Varieties: One Fish, Two Fish, Three, Four, Five Fish.

27. Latent Variable Models And Multivariate Binomial Data

28. Asking About Sex In General Health Surveys: Comparing The Methods And Findings Of The 2010 Health Survey For England With Those Of The Third National Survey Of Sexual Attitudes And Lifestyles

29. Bayesian Continuous Space-Time Model Of Burglaries

30. Tolerance Limits For The Reliability Of Semiconductor Devices Using Longitudinal Data

31. Model-Checking For Regressions: A Local Smoothing-Based Global Smoothing Test

32. Breeding Value Estimation In Partially-Genotyped Populations

33. BIVAS: A Scalable Bayesian Method For Bi-Level Variable Selection

34. Ranking Potential Shoplifters In Real Time

35. Two Stage Approach To Data-Driven Subgroup Identification In Clinical Trials

36. Inverse Regression For Multivariate Functional Data

37. Including Covariate Estimation Error When Predicting Species Distributions: A Simulation Exercise Using Template Model Builder

38. Adjusted Adaptive Index Model For Binary Response

39. Factors Influencing On Growth Of Garments Industry In Bangladesh

40. Comparison Of Exact And Approximate Testing Procedures In Clinical Trials With Multiple Binary Endpoints

41. Multiple Function-On-Function Linear Regression With Application To Weather Forecast Calibration

42. Modelling The Distribution Of Lifetime Using Compound Time-Homogenous Poisson Process

43. Detecting Change-Points In The Stress-Strength Reliability P(X<Y)

44. New Zealand Crime And Victims Survey: Filling The Knowledge Gap

45. Missing Data In Randomised Control Trials: Stepped Multiple Imputation

8. Programme And Abstracts For Wednesday 13th Of December

1. Promoting Your R Package

2. A Smoothing Filter Modelling Approach For Time Series

3. Online Learning For Bayesian Nonparametrics: Weakly Conjugate Approximation

4. Improving The Production Cycle At Stats NZ With RStudio

5. A Max-Type Multivariate Two-Sample Baumgartner Statistic

6. Random Search Global Optimization Using Random Forests

7. gridSVG: Then And Now

8. Probabilistic Outlier Detection And Visualization Of Smart Meter Data

9. The Joint Models For Nonlinear Longitudinal And Time-To-Event Data Using Penalized Splines: A Bayesian Approach

10. R – A Powerful Analysis Tool To Improve Official Statistics In Romania

11. Simultaneous Test For Mean Vectors And Covariance Matrices In High-Dimensional Settings

12. Dimension Reduction For Classification Of High-Dimensional Data By Stepwise SVM

13. Bringing Multimix From Fortran To R

14. Specification Of GARCH Model Under Asymmetric Error Innovations

15. Performance Of Bayesian Credible Interval For Binomial Proportion Using Logit Transformation

16. Statistical Disclosure Control With R: Traditional Methods And Synthetic Data

17. High Dimensional Asymptotics For The Naive Canonical Correlation Coefficient

18. Deep Learning High-Dimensional Covariance Matrices

19. R In Industry – Application On Pipe Renewal Planning

20. Empirical Comparison Of Some Algorithms For Automatic Univariate ARMA Modeling Using RcmdrPlugin.SPSS

21. Bayesian Optimum Warranty Length Under Type-II Unified Hybrid Censoring Scheme

22. Imputation Of The 2016 Economic Census For Business Activity In Japan

23. Applying Active Learning Procedure To Drug Consumption Data

24. R For Everything

25. R Package For New Two-Stage Methods In Forecasting Time Series With Multiple Seasonality

26. Analysis Of Official Microdata Using Secure Statistical Computation System

27. Presenting Flexi, A Statistical Program For Fitting Variance Models

28. Space And Circular Time Log Gaussian Cox Processes With Application To Crime Event Data

29. Cluster-Wise Regression Models Combined By A Quasi-Linear Function

30. Hierarchical Structural Component Analysis Of Gene-Environment Interactions

31. Wavelet-Based Power Transformation Of Non-Gaussian Long Memory Time Series

32. Cross Covariance Estimation For Integration Of Multi-Omics Data

33. Relationships Between Linguistic Characteristics And The Use Of Māori Loanwords In New Zealand English.

34. Transfer Regression And Predictive Distributions

35. An Overview Of The Correspondence Analysis Family

36. Testing For Genetic Associations In Arbitrarily Structured Populations

37. Threshold Determination For The Meteorological Data Quality Control In Korea

38. Regularized Noise-Reduction Methodology For High-Dimensional Data

39. Computation Of Influence Functions For Robust Statistics

40. Adaptive Model Averaging In High-Dimensional Linear Regression

41. Model-Based Clustering For Multivariate Categorical Data With Dimension Reduction

42. Phylogenetic Tree-Based Microbiome Association Test

43. Fitting Additive Hazards Model Using Calibrated Weights For Case-Cohort Data

44. Selecting The Number Of Principal Components

45. Rolling Survival Extrapolation Algorithm For Estimating Life Years Lost Among Subjects Exposed To Long-Term Air Pollution

46. Enhancing The Flexibility Of Regression Modeling By Liquid Association

47. Clusterwise Low-Rank Correlation Analysis Based On Majorization

48. Bayesian Analysis For Fitting Zero-Inflated Count Data With Data Augmentation

49. Towards A Sparse, Scalable, And Stably Positive Definite (Inverse) Covariance Estimator

50. Tick-By-Tick Effect On The Inference Of Ultra-High Frequency Data

51. High Mortality Predictions With Lines Or Curves Fitted To Over-Dispersed Exposure-Mortality Data

52. Lattice Polytope Samplers

53. Statistical Modelling And Analysis Of Cosmic Microwave Background Data

54. Measure Of Departure From Marginal Average Point-Symmetry For Two-Way Contingency Tables With Ordered Categories

55. Sparse Estimates From Dense Precision Matrix Posteriors

56. Dimension Reduction Strategies For Modeling Bi-Clustered High Dimensional Data

57. Investigating Methods To Produce Price Indexes From Big Data

58. Computing Entropies With Nested Sampling

59. Spline-Based Drift Models For High Temperature Operating Life Tests

60. A New Approach To Distribution Free Tests In Contingency Tables

61. A Bayesian Inference For Time Series Via Copula-Based Markov Chain Models

62. Modified Gene Shaving Algorithm - A Dimension REduction And Clustering Method

63. The Potential Of Web Scraping

64. A Simple Method To Construct Confidence Bands In Functional Linear Regression

65. Separation Of Symmetry For Square Contingency Tables With Ordinal Categories

66. Testing For Presence Of Clustering Effect In Multilevel Model With High-Dimensional Predictors

67. Towards A Big Data CPI For New Zealand

9. Programme And Abstracts For Thursday 14th Of December

1. ALTREP: Alternate Representations Of Basic R Objects

2. Penalized Vector Generalized Additive Models

3. A Package For Multiple Precision Floating-Point Computation On R

4. Dissimilarities Between Groups Of Data

5. Comparison Of Tests Of Mean Difference In Longitudinal Data Based On Block Resampling Methods

6. SSREM: A Summary-Statistics-Based Random Effect Model To Estimating Heritability, Co-Heritability And Effect Sizes In GWAS Data Analysis

7. Consistency Of Linear Mixed-Effects Model Selection With Inconsistent Covariance Parameter Estimators

8. An Incomplete-Data Fisher Scoring With An Acceleration Method

9. Interactive Visualization Of Aggregated Symbolic Data

10. Analysis Of Spatial Data With A Gaussian Mixture Markov Random Field Model

11. Forward Selection In Regression Models Based On Robust Estimation

12. Selecting Generalised Linear Models Under Inequality Constraints

13. Improvement Of Computation For Nonlinear Multivariate Methods

14. Feature Selection In High-Dimensional Models With Complex Block Structures

15. Statistical Generalized Derivative Applied To The Profile Likelihood Estimation In A Mixture Of Semiparametric Models

1. Cover

2. Table of contents

Welcomes

Welcome To The Conference From Ian Westbrooke, NZSA President

Welcome to the 2017 Joint Conference of the NZSA and the IASC-ARS,

I bid a special welcome to all who travelled far from countries throughout the Asia-Pacific region and beyond. The New Zealand Statistical Association greatly appreciates the opportunity to share this conference with the International Society for Statistical Computing – Asian Regional Section. The massive growth in interest in data science and analytics demonstrates the need for greater availability of statistical and especially statistical computing expertise and input. This conference provides an opportunity for all of us as statisticians to develop and share the skills we can contribute to the increasing interest and need for our sort of quantitative skills.

It is fitting that we are meeting at Auckland University, the original home of R, and the home of its co-founder, Ross Ihaka. Ross has recently retired, and a theme of this conference is in honour of his huge contributions to our field.

The conference will be opened by the Dean of Science at the University of Auckland, Professor John Hosking, and the President of the IASC-ARS. In-between, we have six outstanding keynote speakers who set the framework for an exciting and diverse scientific program. I invite you all to attend presentations on topics familiar to your research and to explore what is done in other areas.

We have a very healthy number of students and early career researchers among the more than 300 delegates attending, indicating how popular, important and attractive data rich research continues to be in the years to come.

This conference will be a success because of excellent work by many, in particular by Thomas Yee, Ciprian Giurcaneanu and James Curran, who form the local organizing committee. Thanks also to (Chair of the Scientific Program) Ciprian Giurcaneanu, and his team: Rolf Turner and Abhinav Chopra. Thanks finally to Tessa Lloyd-Hagemann, our events coordinator.

I would also thank the four workshop presenters for their substantial contribution to this event before the main conference programme begins.

Last but not least I thank our sponsors: Minitab, The New Zealand Statistical Association, Harmonic Analytics, SAS, The International Society for Bayesian Analysis, Wiley and the Australian and New Zealand Journal of Statistics, for their generous financial support. Many aspects of the conference are only possible because of their help,

Ian Westbrooke, President NZSA

Welcome Address From Jung Jin Lee, IASC-ARS Chair-Person

I appreciate your participation for the Joint Meeting of the 10th Conference of the Asian Regional Section of the International Association for Statistical Computing and the NZ Statistical Association, to be held at Auckland University, home of R, from 10 to 14 December 2017. It brings together eminent statisticians and members of the statistical computing community from Asia, NZ and other continents to present, discuss, promote and disseminate research and its applications. I hope you have a great and enjoyable opportunity for sharing experiences and networking within the R community, as well as enjoying the beautiful and warm ambience of Auckland.

output:
bookdown::gitbook:
includes:
in_header: tblsawinclude.html

Programme At A Glance

Monday

850	**Opening Ceremony**					
910	**R In Times Of Growing User Base And Data Sizes** **Simon Urbanek, AT&T Labs, USA** **098 Lecture Theatre (260-098)**					
1000	***Morning Tea (30 minutes)***					
	098 Lecture Theatre(260-098)	**OGGB4(260-073)**	**OGGB5(260-051)**	**Case Room 2(260-057)**	**Case Room 3(260-055)**	**Case Room 4(260-009)**
1030	***Robust Principal Expectile Component Analysis*** ***Liang-Ching Lin***	***Effect Of Area Level Deprivation On Body Mass Index: Analysis Of NZ Health Surveys*** ***Andrew Adiguna Halim***	***Calendar-Based Graphics For Visualising People's Daily Schedules*** ***Earo Wang***	***Nonparametric Test For Volatility In Clustered Multiple Time Series*** ***Paolo Victor Redondo***	***IGESS: A Statistical Approach To Integrating Individual Level Genotype Data And Summary Statistics In Genome Wide Association Studies*** ***Xiang Wan***	***Author Name Identification For Evaluating Research Performance Of Institutes*** ***Tomokazu Fujino***

1050	A Computational Tool For Detecting Copy Number Variations From Whole Genome And Targeted Exome Sequencing Yu-Chung Wei	Clustering Using Nonparametric Mixtures And Mode Identification Shengwei Hu	Bayesian Curve Fitting For Discontinuous Function Using Overcomplete Representation With Multiple Kernels Youngseon Lee	Estimation Of A Semiparametric Spatiotemporal Models With Mixed Frequency Erniel Barrios	LSMM: A Statistical Approach To Integrating Functional Annotations With Genome-Wide Association Studies Jingsi Ming	A Study Of The Influence Of Articles In The Large-Scale Citation Network Frederick Kin Hing Phoa

	098 Lecture Theatre(260-098)	OGGB4(260-073)	OGGB5(260-051)	Case Room 2(260-057)	Case Room 3(260-055)	Case Room 4(260-009)
1110	Estimating Links Of A Network From Time To Event Data Tso-Jung Yen	Estimation Of A High-Dimensional Covariance Matrix Xiangjie Xue	Innovative Bayesian Estimation In The von Mises Distribution Yuta Kamiya	Evidence Of Climate Change From Nonparametric Change-Point Analysis Angela Nalica	Joint Analysis Of Individual Level Genotype Data And Summary Statistics By Leveraging Pleiotropy Can Yang	An Advanced Approach For Time Series Forecasting Using Deep Learning Balaram Panda
1130		Genetic Map Estimation Using Hidden Markov Models In The Presence Of Partially Observed	A Simple Method For Grouping Patients Based On Historical Doses Shengli Tzeng	Semiparametric Mixed Analysis Of Covariance Model Joseph Ryan Lansangan	Adaptive False Discovery Rate Regression With Application In Integrative Analysis Of Large-Scale Genomic	Structure Of Members In The Organization To Induce Innovation: Quantitatively Analyze The Capability Of The

		Information Timothy Bilton			Data Can Yang	Organization Yuji Mizukami
1150		Vector Generalized Linear Time Series Models Victor Miranda	Local Canonical Correlation Analysis For Multimodal Labeled Data Seigo Mizutani	A Practitioners Guide To Deep Learning For Predictive Analytics On Structured Data Balaram Panda		Clustering Of Research Subject Based On Stochastic Block Model Keisuke Honda
1210	***Group photo, followed by Lunch (1 hour 10 minutes)***					

1320	Zen And The aRt Of Workflow Maintenance Jenny Bryan, University of British Columbia 098 Lecture Theatre (260-098)					
	098 Lecture Theatre(260-098)	OGGB4(260-073)	OGGB5(260-051)	Case Room 2(260-057)	Case Room 3(260-055)	Case Room 4(260-009)
1410	Canonical Covariance Analysis For Mixed Numerical And Categorical Three-Way Three-Mode Data Jun Tsuchida	Variable Selection Algorithms Fangyao Li	Estimating Causal Structures For Continuous And Discrete Variables Mako Yamayoshi	Incorporating Genetic Networks Into Case-Control Association Studies With High-Dimensional DNA Methylation Data Hokeun Sun	Adaptive Model Checking For Functional Single-Index Models Zhenghui Feng	Mobile Learning In Teaching Bioinformatics For Medical Doctors Taerim Lee and Jung Jin Lee
1430	On Optimal Group Testing Designs: Prevalence Estimation, Cost Considerations, And Dilution Effects Shih-Hao Huang	The Use Of Bayesian Networks In Grape Yield Prediction Rory Ellis	Pattern Prediction For Time Series Data With Change Points Satoshi Goto	Test For Genomic Imprinting Effects On The X Chromosome Wing Kam Fung	Fluctuation Reduction Of Value-At-Risk Estimation And Its Applications Shih-Feng Huang	E-Learning Courses On Introductory Statistics Using Interactive Educational Tools Kazunori Yamaguchi
1450	Estimation Of Animal Density From Acoustic	Mixed Models For Complex Survey	Regression With Random Effects For	Genetic Predictors Underlying Long-Term	Bayesian Structure Selection For Vector	Three-Dimensional Data Visualization

	Detections Ben Stevenson	Data Xudong Huang	Analysing Correlated Survival Data: Application To Disease Recurrences Richard Tawiah	Cognitive Recovery Following Mild Traumatic Brain Injury Priya Parmar	Autoregression Model Ray-Bing Chen	Education With Virtual Reality Dae-Heung Jang

	098 Lecture Theatre(260-098)	OGGB4(260-073)	OGGB5(260-051)	Case Room 2(260-057)	Case Room 3(260-055)	Case Room 4(260-009)
1510	Talk Data To Me Lisa Hall	Smooth Nonparametric Regression Under Shape Restrictions Hongbin Guo	Elastic-Band Transform: A New Approach To Multiscale Visualization Guebin Choi	Meta-Analytic Principal Component Analysis In Integrative Omics Application Sunghwan Kim	Flight To Relative Safety: Learning From A No-Arbitrage Network Of Yield Curves Model Of The Euro Area Linlin Niu	
1530	*Afternoon tea (30 minutes)*					
1600	Bayesian Analyses Of Non-Homogeneous Gaussian Hidden Markov Models Shin Sato	Robustness Of Temperature Reconstruction For Past 500 Years Yu Yang	Nonparametric Causal Inference By The Kernel Method Yuchi Matsuoka	A Unified Regularized Group PLS Algorithm Scalable To Big Data	Evaluation Of Spatial Cluster Detection Method Based On All Geographical Linkage	Scoring Rules For Prediction And Classification Challenges Matt Parry

				Pierre Lafaye de Micheaux	Patterns Fumio Ishioka	
1620	Meta-Analysis With Symbolic Data Analysis And Its Application For Clinical Data Ryo Takagi	Real-Time Transit Network Modelling For Improved Arrival Time Predictions Tom Elliott	Visualization And Statistical Modeling Of Financial Big Data Masayuki Jimichi	Sparse Group-Subgroup Partial Least Squares With Application To Genomic Data Matthew Sutton	Genetic Approach And Statistical Approach For Association Study On DNA Data Makoto Tomita	Modeling Of Document Abstraction Using Association Rule Based Characterization Ken Nittono
1640	Bayesian Static Parameter Inference For Partially Observed Stochastic Systems Yaxian Xu	Bayesian Survival Analysis Of Batsmen In Test Cricket Oliver Stevenson	Covariate Discretisation On Big Data Hon Hwang	BIG-SIR A Sliced Inverse Regression Approach For Massive Data Benoit Liquet	Symbolic Data Analytical Approach To Unauthorized-Access Logs Hiroyuki Minami	My Knee Still Hurts; The Statistical Pathway To The Development Of A Clinical Decision Aid Robert Borotkanics

Tuesday

910	Could Do Better ... A Report Card For Statistical Computing Ross Ihaka, University of Auckland 098 Lecture Theatre (260-098)

1000	***Morning Tea (30 minutes)***					
	098 LT	**OGGB4**	**OGGB5**	**Case Room 2**	**Case Room 3**	**Case Room 4**
1030	R&D Policy Regimes In France: New Evidence From A Spatio-Temporal Analysis Benjamin Montmartin	Analysing Scientific Collaborations Of New Zealand Institutions Using Scopus Bibliometric Data Samin Aref	Family Structure And Academic Achievements Of High School Students In Tonga Losana Vao Latu Latu	Analysis Of Multivariate Binary Longitudinal Data: Metabolic Syndrome During Menopausal Transition Geoff Jones	Clustering Of Curves On A Spatial Domain Using A Bayesian Partitioning Model Chae Young Lim	The Uncomfortable Entrepreneurs: Bad Working Conditions And Entrepreneurial Commitment Catherine Laffineur
1050	Spatial Surveillance With Scan Statistics By Controlling The False Discovery Rate Xun Xiao	Statistical Models For The Source Attribution Of Zoonotic Diseases: A Study Of Campylobacteriosis Sih-Jing Liao	Towards An Informal Test For Goodness-Of-Fit Anna Fergusson	Identifying Clusters Of Patients With Diabetes Using A Markov Birth-Death Process Mugdha Manda	Bayesian Temporal Density Estimation Using Autoregressive Species Sampling Models Seongil Jo	How Does The Textile Set Describe Geometric Structures Of Data? Ushio Tanaka
1110	Intensity Estimation Of Spatial Point Processes Based On Area-	Bayesian Inference For Population Attributable Measures Sarah Pirikahu	An Information Criterion For Prediction With Auxiliary Variables Under Covariate	Analysis Of A Brief Telephone Intervention For Problem Gambling And Examining The Impact On Co-Existing	Prior-Based Bayesian Information Criterion Woncheol Jang	

	Aggregated Data Hsin-Cheng Huang		Shift Takahiro Ido	Depression? Nick Garrett		

1130	Geographically Weighted Principal Component Analysis For Spatio-Temporal Statistical Dataset Narumasa Tsutsumida	Dimensionality Reduction Of Multivariate Data For Bayesian Analysis Anjali Gupta	An EWMA Chart For Monitoring Covariance Matrix Based On Dissimilarity Index Longcheen Huwang	Adjusting For Linkage Bias In The Analysis Of Record-Linked Data Patrick Graham	Bayesian Semiparametric Hierarchical Models For Longitudinal Data Analysis With Application To Dose-Response Studies Taeryon Choi	Optimizing Junior Rugby Weight Limits Ankit Patel
1150	Spatial Scan Statistics For Matched Case-Control Data Inkyung Jung	Whitebait In All Its Varieties: One Fish, Two Fish, Three, Four, Five Fish. Bridget Armstrong	Latent Variable Models And Multivariate Binomial Data John Holmes	Asking About Sex In General Health Surveys: Comparing The Methods And Findings Of The 2010 Health Survey For England With Those Of The Third National Survey Of Sexual Attitudes And Lifestyles Philip Prah	Bayesian Continuous Space-Time Model Of Burglaries Paul Brown	Tolerance Limits For The Reliability Of Semiconductor Devices Using Longitudinal Data Thomas Nowak
1210	***Lunch (1 hour 10 minutes)***					

1320	**Session In Memory Of Alastair John Scott** **1320hrs-1530hrs** **Speakers: Professor John Neuhaus (UCSF), Professor Chris Wild, Professor Alan Lee, Professor Thomas Lumley**					
1530	***Afternoon tea (30 minutes)***					
	098 Lecture Theatre(260-098)	**OGGB4(260-073)**	**OGGB5(260-051)**	**Case Room 2(260-057)**	**Case Room 3(260-055)**	**Case Room 4(260-009)**
1600	***Model-Checking For Regressions: A Local Smoothing-Based Global Smoothing Test*** ***Lingzhu Li***	***Breeding Value Estimation In Partially-Genotyped Populations*** ***Alastair Lamont***	***BIVAS: A Scalable Bayesian Method For Bi-Level Variable Selection*** ***Mingxuan Cai***	***Ranking Potential Shoplifters In Real Time*** ***Barry McDonald***	***Two Stage Approach To Data-Driven Subgroup Identification In Clinical Trials*** ***Toshio Shimokawa***	
1620	***Inverse Regression For Multivariate Functional Data*** ***Ci-Ren Jiang***	***Including Covariate Estimation Error When Predicting Species Distributions: A Simulation Exercise***	***Adjusted Adaptive Index Model For Binary***	***Factors Influencing On Growth Of Garments Industry In Bangladesh*** ***Md. Shahidul Islam***	***Comparison Of Exact And Approximate Testing Procedures In Clinical Trials With Multiple Binary***	

		Using Template Model Builder Andrea Havron	Response Ke Wan		Endpoints Takuma Ishihara	
1640	Multiple Function-On-Function Linear Regression With Application To Weather Forecast Calibration Lu-Hung Chen	Modelling The Distribution Of Lifetime Using Compound Time-Homogenous Poisson Process Kien Tran	Detecting Change-Points In The Stress-Strength Reliability P(X<Y) Hang Xu	New Zealand Crime And Victims Survey: Filling The Knowledge Gap Tianying Chu	Missing Data In Randomised Control Trials: Stepped Multiple Imputation Rose Sisk	
1700		NZSA Annual General Meeting 1700 - 1815	IASC General Assembly 1700-1815			
1830	Conference Dinner Owen G Glenn Building Foyer Level 1, 20 Grafton Road 1830-2230					

Wednesday

910	**Professor Michael Ferris, University Of Wisconsin - Madison, USA** **ORSNZ Keynote** **098 Lecture Theatre (260-098)**					
1000	***Morning Tea (30 minutes)***					
	098 Lecture Theatre(260-098)	**OGGB4(260-073)**	**OGGB5(260-051)**	**Case Room 2(260-057)**	**Case Room 3(260-055)**	**Case Room 4(260-009)**
1030	***Promoting Your R Package*** ***Hadley Wickham***	***A Smoothing Filter Modelling Approach For Time Series*** ***Marco Reale***	***Online Learning For Bayesian Nonparametrics: Weakly Conjugate Approximation*** ***Yongdai Kim***	***Improving The Production Cycle At Stats NZ With Rstudio*** ***Gareth Minshall***	***A Max-Type Multivariate Two-Sample Baumgartner Statistic*** ***Hidetoshi Murakami***	***Random Search Global Optimization Using Random Forests*** ***Blair Robertson***
1050	***gridSVG: Then And Now*** ***Paul Murrell***	***Probabilistic Outlier Detection And Visualization Of Smart Meter Data*** ***Rob Hyndman***	***The Joint Models For Nonlinear Longitudinal And Time-To-Event Data Using Penalized Splines: A Bayesian Approach*** ***Thi Thu Huong Pham***	***R – A Powerful Analysis Tool To Improve Official Statistics In Romania*** ***Nicoleta Caragea***	***Simultaneous Test For Mean Vectors And Covariance Matrices In High-Dimensional Settings*** ***Takahiro Nishiyama***	***Dimension Reduction For Classification Of High-Dimensional Data By Stepwise SVM*** ***Elizabeth Chou***
1110	***Bringing Multimix From***	***Specification Of GARCH Model Under Asymmetric Error***	***Performance Of Bayesian Credible Interval For Binomial Proportion Using Logit***	***Statistical Disclosure Control With R: Traditional Methods***	***High Dimensional Asymptotics For The Naive Canonical Correlation***	***Deep Learning High-Dimensional Covariance***

	Fortran To R Murray Jorgensen	Innovations Oyebimpe Adeniji	Transformation Toru Ogura	And Synthetic Data Matthias Templ	Coefficient Mitsuru Tamatani	Matrices Philip Yu

	098 Lecture Theatre(260-098)	OGGB4(260-073)	OGGB5(260-051)	Case Room 2(260-057)	Case Room 3(260-055)	Case Room 4(260-009)
1130	R In Industry – Application On Pipe Renewal Planning Glenn Thomas	Empirical Comparison Of Some Algorithms For Automatic Univariate ARMA Modeling Using Rcmdrplugin.SPSS Dedi Rosadi	Bayesian Optimum Warranty Length Under Type-II Unified Hybrid Censoring Scheme Tanmay Sen	Imputation Of The 2016 Economic Census For Business Activity In Japan Kazumi Wada		Applying Active Learning Procedure To Drug Consumption Data Yuan-Chin Chang
1150	R For Everything Jared Lander	R Package For New Two-Stage Methods In Forecasting Time Series With Multiple Seasonality Shubhabrata Das		Analysis Of Official Microdata Using Secure Statistical Computation System Kiyomi Shirakawa		Presenting Flexi, A Statistical Program For Fitting Variance Models Martin Upsdell
1210	*Lunch (1 hour 10 minutes)*					

1320	Space And Circular Time Log Gaussian Cox Processes With Application To Crime Event Data Alan Gelfand, Duke University 098 Lecture Theatre (260-098)					
	098 Lecture Theatre(260-098)	Case Room 1(260-005)	OGGB5(260-051)	Case Room 2(260-057)	Case Room 3(260-055)	Case Room 4(260-009)
1410	*Cluster-Wise Regression Models Combined By A Quasi-Linear Function* *Kenichi Hayashi*	*Hierarchical Structural Component Analysis Of Gene-Environment Interactions* *Taesung Park*	*Wavelet-Based Power Transformation Of Non-Gaussian Long Memory Time Series* *Kyungduk Ko*	*Cross Covariance Estimation For Integration Of Multi-Omics Data* *Hyungwon Choi*	*Relationships Between Linguistic Characteristics And The Use Of Māori Loanwords In New Zealand English.* *Steven Miller*	*Transfer Regression And Predictive Distributions* *Jinfang Wang*

	098 Lecture Theatre(260-098)	OGGB4(260-073)	OGGB5(260-051)	Case Room 2(260-057)	Case Room 3(260-055)	Case Room 4(260-009)
1430	*An Overview Of The Correspondence Analysis Family* *Eric Beh*	*Testing For Genetic Associations In Arbitrarily Structured*	*Threshold Determination For The Meteorological Data Quality Control In Korea* *Yung-Seop Lee*	*Regularized Noise-Reduction Methodology For High-Dimensional Data* *Kazuyoshi Yata*	*Computation Of Influence Functions For Robust Statistics* *Maheswaran Rohan*	*Adaptive Model Averaging In High-Dimensional Linear Regression* *Tzu-Chang Forrest Cheng*

		Populations *Minsun Song*				
1450	*Model-Based Clustering For Multivariate Categorical Data With Dimension Reduction* *Michio Yamamoto*	*Phylogenetic Tree-Based Microbiome Association Test* *Sungho Won*	*Fitting Additive Hazards Model Using Calibrated Weights For Case-Cohort Data* *Sangwook Kang*	*Selecting The Number Of Principal Components* *Yunjin Choi*	*Rolling Survival Extrapolation Algorithm For Estimating Life Years Lost Among Subjects Exposed To Long-Term Air Pollution* *Jing-Shiang Hwang*	*Enhancing The Flexibility Of Regression Modeling By Liquid Association* *Ker-Chau Li*
1510	*Clusterwise Low-Rank Correlation Analysis Based On Majorization* *Kensuke Tanioka*		*Bayesian Analysis For Fitting Zero-Inflated Count Data With Data Augmentation* *Beomseuk Hwang*	*Towards A Sparse, Scalable, And Stably Positive Definite (Inverse) Covariance Estimator* *Joong-Ho Won*	*Tick-By-Tick Effect On The Inference Of Ultra-High Frequency Data* *Zhi Liu*	*High Mortality Predictions With Lines Or Curves Fitted To Over-Dispersed Exposure-Mortality Data* *John Maindonald*
1530	***Afternoon tea (30 minutes)***					

	098 Lecture Theatre(260-098)	Case Room 4(260-009)	OGGB5(260-051)	Case Room 2(260-057)	Case Room 3(260-055)	Case Room 1(260-005)
1600	***Lattice Polytope Samplers*** ***Martin Hazelton***	***Statistical Modelling And Analysis Of Cosmic Microwave Background Data*** ***Andriy Olenko***	***Measure Of Departure From Marginal Average Point-Symmetry For Two-Way Contingency Tables With Ordered Categories*** ***Kiyotaka Iki***	***Sparse Estimates From Dense Precision Matrix Posteriors*** ***Beatrix Jones***	***Dimension Reduction Strategies For Modeling Bi-Clustered High Dimensional Data*** ***Michael Van Supranes***	***Investigating Methods To Produce Price Indexes From Big Data*** ***Mathew Stansfield***
1620	***Computing Entropies With Nested Sampling*** ***Brendon Brewer***	***Spline-Based Drift Models For High Temperature Operating Life Tests*** ***Thomas Nowak***	***A New Approach To Distribution Free Tests In Contingency Tables*** ***Thuong Nguyen***	***A Bayesian Inference For Time Series Via Copula-Based Markov Chain Models*** ***Li-Hsien Sun***	***Modified Gene Shaving Algorithm - A Dimension Reduction And Clustering Method*** ***Donna Mae Santos***	***The Potential Of Web Scraping*** ***Sam Olivecrona***
1640		***A Simple Method To Construct Confidence Bands In Functional Linear Regression*** ***Masaaki Imaizumi***	***Separation Of Symmetry For Square Contingency Tables With Ordinal Categories*** ***Kouji Tahata***	***Scalable Kernel-Based Variable Selection With Sparsistency*** ***Junhui Wang***	***Testing For Presence Of Clustering Effect In Multilevel Model With High-Dimensional Predictors*** ***Frances Claire San Juan***	***Towards A Big Data CPI For New Zealand*** ***Alan Bentley***

Thursday

910	ALTREP: Alternate Representations Of Basic R Objects Luke Tierney, University of Iowa, USA 098 Lecture Theatre (260-098)				
1000	*Morning Tea (30 minutes)*				
	098 Lecture Theatre(260-098)	**OGGB4(260-073)**	**OGGB5(260-051)**	**Case Room 2(260-057)**	**Case Room 3(260-055)**
1030	*Penalized Vector Generalized Additive Models* *Thomas Yee*	*A Package For Multiple Precision Floating-Point Computation On R* *Junji Nakano*	*Dissimilarities Between Groups Of Data* *Nobuo Shimizu*	*Comparison Of Tests Of Mean Difference In Longitudinal Data Based On Block Resampling Methods* *Hirohito Sakurai*	*SSREM: A Summary-Statistics-Based Random Effect Model To Estimating Heritability, Co-Heritability And Effect Sizes In GWAS Data Analysis* *Jin Liu*
1050	*Consistency Of Linear Mixed-Effects Model Selection With Inconsistent Covariance Parameter Estimators* *Chihhao Chang*	*An Incomplete-Data Fisher Scoring With An Acceleration Method* *Keiji Takai*	*Interactive Visualization Of Aggregated Symbolic Data* *Yoshikazu Yamamoto*	*Analysis Of Spatial Data With A Gaussian Mixture Markov Random Field Model* *Wataru Sakamoto*	*Forward Selection In Regression Models Based On Robust Estimation* *Shan Luo*
1110	*Selecting Generalised Linear Models Under Inequality Constraints* *Daniel Gerhard*	*Improvement Of Computation For Nonlinear Multivariate Methods* *Yuichi Mori*			*Feature Selection In High-Dimensional Models With Complex Block Structures* *Zehua Chen*
1130	*Statistical Generalized Derivative Applied To The Profile Likelihood*				

	Estimation In A Mixture Of Semiparametric Models ***Yuichi Hirose***				
1200	**Closing Ceremony** **098 Lecture Theatre (260-098)**				
1220	***Lunch (1 hour 10 minutes)***				

Am I A Session Chair? Please Read!!!

Monday

Time	Monday 11th					
910	Wing Kam Fung					
	098 Lecture Theatre(260-098)	OGGB4(260-073)	OGGB5(260-051)	Case Room 2(260-057)	Case Room 3(260-055)	Case Room 4(260-009)
1030	*Han-Ming Wu*	*Ciprian Giurcaneanu*	*David Scott*	*Erniel Barrios*	*Jin Liu*	*Junji Nakano*
1320	James Curran					
	098 Lecture Theatre(260-098)	OGGB4(260-073)	OGGB5(260-051)	Case Room 2(260-057)	Case Room 3(260-055)	Case Room 4(260-009)
1410	*Han-Ming Wu*	*Thomas Yee*	*Rolf Turner*	*Priya Parmar*	*Ray-Bing Chen*	*Maxine Pfannkuch*

	098 Lecture Theatre(260-098)	OGGB4(260-073)	OGGB5(260-051)	Case Room 2(260-057)	Case Room 3(260-055)	Case Room 4(260-009)
1600	***Geoff Jones***	***Alan Lee***	***Thomas Lumley***	***Benoit Liquet***	***Hiroyuki Minami***	***Matt Parry***

Tuesday

Time	Tuesday 12th					
910	**Paul Murrell**					
	098 Lecture Theatre(260-098)	**OGGB4(260-073)**	**OGGB5(260-051)**	**Case Room 2(260-057)**	**Case Room 3(260-055)**	**Case Room 4(260-009)**
1030	***Xun Xiao***	***Andrew Balemi***	***Irene Zeng***	***Marie Fitch***	***Jaeyong Lee***	***Catherine Laffineur***
1320	**Chris Triggs**					
	098 Lecture Theatre(260-098)	**OGGB4(260-073)**	**OGGB5(260-051)**	**Case Room 2(260-057)**	**Case Room 3(260-055)**	**Case Room 4(260-009)**
1600	***Hsin-Cheng Huang***	***Russell Millar***	***Philip L.H. Yu***	***Tianying Chu***	***Rose Sisk***	

Wednesday

Time	Wednesday 13th					
	098 Lecture Theatre(260-098)	**OGGB4(260-073)**	**OGGB5(260-051)**	**Case Room 2(260-057)**	**Case Room 3(260-055)**	**Case Room 4(260-009)**
1030	***Paul Murrell***	***Marco Reale***	***Toru Ogura***	***Kazumi Wada***	***Hidetoshi Murakami***	***Blair Robertson***
1320	**Renate Meyer**					
	098 Lecture Theatre(260-098)	**Case Room 1(260-005)**	**OGGB5(260-051)**	**Case Room 2(260-057)**	**Case Room 3(260-055)**	**Case Room 4(260-009)**
1410	***Hiroshi Yadohisa***	***Donguk Kim***	***Chul Eung Kim***	***Johan Lim***	***Steven Miller***	***Jinfang Wang***
	098 Lecture Theatre(260-098)	**Case Room 4(260-009)**	**OGGB5(260-051)**	**Case Room 2(260-057)**	**Case Room 3(260-055)**	**Case Room 1(260-005)**
1600	***Martin Hazelton***	***Andriy Olenko***	***Thuong Nguyen***	***Beatrix Jones***	***Erniel Barrios***	***Alan Bentley***

Thursday

Time	Thursday 14th					
910	**Thomas Yee**					
	098 Lecture Theatre(260-098)	**OGGB4(260-073)**	**OGGB5(260-051)**	**Case Room 2(260-057)**	**Case Room 3(260-055)**	**Case Room 4(260-009)**
1030	***Thomas Yee***	***Yuichi Mori***	***Yoshikazu Yamamoto***	***Masahiro Mizuta***	***Zehua Chen***	

Index By First Name of Author

Ajay Jasra: 088
Alain Vandal: 127
Alan Bentley: 205, 206
Alan Branford: 004
Alan Gelfand: 197
Alastair Lamont: 135
Alexis Comber: 067
Alice Theadom: 085
Amir Bashir: 123
Andrea Havron: 138
Andreea Calude: 137
Andrew Adiguna Halim: 002
Andrew Butcher: 111
Andrew Copas: 193
Andrew Shelling: 085
Andriy Olenko: 160
Angela Nalica: 035
Anjali Gupta: 116
Ankit Patel: 091
Anna Fergusson: 039
Anne Johnson: 193
Anthony Nardone: 193
Antoniade Ciprian Alexandru: 053
Anupama Lakshmanan: 143
Arindam Basu: 002
Balaram Panda: 100, 157
Barry McDonald: 078
Beatrix Jones: 123
Ben Stevenson: 061
Benjamin Montmartin: 025
Benoit Liquet: 019, 092, 140
Beomseuk Hwang: 183
Biswabrata Pradhan: 014
Blair Robertson: 062
Bob Erens: 193
Bogdan Dumitrescu: 038
Brendon Brewer: 080, 126
Brendon McArdle: 198

Bridget Armstrong: 124
Byungyup Kang: 003
Can Yang: 099, 122, 125, 148, 151, 161
Catherine Laffineur: 044
Catherine Mercer: 193
Chae Young Lim: 169
Chaitanya Joshi: 132
Chanatda Somchit: 133
Chi-Hsiang Chu: 049
Chi-Wei Lai: 046
Chihhao Chang: 106
Chika Saka: 149
Ching-Kang Ing: 026
Chloe Robinson: 193
Chris Hansen: 050
Chris Price: 062
Chris Wild: 133
Christopher Triggs: 038, 116
Chul Eung Kim: 178
Ci-Ren Jiang: 184
Ciprian Giurcaneanu: 038
Cord Heuer: 115
Dae-Heung Jang: 182
Dai Ikarashi: 112
Daisuke Miyamoto: 149
Daniel Gerhard: 041, 048
Darfiana Nur: 004, 020
David Borchers: 061
David Friggens: 084
Dedi Rosadi: 059
Dianne Cook: 012
Donna Mae Santos: 104
Earo Wang: 012
Ei-Ji Nakama: 069
Elena Moltchanova: 041
Elizabeth Chou: 075
Elizabeth Fuller: 193
Emma Campbell: 091
Eric Beh: 136
Erniel Barrios: 030, 035, 036, 065, 104, 107
Etsuo Hamada: 139
Fangayo Li: 038

Feifei Chen: 015
Frances Claire San Juan: 107
Frederick Kin Hing Phoa: 187, 190
Fumio Ishioka: 033
Gareth Minshall: 050
Geoff Jones: 017, 115
George Tseng: 180
Glenn Thomas: 141
Granville Tunnicliffe Wilson: 158
Guan-Hua Huang: 155
Guebin Choi: 130
Habib Baluwala: 100
Hadley Wickham: 009
Hang Xu: 072
Hee-Kyung Kim: 179
Hee-Seok Oh: 130
Heng Peng: 151
Hidemine Sekino: 098
Hidetoshi Murakami: 063
Hidetoshi Shimodaira: 147
Hiroe Tsubaki: 098
Hirohito Sakurai: 042
Hiroka Hamada: 128, 190
Hiromi Koh: 166
Hiroshi Yadohisa: 058, 079, 093, 094, 101
Hiroyuki Minami: 087, 095
Hokeun Sun: 010
Hon Hwang: 142
Hongbin Guo: 052
Horst Lewitschnig: 164
Hsin-Cheng Huang: 046
Hyoju Chung: 003
Hyungwon Choi: 166
Hyuntae Kyung: 181
Ingmar Visser: 175
Inkyung Jung: 144
Ivy Liu: 165
Jae Eun Lee: 182
Jaeyong Lee: 037, 174
James Berger: 175
James Curran: 116
Jared Lander: 194

Jennifer Mindell: 193
Jenny Bryan: 200
Jerome Saracco: 092
Jin Liu: 099, 122, 148, 151, 161
Jinfang Wang: 005
Jing-Shiang Hwang: 083
Jingsi Ming: 122, 148, 151
Johan Lim: 166
Johannes Leitner: 164
John Haywood: 158
John Holmes: 191
John Maindonald: 154
Jonathan Marshall: 113
Joong-Ho Won: 172
Joseph Ryan Lansangan: 030, 065, 068, 104, 107
Jun Kawahara: 033
Jun Tsuchida: 079
Jung Jin Lee: 177
Junji Nakano: 043, 069, 089, 159, 190
Kanta Naito: 108
Katsuhiro Omae: 056
Kaye Wellings: 193
Kazeeem Adepoju: 007
Kazumi Wada: 098
Kazunori Yamaguchi: 204
Kazuyoshi Yata: 167
Ke Wan: 074
Keiji Takai: 051
Keisuke Honda: 128, 190
Ken Dodds: 029
Ken Nittono: 150
Kengo Kato: 196
Kenichi Hayashi: 056
Kensuke Tanioka: 058, 073, 074
Ker-Chau Li: 153
Kien Tran: 156
Kiyomi Shirakawa: 112
Kiyotaka Iki: 090
Koji Chida: 112
Koji Kurihara: 033
Kotaro Ohashi: 204
Kouji Tahata: 129

Kouji Yamamoto: 103
Kuhwan Jeong: 003
Kun Yang: 074
Kyungduk Ko: 178
Li-Hsien Sun: 202
Liang-Ching Lin: 022
Lingzhu Li: 189
Linlin Niu: 081
Lisa Hall: 146
Livia Lin Hsuan Chang: 187
Lixing Zhu: 189
Longcheen Huwang: 016
Losana Vao Latu Latu: 170
Louise Ryan: 142
Lu-Hung Chen: 184, 185
Luis Pericchi: 175
Luke Tierney: 199
M. J. Bayarri: 175
Maheswaran Rohan: 117
Mako Yamayoshi: 094
Makoto Aoshima: 167
Makoto Tomita: 040
Marco Reale: 062, 158
Marcos Herrera: 025
Maria Bellringer: 118
Martin Hazelton: 008, 113, 115
Martin Upsdell: 011
Masaaki Imaizumi: 196
Masaaki Taguri: 042
Masahiro Kuroda: 057
Masahiro Mizuta: 087, 095
Masashi Hyodo: 071
Masaya Iizuka: 057
Masayuki Jimichi: 149
Mathew Stansfield: 206, 207
Matt Parry: 134
Matthew Schofield: 029, 060
Matthew Sutton: 019, 140
Matthias Templ: 054
Max Abbott: 118
Maxine Pfannkuch: 039
Mayer Alvo: 072

Md. Shahidul Islam: 109
Meihui Guo: 022
Michael Black: 029
Michael Slyuzberg: 111
Michael Van Supranes: 068
Michiko Watanabe: 204
Michio Yamamoto: 145
Min-Chia Huang: 185
Mingwei Dai: 122, 148, 151, 161
Mingxuan Cai: 122, 148, 151
Minsun Song: 173
Mitsuru Tamatani: 108
Mohammad Sazzad Mosharrof: 109
Mong-Na Lo Huang: 022, 049
Mugdha Manda: 024
Murray Jorgensen: 066
Myungjin Hyun: 179
Nadine Massard: 025
Narumasa Tsutsumida: 067
Nick Garrett: 118
Nicoleta Caragea: 053
Nigel French: 113
Nobuo Shimizu: 043, 089
Olarenwaju Shittu: 007
Oliver Stevenson: 080
Oyebimpe Adeniji: 007
Pam Sonnenberg: 193
Paolo Redondo: 035
Paolo Victor Redondo: 036
Patrick Graham: 162
Paul Bracewell: 091
Paul Brown: 132
Paul Harris: 067
Paul Murrell: 032
Philip L.H. Yu: 072
Philip Prah: 193
Philip Yu: 096
Pierre Lafaye de Micheaux: 019
Priya Parmar: 085
Qing Jiang: 015
Rachel Craig: 193
Ray Bing Chen: 022

Ray-Bing Chen: 049
Raymond Kirk: 002
Richard Barker: 060
Richard Tawiah: 102
Ritwik Bhattacharya: 014
Rob Hyndman: 012, 188
Rob Kydd: 085
Robert Borotkanics: 192
Rodolphe Thiebaut: 140
Rory Ellis: 041
Rose Sisk: 127
Ross Ihaka: 198
Russell Millar: 138
Ryo Kikuchi: 112
Ryo Takagi: 095
Sally Coulson: 116
Sam Olivecrona: 206, 207
Samin Aref: 084
Sangwook Kang: 181
Sarah Pirikahu: 115
Sarah Woodhall: 193
Satoru Hiwa: 058
Satoshi Goto: 101
Satoshi Takahashi: 112
Satoshi Tanaka: 112
Seigo Mizutani: 093
Seongil Jo: 174
Seungyeoun Lee: 171
Shan Luo: 120, 121
Shaun Hendy: 084
Shengli Tzeng: 076
Shengwei Hu: 021
Shigetoshi Hosaka: 005
Shih-Feng Huang: 027, 049
Shih-Hao Huang: 034
Shin Sato: 020
Shinpei Imori: 147
Shinto Eguchi: 056
Shu-Kay Ng: 102
Shubhabrata Das: 143
Shuhei Mano: 037
Sih-Jing Liao: 113

Simon Urbanek: 201
Soazig Clifton: 193
Sojin Ahn: 182
Stephen Joe: 132
Stephen Villejo: 035
Stephen Wright: 142
Steven Miller: 137
Sungho Won: 186
Sunghwan Kim: 180
Sungkyoung Choi: 171
Surajit Ray: 175
Susan Wells: 024
Suzanne Barker-Collo: 085
Suzanne Chambers: 102
Syuichi Nagata: 149
Taerim Lee: 177
Taeryon Choi: 176
Taesung Park: 171
Takahiro Ido: 147
Takahiro Nishiyama: 071
Takemi Yanagimoto: 055, 064
Takuma Ishihara: 103
Tanmay Sen: 014
Thi Thu Huong Pham: 004
Thomas Lumley: 024, 045, 077
Thomas Nowak: 163, 164
Thomas Yee: 031, 133
Thuong Nguyen: 114
Timothy Bilton: 029
Tom Elliott: 077
Tomokazu Fujino: 128
Tomonari Sei: 047
Tomoyuki Hiroyasu: 058
Toru Ogura: 064
Toshinari Kamakura: 055
Toshio Shimokawa: 073, 074
Tso-Jung Yen: 203
Tsuey-Hwa Hu: 083
Tzu-Chang Forrest Cheng: 026
Tzu-Wei Ko: 075
Ushio Tanaka: 047
Valery Feigin: 085

Vera Hofer: 163, 164
Victor Miranda: 031
Virgelio Alao: 065
Vladimir Malabanan: 030
Wataru Sakamoto: 097
Wei-Cheng Hsiao: 026
Wendy Macdowall: 193
Wing Kam Fung: 018
Woncheol Jang: 175
Xiang Wan: 122, 148
Xiangjie Xue: 023
Xin-Hua Wang: 185
Xudong Huang: 045
Xun Xiao: 028
Yaohua Tang: 096
Yaxian Xu: 088
Yogesh Mani Tripathi: 014
Yong Wang: 021, 023, 052
Yongdai Kim: 003
Yoshikazu Yamamoto: 043, 089
Youngin Jo: 174
Youngseon Lee: 037
Yu Yang: 060
Yu-Chung Wei: 155
Yuan-Chin Chang: 105
Yuchi Matsuoka: 139
Yuichi Hirose: 165
Yuichi Mori: 057
Yuji Mizukami: 159
Yukako Toko: 098
Yung-Seop Lee: 179
Yunjin Choi: 168
Yuta Kamiya: 055
Zehua Chen: 120, 121
Zhen Chen: 183
Zhenghui Feng: 015
Zhi Liu: 195
Zhiwu Hong: 081
Zongben Xu: 122

Index by Submission Number

Sub.	Title
002	Effect Of Area Level Deprivation On Body Mass Index: Analysis Of NZ Health Surveys
003	Online Learning For Bayesian Nonparametrics: Weakly Conjugate Approximation
004	The Joint Models For Nonlinear Longitudinal And Time-To-Event Data Using Penalized Splines: A Bayesian Approach
005	Transfer Regression And Predictive Distributions
007	Specification Of GARCH Model Under Asymmetric Error Innovations
008	Lattice Polytope Samplers
009	Promoting Your R Package
010	Incorporating Genetic Networks Into Case-Control Association Studies With High-Dimensional DNA Methylation Data
011	Presenting Flexi, A Statistical Program For Fitting Variance Models
012	Calendar-Based Graphics For Visualising People's Daily Schedules
014	Bayesian Optimum Warranty Length Under Type-II Unified Hybrid Censoring Scheme

015	Adaptive Model Checking For Functional Single-Index Models
016	An EWMA Chart For Monitoring Covariance Matrix Based On Dissimilarity Index
017	Analysis Of Multivariate Binary Longitudinal Data: Metabolic Syndrome During Menopausal Transition
018	Test For Genomic Imprinting Effects On The X Chromosome
019	A Unified Regularized Group PLS Algorithm Scalable To Big Data
020	Bayesian Analyses Of Non-Homogeneous Gaussian Hidden Markov Models
021	Clustering Using Nonparametric Mixtures And Mode Identification
022	Robust Principal Expectile Component Analysis
023	Estimation Of A High-Dimensional Covariance Matrix
024	Identifying Clusters Of Patients With Diabetes Using A Markov Birth-Death Process
025	R&D Policy Regimes In France: New Evidence From A Spatio-Temporal Analysis
026	Adaptive Model Averaging In High-Dimensional Linear Regression
027	Fluctuation Reduction Of Value-At-Risk Estimation And Its Applications
028	Spatial Surveillance With Scan Statistics By Controlling The False Discovery Rate

029	Genetic Map Estimation Using Hidden Markov Models In The Presence Of Partially Observed Information
030	Estimation Of A Semiparametric Spatiotemporal Models With Mixed Frequency
031	Vector Generalized Linear Time Series Models
032	gridSVG: Then And Now
033	Evaluation Of Spatial Cluster Detection Method Based On All Geographical Linkage Patterns
034	On Optimal Group Testing Designs: Prevalence Estimation, Cost Considerations, And Dilution Effects
035	Evidence Of Climate Change From Nonparametric Change-Point Analysis
036	Nonparametric Test For Volatility In Clustered Multiple Time Series
037	Bayesian Curve Fitting For Discontinuous Function Using Overcomplete Representation With Multiple Kernels
038	Variable Selection Algorithms
039	Towards An Informal Test For Goodness-Of-Fit
040	Genetic Approach And Statistical Approach For Association Study On DNA Data
041	The Use Of Bayesian Networks In Grape Yield Prediction
042	Comparison Of Tests Of Mean Difference In Longitudinal Data Based On Block Resampling Methods

043	Dissimilarities Between Groups Of Data
044	The Uncomfortable Entrepreneurs: Bad Working Conditions And Entrepreneurial Commitment
045	Mixed Models For Complex Survey Data
046	Intensity Estimation Of Spatial Point Processes Based On Area-Aggregated Data
047	How Does The Textile Set Describe Geometric Structures Of Data?
048	Selecting Generalised Linear Models Under Inequality Constraints
049	Bayesian Structure Selection For Vector Autoregression Model
050	Improving The Production Cycle At Stats NZ With RStudio
051	An Incomplete-Data Fisher Scoring With An Acceleration Method
052	Smooth Nonparametric Regression Under Shape Restrictions
053	R – A Powerful Analysis Tool To Improve Official Statistics In Romania
054	Statistical Disclosure Control With R: Traditional Methods And Synthetic Data
055	Innovative Bayesian Estimation In The von Mises Distribution
056	Cluster-Wise Regression Models Combined By A Quasi-Linear Function
057	Improvement Of Computation For Nonlinear Multivariate Methods

058	Clusterwise Low-Rank Correlation Analysis Based On Majorization
059	Empirical Comparison Of Some Algorithms For Automatic Univariate ARMA Modeling Using RcmdrPlugin.SPSS
060	Robustness Of Temperature Reconstruction For Past 500 Years
061	Estimation Of Animal Density From Acoustic Detections
062	Random Search Global Optimization Using Random Forests
063	A Max-Type Multivariate Two-Sample Baumgartner Statistic
064	Performance Of Bayesian Credible Interval For Binomial Proportion Using Logit Transformation
065	Semiparametric Mixed Analysis Of Covariance Model
066	Bringing Multimix From Fortran To R
067	Geographically Weighted Principal Component Analysis For Spatio-Temporal Statistical Dataset
068	Dimension Reduction Strategies For Modeling Bi-Clustered High Dimensional Data
069	A Package For Multiple Precision Floating-Point Computation On R
071	Simultaneous Test For Mean Vectors And Covariance Matrices In High-Dimensional Settings
072	Detecting Change-Points In The Stress-Strength Reliability P(X<Y)

073	Two Stage Approach To Data-Driven Subgroup Identification In Clinical Trials
074	Adjusted Adaptive Index Model For Binary Response
075	Dimension Reduction For Classification Of High-Dimensional Data By Stepwise SVM
076	A Simple Method For Grouping Patients Based On Historical Doses
077	Real-Time Transit Network Modelling For Improved Arrival Time Predictions
078	Ranking Potential Shoplifters In Real Time
079	Canonical Covariance Analysis For Mixed Numerical And Categorical Three-Way Three-Mode Data
080	Bayesian Survival Analysis Of Batsmen In Test Cricket
081	Flight To Relative Safety: Learning From A No-Arbitrage Network Of Yield Curves Model Of The Euro Area
083	Rolling Survival Extrapolation Algorithm For Estimating Life Years Lost Among Subjects Exposed To Long-Term Air Pollution
084	Analysing Scientific Collaborations Of New Zealand Institutions Using Scopus Bibliometric Data
085	Genetic Predictors Underlying Long-Term Cognitive Recovery Following Mild Traumatic Brain Injury
087	Symbolic Data Analytical Approach To Unauthorized-Access Logs

088	Bayesian Static Parameter Inference For Partially Observed Stochastic Systems
089	Interactive Visualization Of Aggregated Symbolic Data
090	Measure Of Departure From Marginal Average Point-Symmetry For Two-Way Contingency Tables With Ordered Categories
091	Optimizing Junior Rugby Weight Limits
092	BIG-SIR A Sliced Inverse Regression Approach For Massive Data
093	Local Canonical Correlation Analysis For Multimodal Labeled Data
094	Estimating Causal Structures For Continuous And Discrete Variables
095	Meta-Analysis With Symbolic Data Analysis And Its Application For Clinical Data
096	Deep Learning High-Dimensional Covariance Matrices
097	Analysis Of Spatial Data With A Gaussian Mixture Markov Random Field Model
098	Imputation Of The 2016 Economic Census For Business Activity In Japan
099	SSREM: A Summary-Statistics-Based Random Effect Model To Estimating Heritability, Co-Heritability And Effect Sizes In GWAS Data Analysis
100	A Practitioners Guide To Deep Learning For Predictive Analytics On Structured Data
101	Pattern Prediction For Time Series Data With Change Points

102	Regression With Random Effects For Analysing Correlated Survival Data: Application To Disease Recurrences
103	Comparison Of Exact And Approximate Testing Procedures In Clinical Trials With Multiple Binary Endpoints
104	Modified Gene Shaving Algorithm - A Dimension REduction And Clustering Method
105	Applying Active Learning Procedure To Drug Consumption Data
106	Consistency Of Linear Mixed-Effects Model Selection With Inconsistent Covariance Parameter Estimators
107	Testing For Presence Of Clustering Effect In Multilevel Model With High-Dimensional Predictors
108	High Dimensional Asymptotics For The Naive Canonical Correlation Coefficient
109	Factors Influencing On Growth Of Garments Industry In Bangladesh
111	New Zealand Crime And Victims Survey: Filling The Knowledge Gap
112	Analysis Of Official Microdata Using Secure Statistical Computation System
113	Statistical Models For The Source Attribution Of Zoonotic Diseases: A Study Of Campylobacteriosis
114	A New Approach To Distribution Free Tests In Contingency Tables
115	Bayesian Inference For Population Attributable Measures

116	Dimensionality Reduction Of Multivariate Data For Bayesian Analysis
117	Computation Of Influence Functions For Robust Statistics
118	Analysis Of A Brief Telephone Intervention For Problem Gambling And Examining The Impact On Co-Existing Depression?
120	Forward Selection In Regression Models Based On Robust Estimation
121	Feature Selection In High-Dimensional Models With Complex Block Structures
122	IGESS: A Statistical Approach To Integrating Individual Level Genotype Data And Summary Statistics In Genome Wide Association Studies
123	Sparse Estimates From Dense Precision Matrix Posteriors
124	Whitebait In All Its Varieties: One Fish, Two Fish, Three, Four, Five Fish.
125	Adaptive False Discovery Rate Regression With Application In Integrative Analysis Of Large-Scale Genomic Data
126	Computing Entropies With Nested Sampling
127	Missing Data In Randomised Control Trials: Stepped Multiple Imputation
128	Author Name Identification For Evaluating Research Performance Of Institutes
129	Separation Of Symmetry For Square Contingency Tables With Ordinal Categories
130	Elastic-Band Transform: A New Approach To Multiscale Visualization

132	Bayesian Continuous Space-Time Model Of Burglaries
133	Penalized Vector Generalized Additive Models
134	Scoring Rules For Prediction And Classification Challenges
135	Breeding Value Estimation In Partially-Genotyped Populations
136	An Overview Of The Correspondence Analysis Family
137	Relationships Between Linguistic Characteristics And The Use Of Māori Loanwords In New Zealand English.
138	Including Covariate Estimation Error When Predicting Species Distributions: A Simulation Exercise Using Template Model Builder
139	Nonparametric Causal Inference By The Kernel Method
140	Sparse Group-Subgroup Partial Least Squares With Application To Genomic Data
141	R In Industry – Application On Pipe Renewal Planning
142	Covariate Discretisation On Big Data
143	R Package For New Two-Stage Methods In Forecasting Time Series With Multiple Seasonality
144	Spatial Scan Statistics For Matched Case-Control Data
145	Model-Based Clustering For Multivariate Categorical Data With Dimension Reduction

146	Talk Data To Me
147	An Information Criterion For Prediction With Auxiliary Variables Under Covariate Shift
148	LSMM: A Statistical Approach To Integrating Functional Annotations With Genome-Wide Association Studies
149	Visualization And Statistical Modeling Of Financial Big Data
150	Modeling Of Document Abstraction Using Association Rule Based Characterization
151	BIVAS: A Scalable Bayesian Method For Bi-Level Variable Selection
153	Enhancing The Flexibility Of Regression Modeling By Liquid Association
154	High Mortality Predictions With Lines Or Curves Fitted To Over-Dispersed Exposure-Mortality Data
155	A Computational Tool For Detecting Copy Number Variations From Whole Genome And Targeted Exome Sequencing
156	Modelling The Distribution Of Lifetime Using Compound Time-Homogenous Poisson Process
157	An Advanced Approach For Time Series Forecasting Using Deep Learning
158	A Smoothing Filter Modelling Approach For Time Series
159	Structure Of Members In The Organization To Induce Innovation: Quantitatively Analyze The Capability Of The Organization

160	Statistical Modelling And Analysis Of Cosmic Microwave Background Data
161	Joint Analysis Of Individual Level Genotype Data And Summary Statistics By Leveraging Pleiotropy
162	Adjusting For Linkage Bias In The Analysis Of Record-Linked Data
163	Spline-Based Drift Models For High Temperature Operating Life Tests
164	Tolerance Limits For The Reliability Of Semiconductor Devices Using Longitudinal Data
165	Statistical Generalized Derivative Applied To The Profile Likelihood Estimation In A Mixture Of Semiparametric Models
166	Cross Covariance Estimation For Integration Of Multi-Omics Data
167	Regularized Noise-Reduction Methodology For High-Dimensional Data
168	Selecting The Number Of Principal Components
169	Clustering Of Curves On A Spatial Domain Using A Bayesian Partitioning Model
170	Family Structure And Academic Achievements Of High School Students In Tonga
171	Hierarchical Structural Component Analysis Of Gene-Environment Interactions
172	Towards A Sparse, Scalable, And Stably Positive Definite (Inverse) Covariance Estimator
173	Testing For Genetic Associations In Arbitrarily Structured Populations

174	Bayesian Temporal Density Estimation Using Autoregressive Species Sampling Models
175	Prior-Based Bayesian Information Criterion
176	Bayesian Semiparametric Hierarchical Models For Longitudinal Data Analysis With Application To Dose-Response Studies
177	Mobile Learning In Teaching Bioinformatics For Medical Doctors
178	Wavelet-Based Power Transformation Of Non-Gaussian Long Memory Time Series
179	Threshold Determination For The Meteorological Data Quality Control In Korea
180	Meta-Analytic Principal Component Analysis In Integrative Omics Application
181	Fitting Additive Hazards Model Using Calibrated Weights For Case-Cohort Data
182	Three-Dimensional Data Visualization Education With Virtual Reality
183	Bayesian Analysis For Fitting Zero-Inflated Count Data With Data Augmentation
184	Inverse Regression For Multivariate Functional Data
185	Multiple Function-On-Function Linear Regression With Application To Weather Forecast Calibration
186	Phylogenetic Tree-Based Microbiome Association Test
187	A Study Of The Influence Of Articles In The Large-Scale Citation Network

188	Probabilistic Outlier Detection And Visualization Of Smart Meter Data
189	Model-Checking For Regressions: A Local Smoothing-Based Global Smoothing Test
190	Clustering Of Research Subject Based On Stochastic Block Model
191	Latent Variable Models And Multivariate Binomial Data
192	My Knee Still Hurts; The Statistical Pathway To The Development Of A Clinical Decision Aid
193	Asking About Sex In General Health Surveys: Comparing The Methods And Findings Of The 2010 Health Survey For England With Those Of The Third National Survey Of Sexual Attitudes And Lifestyles
194	R For Everything
195	Tick-By-Tick Effect On The Inference Of Ultra-High Frequency Data
196	A Simple Method To Construct Confidence Bands In Functional Linear Regression
197	Space And Circular Time Log Gaussian Cox Processes With Application To Crime Event Data
198	Could Do Better … A Report Card For Statistical Computing
199	ALTREP: Alternate Representations Of Basic R Objects
200	Zen And The aRt Of Workflow Maintenance
201	R In Times Of Growing User Base And Data Sizes

202	A Bayesian Inference For Time Series Via Copula-Based Markov Chain Models
203	Estimating Links Of A Network From Time To Event Data
204	E-Learning Courses On Introductory Statistics Using Interactive Educational Tools
205	Towards A Big Data CPI For New Zealand
206	Investigating Methods To Produce Price Indexes From Big Data
207	The Potential Of Web Scraping

Programme And Abstracts For Monday 11th Of December

Keynote: Monday 11th 9:10 098 Lecture Theatre (260-098)

R In Times Of Growing User Base And Data Sizes

Simon Urbanek
AT&T Labs

Abstract: R has been historically used mainly on single machines, the analyst performing both analysis and visualization locally. However, the flexible abstraction of graphics in R and its extensibility makes R a great tool to be used remotely and across large clusters. The sizes of datasets as well as the popularity of R have created a demand for extending R's capabilities beyond single machine. In this talk we will illustrate how R can be used by many users in a collaborative open-source RCloud environment to share data analyses, visualizations and results openly. The design also allows scaling across many instances. At the same time this environment can be combined with distributed computing to scale not only with the number of users but also with the size of datasets. In the second part of the talk we will show several approaches how R can be used very efficiently for Big Data analytics at scale leveraging the Hadoop ecosystem. We will start with hmr - a faster way to use the map/reduce framework from R, introduce ROctopus which allows us to perform arbitrary operations on large data without the constraints of a map/reduce framework and show a general framework for developing and using models in R that can leverage distributed systems. We will illustrate the use of the approaches on real dataset and a large cluster.

Monday 11th 10:30 098 Lecture Theatre (260-098)

Robust Principal Expectile Component Analysis

Liang-Ching Lin[1], Ray Bing Chen[1], Mong-Na Lo Huang[2], and Meihui Guo[2]
[1]National Cheng Kung University
[2]National Sun Yat-sen University

Abstract: Principal component analysis (PCA) is widely used in dimensionality reduction for high-dimensional data. It identifies principal components by sequentially maximizing the component score variance around the mean. However, in many applications, one is interested in capturing the tail variations of the data rather than variation around the center. To capture the tail characteristics, Tran et al. (2016), based on an asymmetric L2L_2 norm, proposed principle expectile components (PECs). In this study, we introduce a new method called Huber-type principal expectile component (HPEC) using an asymmetric Huber norm to produce robust PECs. The statistical properties of HPECs are derived, and a derivative free optimization approach, particle swarm optimization (PSO), is used to find HPECs. As a demonstration, HPEC is applied to real and simulated data with encouraging results.

Keywords: asymmetric norm, expectile, Huber’s criterion, particle swarm optimization, principle component

References:

Tran, N. M., Burdejová, P., Osipenko, M. and Hárdle, W. K. (2016). *Principal Component Analysis in an Asymmetric Norm.* SFB 649 Discussion Paper 2016-040, Sonderforschungsbereich 649, Humboldt Universitát zu Berlin, Germany.

Monday 11th 10:30 OGGB4 (260-073)

Effect Of Area Level Deprivation On Body Mass Index: Analysis Of NZ Health Surveys

Andrew Adiguna Halim, Arindam Basu, and Raymond Kirk
Unversity of Canterbury

Abstract: Obesity is a growing public health problem in New Zealand but the trends of its determinants are unclear. We obtained the confidentialised unit record files (CURF) of the New Zealand Health Surveys (NZHS) from the Statistics New Zealand, containing multiple sets of anonymised individual level data from 2002/03 to 2014/15. We assessed the association between deprivation quintile and compliance with the dietary guideline, and the prevalence of overweight/obesity. For adults, we converted Body Mass Index (BMI) variable into tertiles. Then we regressed the BMI tertiles on deprivation level, ethnicity, age, sex, physical activity, education, smoking status, fruit guideline, vegetable guideline, and household income variables using stepwise ordinal logistic regression with complex survey design. We regressed the BMI categories on deprivation level, ethnicity, age, sex, household income, education, fruit guideline, vegetable guideline, soft drink consumption, and fast food consumption in the child data. We found that people living in the highest deprivation quintile were more likely to be in the higher BMI tertile in adults and BMI category in children compared with those living in the lowest deprivation quintile after adjusting for other confounding variables. For adults and children the ORs (95% CI) were 1.349 (95% CI: 1.240-1.468, p<0.001) and 1.803 (95% CI: 1.531-2.125, p<0.001) respectively. In contrast, the ORs (95% CI) for meeting the fruit and vegetable guidelines in adults were 0.968 (95% CI: 0.933-1.005, p: 0.088) and 1.029 (95%CI: 0.988-1.072, p: 0.172) respectively. The ORs (95% CI) for meeting the fruit and vegetable guidelines in children were 0.931 (95% CI: 0.843-1.029, p: 0.164) and 0.994 (95% CI: 0.908-1.088, p: 0.893) respectively. These results suggest that deprivation independently influences BMI, and the effect of meeting dietary guidelines are confounded by deprivation.

Keywords:

obesity, BMI, dietary guideline, deprivation, r statistics, proportional odds regression, survey complex design

Monday 11th 10:30 OGGB5 (260-051)

Calendar-Based Graphics For Visualising People's Daily Schedules

Earo Wang, Dianne Cook, and Rob Hyndman
Monash University

Abstract: This paper describes a frame_calendar function that organises and displays temporal data, collected on sub-daily resolution, into a calendar layout. Calendars are broadly used in society to display temporal information, and events. The frame_calendar uses linear algebra on the date variable to create the layout. It utilises the grammar of graphics to create the plots inside each cell, and thus synchronises neatly with **ggplot2** graphics. The motivating application is studying pedestrian behaviour in Melbourne, Australia, based on counts which are captured at hourly intervals by sensors scattered around the city. Faceting by the usual features such as day and month, was insufficient to examine the behaviour. Making displays on a monthly calendar format helps to understand pedestrian patterns relative to events such as work days, weekends, holidays, and special events. The layout algorithm has several format options and variations. It is implemented in the R package **sugrrants**.

Keywords: data visualisation, statistical graphics, time series, R package, grammar of graphics

References:

Van Wijk JJ, Van Selow ER (1999). Cluster and Calendar Based Visualization of Time Series Data. In *Information Visualization, 1999.(Info Vis' 99) Proceedings*. 4–9.

Wickham H (2009). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, New York, NY.

Wickham H, Hofmann H, Wickham C, Cook D (2012). Glyph-maps for Visually Exploring Temporal Patterns in Climate Data and Models. *Environmetrics*, **23**(5), 382–393.

Monday 11th 10:30 Case Room 2 (260-057)

Nonparametric Test For Volatility In Clustered Multiple Time Series

Paolo Victor Redondo and Erniel Barrios
University of the Philippines Diliman

Abstract: We proposed a test for volatility in clustered multiple time series based on sieve bootstrap. Clustering of observations is intended to capture contagion effect in multiple time series data, assumed to be present in the data generating process where the test is based from. We designed a simulation study to evaluate the test procedure. The method is further illustrated using data on global stock prices and rice production among Asian countries. The test is potentially robust to some distributional assumption but is possibly affected by the nature of volatility.

Keywords:

multiple time series; volatility; nonparametric test; Sieve Bootstrap

IGESS: A Statistical Approach To Integrating Individual Level Genotype Data And Summary Statistics In Genome Wide Association Studies

Mingwei Dai[1], Jingsi Ming[2], Mingxuan Cai[2], Jin Liu[3], Can Yang[4], Xiang Wan[2], and Zongben Xu[1]
[1]Xi'an Jiaotong University
[2]Hong Kong Baptist University
[3]Duke-NUS Medical School
[4]Hong Kong University of Science and Technology

Abstract: Recent genome-wide association studies (GWAS) suggests that a complex phenotype is often affected by many variants with small effects, known as "polygenicity". Tens of thousands of samples are often required to ensure statistical power of identifying these variants with small effects. In this study, we propose a statistical approach, IGESS, to increasing statistical power of identifying risk variants and improving accuracy of risk prediction by integrating individual level genotype data and summary statistics. An efficient algorithm based on variational inference is developed to handle genome-wide-scale analysis. Through comprehensive simulation studies, we demonstrated the advantages of IGESS over the methods which take either individual level data or summary statistics data as input. We applied IGESS to perform integrative analysis of Crohn's Disease from WTCCC and summary statistics from other studies. IGESS was able to significantly increase statistical power of identifying risk variants and improve risk prediction accuracy.

Keywords:

GWAS, functional annotations, variational inference

Monday 11th 10:30 Case Room 4 (260-009)

Author Name Identification For Evaluating Research Performance Of Institutes

Tomokazu Fujino[1], Keisuke Honda[2], and Hiroka Hamada[2]
[1]Fukuoka Women's University
[2]Institute of Statistical Mathematics

Abstract: In this paper, we propose a new framework to extract a complete list of the articles written by researchers who belong to a specific research or educational institute from an academic document database such as Web of Science and Scopus. In this framework, it is necessary to perform author name identification because the query for the database is based on the author's name to extract documents written before the author comming to the current institute. The framework is based on the latent dirichlet allocation (LDA), which is a kind of topic modeling, and some techniques and indices such as synonym retrieval and inverse document frequency (IDF) are used for enhancing the framework.

Keywords: Institutional Research, Topic Modeling, Latent Dirichlet Allocation

References:

Tang, L. and Walsh, J. P. (2010). Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics*, 84(3), 763–784.

Strotmann, A., Zhao, D. and Bubela, T. (2009). Author name disambiguation for collaboration network analysis and visualization. *Proc. American Society for Information Science and Technology*, 46(1), 1–20.

Soler, J. M. (2007). Separating the articles of authors with the same name. *Scientometrics*, 72(2), 281–290.

Monday 11th 10:50 098 Lecture Theatre (260-098)

A Computational Tool For Detecting Copy Number Variations From Whole Genome And Targeted Exome Sequencing

Yu-Chung Wei[1] and Guan-Hua Huang[2]
[1]Feng Chia University
[2]National Chiao Tung University

Abstract: Copy number variations (CNVs) are genomic structural mutations with abnormal gene fragment copies. Current CNV detection algorithms for next generation sequencing (NGS) are developed for specific genome targets, including whole genome sequencing and targeted exome sequencing based on the differently data types and corresponding assumptions. Many whole genome tools assume the continuity of search space and reads uniform coverage across the genome. However, these assumptions break down in the exome capture because of discontinuous segments and exome specific functional biases. In order to develop a method adapting to both data types, we specify the large unconsidered genomic fragments as gaps to preserve the truly location information. A Bayesian hierarchical model was built and an efficient reversible jump Markov chain Monte Carlo inference algorithm was utilized to incorporate the gap information. The performance of gap settings for the Bayesian procedure was evaluated and compared with competing approaches using both simulations and real data.

Keywords:

Bayesian inference, Bioinformatics, copy number variation, next generation sequencing, reversible jump Markov chain Monte Carlo

Clustering Using Nonparametric Mixtures And Mode Identification

Shengwei Hu and Yong Wang
University of Auckland

Abstract: Clustering aims to partition a set of observations into a proper number of clusters with similar objects allocated to the same group. Current partitioning methods mainly include those based on some measure of distance or probability distribution. Here we propose a mode-based clustering methodology motivated via density estimation and mode identification procedures. The idea is to estimate the data-generating probability distribution using a nonparametric density estimator and then locate the modes of the density obtained. In the nonparametric mixture models, each mode and the observations ascend to it correspond to a single cluster. Thus, the problem of determining the number of clusters can be recast as a mode merging problem. A criterion of measuring the separability between modes is also addressed in this work. The most similar modes would be merged sequentially until the optimal number of clusters is reached. The performance of the proposed method is investigated on both simulated and real datasets.

Keywords: Clustering, Nonparametric mixtures, Mode identification

References:

Wang, X. and Wang, Y.: *Nonparametric multivariate density estimation using mixtures*. Stat. Comput. **25**, 349–-364 (2015).

Li, J., Ray S. and Lindsay B.G.: *A nonparametric statistical approach to clustering via mode identification*. Journal of Machine Learning Research. **8**, 1687–-1723 (2007).

Bayesian Curve Fitting For Discontinuous Function Using Overcomplete Representation With Multiple Kernels

Youngseon Lee[1], Shuhei Mano[2], and Jaeyong Lee[1]
[1]Seoul National University
[2]Institute of Statistical Mathematics

Abstract: We propose a new Bayesian methodology for estimating discontinuous functions. In this model, the estimated function is expressed by the overcomplete representation with multiple kernels. Therefore, the complex shape of functions can be expressed by the much smaller number of parameters due to the nature of the sparseness. It does not need any assumptions about the location of discontinuities, the smoothness of the function, the number of features. The form of the function taking all of these into account is determined naturally by the random Levy measure. Simulation data and real data analysis show that this model is suitable for fitting discontinuous functions. We also proved theoretical properties about the support of the function space having jumps in this paper.

Keywords: Bayesian, nonparametric regression, discontinuous curve fitting, overcomplete, multiple kernel, Levy random field

References:

Chu, J. H., Clyde, M. A., and Liang, F. (2009). Bayesian function estimation using continuous wavelet dictionaries, *Statistica Sinica*, 1419–1438

Clyde, M. A., and Wolpert, R. L. (2007). Nonparametric function estimation using overcomplete dictionaries, *Bayesian Statistics*, **8**, 91–114.

Green, Peter J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, **82(4)**, 711–732.

Khinchine, Alexander Ya and Lévy, Paul (1936). Sur les lois stables, *CR Acad. Sci. Paris*, **202**, 374–376.

Müller, P., and Quintana, F. A. (2004). Nonparametric Bayesian data analysis, *Statistical science*, 95–110

Pillai, N. S., Wu, Q., Liang, F., Mukherjee, S., and Wolpert, R. L. (2007). Characterizing the function space for Bayesian kernel models, *Journal of Machine Learning Research*, **8**, 1769–1797.

Qiu, Peihua (2011). *Jump Regression Analysis*. Springer.

Wolpert, R. L., Clyde, M. A., and Tu, C. (2011). Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels, The *Annals of Statistics*, 1916–1962.

Estimation Of A Semiparametric Spatiotemporal Models With Mixed Frequency

Vladimir Malabanan, Erniel Barrios, and Joseph Ryan Lansangan
University of the Philippines Diliman

Abstract: A semiparametric spatiotemporal model is postulated with data measured at varying frequency. The model optimizes utilization of information from variables measured at higher frequency by estimating its nonparametric effect on the response through the backfitting algorithm. Simulation studies support the optimality of the model over simple generalized additive model with aggregation of high frequency data. The method is then used in analyzing the spatiotemporal dynamics of corn yield based on some remotely-sensed data as covariates.

Keywords:

spatiotemporal model, semiparametric model, backfitting, mixed frequency

Monday 11th 10:50 Case Room 3 (260-055)

LSMM: A Statistical Approach To Integrating Functional Annotations With Genome-Wide Association Studies

Jingsi Ming[1], Mingwei Dai[2], Mingxuan Cai[1], Xiang Wan[1], Jin Liu[3], and Can Yang[4]
[1]Hong Kong Baptist University
[2]Xi'an Jiaotong University
[3]Duke-NUS Medical School
[4]Hong Kong University of Science and Technology

Abstract: Thousands of risk variants underlying complex phenotypes have been identified in genome-wide association studies (GWAS). However, there are two major challenges towards fully characterizing the biological basis of complex diseases. First, many complex traits are suggested to be highly polygenic, whereas a large proportion of risk variants with small effects remains unknown. Second, the functional roles of the majority of GWAS hits in the non-coding region is largely unclear. In this paper, we propose a latent sparse mixed model (LSMM) to address the challenges by integrating functional annotations with summary statistics from GWAS. An efficient variational expectation-maximization (EM) algorithm is developed. We conducted comprehensive simulation studies and then applied it to 30 GWAS of complex phenotypes integrating 9 genic annotation categories and 127 tissue-specific functional annotations from the Roadmap project. The results demonstrate that LSMM is not only able to increase the statistical power to identify risk variants, but also provide a deeper understanding of genetic architecture of complex traits by detecting relevant functional annotations.

Keywords:

GWAS, functional annotations, variational inference

A Study Of The Influence Of Articles In The Large-Scale Citation Network

Frederick Kin Hing Phoa[1] and Livia Lin Hsuan Chang[2]
[1]Academia Sinica
[2]Institute of Statistical Mathematics

Abstract:

Nowadays there are many research metrics at the author-, article-, journal-levels, like the impact factors and many others. However, none of them possess a universally meaningful interpretation on the research influence at all levels, not mentioning that many are subject-biased and consider neighboring relations only. In this work, we introduce a new network-based research metric called the network influence. It utilizes all information in the whole network and it is universal to any levels. Due to its statistical origin, this metric is computationally efficient and statistically interpretable even if one applies it to a large-scale network. This work demonstrates the analysis of networks via network influence using a large-scale citation database called the Web of Science. By just considering the articles among statistics community in 2005-2014, the network influence of all articles are calculated and compared, resulting in a top-ten important articles that are slightly different from the list via impact factors. This metric can be easily extended to author citation network and many similar networks embedded in the Web of Science.

Estimating Links Of A Network From Time To Event Data

Tso-Jung Yen
Academia Sinica

Abstract:

In this paper we develop a statistical method for identifying links of a network from time to event data. This method models the hazard function of a node conditional on event time of other nodes, parameterizing the conditional hazard function with the links of the network. It then estimates the hazard function by maximizing a pseudo partial likelihood function with parameters subject to a user-specified penalty function and additional constraints. To make such estimation robust, it adopts a prespecified risk control on the number of false discovered links by using the Stability Selection method. Simulation study shows that under this hybrid procedure, the number of false discovered links is tightly controlled while the true links are well recovered. We apply our method to estimate a political cohesion network that drives donation behavior of 146 firms from the data collected during the 2008 Taiwanese legislative election. The results show that firms affiliated with elite organizations or firms of monopoly are more likely to diffuse donation behavior. In contrast, firms belonging to technology industry are more likely to act independently on donation.

Monday 11th 11:10 OGGB4 (260-073)

Estimation Of A High-Dimensional Covariance Matrix

Xiangjie Xue and Yong Wang
University of Auckland

Abstract: The estimation of covariance or precision (inverse covariance) matrices plays a prominent role in multivariate analysis. The usual estimator, the sample covariance matrix, is known to be unstable and ill-conditioned in high-dimensional setting. In the past two decades, various methods have been developed to give a stable and well-conditioned estimator and they have their own advantages and disadvantages. We will review some of the most popular methods and describe a new method to estimate the correlation matrix and hence the covariance matrix using the empirical Bayes method. Similar to many element-wise methods in the literature, we also assume that the elements in a correlation matrix are independent of each other. We use the fact that the elements in a sample correlation matrix can be approximated by the same one-parameter normal distribution with unknown means , along with the non-parametric maximum likelihood estimation to give a new estimator of the correlation matrix. Preliminary simulation results show that the new estimator has some advantages over various thresholding methods in estimating sparse covariance matrices.

Keywords: Big Data, Multivariate Analysis, Statistical Inference

References:

Efron, B., 2010. *Correlated zz-values and the accuracy of large-scale statistical estimates*. J Am Stat Assoc **105**, 1042 - 1055.

Fan, J., Liao, Y., Liu, H., 2016. *An overview of the estimation of large covariance and precision matrices*. Econometrics Journal **19**, C1 - C32.

Wang, Y., 2007. *On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution*. Journal of the Royal Statistical Society: Series B **69**, 185 - 198.

Monday 11th 11:10 OGGB5 (260-051)

Innovative Bayesian Estimation In The von Mises Distribution

Yuta Kamiya[1], Toshinari Kamakura[1], and Takemi Yanagimoto[2]
[1]Chuo University
[2]Institute of Statistical Mathematics

Abstract: In spite of recent growing interest in applying the von-Mises distribution to circular data in various scientific fields, researches on the parameter estimation are surprisingly sparse. The standard estimators are the MLE and the maximum marginal likelihood estimator (Schou 1978). Although Bayesian estimators are promising, it looks that they have not been fully developed. We propose the posterior mean of the canonical parameter, instead of the mean parameter, under the reference prior. This estimator satisfies an optimality property, and performs favorably for wide ranges of true parameters. Extensive simulation studies yield that the risks of the proposed estimator are significantly small, compared with the existing estimators. An interesting finding is that the estimating function for the dispersion parameter behaves reasonably. Notable advantages of the present approach are its straightforward extensions to various procedures, including Bayesian estimator under an informative prior based on the reference prior. The proposed estimator is examined by applying to practical datasets.

Keywords: von-Mises distribution, bayesian estimation, canonical parameter

References:

Fisher, Nicholas I. *Statistical analysis of circular data.* Cambridge University Press, 1995.

Schou, Geert. “Estimation of the concentration parameter in von Mises–Fisher distributions.” Biometrika 65.2 (1978): 369-377.

Monday 11th 11:10 Case Room 2 (260-057)

Evidence Of Climate Change From Nonparametric Change-Point Analysis

Angela Nalica, Paolo Redondo, Erniel Barrios, and Stephen Villejo
University of the Philippines Diliman

Abstract: Suppose that the time series data is sufficiently explained by a model, e.g., autoregressive model, transfer function model. A change-point is considered to exist if any of the model parameters is substantially different in two or more regimes. We proposed a test for existence of a change-point (assuming that location of the change is known) based on nonparametric bootstrap. The method is used in verifying whether the southern oscillation index exhibits change-point which is taken as an evidence of climate change. There is indeed an evidence of climate change in the period.

Keywords:

change-point analysis, block bootstrap, southern oscillation index (SOI)

Joint Analysis Of Individual Level Genotype Data And Summary Statistics By Leveraging Pleiotropy

Mingwei Dai[1], Jin Liu[2], and Can Yang[3]
[1]Xi'an Jiaotong University
[2]Duke-NUS Medical School
[3]Hong Kong University of Science and Technology

Abstract: Results from Genome-wide association studies (GWAS) suggest that a complex phenotype is often affected by many variants with small effects, known as "polygenicity". Tens of thousands of samples are often required to ensure statistical power of identifying these variants with small effects. However, it is often the case that a research group can only get approval for the access to individual-level genotype data with a limited sample size (e.g., a few hundreds or thousands). Meanwhile, pleiotropy is a pervasive phenomenon in genetics whereby a DNA variant influences multiple traits, and summary statistics for genetically related traits (e.g., autoimmune diseases or psychiatric disorders) are becoming publicly available. The sample sizes associated with the summary statistics data sets are usually quite large. How to make the most efficient use of existing abundant data resources largely remains an open problem.

In this study, we propose a statistical approach, LEP, to increasing statistical power of identifying risk variants and improving accuracy of risk prediction by integrating individual level genotype data and summary statistics by veraging leiotropy. An efficient algorithm based on variational inference is developed to handle the genome-wide analysis. Through comprehensive simulation studies, we demonstrated the advantages of LEP over the methods which take either individual-level data or summary statistics data as input. We applied LEP to perform integrative analysis of several auto-immune diseases from WTCCC and summary statistics from other studies. LEP was able to significantly increase the statistical power of identifying risk variants and improve the risk prediction accuracy by jointly analyzing autoimmune diseases.

Keywords: GWAS, pleiotropy, polygenicity, summary statistics, variational inference

References:

Solovieff N, Cotsapas C, Lee P H, et al. (2013) Pleiotropy in complex traits: challenges and strategies In: *Nature reviews. Genetics* 14(7): 483.

Carbonetto P, Stephens M. (2012) Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies In: *Bayesian analysis* 7(1): 73-108.

Chung D, Yang C, Li C, et al. (2014). GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation In: *PLoS genetics*

Dai M, Ming J, Cai M, et al. (2017). IGESS: a statistical approach to integrating individual-level genotype data and summary statistics in genome-wide association studies. In: *Bioinformatics*

Monday 11th 11:10 Case Room 4 (260-009)

An Advanced Approach For Time Series Forecasting Using Deep Learning

Balaram Panda
Inland Revenue Department

Abstract: Time series forecasting is a decade-long research and which is being evolving day by day. Due to the recent advancement is deep learning technique many of the complex problems have been solved using deep learning. Deep learning techniques have shown tremendous better performance in supervised learning problem. One of the reasons for this success is the ability of deep feedforward network methods to learn multiple feature interaction for a single instance. However, the time-dependent nature not being captured by deep feedforward network till the evolution of RNN(recurrent neural network) and LSTM(long short term memory) network architecture. This paper reveals the success of LSTM time series in comparison with ARIMA and other standard approaches for time series modeling. A sensitivity analysis is also conducted to explore the effect of hyper parameter tuning on LSTM model to reduce the time series forecasting error. We also derive practical advice from our empirical results for those interested in getting most out of LSTM time series for modern time series forecasting.

Keywords: Deep Learning, Time Series, LSTM, RNN

References:

Längkvist, Martin, Lars Karlsson, and Amy Loutfi. “A review of unsupervised feature learning and deep learning for time-series modeling.” Pattern Recognition Letters 42 (2014): 11-24.

Zheng, Yi, et al. “Time series classification using multi-channels deep convolutional neural networks.” International Conference on Web-Age Information Management. Springer, Cham, 2014.

Monday 11th 11:30 OGGB4 (260-073)

Genetic Map Estimation Using Hidden Markov Models In The Presence Of Partially Observed Information

Timothy Bilton[1,2], Matthew Schofield[1], Ken Dodds[2], and Michael Black[1]
[1]University of Otago
[2]AgResearch

Abstract: A genetic linkage map shows the relative position of and genetic distance between markers, positions of the genome that exhibit variation, and underpins the study of species' genomes in a number of scientific applications. Genetic maps are constructed by tracking the transmission of genetic information from individuals to their offspring, which is frequently modelled using a hidden Markov model (HMM) since only the expression and not the transmission of genetic information is observed. However, constructing genetic maps with data generated using the latest sequencing technology is complicated by the fact that some observations are only partially observed which, if unaccounted for, typically results in inflated estimates. We extend the HMM to model partially observed information by including an additional layer of latent variables. In addition, we investigate several different approaches for computing confidence intervals of the genetic map estimates obtained from the extended HMM. Results show that our model is able to produce accurate genetic map estimates, even in situations where a large proportion of the data is only partially observed. Our methodology has been implemented in the R package GusMap.

Keywords:

hidden Markov models, linkage mapping, partially observed data, confidence intervals

Monday 11th 11:30 OGGB5 (260-051)

A Simple Method For Grouping Patients Based On Historical Doses

Shengli Tzeng
China Medical University

Abstract: Monitoring dose patterns over time helps physicians and patients learn more about metabolic change, disease evolution, etc. One way to turn such longitudinal data into clinically useful information is through cluster analysis, which aims to separate the "profiles of doses" among patients into homogeneous subgroups. Different doses patterns reflect heterogeneity in patients' characteristics and effectiveness of therapy. However, not all patients were prescribed at regular time points, and missing values seems ubiquitous if one aligns records at distinct time points. Moreover, a few outliers may heavily influence the estimation for within and/or between variations of clusters, making the distinction among clusters blurred. In this study, a simple method based on a novel pairwise dissimilarity is proposed, which also serves as a screen tool to detect potential outliers. We use smoothing splines, handling data observed either at regular or irregular time points, and measure the dissimilarity between patients based on pairwise varying curve estimates with commutation of smoothing parameters. It takes into account the estimation uncertainty and is not strongly affected by outliers. The effectiveness of our proposal is shown by simulations comparing it to other dissimilarity measures and by a real application to methadone dosage maintenance levels.

Keywords: Clustering, longitudinal data, smoothing splines, outliers

References:

Lin, Chien-Ju, Christian Hennig, and Chieh-Liang Huang. (2016). Clustering and a dissimilarity measure for methadone dosage time series. In *Analysis of Large and Complex Data*, 31-41. Springer, Switzerland.

Semiparametric Mixed Analysis Of Covariance Model

Virgelio Alao, Erniel Barrios, and Joseph Ryan Lansangan
University of the Philippines Diliman

Abstract: A semiparametric mixed analysis of covariance model is postulated and estimated using the two procedures: based on an imbedded restricted maximum likelihood (REML) and nonparametric regression (smoothing splines) estimation into the backfitting framework (ARMS); and infusing bootstrap into the ARMS (B-ARMS). The heterogeneous effect of covariates across the groups is postulated to affect the response through a nonparametric function to mitigate overparameterization. Using simulation studies, we exhibited the capability of the postulated model (and estimation procedures) in increasing predictive ability and stabilizing variance components estimates even for small sample size and with minimal covariate effect, and regardless of whether the model is correctly specified or there is misspecification error.

Keywords:

mixed ANCOVA model, nonparametric regression, backfitting, bootstrap, random effects, variance components

Monday 11th 11:30 Case Room 3 (260-055)

Adaptive False Discovery Rate Regression With Application In Integrative Analysis Of Large-Scale Genomic Data

Can Yang
Hong Kong University of Science and Technology

Abstract: Recent international projects, such as the Encyclopedia of DNA Elements (ENCODE) project, the Roadmap project and the Genotype-Tissue Expression (GTEx) project, have generated vast amounts of genomic annotation data, e.g., epigenome and transcriptome. There is great demanding of effective statistical approaches to integrate genomic annotations with the results from genome-wide association studies. In this talk, we introduce a statistical framework, named AdaFDR, for integrating multiple annotations to characterize functional roles of genetic variants that underlie human complex phenotypes. For a given phenotype, AdaFDR can adaptively incorporates relevant annotations for prioritization of genetic risk variants, allowing nonlinear effects among these annotations, such as interaction effects between genomic features. Specifically, we assume that the prior probability of a variant associated with the phenotype is a function of its annotations F(X)F(X), where XX is the collection of the annotation status and F(X)F(X) is an ensemble of decision trees, i.e., F(X)=∑kfk(X)F(X) = \sum_k f_k(X) and fk(X)f_k(X) is a shallow decision tree. We have developed an efficient EM-Boosting algorithm for model fitting, where a shallow decision tree grows in a gradient-Boosting manner (Friedman J. 2001) at each EM-iteration. Our framework inherits the nice property of gradient boosted trees: (1) The gradient accent property of the Boosting algorithm naturally guarantees the convergence of our EM-Boosting algorithm. (2) Based on the fitted ensemble F^(X)\hat{F}(X), we are able to rank the importance of annotations, measure the interaction among annotations and visualize the model via partial plots (Friedman J. 2008). Using AdaFDR, we performed integrative analysis of genome-wide association studies on human complex phenotypes and genome-wide annotation resources, e.g., Roadmap epigenome. The analysis results revealed interesting regulatory patterns of risk variants. These findings deepen our understanding of genetic architectures of complex phenotypes. The statistical framework developed here is also broadly applicable to many other areas for integrative analysis of rich data sets.

Keywords: False Discovery Rate, integrative analysis, functional annotation, genomic data

References:

Friedman, Jerome H (2001). Greedy function approximation: a gradient boosting machine, *Annals of statistics*, **29:5**,1189–1232.

Jerome H. Friedman and Bogdan E. Popescu (2008) Predictive Learning via Rule Ensembles *The Annals of Applied Statistics*, **2:3**, 916–954

Monday 11th 11:30 Case Room 4 (260-009)

Structure Of Members In The Organization To Induce Innovation: Quantitatively Analyze The Capability Of The Organization

Yuji Mizukami[1] and Junji Nakano[2]
[1]Nihon University
[2]Institute of Statistical Mathematics

Abstract: Innovation is the act of creating new value by using “new connection”, “new point of view”, “new way of thinking”, “new usage method” (Schumpeter 1912). In recent years, the promotion of the Innovation has been strongly encouraged. In the field of research, attempts are also being made to create new value through connection between those fields. Moreover, along with the move to promote integration among these research fields, research is being conducted to grasp and promote the degree of them. In this research, for the purpose of providing indices for measuring the degree of them, we show indices quantitatively indicating the degree of fusion in different fields and the distance between the fields. Also, we have try to present indices for grasping the whole image based on the random graph.

Keywords: Research Metrix, Institute Research, Co-author analysis

References:

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J. and Börner, K. (2011). *Approaches to understanding and measuring interdisciplinary scientific research: A review of the literature, Journal of Informetrics*. Vol. 5, No. 1, pp. 14-26.

Mizukami, Y., Mizutani, Y., Honda, K., Suzuki, S., Nakano, J. (2017). *An International Research Comparative Study of the Degree of Cooperation between disciplines within mathematics and mathematical sciences, Behaviormetrika*, **1**, 19 pages, On-line.

Monday 11th 11:50 OGGB4 (260-073)

Vector Generalized Linear Time Series Models

Victor Miranda and Thomas Yee
University of Auckland

Abstract: Since the introduction of the ARMA class in the early 1970s many time series (TS) extensions have been proposed, e.g., vector ARMA and GARCH-type models for heteroscedasticity. The result has been a plethora of models having pockets of substructure but little overriding framework. In this talk we propose a class of TS models called Vector Generalized Linear Time Series Models (VGLTSM), which can be thought of as multivariate generalized linear models directed towards time series data. The crucial VGLM ideas are constraint matrices, vector responses and covariate-specific linear predictors, and estimation by iteratively reweighted least squares and Fisher scoring. The only addition to the VGLM framework is a log-likelihood that depends on past values. We show how several popular sub-classes of TS models are accommodated as special cases of VGLMs, as well as new work that broadens TS modelling even more. Algorithmic details of its implementation in , and properties such as stationarity, parameters depending on covariates, expected information matrices and cointegrated TS are surveyed.

Keywords: VGLM, time series, Fisher scoring.

References:

Yee, T. W. (2015) *Vector Generalized Linear and Additive Models: With an Implementation in R.* New York, USA: Springer.

Monday 11th 11:50 OGGB5 (260-051)

Local Canonical Correlation Analysis For Multimodal Labeled Data

Seigo Mizutani and Hiroshi Yadohisa
Doshisha University

Abstract: In supervised learning, canonical correlation analysis (CCA) is widely used for dimension reduction problems. When using dimension reduction methods, researchers should always aim to preserve the data structure in a low dimensional space. However, if the obtained data are assumed to be multimodal labeled data, that is, each cluster can be subdivided into several latent clusters, CCA is rarely able to preserve the data structure in a low dimensional space.

In this study, we propose local CCA (LCCA) for multimodal labeled data. This method is based on local Fisher discriminant analysis (LFDA) (Sugiyama, 2007). We do not employ the same local covariance matrix of the explanatory variables as under LFDA, which uses a local between-group variance matrix and a local within-group variance matrix. Instead, in our proposed method, we use a covariance matrix of the explanatory variables as well as a weighted affinity matrix. The usefulness of LCCA in data visualization and clustering is then demonstrated by simulation studies.

Keywords: Supervised learning, Dimension reduction, Local Fisher discriminant analysis (LFDA), Weighted affinity matrix

References:

Sugiyama, M. (2007). Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *Journal of Machine Learning Research*, **8**, 1027-1061.

Hastie, T. and Buja, A. and Tibshirani, R. (1995) Penalized discriminant analysis., 73-102.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, **28**, 321-377.

Monday 11th 11:50 Case Room 2 (260-057)

A Practitioners Guide To Deep Learning For Predictive Analytics On Structured Data

Balaram Panda and Habib Baluwala
Inland Revenue Department

Abstract: Recently, deep learning techniques have shown remarkably strong performance in problems involving unstructured data (ex. text, image, and video). One of the reasons for this success is the ability of deep learning methods to learn multiple levels of abstraction and feature interaction. However, the advantages of using deep learning techniques for structured/ event/transactional data has not been studied in detail. The purpose of this paper is to review the advantages and limitations of using deep feed forward networks on structured data. This is achieved by comparing the performance of deep feed forward networks with conventional machine learning techniques applied on a large structured dataset for classification problem. The paper also describes methodologies for optimizing the deep feed forward networks to achieve better accuracy and different approaches to reduce over fitting for deep feed forward network. A sensitivity analysis is conducted to explore the effect of hyper parameter tuning on model performance. We also derive practical advice from our extensive empirical results for those interested in getting most out of deep feed forward networks for real world settings.

Keywords: Deep Learning, deep feed forward networks, machine learning, R, Tensorflow, Python

References:

Bengio, Yoshua. “Learning deep architectures for AI.” Foundations and trends® in Machine Learning 2.1 (2009): 1-127.

Goodfellow, Ian J., et al. “Maxout networks.” arXiv preprint arXiv:1302.4389 (2013).

Monday 11th 11:50 Case Room 4 (260-009)

Clustering Of Research Subject Based On Stochastic Block Model

Hiroka Hamada[1], Keisuke Honda[1], Frederick Kin Hing Phoa[2], and Junji Nakano[1]
[1]Institute of Statistical Mathematics
[2]Academia Sinica

Abstract: In this paper, we propose a new clustering method to measure influence of papers in all areas of science. To see structure of entire relationship we apply stochastic block model (SBM) on big scale citation network data. SBM generates a matrix which divides several blocks which represent relationship among research fields. We show this matrix can be used to visual exploratory analysis. When lists of papers are mapped this matrix we can get useful information by varied locations in visually. Elastic Map is used as dimension reduction method to calculate scalar value onto onto the corresponding principal points of each papers. We demonstrate that this projection score is can be used to evaluate divergence impact of papers across all field. To illustrate one application of our method, we analyze 450k+ articles published between 1981 and 2016 Web of Science data. In this beta version of our system, Edward, probabilistic programming language is used for estimation of SBM parameters and calculation of divergence score of papers.

Keywords: Institutional Research, Stochastic Block Model, Elastic Map

References:

Nowicki,K. and Snijders,T. (2001). Estimation and prediction for stochastic block structures. *Journal of the American Statistical Association*, 96, 1077–1087.

Gorban,A. and Zinovyev,A. (2005). Elastic Principal Graphs and Manifolds and their Practical Applications. *Computing*, 75(4), 359–379.

Tran,D., Kucukelbir,A., Dieng, A.B., Rudolph,M., Liang,D. and Blei,D.M. (2016). Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*.

Keynote: Monday 11th 13:20 098 Lecture Theatre (260-098)

Zen And The aRt Of Workflow Maintenance

Jenny Bryan
University of British Columbia

Abstract:

My talk deals with two related themes: the ongoing discussion of "data science vs. statistics" and the importance of developing your own data analysis workflow. These topics are related in my mind because I believe, as academic statisticians, we have an unnecessarily narrow view of our discipline. The "data science vs statistics" debate brings this to a head, because the desire to join and even lead data science initiatives provides an incentive to broaden our mandate. What if we embraced the development and teaching of tools -- both mental and digital -- that address the entire data analysis process? We'll conclude with something very concrete: a tour of semi-recent developments in the R ecosystem, aimed at people who want to make their workflow more productive and less aggravating.

Monday 11th 14:10 098 Lecture Theatre (260-098)

Canonical Covariance Analysis For Mixed Numerical And Categorical Three-Way Three-Mode Data

Jun Tsuchida and Hiroshi Yadohisa
Doshisha University

Abstract: Three-mode three-way data (objects ×\times variable ×\times conditions) have been observed in many areas of research. For example, panel data often include values for the same objects and variables at different times. Given two three-mode three-way data sets, we often investigate two types of factors: common factors, which show the relationships between the two data sets, and unique factors, which represent the uniqueness of each data set. In light of this, canonical covariance analysis has been proposed. However, these datasets often have numerical and categorical variables simultaneously. Many multivariate methods for two three-mode thee-way data sets assume that the data has numerical variables only. To overcome this problem, we propose three-mode three-way canonical covariance analysis with numerical and categorical variables. We use an optimal scaling method (for example, Yong (1987)) for the quantification of categorical data because the values of a categorical variable could not be compared with the value of a numerical variable.

Keywords: Alternative least squares, Dimensional reduction, Optimal scaling, Quantification method

References:

Young, F. W. (1981). Quantitative analysis of qualitative data. *Psychometrika*, **46**, pp. 357–388

Variable Selection Algorithms

Fangayo Li[1], Christopher Triggs[1], Bogdan Dumitrescu[2], and Ciprian Giurcaneanu[1]
[1]University of Auckland
[2]University Politehnica of Bucharest

Abstract: The matching pursuit algorithm (MPA) is an efficient solution for high dimensional variable selection (Bühlmann and van de Geer, 2011). There is, however, no widely accepted stopping rule for MPA. (Li et al., 2017) have given novel stopping rules based on information theoretic criteria (ITC). All of these ITC are based on the degrees of freedom (df) of the hat matrix which maps the data vector to its estimate. We derive some properties of the hat matrix when MPA is used. These allow us to give an upper bound on the possible increase in df between successive MPA iterations. A simulation study with data generated from different models compares the mean integrated square error of the different ITC and cross validation (Sancetta, 2016).

Keywords: Matching pursuit algorithm, degrees of freedom, hat matrix

References:

A.Sancetta (2016). *Greedy algorithms for prediction*. Bernoulli, vol. 22, pp. 1227 - 1277.

P.Bühlmann and S.van de Geer (2011). *Statistics for high-dimensional data. Methods, theory and applications*. Springer Science & Business Media.

F.Li, C.Triggs, B.Dumitrescu, and C.D.Giurcăneanu (2017). *On the number of iterations for the matching pursuit algorithm* . Proceedings of the 25th European Signal Processing Conference (EUSIPCO), pp. 191 - 195. (to appear)

Estimating Causal Structures For Continuous And Discrete Variables

Mako Yamayoshi and Hiroshi Yadohisa
Doshisha University

Abstract: Structural equation models have been used extensively for continuous variable data to find causal structures. In such a framework, the Linear Non- Gaussian Acyclic Model (LiNGAM) could enable finding a whole causal model (Shimizu et al., 2006). However, in many desciplines, the data include both continuous and discrete variables. LiNGAM could fail to capture the actual causal relationship for such data because it handles both discrete and continuous variables as continuous. Therefore, it is necessary to improve the estimation method for causal structures in such conditions.

In this study, we propose a method to find causal structures for continuous and discrete variables. To overcome the problems of the existing method, we use the Link function. Using simulation studies, we show that the proposed method performs more efficiently for data that includes continuous and discrete variables.

Keywords: Causal direction, Latent variables, Link function, SEM, LiNGAM

References:

Barnett, J.A., Payne, R.W. and Yarrow, D. (1990). *Yeasts: Characteristics and identification: Second Edition.* Cambridge: Cambridge University Press.

S. Shimizu, P.O. Hoyer, A. Hyvärinen, and A. Kerminen (2006). A linear non-Gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, vol. 7, pp. 2003-2030.

(ed.) Barnett, V., Payne, R. and Steiner, R. (1995). *Agricultural Sustainability: Economic, Environmental and Statistical Considerations*. Chichester: Wiley.

Payne, R.W. (1997). *Algorithm AS314 Inversion of matrices Statistics*, **46**, 295–298.

Payne, R.W. and Welham, S.J. (1990). A comparison of algorithms for combination of information in generally balanced designs. In: *COMPSTAT90 Proceedings in Computational Statistics*, 297–302. Heidelberg: Physica-Verlag.

Monday 11th 14:10 Case Room 2 (260-057)

Incorporating Genetic Networks Into Case-Control Association Studies With High-Dimensional DNA Methylation Data

Hokeun Sun
Pusan National University

Abstract: In human genetic association studies with high-dimensional microarray data, it has been well known that statistical methods utilizing prior biological network knowledge such as genetic pathways and signaling pathways can outperform other methods that ignore genetic network structures. In recent epigenetic research on case-control association studies, relatively many statistical methods have been proposed to identify cancer-related CpG sites and the corresponding genes from high-dimensional DNA methylation data. However, most of existing methods are not able to utilize genetic networks although methylation levels among linked genes in the networks tend to be highly correlated with each other. In this article, we propose new approach that combines independent component analysis with network-based regularization to identify outcome-related genes for analysis of high-dimensional DNA methylation data. The proposed approach first captures gene-level signals from multiple CpG sites using independent component analysis and then regularizes them to perform gene selection according to given biological network information. In simulation studies, we demonstrated that the proposed approach overwhelms other statistical methods that do not utilize genetic network information in terms of true positive selection. We also applied it to the 450K DNA methylation array data of the four breast invasive carcinoma cancer subtypes from The Cancer Genome Atlas (TCGA) project.

Keywords:

Independent component analysis, network-based regularization, genetic network, DNA methylation, high-dimensional data

Monday 11th 14:10 Case Room 3 (260-055)

Adaptive Model Checking For Functional Single-Index Models

Feifei Chen[1], Qing Jiang[2], and Zhenghui Feng[3]
[1]Renmin University
[2]Beijing Normal University
[3]Xiamen University

Abstract: In this paper, a model-adaptive test statistic is proposed to do model checking for functional single-index models. Dimension reduction methods are included to handle the curse of dimensionality. The test statistic consists of two parts: the first term is a naive one, and the second term is adaptive to the model as if the model were univariate. It is consistent and can detect local alternative at a fast rate. Monte Carlo method is used to find the critical value under null hypothesis. Simulation studies show the performance of our proposed test procedure.

Keywords:

Functional single-index models, dimension reduction, model checking

Mobile Learning In Teaching Bioinformatics For Medical Doctors

Taerim Lee[1] and Jung Jin Lee[2]
[1]Korea National Open University
[2]Soongsil University

Abstract:

This paper describes an implementation of mobile learning initiative in Bioinformatics Training & Education Center (BITEC) for medical doctors supported from Ministry of Welfare and Public Health Korea. This project was initiated by Dept. of Bioinformatics & Statistics KNOU and Dept. of Medical Informatics of SNU Medical College for training medical doctors. The high penetration rates of mobile phone subscriptions and the rapid growing of handheld users show that it is viable for making inroads towards the usage of mobile devices as an alternative learning mode for distance learning. The mobile learning initiative is aimed to encourage learning and interactions in distance learning communities aiming to bridge the transactional distances faced by the learners and adopt mobility as the key tool in Bioinformatics courses delivery. The BITEC m-Learning initiative focuses on introducing Bioinformatics using handheld devices to be made easily accessible through the use of mobile devices for ODL Bioinformatics learners who are very busy medical doctors in ubiquitous learning environment. The m-Learning approach is considered as a learning alternative to support distance learners mainly working doctors and medical researchers in Korea. This research paper discusses the implementation of the mobile e-Book approach which has better affordable, accessible and flexible educational media.

On Optimal Group Testing Designs: Prevalence Estimation, Cost Considerations, And Dilution Effects

Shih-Hao Huang
Academia Sinica

Abstract: Group testing has been used for decades to estimate the prevalence of a rare disease when samples from multiple subjects can be pooled and tested as a group. A group testing design is specified by the support points (distinct group sizes) and their corresponding frequencies. In this series of works, we construct locally optimal approximate designs for group testing with uncertain error rates, where the goal is to maximize the precision of the prevalence estimate. We also provide a guaranteed algorithm based on the approximate theory for constructing exact designs for practical use. Our simulated examples based on a Chlamydia study in the United States show that the proposed design outperforms competing designs, and its performance is quite stable to the working parameters. We then extend the framework to accommodate two features likely to be encountered in real-world studies. We develop optimal budgeted-constrained designs, where both subjects and tests incur costs, and the error rates of the the assay are linked to the group sizes, allowing dilution effects to reduce the test performance. (Work done jointly with M.-N. L. Huang, K. Shedden, and W. K. Wong.)

Keywords: Budget-constrained design, dilution effect, DsD_s-optimality, group testing, sensitivity, specificity

References:

Huang, S.-H., Huang, M.-N. L., Shedden, K. and Wong, W. K. (in press). Optimal group testing designs for estimating prevalence with uncertain testing errors. *Journal of the Royal Statistical Society: Series B.* DOI: 10.1111/rssb.12223.

Huang, S.-H., Huang, M.-N. L. and Shedden, K. (manuscript). Cost considerations for efficient group testing studies.

The Use Of Bayesian Networks In Grape Yield Prediction

Rory Ellis, Daniel Gerhard, and Elena Moltchanova
University of Canterbury

Abstract: The requirement for predictions to be made earlier in the growing season has become more important, as the opportunity to plan for the wine production and export earlier in the season becomes desirable. The issue with this is there is less information available to those wishing to make early predictions. The analysis in this paper implements a double sigmoidal curve to model the grape growth over the growing season, as this is typically used in agriculture.

In order to conduct prediction in this study, a Bayesian Network is considered. This allows the opportunity to consider the knowledge of experts in the field, where grape growers would know the growth behaviour of the grapes, as well as using new data to update the Bayesian Network. This information is then implemented in the form of priors, which involves estimating the parameters of the aforementioned double sigmoidal model. Sensitivity Analysis is done in this research, which looks at the impact of prior assumptions (or lack thereof) from experts. Examinations are also made of the value of adding information to the model, as it can be determined whether the precision in the predictions improves as a result of adding data. The results in this analysis are based off simulation studies.

Monday 11th 14:30 OGGB5 (260-051)

Pattern Prediction For Time Series Data With Change Points

Satoshi Goto and Hiroshi Yadohisa
Doshisha University

Abstract: Recently, there have been various types of time series data, such as daily stock prices and Web-click logs, that have complicated the structure. In several cases, because of the complexity, time series data cannot satisfy the stationary process assumption. REGIMECAST (Matsubara and Sakura, 2016) has been proposed as a method to forecast time series data. It is useful for capturing changes in time series patterns and representing the non-linear system. However, it cannot adequately represent time series data after radical changes. Generally, radical changes in time series data can be detected using existing methods, such as change-point detection and anomaly detection. These methods are rarely used for forecasting time series data, although these data often show different behaviors after radical changes.

In this study, we propose a method that can forecast future time series data after events involving radical changes. The method has two features: i) appropriate pattern discovery, as it recognizes the appropriate learning section with change-point detection, and ii) flexible representation, as it represents non-stationary processes with a non-linear state space model. We also provide empirical examples using a variety of real datasets.

Keywords: anomaly detection, change-point detection, non-linear state space model, pattern discovery, REGIMECAST

References:

Y. Matsubara and Y. Sakurai (2016). Regime shifts in streams: Real-time forecasting of co-evolving time sequences, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13–17, 2016.

Monday 11th 14:30 Case Room 2 (260-057)

Test For Genomic Imprinting Effects On The X Chromosome

Wing Kam Fung
Unversity of Hong Kong

Abstract: Genomic imprinting is an epigenetic phenomenon that the expression of an allele copy depends on its parental origin. This mechanism has been found to play an important role in many diseases. Methods for detecting imprinting effects have been developed primarily for autosomal markers. However, no method is available in the literature to test for imprinting effects on the X chromosome. Therefore, it is necessary to suggest methods for detecting such imprinting effects. In this talk, the parental-asymmetry test on X the chromosome (XPAT) is first developed to test for imprinting for qualitative traits in the presence of association, based on family trios each with both parents and their affected daughter. Then, we propose 1-XPAT to tackle parent-daughter pairs, each with one parent and his/her affected daughter. By simultaneously considering family trios and parent-daughter pairs, C-XPAT is constructed to test for imprinting. Further, we extend the proposed methods to accommodate complete (with both parents) and incomplete (with one parent) nuclear families having multiple daughters of which at least one is affected. Simulations are conducted to assess the performance of the proposed methods under different settings. Simulation results demonstrate that the proposed methods control the size well, irrespective of the inbreeding coefficient in females being zero or nonzero. By incorporating incomplete nuclear families, C-XPAT is more powerful than XPAT using only complete nuclear families. For practical use, these proposed methods are applied to analyze the rheumatoid arthritis data.

Keywords:

Imprinting effects, X chromosome, qualitative traits, nuclear family

Fluctuation Reduction Of Value-At-Risk Estimation And Its Applications

Shih-Feng Huang
National University of Kaohsiung

Abstract: Value-at-Risk (VaR) is a fundamental tool for risk management and is also associated with the capital requirements of banks. Banks need to adjust their capital levels for satisfying the Basel Capital Accord. On the other hand, managements do not like to change the capital levels too often. To achieve a balance, this study proposes an approach to reduce the fluctuation of VaR estimation. The first step is to fit a time series model to the underlying asset returns and obtain the conventional VaR process. A new VaR (NVaR) estimation of the conventional VaR process is then determined by applying change-point detection algorithms and a proposed combination scheme. The capital levels computed from the NVaR process are capable of satisfying the Basel Accord and reducing the fluctuation of capital levels simultaneously. To apply the proposed method to the calculation of future capital requirements, an innovative approach for NVaR prediction is also proposed by incorporating the concept of CUSUM control charts. The return processes of 30 companies on the list of S&\&P 500 from 2005 to 2016 are employed for our empirical investigation. Numerical results indicate that the proposed NVaR prediction is capable of satisfying the Basel Accord and reducing the fluctuation of capital requirements simultaneously by using a comparable average amount of capital requirements to the conventional VaR estimator.

Keywords:

Capital requirement, change point detection, CUSUM control chart, fluctuation reduction, Value-at-Risk

E-Learning Courses On Introductory Statistics Using Interactive Educational Tools

Kazunori Yamaguchi[1], Kotaro Ohashi[1], and Michiko Watanabe[2]
[1]Rikkyo University
[2]Keio University

Abstract:

This paper introduces e-learning courses for principles and methods of introductory statistics, which is developing for undergarduate students. The course consists of the following contents; Usages and linkages to the official statistics in Japan, videos titled statistics for daily life, and interactive learning contents. For this course, we have developed Japanese versions of interactive Java applets for understanding statistical concepts and a tool for the simulation and data analysis. We expect that combination of these tools and e-learning contents not only makes students easy to understand basic concepts of statistics but also motivate students to study statistics.

Estimation Of Animal Density From Acoustic Detections

Ben Stevenson[1] and David Borchers[2]
[1]University of Auckland
[2]University of St Andrews

Abstract: Estimating the density of animal populations is of central importance in ecology, with practical applications that affect decision making in the fields of wildlife management, conservation, and beyond. For species that vocalise, surveys using acoustic detectors such as microphones, hydrophones, or human observers can be vastly cheaper than traditional surveys that physically capture or visually detect animals. In this talk I describe a spatial capture-recapture approach to estimate animal density from acoustic surveys and present a software implementation in the R package ascr, with examples applied to populations of frogs, gibbons, and whales.

Keywords:

Ecological statistics, mark-recapture, point process

Mixed Models For Complex Survey Data

Xudong Huang and Thomas Lumley
University of Auckland

Abstract: I want to fit a mixed model to a population distribution, but I have data from a complex (multistage) sample. The sampling is informative, that is, the model holding for the population is different from the model holding for the (biased) sample. Ignoring the sampling design and just fitting the mixed model to the sample distribution will lead to biased inference. Although both the model and sampling involve "clusters", the model clusters and sample clusters need not be the same. I will use a pairwise composite likelihood method to estimate the parameters of the population model under this setting. In particular, consistency and asymptotic normality can be established. Variance estimation in this problem is challenging. I will talk about a variance estimator and how to show it is consistent.

Keywords: Mixed model, Complex sampling, Pairwise composite likelihood

References:

Yi, G. , Rao, J. and Li, H.(2016). *A weighted composite likelihood approach for analysis of survey data under two-level models.* Statistica Sinica, 2016, 26, 569-587

Monday 11th 14:50 OGGB5 (260-051)

Regression With Random Effects For Analysing Correlated Survival Data: Application To Disease Recurrences

Richard Tawiah, Suzanne Chambers, and Shu-Kay Ng
Griffith University

Abstract: Correlated failure time data arise in many biomedical studies, due to multiple occurrences of the same disease in an individual patient. To account for this correlation phenomenon, we formulate a random effect (frailty) survival model with an autoregressive (AR) covariance structure and adopt the generalized linear mixed model (GLMM) methodology for estimation of regression and variance component parameters. A more general case of the problem is also considered via a multilevel random effect approach where the correlation of survival times is induced by a hierarchical clustering structure, such as the appearances of repeated failures in patients from the same hospital in a multicentre clinical trial setting. Our modelling problem is used to investigate prognostic and treatment effects on disease relapses in two data sets, (1) tumour recurrences in bladder cancer patients and (2) recurrent infections in children with chronic granulomatous disease (CGD). Using the first data set, the effect of treatment thiotepa was found to be insignificant but demonstrated an effect in reducing tumour recurrences with adjusted hazard ratio (AHR) of 0.58 (95% CI: 0.29-1.16, p=0.124). The initial number of tumours (AHR: 1.26, 95% CI: 1.08-1.47, p=0.004) had significant positive effect but the effect of the size of the largest initial tumour was insignificant. In the case of the CGD data, treatment gamma interferon showed a significant decreasing effect (AHR: 0.27, 95% CI: 0.13-0.56, p$<$0.001) on the incidence of recurrent infections. In addition, age effect was significant (AHR: 0.90, 95% CI: 0.81-1.0, p=0.042). Pattern of inheritance, height, weight, sex, use of corticosteroids and prophylactic antibiotics did not exhibit significant association with recurrent infections. The appropriateness of our modelling methodology is investigated in a simulation study. The simulation results show that parameters are satisfactorily estimated in the special case where AR random effect is merely used. However, in the multilevel context bias in the variance parameter of random hospital effect increases as the true magnitude of variation in hospital effects increases.

Keywords: Frailty model, random effect, correlated survival times, recurrent event, GLMM, bladder cancer, CGD

Genetic Predictors Underlying Long-Term Cognitive Recovery Following Mild Traumatic Brain Injury

Priya Parmar[1], Rob Kydd[2], Andrew Shelling[2], Suzanne Barker-Collo[2], Alice Theadom[1], and Valery Feigin[1]
[1]Auckland University of Technology
[2]University of Auckland

Abstract: Traumatic Brain Injury (TBI) is a major cause of death and disability. While moderate and severe forms of TBI develop the most significant impairments even mild TBI may be followed by persisting post-concussion symptoms, neurocognitive problems and mental health disorders such as anxiety. Cognitive impairments can impact on all areas of an individual's work, home and social life and are important to understand and predict overall recovery. These outcomes may be in part be determined by genetic variants that influence the molecular and physiological response of the brain to damage, as well as determining pre-injury reserve and vulnerability to co-morbidities.

A number of studies have examined the relationship between genetic variants and outcomes following TBI. Most have examined groups with moderate to severe injury and are limited by small sample sizes, selection biases, failed to correct for ethnic factors, and have evaluated outcomes at various time points, making comparison between studies difficult.

Using the population-based study of TBI in NZ (BIONIC) we analysed the association between cognitive outcomes with 18 genetic markers (SNPs-single nucleotide polymorphisms) from 12 genes previously studied in relation with TBI; FAAH, GAD1, WWC1, CHMR2, ANKK1, BDNF, NGB, BCL2, APOE, S100B, HMOX1 and COMT in a sample of 183 European and 76 Maori adults. We used the CNS-Vital Signs (computerised neurocognitive test battery) to provide 11 measures of cognitive functioning, memory and attention collected at baseline, 1-, 6-, 12- and 48 months post-injury.

ANCOVA models were used to identify the association between time, SNP (modelled as major, heterozygous and minor alleles) and SNP by time effect for each CNS-Vital Signs outcome. Statistically significant findings were observed in both European and Maori samples for being associated with the same CNS-Vital Signs outcome for rs8191992 (CHMR2), rs4680 (rs4680), rs2071746 (HMOX1) and rs17071145 (WWC1).

A linear mixed effects model was utilised to analyse each individual's natural cognitive recovery trajectory over time. The individuals' age, gender, whether or not this was their first TBI, the severity level of the mild TBI (low, medium or high) and SNP were all included in the model as covariates.

Regression analyses identified the following SNPs to be statistically associated with several CNS-Vital Signs outcomes; rs8191992 (CHMR2) was shown to be associated with attention, neurocognition, composite memory, executive functioning as well as processing and psychomotor speed in Europeans.

Whilst rs3798178 (GAD1) was associated with two domains of attention, neurocognition and three domains of memory (composite, visual and working) in Maori. We found rs3791879 was associated with increased attention and neurocognition in our European sample.

Furthermore, the minor alleles of rs11604671 (ANKK1) were associated with poorer cognitive recovery (compared to those with homozygous major alleles) for two domains of attention, executive functioning, processing speed, social acuity and working memory over time in Maori. We found rs11604671 was associated with reduced executive functioning and processing speed in our European sample

Unlike other genetic studies on TBI patients, our study investigated several different genetic variants in a larger ethnically diverse population sample of individuals with primarily mild TBI. Although our findings agreed with previous literature for genetic associations for cognitive recovery post-injury, for the first time, we were able to identify ethnic differences in specific genetic markers determining specific cognitive outcomes in European and Maori people with TBI. Further large TBI population based cohort studies are warranted to replicate these genetic associations, both locally and globally in order to better understand the differences underlying an individual's outcome trajectory and inform more effective treatment strategies.

Bayesian Structure Selection For Vector Autoregression Model

Chi-Hsiang Chu[1], Mong-Na Lo Huang[1], Shih-Feng Huang[2], and Ray-Bing Chen[3]
[1]National Sun Yat-sen University
[2]National University of Kaohsiung
[3]National Cheng Kung University

Abstract: Vector autoregression (VAR) model is powerful in economic data analysis because it can be used to analyze several different time series data simultaneously. However, in VAR model, we need to deal with the huge coefficient dimensionality and it would be caused some computational problems for coefficient inference. To reduce the dimensionality, we could take some model structures into account based on the prior knowledge. In this paper, several group structures of the coefficient matrices are considered. Due to different types of VAR structures, corresponding MCMC algorithms are proposed to generate posterior samples for making inference of the structure selection. Simulation studies and a real example are used to show the performances of the proposed Bayesian approaches.

Keywords:

Bayesian variable selection, time series, universal grouping, segmentized grouping

Three-Dimensional Data Visualization Education With Virtual Reality

Dae-Heung Jang, Jae Eun Lee, and Sojin Ahn
Pukyong National University

Abstract: A variety of data visualization methods are utilizing to analyze huge amount of data. Among various methods, a three-dimensional image requires the rotation of the image to show stereo image on the two-dimensional screen. This study discusses data visualization education of two methods (static method and dynamic method) which make it possible to analyze the construct of stereo image to improve the restriction of the three-dimensional image display with virtual reality. This investigation can be useful to explore three-dimensional data structure more clearly.

Keywords: Data visualization education, Virtual reality, Stereo image, R package

References:

Bowman, A. (2015). *rpanel*: Simple interactive controls for R using the tcltk library. R package version 1.1-3.

Campos, M. M. (2007). Way Cooler: PCA and Visualization Linear Algebra in the Oracle Database 2, http://oracledmt.blogspot.kr/2007/06/way-cooler-pca-and-visualization-linear.html.

Ligges, U. (2017). *scatterplot3d*: 3D Scatter Plot. R package version 0.3-38.

Murdoch, D. (2017). *rgl*: 3D Visualization Using OpenGL. R package version 0.97.0.

Myers, R. H., Montgomery, D. C. and Anderson-Cook, C. M. (2016). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments, 4th ed*, Wiley, New York.

Ripley, B. (2016). *MASS*: Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-47.

Sarkar, D. (2016). *lattice*: Trellis Graphics for R. R package version 0.20-35.

Soetaert, K. (2016). *plot3D*: Plotting Multi-Dimensional Data. R package version 1.1.

Wolf, H. P. (2015). *aplpack*: Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, plotsummary, plothulls, and some slider functions. R package version 1.3.0.

http://astrostatistics.psu.edu/datasets/SDSS quasar.html.

http://forbes.com/mlb valuations/list.

http://gartner.com/newsroom/id/3412017.

Monday 11th 15:10 098 Lecture Theatre (260-098)

Talk Data To Me

Lisa Hall
Fonterra

Abstract:

For some scientists, seeing a statistician in the corridor is the equivalent of coming face to face with the grim reaper. They fear being beaten over the head with a Stat101 textbook as you try to hammer into them the importance of replication, randomisation or an appropriate sample size. Many scientists would rather blunder through on their own than admit to the statistician that they don't understand the ins-and-outs of ANOVA. So we end up with peer reviewed research with badly interpreted p-values and underwhelming Excel graphs despite the existence of more elegant solutions. As ambassadors of statistical rigour, we are the ones who can turn this around. When the scientists won't come to us, how can we improve data usage short of chasing scientists down corridors with textbooks? I will give examples of how to Talk Data with scientists to encourage their statistical awakening. In the right context, even the most puritanical scientists can become more comfortable Talking Data. They may even find they enjoy it!

Smooth Nonparametric Regression Under Shape Restrictions

Hongbin Guo and Yong Wang
University of Auckland

Abstract: Shape-restricted regression, in particular under isotonicity and convexity(concavity) constraints, has many practical applications. Traditional nonparametric methods to the problem using least squares or maximum likelihood result in discrete step functions or nonsmooth piecewise linear functions, which are unsatisfactory both predictively and visually. In this talk, we describe a new, smooth, nonparametric estimator under the above-mentioned shape restrictions. In particular, the discrete measures that are inherent in the previous estimators are replaced with continuous ones. A new algorithm that can rapidly find the corresponding estimate will also be presented. Numerical studies show that the new estimator outperforms major existing methods in almost all cases.

Keywords: Nonparametric regression, smooth, shape restriction, convex, monotonic

References:

Groeneboom, P., Jongbloed, G. and Wellner, A. (2001). Estimation of a Convex Function: Characterizations and Asymptotic Theory. *Ann. Statist*. **29**(6), 1653–1698.

Wang, Y. (2007). On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **69** (2), 185–198.

Meyer, M.(2008). Inference using shape-restricted regression splines. *Ann. Appl. Stat.* **2**(3), 1013–1033.

Elastic-Band Transform: A New Approach To Multiscale Visualization

Guebin Choi and Hee-Seok Oh
Seoul National University

Abstract: This paper presents a new transformation technique for multiscale visualization of one-dimensional data such as time series and functional data under the concept of the scale-space approach. The proposed method uses a range of regular observations (eye scanning) with varying intervals. The results, termed 'elastic-band transform' can be considered as a collection of observations over different intervals of viewing. It is motivated by a way that human looks at an object such as a sequence of data repeatedly in order to overview a global structure of it as well as find some specific features of it. Some measures based on elastic-bands are discussed for describing characteristics of data, and two-dimensional visualizations induced by the measures are developed for understanding and detecting important structures of data. Furthermore, some statistical applications are studied.

Keywords: Transformation; Visualization; Decomposition; Filter; Time Series

References:

Chaudhuri, P. and Marron, J. S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, **94**, 807–823.

Donoho, D. L., and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425-455.

Dragomiretskiy, K. and Zosso, D. (2014). Variational mode decomposition. *IEEE Transactions on Signal Processing*, **62**, 531–544.

Erästö, P. and Holmström, L. (2005). Bayesian multiscale smoothing for making inferences about features in scatter plots. *Journal of Computational and Graphical Statistics*, **14**, 569–589.

Fryzlewicz, P. and Oh, H.-S. (2011). Thick pen transformation for time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 499–529.

Hannig, J. and Lee, T. C. M. (2006). Robust SiZer for exploration of regression structures and outlier detection. *Journal of Computational and Graphical Statistics*, **15**, 101–117.

Hannig, J., Lee, T. and Park, C. (2013). Metrics for SiZer map comparison. *Stat*, **2**, 49–60.

Holmström, L. (2010a). BSiZer. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**, 526–534.

Holmströma, L. (2010b). Scale space methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**,150–159.

Holmströma, L. and Pasanena, L. (2017). Statistical scale space methods. *International Statistical Review*, **85**, 1–30.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., ... & Liu, H. H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, **454**, 903–995.

Lindeberg, T. (1994). *Scale-Space Theory in Computer Vision*, Springer Science & Business Media, New York.

Park, C., Hannig, J. and Kang, K. H. (2009). Improved SiZer for time series. *Statistica Sinica*, **19**, 1511–1530.

Park, C, Lee, T. C. and Hannig, J. (2010). Multiscale exploratory analysis of regression quantiles using quantile SiZer. *Journal of Computational and Graphical Statistics*, **19**, 497–513.

Rioul, O. and Vetterli, M. (1991). Wavelets and signal processing. *IEEE Signal Processing Magazine*, 8(LCAV-ARTICLE-1991-005), 14–38.

Vogt, M., & Dette, H. (2015). Detecting gradual changes in locally stationary processes. The Annals of Statistics, 43(2), 713-740.

Meta-Analytic Principal Component Analysis In Integrative Omics Application

Sunghwan Kim[1] and George Tseng[2]
[1]Keimyung University
[2]University of Pittsburgh

Abstract: With the prevalent usage of microarray and massively parallel sequencing, numerous high-throughput omics datasets have become available in the public domain. Integrating abundant information among omics datasets is critical to elucidate biological mechanisms. Due to the high- dimensional nature of the data, methods such as principal component analysis (PCA) have been widely applied, aiming at effective dimension reduction and exploratory visualization. In this paper, we combine multiple omics datasets of identical or similar biological hypothesis and introduce two variations of meta-analytic framework of PCA, namely MetaPCA. Regularization is further incorporated to facilitate sparse feature selection in MetaPCA. We apply MetaPCA and sparse MetaPCA to simulations, three transcriptomic meta-analysis studies in yeast cell cycle, prostate cancer, mouse metabolism, and a TCGA pan-cancer methylation study. The result shows improved accuracy, robustness and exploratory visualization of the proposed framework.

Keywords: principal component analysis, meta-analysis, omics data

References:

Flury (1984) *Common principal components in k groups.* Journal of the American Statistical Association, 79, 892–898.

Krzanowski (1979) *Between-groups comparison of principal components.* Journal of the American Statistical Association, 74, 703–707

Flight To Relative Safety: Learning From A No-Arbitrage Network Of Yield Curves Model Of The Euro Area

Zhiwu Hong[1] and Linlin Niu[2]
[1]HKUST Business School
[2]Xiamen University

Abstract:

This paper develops a no-arbitrage network of yield curves model for the euro area to study the joint dynamics of and risk transmission across yield curves of different economies. The model nests 1+M yield curves of a benchmark country and M periphery countries with 3+(2+M) state factors. The benchmark curve is spanned by three yield factors, which are also common basis for all curves. Periphery yield spreads are spanned by three factors, including two common risk factors related to market liquidity risk and common sovereign credit risk, and a country-specific risk factor. Parsimony can be guaranteed as M increases with additional yield curve data, and factors can be strongly identified with structural restrictions under the no-arbitrage conditions. We estimate a 1+5 yield curves model for Germany and GIIPS countries using weekly data from 2009 to 2016. The results show that, the German yields enjoy a 'flight to liquidity' effect under a liquidity shock, which pushes down German yields while driving up periphery spreads. However, in response to a shock of common sovereign credit risk, yields of all countries, including Germany, tend to increase immediately and persistently. The weaker is a country's economic fundamental, the higher its risk exposure to shocks. Though spillover effects among periphery curves are generally positive, when a Greek shock bursts, the Italian risk factor may be temporarily mitigated due to its relative stronger economy. The network model can be adapted with a time-varying parameter VAR to monitor evolving contributions of factors.

Bayesian Analyses Of Non-Homogeneous Gaussian Hidden Markov Models

Shin Sato and Darfiana Nur
Flinders University

Abstract: We investigate a non-homogeneous Gaussian hidden Markov model where the model assumes the transition probabilities between the hidden states depend on each discrete-time. The methodology of the statistical inference for the model follows the Bayesian approach implementing the Markov chain Monte Carlo (MCMC) methods for parameter estimation. The methods include: the Metropolis-Hastings, the delayed rejection Metropolis-Hastings, the multiple-try Metropolis-Hastings, and the adaptive Metropolis algorithms.

For simulation studies, we have successfully implemented all the algorithms proposed on the simulated data set that was investigated by Diebold et al. (1994), although we had been faced with the difficulties of estimating each parameter due to the large noises in the data. For a case study, the model was implemented on a data set of the monthly US 3-month treasury bill rates with six financial exogenous variables in which the settings are identical to that of Meligkotsidou and Dellaportas's (2011), except for the algorithm.

Keywords: Non-homogeneous hidden Markov model, Bayesian inference, Markov chain Monte Carlo methods, Metropolis-Hastings algorithms

References:

Diebold, F.X., Lee, J.-H., and Weinbach, G.C. (1994). Regime switching with time-varying transition probabilities. *Business Cycles: Durations, Dynamics, and Forecasting*, 144–165.

Spezia, L. (2006). Bayesian analysis of non-homogeneous hidden markov models. *Journal of Statistical Computation and Simulation*, **76**(8), 713–725.

Monday 11th 16:00 OGGB4 (260-073)

Robustness Of Temperature Reconstruction For Past 500 Years

Yu Yang, Matthew Schofield, and Richard Barker
University of Otago

Abstract: Temperature reconstruction is vital to studies of climate change. Instrumental records are only available back to 19th century, too short to describe changes that occur over hundreds or thousands of years. Fortunately, nature environmental clues (such as tree rings, pollens and ice cores) can be pieced together to reconstruct unrecorded temperatures. We use tree-ring width to study summer temperature in Northern Sweden for past 500 years. Previous work has shown the predictions to be sensitive to model assumptions. We gain a new insight into this problem by attempting to separately estimate aspects of the process that are robustly estimated. One of these are the years in which the climate is colder or warmer than recent observations. We implement this by considering hidden Markov models on the partially observed temperature series. The model is fitted using Hamiltonian Monte Carlo in Stan.

Keywords: temperature reconstruction, robust estimator, hidden Markov model, Bayesian analysis

References:

Schofield, M. R., Barker, R. J., Gelman, A., Cook, E. R., and Briffa, K. R. (2016). A model-based approach to climate reconstruction using tree-ring data. *Journal of the American Statistical Association*, 111(513), 93-106.

Nonparametric Causal Inference By The Kernel Method

Yuchi Matsuoka and Etsuo Hamada
Osaka University

Abstract: Rubin causal model is a statistical model to estimate the effect of a treatment on the outcome based on the framework of potential outcomes. To estimate a causal effect based on Rubin causal model, propensity score plays a central role. In particular, matching and weighting methods like Inverse Probability Weighted Estimator (IPWE) and Doubly-Robust estimator based on the estimated propensity score are widely used. Despite its popularity, it was pointed out that model misspecification of the propensity score can result in substantial bias of the resulting estimators of a causal effect and potential outcomes. It is possible to estimate propensity score in nonparametric ways or machine learning methods to avoid model misspecification. However, it doesn't work well in most situations due to following reasons: 1) Curse of dimensionality. 2) They only aim at an accuracy of classification and don't optimize the covariate balancing. To overcome the problems above, we propose a new estimator of propensity score using kernel mean embeddings of conditional distributions. Although our proposal is completely nonparametric, our estimator has a dimensionality-independent rate of convergence. Using kernel measures of conditional independence for model selection, our estimator can also correct the bias that arises from the imbalance of covariates. In numerical simulations, we confirm that our method can reduce the bias in misspecified settings. We also describe several asymptotic properties of our estimator.

Keywords:

Rubin causal model, Propensity score, Kernel method, Kernel mean embedding, Hilbert-Schmidt Independence Criterion

Monday 11th 16:00 Case Room 2 (260-057)

A Unified Regularized Group PLS Algorithm Scalable To Big Data

Pierre Lafaye de Micheaux[1], Benoit Liquet[2], and Matthew Sutton[2]

[1]University of New South Wales

[2]Queensland University of Technology

Abstract: Partial Least Squares (PLS) methods have been heavily exploited to analyse the association between two blocs of data. These powerful approaches can be applied to data sets where the number of variables is greater than the number of observations and in presence of high collinearity between variables. Different sparse versions of PLS have been developed to integrate multiple data sets while simultaneously selecting the contributing variables. Sparse modelling is a key factor in obtaining better estimators and identifying associations between multiple data sets. The cornerstone of the sparsity version of PLS methods is the link between the SVD of a matrix (constructed from deflated versions of the original matrices of data) and least squares minimisation in linear regression. We present here an accurate description of the most popular PLS methods, alongside their mathematical proofs. A unified algorithm is proposed to perform all four types of PLS including their regularised versions. Various approaches to decrease the computation time are offered, and we show how the whole procedure can be scalable to big data sets.

Keywords: Big data, High dimensional data, Lasso Penalties, Partial Least Squares, Sparsity, SVD

References:

Lafaye de Micheaux, P., Liquet, B. & Sutton, M. (2017), *A Unified Parallel Algorithm for Regularized Group PLS Scalable to Big Data*, ArXiv e-prints .

Liquet, B., Lafaye de Micheaux, P., Hejblum, B. & Thiebaut, R. (2016), *Group and sparse group partial least square approaches applied in genomics context*, Bioinformatics 32, 35-42.

Monday 11th 16:00 Case Room 3 (260-055)

Evaluation Of Spatial Cluster Detection Method Based On All Geographical Linkage Patterns

Fumio Ishioka[1], Jun Kawahara[2], and Koji Kurihara[1]
[1]Okayama University
[2]Nara Institute of Science and Technology

Abstract: Currently, it is becoming easier to analyze the various types of spatial data and express them visually on a map. However, it is still difficult to estimate the location of spatial clusters based on statistical evidence. The spatial scan statistic (Kulldorff 1997), which is based on the idea of maximizing the likelihood of cluster, has been widely used for spatial cluster detection method. It is important how effectively and efficiently we find a cluster whose likelihood is high, and to find such a cluster, some scan approaches are proposed. However, most of them are limited in the shape of a detected cluster, or need an unrealistic computational time if the data size is too large. The zero-suppressed binary decision diagram (ZDD) (Minato, 1993), one approach to frequent item set mining, enables us to extract all of the potential cluster areas at a realistic computational cost. In this study, we try a new way of spatial cluster detection method to detect a cluster with truly highest likelihood by applying the ZDD, and by using them, we compare and evaluate the performance of the existing scan methods.

Keywords: Spatial cluster, Spatial scan statistic, ZDD

References:

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, **26**, 1481–1496.

Minato, S. (1993). Zero-suppressed BDDs for set manipulation in combinatorial problems. *In: Proceedings of the 30th ACM/IEEE Design Automation Conference*, 272–277.

Scoring Rules For Prediction And Classification Challenges

Matt Parry
University of Otago

Abstract: Prediction and classification challenges have become an exciting and useful feature of the statistical and machine learning community. For example, Good Judgement Open asks forecasters to predict the probability of particular world events, and Kaggle.com regularly sets classification challenges. Challenge organizers typically publish a ranked list of the leading submissions and, ultimately, announce the winner of the challenge. However, in order for such a competition to be considered worth entering, the challenge organizers must be seen to evaluate the submissions in a fair and open manner. Scoring rules were devised precisely to solve this problem. Crucially, *proper* scoring rules elicit honest statements of belief about the outcome. If the challenge organizers use a proper scoring rule to evaluate submissions, a competitor's expected score under their true belief will be optimized by actually quoting that belief to the organizers. A proper scoring rule therefore rules out any possibility of a competitor gaming the challenge. We discuss a class of proper scoring rules called linear scoring rules that are specifically adapted to probabilistic binary classification. When applied in competition situations, we show that all linear scoring rules essentially balance the needs of organizers and competitors. We also develop scoring rules to score a sequence of predictions that are targeting a single outcome. These scoring rules discount predictions over time and appropriately weight prediction updates.

Keywords: Probabilistic forecast, sequence, prequential principle, discounting

References:

Parry, M. (2016). *Linear scoring rules for probabilistic binary classification*. Electronic Journal of Statistics, 10 (1), 1596–1607.

Monday 11th 16:20 098 Lecture Theatre (260-098)

Meta-Analysis With Symbolic Data Analysis And Its Application For Clinical Data

Ryo Takagi, Hiroyuki Minami, and Masahiro Mizuta
Hokkaido University

Abstract: We discuss a method of meta-analysis based on symbolic data analysis (SDA). Meta-analysis, mainly used in social and medical science, is a statistical method of combining scientific studies to obtain quantitative results and provides a high level of evidence. Differences between the studies are caused by heterogeneity between the studies. It is useful to detect relationship among scientific studies. A target of analysis on SDA is *concept*, a set of individuals. We apply SDA to meta-analysis. In other words, we regard scientific studies as concepts. For example, symbolic clustering or symbolic MDS are useful to preprocess the scientific studies in meta-analysis. In this study, we propose a new approach based on SDA for meta-analysis and show the results of the proposed approach using clinical datasets.

Keywords: symbolic clustering, symbolic MDS, concept in SDA

References:

Edwin Diday and Monique Noirhomme-Fraiture. (2008). *Symbolic data analysis and the SODAS software.* John Wiley & Sons, Ltd.

David Edward Matthews and Vernon Todd Farewell. (2015). *Using and understanding medical statistics* (5th, revised and extended edition). Karger Publishers.

Monday 11th 16:20 OGGB4 (260-073)

Real-Time Transit Network Modelling For Improved Arrival Time Predictions

Tom Elliott and Thomas Lumley
University of Auckland

Abstract: The growing availability of GPS tracking devices means that public transport passengers can now check on the real-time location of their bus from their mobile phone, helping them to decide when to leave home, and once at the stop, how long until the bus arrives. A side effect of this technology is that statistical models using vehicle location data to predict arrival times have taken a “back seat” in preference for methods that are simpler and faster, but less robust. Auckland Transport, who operate our local public transport network, demonstrate this: the estimated arrival time (ETA) of a bus at a stop is simply the time until scheduled arrival, plus the delay at the bus’ most recently visited stop. The most evident problem with this approach is that intermediate stops, traffic lights, and road congestion—all of which affect ETAs—are not considered. We have been developing a modelling framework consisting of (1) a vehicle state model to infer parameters, such as speed, from a sequence of GPS positions; (2) a transit network model that uses information from the vehicle model to estimate traffic conditions along roads in the network; and (3) a predictive model combining vehicle and transit network states to predict arrival times. Since multimodality is common—for example a bus may or may not stop at a bus stop or traffic lights—we are using a particle filter to estimate vehicle state, which makes no assumptions about the shape of the distribution, and allows for a more intuitive likelihood function. While this provides a very flexible framework, it is also a computationally intensive one, so computational demands need to be considered to ensure it will be viable as a real-time application for providing passengers with improved, and hopefully reliable, arrival time information.

Keywords:

transit, real-time, particle filter

Monday 11th 16:20 OGGB5 (260-051)

Visualization And Statistical Modeling Of Financial Big Data

Masayuki Jimichi[1], Daisuke Miyamoto[2], Chika Saka[1], and Syuichi Nagata[1]
[1]Kwansei Gakuin University
[2]Nara Institute of Science and Technology

Abstract: In this work, we manipulate financial big data of world-wide companies by R packages SparkR and sparklyr, and treate data visualization (e.g. Ihaka, 2017; Unwin, 2015) and statistical model (e.g. Chambers and Hastie, 1991) based on exploratory data analysis (Tukey, 1977) with R. The dataset is extracted from the database systems by Bureau van Dijk KK which contains information on over 80,000 listed companies. We find that a log-skew-t linear model (e.g. Azallini and Capitanio, 2014) is very useful for explaining sales by employees and assets.

Keywords: Financial Big Data, Data Visualization, Statistical Modeling, Log-skew-t Linear Model, SparkR, sparklyr

References:

Azzalini, A. with the collaboration of Capitanio, A. (2014). *The Skew-Normal and Related Families*. Cambridge University Press. Institute of Mathematical Statistics Monographs.

Chambers, J. M. and Hastie, T. J. ed. (1991). *Statistical Models in S*. Chapman and Hall/CRC.

Ihaka, R. (2017). Lecture Notes. https://www.stat.auckland.ac.nz/~ihaka/?Teaching

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Co.

Unwin, A. (2015). *Graphical Data Analysis with R*, Chapman and Hall/CRC.

Sparse Group-Subgroup Partial Least Squares With Application To Genomic Data

Matthew Sutton[1], Benoit Liquet[1], and Rodolphe Thiebaut[2]
[1]Queensland University of Technology
[2]Inria, SISTM, Talence and Inserm, U1219, Bordeaux, Bordeaux University, Bordeaux and Vaccine Research Institute

Abstract: Integrative analysis of high dimensional omics datasets has been studied by many authors in recent years. By incorporating prior known relationships among the variables, these analyses have been successful in elucidating the relationships between different sets of omics data. In this article, our goal is to identify important relationships between genomic expression and cytokine data from an HIV vaccination trial. We proposed a flexible Partial Least Squares technique which incorporates group and subgroup structure in the modelling process. Our new methodology expands on previous work, by accounting for both grouping of genetic markers (e.g. genesets) and temporal effects. The method generalises existing sparse modelling techniques in the PLS methodology and establishes theoretical connections to variable selection methods for supervised and unsupervised problems. Simulation studies are performed to investigate the performance of our methods over alternative sparse approaches. Our method has been implemented in a comprehensive R package called sgsPLS.

Keywords: genomics, group variable selection, latent variable modelling, partial least squares, singular value decomposition

References:

Chaussabel, D. et al. (2008). A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity* **29,** 150–164.

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72,** 3–25.

Garcia, T. P., Muller, S., Carroll, R., and Walzem, R. (2014). Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data. *Bioinformatics* **30,** 35–42.

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., and Tegnér, J. (2014). Data integration in the era of omics: current and future challenges. *BMC Systems Biology* **8,** 1–10.

Hejblum, B., Skinner, J., and Thièbaut, R. (2015). Time-course gene set analysis for longitudinal gene expression data. *PLOS Computational Biology* **11,** 1–21.

Le Cao, K., Rossouw, D., Robert-Granie, C., and Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol Biol* **7,** 37.

Lèvy, Y., Thièbaut, R., Montes, M., Lacabaratz, C., Sloan, L., King, B., Pèrusat, S., Harrod, C., Cobb, A., Roberts, L., Surenaud, M., Boucherie, C., Zurawski, S., Delaugerre, C., Richert, L., Chêne, G., Banchereau, J., and Palucka, K. (2014). Dendritic cell-based therapeutic vaccine elicits polyfunctional hiv-specific t-cell immunity associated with control of viral load. *European Journal of Immunology* **44,** 2802–2810.

Lin, D., Zhang, J., Li, J., Calhoun, V., Deng, H., and Wang, Y. (2013). Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinformatics* **14,** 1–16.

Liquet, B., de Micheaux, P. L., Hejblum, B., and Thiébaut, R. (2016). Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics* **32,** 35–42.

Nowak, G., Hastie, T., Pollack, J., and Tibshirani, R. (2011). A fused lasso latent feature model for analyzing multisample acgh data. *Biostatistics* **12,** 776–791.

Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical applications in genetics and molecular biology* **8,** Article 1.

Rosipal, R. and Krämer, N. (2006). Overview and recent advances in partial least squares. *Subspace, Latent Structure and Feature Selection*

Safo, S. E., Li, S., and Long, Q. (2017). Integrative analysis of transcriptomic and metabolomic data via sparse canonical correlation analysis with incorporation of biological information. *Biometrics* .

Simon, N., Friedman, J., Tibshirani, R., and Hastie, T. (2013). A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics* **22,** 213–245.

Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K., Grill, J., and Frouin, V. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics* **15,** 569–583.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58,** 267–288.

Witten, D., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10,** 515–534.

Monday 11th 16:20 Case Room 3 (260-055)

Genetic Approach And Statistical Approach For Association Study On DNA Data

Makoto Tomita
Tokyo Medical and Dental University

Abstract: Genomic information such as genome-wide association analysis (GWAS) in DNA data is very large, however if the sample size corresponding to it is not enough, as an idea to solve, the author considers by a statistical approach and a genetic approach. The former will be briefly introduced, and the latter will be mainly explained. Basically, the method of focusing genome information becomes the center of presentation.

Keywords: genome wide association study, linkage disequilibrium, statistical power

References:

Tomita, M. (2013). Focusing Approach Using LD Block and Association Study with Haplotype Combination on DNA Data, In: *Proceedings 2013 Eleventh International Conference on ICT and Knowledge Engineering*, 5–10. Bangkok: IEEE Conference #32165.

Tomita, M. (2015). Haplotype estimation, haplotype block identification and statistical analysis for DNA data, In: *Conference Program and Book of Abstracts, Conference of the International Federation of Classication Societies (IFCS-2015)*, 227–228, Bologna.

Tomita, M., Hatsumichi, M. and Kurihara, K. (2008). *Computational Statistics and Data Analysis*, **52**(4), 1806–1820.

Tomita, M., Hashimoto, N. and Tanaka, Y. (2011). *Computational Statistics and Data Analysis*, **55**(6), 2104–2113.

Tomita, M., Kubota, T. and Ishioka, F. (2015). *PLoS ONE*, **10**(7), e0127358.

Monday 11th 16:20 Case Room 4 (260-009)

Modeling Of Document Abstraction Using Association Rule Based Characterization

Ken Nittono
Hosei University

Abstract: The importance of systems enabling us to extract useful information from enormous text data produced every day along with our social activities in organizations or on the internet and utilize the information immediately and efficiently have been increasing. In this research, an analyzing method which extracts essential parts from a huge document set utilizing association rule analysis as a data mining method is introduced. The method detects typical combinations of terms involved in contexts and regards them as the characterization of text data and it is also combined with information retrieval methods for the sake of further selection as some parts of the essential contexts. This method is considered to enhance its ability of detection for particular contexts that contain some topics and include moderately distributed terminologies. And implementation of the system is discussed in order for utilizing the abstracted documents efficiently as some sort of knowledge such as collective intelligence. An approach for linkage with R is also mentioned in the phase of the implementation of the model.

Keywords: Association rule, Text mining, Big data, Information retrieval

References:

Agrawal, R. Imielinski, T. and Swami, A. (1993). *Mining association rules between sets of items in large databases*, Proceedings of the ACM SIGMOD Washington, D.C, 207–216.

Nittono, K. (2013). *Association rule generation and mining approach to concept space for collective documents*, Proceedings of the 59th World Statistics Congress of the International Statistical Institute, pp. 5515–5520.

Bayesian Static Parameter Inference For Partially Observed Stochastic Systems

Yaxian Xu and Ajay Jasra
National University of Singapore

Abstract: We consider Bayesian static parameter estimation for partially observed stochastic systems with discrete-time observations. This is a very important problem, but is very computationally challenging as the associated posterior distributions are highly complex and one has to resort to discretizing the associated probability law of the underlying stochastic system and advanced Markov chain Monte Carlo (MCMC) techniques to infer the parameters. We are interested in the situation where the discretization is in multiple dimensions. For instance, for partially observed stochastic partial differential equations (SPDEs), where dicretization is in both space and time. In such cases, multi-index Monte Carlo (MIMC) is known to have the potential to reduce the computational cost for a prescribed level of error, relative to i.i.d. sampling from the most precise discretization. We demonstrate how MCMC and particularly particle MCMC can be used in the multi-index framework for Bayesian static parameter inference for the above-mentioned models. The main idea involves constructing an approximate coupling of the posterior density of the joint on the parameter and hidden space and then correcting by an importance sampling method. Our method is illustrated numerically to be preferable for inference of parameters for a partially observed SPDE.

Keywords: Multi-index Monte Carlo, Markov chain Monte Carlo, stochastic partial differential equations

References:

Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.

Haji-Ali, A. L., Nobile, F. & Tempone, R. (2016). Multi-Index Monte Carlo: When sparsity meets sampling. *Numerische Mathematik*, 132, 767–806.

Bayesian Survival Analysis Of Batsmen In Test Cricket

Oliver Stevenson and Brendon Brewer
University of Auckland

Abstract: It is widely accepted that in the sport of cricket, batting is more difficult early in a player's innings, but becomes easier as a player familiarizes themselves with the local conditions. Here we develop a Bayesian survival analysis method to predict and quantify the Test Match batting abilities for international cricketers, at any stage of a player's innings. The model is applied in two stages, firstly to individual players, allowing us to quantify players' initial and equilibrium batting abilities, and the rate of transition between the two. The results indicate that most players begin a Test match innings batting with between a quarter and a half of their potential batting ability. The model is then implemented using a hierarchical structure, providing us with more general inference concerning a selected group of opening batsmen from New Zealand. Using this hierarchical structure we are able to make predictions for the batting abilities of the next opening batsman to debut for New Zealand. These results are considered in conjunction with other performance based metrics, allowing us to identify players who excel in the role of opening the batting, which has practical implications in terms of batting order and team selection policy.

Keywords: Bayesian survival analysis, hierarchical modelling, cricket

References:

Stevenson, O.G. and Brewer, B.J. (2017). Bayesian survival anaylsis of opening batsmen in Test cricket *Journal of Quantitative Analysis in Sports*, *13*(1), 25-36.

Covariate Discretisation On Big Data

Hon Hwang[1], Stephen Wright[2], and Louise Ryan[1]
[1]University of Technology Sydney
[2]Australian Red Cross Blood Service

Abstract: Distributed Computing Systems such as Hadoop and Spark allow statistical analysis to be performed on arbitrary large datasets. However, when performing statistical analysis on these systems, the data communication between the nodes of a distributed computing system can become a major performance bottleneck. In this work, we outline a novel combination of statistical and computation techniques to address this issue. We first apply data reduction technique such as coarsening (interval-censoring) on large data sets using a distributed computing system. We then perform statistical analysis on the coarsened data. However, performing analysis using coarsened data potentially introduces biases in the results. To address this, we use the Expectation-Maximisation (EM) algorithm to recover the complete (non-coarsened) data model. Our work draws on methods for the analysis of data involving coarsened co-variates using EM by methods of weights. We explore different coarsening strategies (e.g., rounding, quantile and quintile) and discuss how our methods can scale to very large data settings. Through simulation studies, we find our method works especially well when data is coarsened from a wide interval, where there are more loss of information. Compared with naïvely using the coarsened data, our method is able to estimate regression coefficients that are closer to estimates obtained from using the complete data. In addition, the standard errors from our method reflect more accurately the uncertainty arising from using coarsened data.

Keywords:

EM algorithm, coarsened data, regression, big data

Monday 11th 16:40 Case Room 2 (260-057)

BIG-SIR A Sliced Inverse Regression Approach For Massive Data

Benoit Liquet[1] and Jerome Saracco[2]
[1]Queensland University of Technology
[2]University of Bordeaux

Abstract: In a massive data setting, we focus on a semiparametric regression model involving a real dependent variable YY and a pp-dimensional covariable XX. This model includes a dimension reduction of X via an index X′βX'\beta. The Effective Dimension Reduction (EDR) direction β\beta cannot be directly estimated by the Sliced Inverse Regression (SIR) method due to the large volume of the data. To deal with the main challenges of analysing massive datasets which are the storage and computational efficiency, we propose a new SIR estimator of the EDR direction by following the "divide and conquer" strategy. The data is divided into subsets. EDR directions are estimated in each subset which is a small dataset. The recombination step is based on the optimisation of a criterion which assesses the proximity between the EDR directions of each subset. Computations are run in parallel with no communication among them. The consistency of our estimator is established and its asymptotic distribution is given. Extensions to multiple indices models, qq-dimensional response variable and/or SIRα_{\alpha}-based methods are also discussed. A simulation study using our edrGraphicalTools R package shows that our approach enables us to reduce the computation time and conquer the memory constraint problem posed by massive datasets. A combination of foreach and bigmemory R packages are exploited to offer efficiency of execution in both speed and memory. Results are visualised using the bin-summarise-smooth approach through the bigvis R package. Finally, we illustrate our proposed approach on a massive airline data set.

Keywords: High performance computing, Effective Dimension Reduction (EDR), Parallel programming, R software, Sliced Inverse Regression (SIR)

References:

Liquet, B., & Saracco, J. (2016), *BIG-SIR a Sliced Inverse Regression Approach for Massive Data*, Statistics and Its Interface. Vol 9, 509-520.

Monday 11th 16:40 Case Room 3 (260-055)

Symbolic Data Analytical Approach To Unauthorized-Access Logs

Hiroyuki Minami and Masahiro Mizuta
Hokkaido University

Abstract: We have been annoyed by tons of unwilling accesses in many Internet applications including SSH (Secure SHell) known as a typical remote access tool and E-mail delivery protocols. An attacked server put a report according to the configuration and the log files have grown day by day.

Bad accesses might be caused by computer virus and so-called *zombie*, a hi-jacked computer. We assume that the actions would have their own trends. For example, we sometimes find that a few attacks come simultaneously from only 1 site, however, we also find several attacks from a set of the sites within 1 minute or 1 day. The IP-Addresses might be variable, however, within the assigned range. It suggests that the victimizer is just one but gets an IP-Address allocation so many times.

To analyze the log files and give an interpretation to them, we introduce Symbolic Data Analysis (SDA) to adopt its main idea *concept*. If we configure an appropriate *concept* whose elements (*individuals* in SDA) are IP-Address, port-numbers and attack time span, we can reveal some relationship between *concepts* and classify them into perspective. The results would give us some useful information to protect our Internet environment.

We discuss how we get them and the interpretation appropriately through some practical examples.

Keywords: Invalid network access, Firewall, Massive Data Analysis

References:

Collins, M. (2014). *Network Security Through Data Analysis*. O'Reilly.

Minami, H. and Mizuta, M. (2016). A study on the Analysis of the refused logs by Internet Firewall. *Proceedings of 2016 International Conference for JSCS 30th Anniversary in Seattle*.

Monday 11th 16:40 Case Room 4 (260-009)

My Knee Still Hurts; The Statistical Pathway To The Development Of A Clinical Decision Aid

Robert Borotkanics
Auckland University of Technology

Abstract: Total knee arthroplasty (TKA) is considered an effective intervention to improve physical function and reduce joint pain in those with end stage knee arthritis, yet up to 34 What is reported herein is the methodological approaches applied to tease out the various nuances of developing such a clinical decision aid. By way of summary, a series of logistic regression models were developed and refined to identify predictors chronic postoperative pain, where pain was reported using the numerical pain rating scale (NRS). Self-reported NRS pain was dichotomized based on functional status using a defined statistical approach. Multivariate models were developed using a stepwise selection approach, accounting for interaction and collinearity. The effect of changing collinearity thresholds on information criterion is illustrated. The sensitivity and specificity were calculated, along with receiver operating characteristic (ROC) analyses for each logistic model. Final models were chosen by a combination of superior area under the curve (AUC) and Akaike Information Criterion (AIC). The stability of β coefficients across the top performing models is reported, along with model goodness-of-fit using the Hosmer and Lameshow methodology. Cut-point analyses are reported on models performances, including the effect of changing pain thresholds on model accuracy. Finally, the conversion of a superior logistic model into a probabilistic function, of potential utility for clinicians is illustrated.

Keywords:

total knee arthroplasty; logistic regression; clinical decision support

Programme And Abstracts For Tuesday 12th Of December

Keynote: Tuesday 12th 9:10 098 Lecture Theatre (260-098)

Could Do Better ... A Report Card For Statistical Computing

Ross Ihaka and Brendon McArdle
University of Auckland

Abstract: Since the introduction of R, research in Statistical Computing has plateaued. Although R is, at best, a stop-gap system, there appears to be very little active research on creating better computing environments for Statistics.

When work on R commenced there were a multitude of software systems for statistical data analysis in use and under development. There was friendly competition and collaboration between developers. While R can be seen as providing a useful unification for users, its success and dominance can be viewed as now holding back research and the development of new systems.

In this talk we'll examine what might be behind this and also look at some research aimed at exploring some of the design space for new systems. The aim is to show constructively that new work in the area is still possible.

R&D Policy Regimes In France: New Evidence From A Spatio-Temporal Analysis

Benjamin Montmartin[1], Marcos Herrera[2], and Nadine Massard[3]
[1]GREDEG CNRS
[2]CONICET
[3]GAEL

Abstract: Using a unique database containing information on the amount of R&D tax credits and regional, national and European subsidies received by firms in French NUTS3 regions over the period 2001-2011, we provide new evidence on the efficiency of R&D policies taking into account spatial dependency across regions. By estimating a spatial Durbin model with regimes and fixed effects, we show that in a context of yardstick competition between regions, national subsidies are the only instrument that displays total leverage effect. For other instruments internal and external effects balance each other resulting in insignificant total effects. Structural breaks corresponding to tax credit reforms are also revealed.

Keywords: Additionality, French policy mix, Spatial panel, Structural break

References:

Pesaran, M. H. (2007). A simple panel unit root test in the presence of cross-section dependence In: *Journal of Applied Econometrics*, **22**, 265–312.

Hendry, D. F. (1979). Predictive failure and econometric modelling in macroeconomics: The transactions demand for money. In: *P. Ormerod (Ed.), Economic Modelling: Current Issues and Problems in Macroeconomic Modelling in the UK and the US*, **9**, 217–242. Heinemann Education Books, London.

Tuesday 12th 10:30 OGGB4 (260-073)

Analysing Scientific Collaborations Of New Zealand Institutions Using Scopus Bibliometric Data

Samin Aref[1], David Friggens[2], and Shaun Hendy[1]
[1]University of Auckland
[2]Ministry of Business Innovation & Employment

Abstract: Scientific collaborations are among the main enablers of development in small national science systems. Although analysing scientific collaborations is a well-established subject in scientometrics, evaluations of collaborative activities of countries remain speculative with studies based on a limited number of fields or using data too inadequate to fully represent collaborations at a national level. This study provides a unique view on the collaborative aspect of scientific activities in New Zealand. We perform a quantitative study based on all Scopus publications in all subjects for over 1500 New Zealand institutions over a period of 6 years to generate an extensive mapping of New Zealand scientific collaborations. The comparative results reveal the levels of collaboration between New Zealand institutions and business enterprises, government institutions, higher education providers, and private not for profit organisations in 2010-2015. Constructing a collaboration network of institutions, we observe a power-law distribution indicating that a small number of New Zealand institutions account for a large proportion of national collaborations. Network centrality measures are deployed to identify the most influential institutions of the country in terms of scientific collaboration. We also provide comparative results on 15 universities and crown research institutes based on 27 subject classifications. This study was based on Scopus custom data and supported by the Te Pūnaha Matatini internship program at Ministry of Business, Innovation & Employment.

ArXiv preprint link: https://arxiv.org/pdf/1709.02897

Keywords:

Big data modelling, Scientific collaboration, Scientometrics, Network analysis, Scopus, New Zealand

Tuesday 12th 10:30 OGGB5 (260-051)

Family Structure And Academic Achievements Of High School Students In Tonga

Losana Vao Latu Latu
University of Canterbury

Abstract: In this study we examine how family structure affects the academic achievement of students at the secondary level of education age in Tonga. It is a comparative study aiming to find out whether there is a significant difference between the academic achievements of students from a traditional family and those from a non-traditional family. We define a Tongan traditional family as being two biological parents (or adoptive parents from birth), one male and one female where as non-traditional family can be a single parent family, or the student has no parent present (for example they are staying with relatives or friends). In our study we are looking at what are the key drivers of success and trying to understand the relationship between academic achievements and family structure. We hope the study will provide evidence-based information to aid the administrators, other educators and parents to adopt the best practices and actions for the students. The target population for this study is the high school students age 13 to 18 in Tonga. The study is limited to the high schools in the main island of Tonga- Tongatapu which has 12 high schools where two high schools are government schools and the others are private schools run by different religions. In April we surveyed 360 students, 60 from each of 6 high schools, and present here our preliminary results.

Keywords:

Education, policy, stratified sampling

Analysis Of Multivariate Binary Longitudinal Data: Metabolic Syndrome During Menopausal Transition

Geoff Jones
Massey University

Abstract: Metabolic syndrome (MetS) is a major multifactorial condition that predisposes adults to type 2 diabetes and cardiovascular disease. It is defined as having at least three of five cardiometabolic risk components: 1) high fasting triglyceride level, 2) low high-density lipoprotein (HDL) cholesterol, 3) elevated fasting plasma glucose, 4) large waist circumference (abdominal obesity) and 5) hypertension. In the US Study of Women's Health Across the Nation (SWAN), a 15-year multi-centre prospective cohort study of women from five racial/ethnic groups, the incidence of MetS increased as midlife women underwent the menopausal transition (MT). A model is sought to examine the interdependent progression of the five MetS components and the influence of demographic covariates.

Keywords:

Multivariate binary data, longitudinal analysis, metabolic syndrome

Clustering Of Curves On A Spatial Domain Using A Bayesian Partitioning Model

Chae Young Lim
Seoul National University

Abstract: We propose a Bayesian hierarchical model for spatial clustering of the high-dimensional functional data based on the effects of functional covariates. We couple the functional mixed-effects model with a generalized spatial partitioning method for: (1) identifying subregions for the high-dimensional spatio-functional data; (2) improving the computational feasibility via parallel computing over subregions or multi-level partitions; and (3) addressing the near-boundary ambiguity in model-based spatial clustering techniques. The proposed model extends the existing spatial clustering techniques to produce spatially contiguous partitions for spatio-functional data. The model successfully captured the regional effects of the atmospheric and cloud properties on the spectral radiance measurements. This elaborates the importance of considering spatially contiguous partitions for identifying regional effects and small-scale variability.

Keywords:

spatial clustering, Bayesian wavelets, Voronoi tessellation, functional covariates

The Uncomfortable Entrepreneurs: Bad Working Conditions And Entrepreneurial Commitment

Catherine Laffineur
Université Côte d'Azur, GREDEG-CNRS

Abstract: In contrast to previous model dividing necessity entrepreneurs as individuals facing push factors due to lack of employment, we consider the possibility of push factors faced by employed individuals (Folta et al. (2010)). The theoretical model yields distinctive predictions relating occupation characteristics and the probability of entry into entrepreneurship. Using PSED and ONET data, we investigate how the characteristics of individuals? primary occupations affect nascent entrepreneurs? effort put into venture creation. The empirical evidences show that necessity entrepreneurs are not only confined to unemployed individuals. We find compelling evidence that individuals facing arduous working conditions (e.g. stressful environment and physical tiredness) have a higher likelihood of entering and succeeding in self-employment than others. Contrariwise, individuals who experience high degree of self-realization, independence and responsibility in the workplace are less committed to their business than individuals exposed to arduous working conditions. These findings have strong implication for how we interpret and analyze necessity entrepreneurs and provide novel insights into the role of occupational experience in the process of venture emergence.

Keywords: Entrepreneurship, Motivation, Occupational characteristics, Employment choice.

References:

Folta, T. B., Delmar, F., & Wennberg, K. 2010. Hybrid entrepreneurship. *Management Science*, 56(2), 253-269.

Spatial Surveillance With Scan Statistics By Controlling The False Discovery Rate

Xun Xiao
Massey University

Abstract: In this paper, I investigate a false discovery approach based on spatial scan statistics to detect the spatial disease clusters in a geographical region proposed by Li et al. (2016). The incidence of disease is assumed to follow an inhomogeneous Poisson model discussed in Kulldorff (1997). I show that, though spatial scan statistics are highly correlated, the simple Banjamini-Hochberg (linear step-up) procedure can control the false discovery rate of them by proving that the multivariate Poisson distribution satisfies the PRDS condition (positive regression dependence on a subset) in Benjamini and Yekutieli (2001).

Keywords: False Discovery Rate, Poisson Distribution, PRDS, Spatial Scan Statistics

References:

Benjamini, Y. and Yekutieli, D. (2001). *The control of the false discovery rate in multiple testing under dependency*, Annals of Statistics, **29**(4), 1165–1188.

Kulldorff, M. (1997). *A spatial scan statistic*, Communications in Statistics-Theory and Methods **26**(6), 1481–1496.

Li, Y., Shu, L., and Tsung, F. (2016). *A false discovery approach for scanning spatial disease clusters with arbitrary shapes*, IIE transactions, **48**(7), 684–698.

Statistical Models For The Source Attribution Of Zoonotic Diseases: A Study Of Campylobacteriosis

Sih-Jing Liao, Martin Hazelton, Jonathan Marshall, and Nigel French
Massey University

Abstract: Preventing and controlling zoonoses with a public health policy depends on the knowledge scientists have about the transmitted pathogens. Modelling jointly the epidemiological data and genetic information provides a methodology for tracing back the source of infection. However, this creates difficulties in assessing genetic efforts behind models of the final statistical inferences due to increased model complexity. To explore the genetic effects in the joint model, we develop a genetic free model and compare it to the joint model. We apply the two models to a recent campylobacteriosis study to estimate the attribution probability for each source. A spatial covariate is also considered in the models in order to investigate the effect of the level of rurality on the source attributions. Comparing the attributions generated by the two models, we find that: i) the genetic information integrated in the joint model gives a little more precise inference to the sparse cases observed in highly rural areas than the genetic free model; ii) on the logit scale, source attribution probabilities follow linear trends against level of rurality; and iii) poultry is the dominant source of campylobacteriosis in urban centres, whereas ruminants are the most attributable source when in rural areas.

Keywords: source attribution, *Campylobacter*, multinomial model, Dirichlet prior, HPD interval, DIC

References:

Bronowski, C., James, C.E. and Winstanley, C. (2014). Role of environmental survival in transmission of *Campylobacter jejuni*. *FEMS Microbiol Lett.*, **356**(1) 8–19.

Dingle, K.E., Colles, F.M., Wareing, D.R., Ure, R., Fox, A.J., Bolton, F.E., Bootsma, H.J., Willems, R.J. and Maiden, M.C. (2001). Multilocus sequence typing system for *Campylobacter jejuni*. *J Clin Microbiol*, **39**(1):14–23.

Marshall, J.C. and French, N.P. (2015). Source attribution January to December 2014 of human *Campylobacter jejuni* cases from the Manawatu. *Technical Report*.

Wilson, D.J., Gabriel, E., Leatherbarrow, A.J., Cheesbrough, J., Gee, S., Bolton, E., Fox, A., Fearnhead, P., Hart, C.A. and Diggle, P.J. (2008). Tracing the source of campylobacteriosis. *PLoS Genet*, **4**(9):e1000203.

Wagenaar, J.A., French, N.P. and Havelaar, A.H. (2013). Preventing *Campylobacter* at the source: why is it so difficult? *Clin Infect Dis*, **57**(11):1600–1606.

Biggs, P.J., Fearnhead, P., Hotter, G., Mohan, V., Collins-Emerson, J., Kwan, E., Besser, T.E., Cookson, A., Carter, P.E. and French, N.P. (2011). Whole-genome comparison of two *Campylobacter jejuni* isolates of the same sequence type reveals multiple loci of different ancestral lineage. *PLoS One*, **6**(11):e27121.

Tuesday 12th 10:50 OGGB5 (260-051)

Towards An Informal Test For Goodness-Of-Fit

Anna Fergusson and Maxine Pfannkuch
University of Auckland

Abstract:

Informal approaches to goodness-of-fit tests often involve examining the visual fit of the model to data ’by eye’. Such approaches are problematic for Year 13 and undergraduate students and teachers from a pedagogical perspective as key aspects such as sample size, the number of categories and expected variation of sample proportions are difficult to consider. In formal tests for goodness-of-fit a test statistic is used in reference to its sampling distribution to decide if the model distribution can be rejected. In general, a numeric test statistic does not have an obvious graphical representation within the data itself. This talk presents a new informal goodness-of-fit test that uses a simulation-based modelling tool. Drawing on ideas from graphical inference, the proposed test does not use numerical test statistics but plots as test statistics. Comparisons of performance demonstrate that the proposed test leads to similar decisions about the fit of the model distribution as the chi square goodness-of-fit test. A research study with Year 13 teachers indicated that there could be pedagogical benefits of using this informal goodness-of-fit test in terms of introducing important modelling and hypothesis test concepts.

Identifying Clusters Of Patients With Diabetes Using A Markov Birth-Death Process

Mugdha Manda, Thomas Lumley, and Susan Wells
University of Auckland

Abstract: Estimating disease trajectories has increasingly become more essential to clinical practitioners to administer effective treatment to their patients. A part of describing disease trajectories involves taking patients' medical histories and sociodemographic factors into account and grouping them into similar groups, or clusters. Advances in computerised patient databases have paved a way for identifying such trajectories in patients by recording a patient's medical history over a long period of time (longitudinal data): we studied data from the PREDICT-CVD dataset, a national primary-care cohort from which people with diabetes from 2002-2015 were identified through routine clinical practice. We fitted a Bayesian hierarchical linear model with latent clusters to the repeated measurements of HbA1c_1c and eGFR, using the Markov birth-death process proposed by Stephens (2000) to handle the changes in dimensionality as clusters were added or removed.

Keywords: Diabetes management, longitudinal data, Markov chain Monte Carlo, birth-death process, mixture model, Bayesian analysis, latent clusters, hierarchical models, primary care, clinical practice

References:

Stephens, M. (2000). Bayesian Analysis of Mixture Models with an Unknown Number of Components - An Alternative to Reversible Jump Methods. In: *The Annals of Statistics*, 28(1), 40-74.

Tuesday 12th 10:50 Case Room 3 (260-055)

Bayesian Temporal Density Estimation Using Autoregressive Species Sampling Models

Youngin Jo[1], Seongil Jo[2], and Jaeyong Lee[3]
[1]Kakao Corporation
[2]Chonbuk National University
[3]Seoul National University

Abstract: We propose a Bayesian nonparametric (BNP) model, which is built on a class of species sampling models, for estimating density functions of temporal data. In particular, we introduce species sampling mixture models with temporal dependence. To accommodate temporal dependence, we define dependent species sampling models by modeling random support points and weights through an autoregressive model, and then we construct the mixture models based on the collection of these dependent species sampling models. We propose an algorithm to generate posterior samples and present simulation studies to compare the performance of the proposed models with competitors that are based on Dirichlet process mixture models. We apply our method to the estimation of densities for the price of apartment in Seoul, the closing price in Korea Composite Stock Price Index (KOSPI), and climate variables (daily maximum temperature and precipitation) of around the Korean peninsula.

Keywords: Autoregressive species sampling models; Dependent random probability measures; Mixture models; Temporal structured data

Acknowledgements:

This work is a part of the first author's Ph.D. thesis at Seoul National University. Research of Seongil Jo was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A3B03035235). Research of Jaeyong Lee was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0030811).

Tuesday 12th 10:50 Case Room 4 (260-009)

How Does The Textile Set Describe Geometric Structures Of Data?

Ushio Tanaka[1] and Tomonari Sei[2]
[1]Osaka Prefecture University
[2]Unversity of Tokyo

Abstract: The textile set is defined from the textile plot proposed by Kumasaka and Shibata (2007, 2008), which is a powerful tool for visualizing high dimensional data. The textile plot is based on a parallel coordinate plot, where the ordering, locations and scales of each axis are simultaneously chosen so that all connecting lines, each of which signifies an observation, are aligned as horizontally as possible. The textile plot transforms a data matrix in order to delineate a parallel coordinate plot. Using the geometric properties of the textile set derived by Sei and Tanaka (2015), we show that the textile set describes an intrinsically geometric structures of data.

Keywords: Parallel coordinate plot, Textile set, Differentiable manifold

References:

Kumasaka, N. and Shibata, R. (2007). The Textile Plot Environment, *Proceedings of the Institute of Statistical Mathematics*, **55**, 47–68.

Kumasaka, N. and Shibata, R. (2008). High-dimensional data visualisation: The textile plot, *Computational Statistics and Data Analysis*, **52**, 3616–3644.

Sei, T. and Tanaka, U. (2015). Geometric Properties of Textile Plot: *Geometric Science of Information*, *Lecture Notes in Computer Science*, **9389**, 732–739.

Intensity Estimation Of Spatial Point Processes Based On Area-Aggregated Data

Hsin-Cheng Huang and Chi-Wei Lai
Academia Sinica

Abstract: We consider estimation of intensity function for spatial point processes based on area-aggregated data. A standard approach for estimating the intensity function for a spatial point pattern is to use a kernel estimator. However, when data are only available in a spatially aggregated form with the numbers of events available in geographical subregions, traditional methods developed for individual-level event data become infeasible. In this research, a kernel-based method will be proposed to produce a smooth intensity function based on aggregated count data. Some numerical examples will be provided to demonstrate the effectiveness of the proposed method.

Keywords:

Area censoring, inhomogeneous spatial point processes, kernel density estimation

Bayesian Inference For Population Attributable Measures

Sarah Pirikahu, Geoff Jones, Martin Hazelton, and Cord Heuer
Massey University

Abstract:

Epidemiologists often wish to determine the population impact of an intervention to remove or reduce a risk factor. Population attributable type measures, such as the population attributable risk (PAR) and population attributable fraction (PAF), provide a means of assessing this impact, in a way that is accessible for a non-statistical audience. To apply these concepts to epidemiological data, the calculation of estimates and confidence intervals for these measures should take into account the study design (cross-sectional, case-control, survey) and any sources of uncertainty (such as measurement error in exposure to the risk factor). We provide methods to produce estimates and Bayesian credible intervals for the PAR and PAF from common epidemiological study types and assess the Frequentist properties. The model is then extended by incorporating uncertainty due to the use of imperfect diagnostic tests for disease or exposure. The resulting model can be non-identifiable, causing convergence problems for common MCMC samplers, such as Gibbs and Metropolis-Hastings. An alternative importance sampling method performs much better for these non-identifiable models and can be used to explore the limiting posterior distribution. The data used to estimate these population attributable measures may include multiple risk factors in addition to the one being considered for removal. Uncertainty regarding the distribution of these risk factors in the population affects the inference for PAR and PAF. To allow for this we propose a methodology involving the Bayesian bootstrap. We also extend the analysis to allow for complex survey designs with unequal weights, stratification and clustering.

An Information Criterion For Prediction With Auxiliary Variables Under Covariate Shift

Takahiro Ido[1], Shinpei Imori[1,2], and Hidetoshi Shimodaira[2,3]
[1]Osaka University
[2]RIKEN Center for Advanced Intelligence Project (AIP)
[3]Kyoto University

Abstract: It is beneficial for modeling data of interest to exploit secondary information. The secondary information is called auxiliary variables, which may not be observed in testing data because they are not of primary interest. In this paper, we incorporate the auxiliary variables into a framework of supervised learning. Furthermore, we consider a covariate shift situation that allows a density function of covariates to change between testing and training data. It is known that the Maximum Log-likelihood Estimate (MLE) is not a good estimator under model misspecification and the covariate shift. This problem can be resolved by the Maximum Weighted Log-likelihood Estimate (MWLE).

When we have multiple candidate models, it needs to select the best candidate model where its optimality is measured by the expected Kullback-Leibler (KL) divergence. The Akaike information criterion (AIC) is a well known criterion based on the KL divergence and using the MLE. Therefore, its validity is not guaranteed when the MWLE is used under the covariate shift. An information criterion under the covariate shift was proposed in Shimodaira (2000, JSPI) but this criterion does not take use of the auxiliary variables into account. Hence, we resolve this problem by deriving a new criterion. In addition, simulations are conducted to examine the improvement.

Keywords:

Auxiliary variables; Covariate shift; Information criterion; Kullback-Leibler divergence; Misspecification; Predictions.

Analysis Of A Brief Telephone Intervention For Problem Gambling And Examining The Impact On Co-Existing Depression?

Nick Garrett, Maria Bellringer, and Max Abbott
Auckland University of Technology

Abstract:

This study investigated the outcomes of a brief telephone intervention for problem gambling. A total of 150 callers were recruited and followed for 36 months. After giving consent, participants received a baseline assessment followed by a manualised version of the helpline's standard care. Eight-six percent of participants were re-assessed at three months, 79Depression is found to often be associated with problem gambling behaviour, and analysis was undertaken to examine the impact of a brief telephone intervention for problem gambling on rates of depression using logistic regression. At baseline depression was found to be associated with gender, problem gambling risk (PGSI), and deprivation (NZiDep). A multiple variable model found that PGSI and mental health medication best explained depression at baseline. A repeated measures logistic regression utilising all 36 months of data found that PGSI, NZiDep, and mental health medication were the best variables to explain the change over time. Conclusion was that the intervention's impact on problem gambling behaviour also changed depression rates, however deprivation and mental health medication also contributed.

Prior-Based Bayesian Information Criterion

M. J. Bayarri[1], James Berger[2], Woncheol Jang[3], Surajit Ray[4], Luis Pericchi[5], and Ingmar Visser[6]
[1]University of Valencia
[2]Duke University
[3]Seoul National University
[4]University of Glasgow
[5]University of Puerto Rico
[6]University of Amsterdam

Abstract: We present a new approach to model selection and Bayes factor determination, based on Laplace expansions (as in BIC), which we call Prior-based Bayes Information Criterion (PBIC). In this approach, the Laplace expansion is only done with the likelihood function, and then a suitable prior distribution is chosen to allow exact computation of the (approximate) marginal likelihood arising from the Laplace approximation and the prior. The result is a closed-form expression similar to BIC, but now involves a term arising from the prior distribution (which BIC ignores) and also incorporates the idea that different parameters can have different effective sample sizes (whereas BIC only allows one overall sample size nn). We also consider a modification of PBIC which is more favorable to complex models.

Keywords:

Bayes factors, model selection, Cauchy priors, consistency, effective sample size, Fisher information, Laplace expansions, robust priors

Geographically Weighted Principal Component Analysis For Spatio-Temporal Statistical Dataset

Narumasa Tsutsumida[1], Paul Harris[2], and Alexis Comber[3]
[1]Kyoto University
[2]Rothamsted Research
[3]Univerisity of Leeds

Abstract: Spatio-temporal statistical datasets are becoming widely available for social, ecomonic, and environmental researches, however it is often difficult to summarize it and undermine hidden spatial/temporal patterns due to its complexity. Geographically weighted principal component analysis (GWPCA), which uses a moving window or kernel and applies localized PCAs over geographical scape, may be worth to do it, while to optimize kernel bandwidth size and to determine the number of component to retain (NCR) were the most concern (Tsutsumida et al (2017)). In this research we determine both of them together simultaneously so as to minimize leave-one-out residual coefficient of variation of GWPCA with changing bandwidth size and NCR. As a case study we use annual goat population statistics across 341 administrative units in Mongolia in 1990-2012, and show spatiotemporal variations in data, especially influenced by natural disasters.

Keywords: Geographically weighted model, Spatio-temporal data, Parameter optimization

References:

Tsutsumida N., P. Harris, , A. Comber. 2017. The Application of a Geographically Weighted Principal Component Analysis for Exploring Twenty-three Years of Goat Population Change across Mongolia. *Annals of the American Association of Geographers*, **107(5)**, 1060–1074.

Tuesday 12th 11:30 OGGB4 (260-073)

Dimensionality Reduction Of Multivariate Data For Bayesian Analysis

Anjali Gupta[1], James Curran[1], Sally Coulson[2], and Christopher Triggs[1]
[1]University of Auckland
[2]ESR

Abstract:

In 2004, Aitken and Lucy published an article detailing a two-level likelihood ratio for multivariate trace evidence. This model has been adopted in a number of forensic disciplines such as the interpretation of glass, drugs (MDMA), and ink. Modern instrumentation is capable of measuring many elements in very low quantities and, not surprisingly, forensic scientists wish to exploit the potential of this extra information to increase the weight of this evidence. The issue, from a statistical point of view, is that the increase in the number of variables (dimension) in the problem leads to increased data demand to understand both the variability within a source, and in between sources. Such information will come in time, but usually we don’t have enough. One solution to this problem is to attempt to reduce the dimensionality through methods such as principal component analysis. This practice is quite common in high dimensional machine learning problems. In this talk, I will describe a study where we attempt to quantify the effects of this this approach on the resulting likelihood ratios using data obtained from SEM-EDX instrument.

An EWMA Chart For Monitoring Covariance Matrix Based On Dissimilarity Index

Longcheen Huwang
National Tsing Hua University

Abstract: In this talk, we propose an EWMA chart for monitoring covariance matrix based on the dissimilarity index of two matrices. It is different from the conventional EWMA charts for monitoring covariance matrix which are either based on comparing the sum or product or both of the eigenvalues of the estimated EWMA covariance matrix with those of the IC covariance matrix. The proposed chart essentially monitors covariance matrix by comparing the individual eigenvalues of the estimated EWMA covariance matrix with those of the estimated covariance matrix from the IC phase I data. We evaluate the performance of the proposed chart by comparing it with the best existing chart under the multivariate normal process. Furthermore, to prevent the control limit of the proposed EMMA chart using the limited IC phase I data from having extensively excessive false alarms, we use a bootstrap method to adjust the control limit to guarantee that the proposed chart has the actual IC average run length not less than the nominal one with a certain probability. Finally, we use an example to demonstrate the applicability and implementation of the proposed chart.

Keywords: Average run length, dissimilarity index, EWMA; out-of-control

References:

Hawkins, D.M. and Maboudou-Tchao E.M. (2008). Multivariate exponentially weighted moving covariance matrix. *Technometrics*, **50**, 155-166.

Kano, M., Hasebe, S. and Hashimoto, I. (2002). Statistical process monitoring based on dissimilarity of process data. *AIChE Journal*, **48**, 1231-1240.

Adjusting For Linkage Bias In The Analysis Of Record-Linked Data

Patrick Graham
Stats NZ and Bayesian Research

Abstract: Data formed from record-linkage of two or more datasets are an increasingly important source of data for public health and social science research. For example, a study cohort may be linked to administrative data in order to add outcome or covariate information to data collected directly from study participants. However, regardless of the linkage method, it is often the case that not all records are linked. Further, linkage rates usually vary with characteristics of analytical interest and this differential linkage can bias analyses restricted just to linked records. While linked records have full outcome and covariate information, unlinked records exhibit "block-missingness" whereby the values for the entire block of variables contained in the file that is linked to are missing for unlinked records. Similar missing data structures occur in other contexts, including panel studies when participants decline participation in one or more study waves. In this paper, I consider the problem of adjusting for linkage bias from both Bayesian and frequentist perspectives. A basic distinction is whether analysis is based on all available data or just the linked cases. The Bayesian perspective leads to the former option and to Gibbs sampling and multiple imputation as reasonable methods. Basing analysis only on the linked cases seems to require a frequentist perspective and leads to inverse probability of linkage weighting and conditional maximum likelihood as reasonable approaches. The implications of the assumption of ignorable linkage also differ somewhat between the approaches. A simulation investigation confirms that, assuming ignorable linkage given observed data, multiple imputation, conditional maximum likelihood and inverse probability of linkage weighting all succeed in adjusting for linkage bias and achieve nominal interval coverage rates. Conditional maximum likelihood is slightly more efficient than inverse probability of linkage weighting and that multiple imputation can be more efficient than conditional maximum likelihood. Extensions to the case of non-ignorable linkage are also considered.

Keywords:

Record linkage, Missing data, Bayesian inference, Gibbs sampler, Multiple imputation

Bayesian Semiparametric Hierarchical Models For Longitudinal Data Analysis With Application To Dose-Response Studies

Taeryon Choi
Korea University

Abstract: In this work, we propose semiparametric Bayesian hierarchical additive mixed effects models for analyzing either longitudinal data or clustered data with applications to dose-response studies. In the semiparametric mixed effects model structure, we estimate nonparametric smoothing functions of continuous covariates by using a spectral representation of Gaussian processes and the subject-specific random effects by using Dirichlet process mixtures. In this framework, we develop semiparametric mixed effects models that include normal regression and quantile regressions with or without shape restrictions. In addition, we deal with the Bayesian nonparametric measurement error models, or errors-in-variable regression models, using Fourier series and Dirchlet process mixtures, in which the true covariate is not observable, but the surrogate of the true covariate, is only observed. The proposed methodology is compared with other existing approaches to additive mixed models in simulation studies and benchmark data examples. More importantly, we consider a real data application for dose-response analysis, in which measurement errors and shape constraints in the regression functions need to be incorporated with inter-study variability.

Keywords:

Cadmium toxicity, Cosine series, Dose-response study, Hierarchical Model, Measurement errors, Shape restriction

Optimizing Junior Rugby Weight Limits

Emma Campbell, Ankit Patel, and Paul Bracewell
DOT Loves Data

Abstract: The New Zealand rugby community is aware of safety issues within the junior game and has applied weight limits for each tackle grade to minimize injury risk. However, for heavier children this can create an uncomfortable situation as they may no longer be playing with their peer group. The study evaluated almost 13,000 observations from junior rugby players across three seasons (2015-2017) using data supplied by Wellington Rugby. To protect privacy, the data was structured so that an individual could not be readily identified but could be tracked across seasons to determine churn. As data for several consecutive seasons was available, we could determine the likelihood of a junior player returning the following season and isolate the drivers of this behaviour. Applying a logistic regression and repeated measures analysis the study determined if children who are over the specified weight limit for their age group are more likely to leave the game. Furthermore, assuming the importance of playing with peers, the study identified the impact of age in relation to the date-of-birth cut-off of January 1st. This is of interest given that a child playing above their age-weight grade could be competing against individuals three school years above them. The study primarily focuses on determining the optimal age-weight bands while the secondary focus is on determining the likelihood of a junior Wellington rugby player returning the following season and isolating the drivers of this behaviour.

Keywords:

Logistic regression, repeated measures, player retention, optimization

Spatial Scan Statistics For Matched Case-Control Data

Inkyung Jung
Yonsei University College of Medicine

Abstract: Spatial scan statistics are widely used for cluster detection analysis in geographical disease surveillance. While the method has been developed for various types of data such as binary, count and continuous data, spatial scan statistics for matched case-control data, which often arise in spatial epidemiology, have not been considered yet. In this paper, we propose two spatial scan statistics for matched case-control data. The proposed test statistics properly consider the correlations between matched pairs. We evaluate statistical power and cluster detection accuracy of the proposed methods through simulations comparing with the Bernoulli-based method. We illustrate the methods with the use of a real data example.

Keywords:

Spatial epidemiology, cluster detection, SaTScan, McNemar test, conditional logistic regression

Tuesday 12th 11:50 OGGB4 (260-073)

Whitebait In All Its Varieties: One Fish, Two Fish, Three, Four, Five Fish.

Bridget Armstrong
University of Canterbury

Abstract:

There are five species of fishes of the genus Galaxias that make up whitebait catches in New Zealand, although one species (G. maculatus) makes up >90% of the catch. Whitebait are immature post-larval fish that have yet to develop the distinctive morphological traits of adults. However, in their tiny stages as whitebait the five species are difficult to tell apart. There are also distinct spatial (rivers) and temporal (different months in the whitebait fishing season) differences among the species and even within species. To manage the fishery better it is necessary to identify regional differences in the species composition of catches, which is difficult because of the time and effort required to sample catches and identify species morphologically or genetically. In my study, I will use a recently compiled database comprising 17,000 entries of whitebait samples, species composition, and variability to develop a statistical model to predict the likelihood of species-to-species composition of catches throughout New Zealand. This probabilistic model could potentially be a powerful tool in the fishery and conservation of whitebait species, some of which are considered to be threatened.

Latent Variable Models And Multivariate Binomial Data

John Holmes
University of Otago

Abstract: A large body of work has been devoted to latent variable models applicable to multivariate binary data. However little work has been put into extending these models to cases where the observed data is multivariate binomial. In this paper, we will first show that models that use either a logit or probit link function, offer the same level of modelling flexibility in the binary case, but only the logit link fits into a data augmentation approach that compactly extends from binary to binomial. Secondly, we will demonstrate that multivariate binomial data provides greater flexibility in how the link function can be represented. Lastly, we will consider properties of the implied distribution of latent probabilities under a logit link.

Keywords: Multivariate binomial data, principal components/factor analysis, item response theory, link functions, logit-normal distributions

References:

(ed.) Bartholomew, D. J. and Knott, M. and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*. Chichester: John Wiley & Sons.

Johnson, N.L. (1949). Systems of Frequency Curves Generated by Methods of Translation. *Biometrika*, **36**, 149–276.

Polson, N. G. and Scott, J. G. and Windle, J. (2013). Bayesian inference for logistic models using Pólya-gamma latent variables. *Journal of the American Statistical Association*, **108**, 1339–1349.

Tuesday 12th 11:50 Case Room 2 (260-057)

Asking About Sex In General Health Surveys: Comparing The Methods And Findings Of The 2010 Health Survey For England With Those Of The Third National Survey Of Sexual Attitudes And Lifestyles

Philip Prah[1], Anne Johnson[2], Soazig Clifton[2], Jennifer Mindell[2], Andrew Copas[2], Chloe Robinson[3], Rachel Craig[3], Sarah Woodhall[2], Wendy Macdowall[4], Elizabeth Fuller[3], Bob Erens[2], Pam Sonnenberg[2], Kaye Wellings[4], Catherine Mercer[2], and Anthony Nardone[5]
[1]Auckland University of Technology
[2]University College London
[3]NatCen
[4]London School of Hygiene & Tropical Medicine
[5]Public Health England

Abstract:

Including questions about sexual health in the annual Health Survey for England (HSE) provides opportunities for regular measurement of key public health indicators, augmenting Britain's decennial National Survey of Sexual Attitudes and Lifestyles (Natsal). However, contextual and methodological differences may limit comparability of the findings. For instance both surveys used self-completion for administering sexual behaviour questions but this was via computer-assisted self-interview (CASI) in Natsal-3 and a pen-and-paper questionnaire in HSE 2010. We examine the extent of these differences between HSE 2010 and Natsal-3 (undertaken 2010-2012) and investigate their impact on parameter estimates. For inclusion to this study, we restricted participants to men and women in the 2010 HSE (n = 2,782 men and 3,588 women) and Natsal-3 (n = 4,882 men and 6,869 women) aged 16-69 years and resident in England. We compared their demographic characteristics, the amount of non-response to, and estimates from, sexual health questions. We used complex survey analysis to take into account stratification, clustering, and weighting of the data in each survey. Logistic regression was used to measure the extent to which sexual health estimates differ in HSE 2010 relative to Natsal-3, with multivariable models to adjust for significant demographic confounders. Additionally, investigated age-group interactions to see if differences between the surveys varied by age. The surveys achieved similar response rates, both around 60While a

relatively high response to sexual health questions in HSE 2010 demonstrates the feasibility of asking such questions in a general health survey, differences with Natsal-3 do exist. These are likely due to the HSE's context as a general health survey and methodological limitations such as its current use of pen-and-paper questionnaires.

Bayesian Continuous Space-Time Model Of Burglaries

Chaitanya Joshi, Paul Brown, and Stephen Joe
University of Waikato

Abstract: Building a predictive model of crime with good predictive accuracy has a great value in enabling efficient use of policing resources and reduction in crime. Building such models is not straightforward though due to the dynamic nature of the crime process. The crime not only evolves over both space and time, but is also related to several complex socio-economic factors, not all of which can be measured directly and accurately. The last decade or more has seen a surge in the effort to model crime more accurately. Many of the models developed so far have failed to capture the crime with a great degree of accuracy. The main reasons could be that all these models discretise the space using grid cells and that they are spatial, not spatio-temporal. We fit a log Gaussian Cox process model using the INLA-SPDE approach. This not only allows us to capture crime as a process continuous in both space and time, but also allows us to include socio-economic factors as well as the 'near repeat' phenomenon. In this talk, we will discuss the model building process and the accuracy achieved.

Keywords:

Bayesian spatio-temporal model, INLA-SPDE, predicting crime

Tolerance Limits For The Reliability Of Semiconductor Devices Using Longitudinal Data

Vera Hofer[1], Johannes Leitner[1], Horst Lewitschnig[2], and Thomas Nowak[1]
[1]University of Graz
[2]Infineon Technologies Austria AG

Abstract: Especially in the automotive industry, semiconductor devices are key components for the proper functioning of the entire vehicle. Therefore, issues concerning the reliability of these components are of crucial importance to manufacturers of semiconductor devices.

In this quality control task, we consider longitudinal data from high temperature operating life tests. Manufacturers then need to find appropriate tolerance limits for their final electrical product tests, such that the proper functioning of their devices is ensured. Based on these datasets, we compute tolerance limits that could then be used by automated test equipment for the ongoing quality control process. Devices with electrical parameters within their respective tolerance limits can successfully finish the production line, while all other devices will be discarded. In calculating these tolerance limits, our approach consists of two steps: First, the observed measurements are transformed in order to capture measurement biases and gauge repeatability and reproducibility. Then, in the second step, we compute tolerance limits based on a multivariate copula model with skew normal distributed margins. In order to solve the resulting optimization problem, we propose a new derivative-free optimization procedure.

The capability of the model is demonstrated by computing optimal tolerance limits for several drift patterns that are expected to cover a wide range of scenarios. Based on these computations, we show the resulting yield losses and analyze the performance of the tolerance limits a large simulation study.

Acknowledgment

This work was supported by the ECSEL Joint Undertaking under grant agreement No. 662133 - PowerBase. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and Austria, Belgium, Germany, Italy, Netherlands, Norway, Slovakia, Spain and United Kingdom.

Keywords:

quality control, tolerance limits, copulas, skew normal distribution

Tuesday 12th 16:00 098 Lecture Theatre (260-098)

Model-Checking For Regressions: A Local Smoothing-Based Global Smoothing Test

Lingzhu Li and Lixing Zhu
Hong Kong Baptist University

Abstract: As the two kinds of methods for model specification problem, local smoothing tests and global smoothing tests exhibit different characteristics. Compared with global smoothing tests, local smoothing tests can only detect local alternatives distinct from the null hypothesis at a much slower rate when the dimension of predictor vector is high, but can be more sensitive to high-frequency alternatives. We suggest a projection-based test that builds a bridge between the local and global smoothing methodologies to benefit from their own advantages. The test construction is based on a kernel estimation-based local smoothing method and the resulting test becomes a distance-based global smoothing test. A closed-form expression of the test statistic is derived and the asymptotic properties are investigated. Simulations and a real data analysis are conducted to evaluate the performance of the test in finite sample cases.

Keywords: Global smoothing test, projection-based methods, local smoothing test

References:

Zheng, J. X. (1996). *Journal of Econometrics: A consistent test of functional form via nonparametric estimation techniques*, **75(2)**, 263–289.

Bierens, H. J. (1982). *Journal of Econometrics: Consistent model specification tests*, **20**, 105-134.

Lavergne, P. and Patilea, V. (2012). *Journal of business & economic statistics: One for all and all for one: regression checks with many regressors.* **30(1)**, 41–52. Taylor & Francis Group.

Tuesday 12th 16:00 OGGB4 (260-073)

Breeding Value Estimation In Partially-Genotyped Populations

Alastair Lamont
University of Otago

Abstract: In livestock, a primary goal is the identification of individuals' breeding values - a measure of their genetic worth. This identification can be used to aid with selective breeding, but is non trivial due to how large data can be.

Measured traits are typically modelled as being caused by both breeding values and also environmental fixed effects. An efficient method for fitting this model was developed by Henderson (1984), based upon generalized least squares. This method could be applied to data where the pedigree - how each animal was related to one another - was fully known.

Improvements in technology have allowed the genetic information of an animal to be directly measured. These measurements can be taken very early in life, with the goal of informing selective breeding faster and more efficiently. Meuwissen (2001) adapted the standard model to incorporate genetic data, and additionally developed multiple fitting methods for this model.

Modern datasets are frequently only partially genotyped. The methods of Meuwissen cannot be used for these data, as they are only applicable to populations in which every individual is gentoyped. Modern fitting approaches aim to make use of the available genetic information without requiring all individuals be genotyped.

These approaches tend to either impute or average over missing genotype data, which can affect the overall accuracy of breeding value estimation. We are developing an alternative which instead incorporates missing data within the model, rather than having to adapt fitting approaches to accommodate it.

Preliminary results suggest that approaching fitting is this way can lead to improved accuracy of estimation in certain situations.

Tuesday 12th 16:00 OGGB5 (260-051)

BIVAS: A Scalable Bayesian Method For Bi-Level Variable Selection

Mingxuan Cai[1], Mingwei Dai[2], Jingsi Ming[1], Jin Liu[3], Can Yang[4], and Heng Peng[1]
[1]Hong Kong Baptist University
[2]Xi'an Jiaotong University
[3]Duke-NUS Medical School
[4]Hong Kong University of Science and Technology

Abstract:

In this paper we propose a bi-level variable selection approach, Bivas, for linear regression under the Bayesian framework. This model assumes that each variable is assigned to a pre-specified group where only a subset of the groups truly contribute to the response variable. Besides, within the active groups, there are only a small number of variables are important. A hierarchical formulation is adopted to mimic this pattern, where the spike-slab prior is put on both individual variable level and group level. A computationally efficient algorithm is developed using variational inference. Both simulation studies and real examples are analyzed, through which we illustrate the advantages of our method for both variable selection and parameter estimation under certain conditions.

Ranking Potential Shoplifters In Real Time

Barry McDonald
Massey University

Abstract: A company with a focus on retail crime prevention brought to MINZ (Mathematics in Industry in New Zealand) the task of "*Who is most likely to offend in my store, now*". The company supplied an anonymised set of data on incidents and offenders. The task, for the statisticians and mathematicians involved, was to try to find ways to use the data to nominate, say, the top ten likely offenders for any particular store and any particular time, using up-to-the-minute information (real time). The problem was analogous to finding a regression model when every row of data has response identically 1 (an incident), and for many places and times there is no data. This talk will describe how the problem was tackled.

Keywords:

Retail crime, ranking, ZINB, regression, real time

Two Stage Approach To Data-Driven Subgroup Identification In Clinical Trials

Toshio Shimokawa and Kensuke Tanioka
Wakayama Medical University

Abstract: A personalized medicine have been improved through the statistic analysis of Big data such as registry data. In these researches, subgroup identification analysis have been focused on. The purpose of the analysis is detecting subgroup such that the efficacy of the medical treatment is effective based on predictive factors for the treatment.

Foster et al., (2011) proposed the subgroup identification method based on two stage approach, called Virtual Twins (VT) method. In the first stage of VT, the difference of treatment effect between treatment group and control group is estimated by Random Forest. In the second stage, responders are identified by using CART, where the estimated these differences are set as the predictor variables.

However, the prediction accuracy of RandomForest tends to be lower than that of Boosting. Therefore, generalized boosted model (Ridgeway, 2006) is adopted in the first step. In addition to that, the number of rules tend to be large in the second step when CART is used. In this paper, we adopt a priori algorithm as the same way of SIDES(Lipkovich et al., 2011).

Keywords: A priori algorithm, boosting, personalized medicine

References:

Forster, J.C., Taylor, J.M.G and Ruberg, S.J. (2011). *Subgroup identification from randomized clinical trial data.* Stat.Med, **30**, 2867-2880.

Lipkovich, I., Dmitrienko, A., Denne, J. and Enas, G. (2011). *Subgroup identification based on differential effect search-recursive partitioning method for establishing response to treatment in patient subpopulations*. Stat.Med, **30**, 2601-2880.

Ridgeway, G. (2006).Gbm: Generalized boosted regression models. R package version 1.5-7. Available at http://www.i-pensieri.com/gregr/gbm.shtml.

Inverse Regression For Multivariate Functional Data

Ci-Ren Jiang[1] and Lu-Hung Chen[2]
[1]Academia Sinica
[2]National Chung Hsing University

Abstract: Inverse regression is an appearing dimension reduction method for regression models with multivariate covariates. Recently, it has been extended to the cases with functional or longitudinal covariates. However, the extensions focus on one functional/longitudinal covariate only. In this work, we extend functional inverse regression to the cases with multivariate functional covariates. The asymptotical properties of the proposed estimators are investigated. Simulation studies and data analysis are also provided to demonstrate the performance of our method.

Keywords:

Multidimensional/Multivariate Functional Data Analysis, Inverse Regression, Parallel Computing, Smoothing

Tuesday 12th 16:20 OGGB4 (260-073)

Including Covariate Estimation Error When Predicting Species Distributions: A Simulation Exercise Using Template Model Builder

Andrea Havron and Russell Millar
University of Auckland

Abstract: Ecological managers often require knowledge about species distributions across a spatial region in order to facilitate best management practices. Statistical models are frequently used to infer relationships between species observations (eg. presence, abundance, biomass, etc.) and environmental covariates in order to predict values at unobserved locations. Issues remain for situations where covariate information is not available for a predictive location. In these cases, spatial maps of covariates are often generated using tools such as kriging; however, the uncertainties from this statistical estimation are not carried through to the final species distribution map. New advances in spatial modelling using the automated differentiation software, Template Model Builder, allow both the spatial process of the environmental covariates and the observations to be modelled simultaneously by maximizing the marginal likelihood of the fixed effects with a Laplace approximation after integrating out the random spatial effects. This method allows for the uncertainty of the covariate estimation process to be included in the standard errors of final predictions as well as any derived quantities, such as total biomass for a spatial region. We intend to demonstrate this method and compare our predictions to those from a model where regional covariate information is supplied from a kriging model.

Keywords: spatial model, predicting covariates, Template Model Builder

References:

Kristensen, K.,Nielsen, A., Berg, C.W., Skuag, H. and Bell, B. (2015). TMB: Automatic Differentiation and Laplace Approximation. In: *Journal of Statistical Software*,**70**, 1–21.

Tuesday 12th 16:20 OGGB5 (260-051)

Adjusted Adaptive Index Model For Binary Response

Ke Wan[1], Kensuke Tanioka[1], Kun Yang[2], and Toshio Shimokawa[1]
[1]Wakayama Medical University
[2]Southwest Jiaotong University

Abstract: In questionnaire surveys, multiple regression analysis is usually used to evaluate influence factors. In addition to that, data mining methods such as Classification and Regression Trees (Breiman et al., 1984) are also used. In the research for tourism studies, it is difficult to contribute the policies for landscape or buildings from the results. In this paper, we call these factors " uncontrollable exploratory variables". On the other hands, the polices for amounts of garbages or inhabitant consciousness can be contributed from the results. We call these factors "controllable exploratory variables". The purpose of this report is grading for each subject which is conducted based on controllable exploratory variables with adjusting the effects of uncontrollable exploratory variables. Concretely, we modified the AIM method (Tian and Tibshirani, 2010) and conduct gradings based on the sum of the production rules for controllable exploratory variables with adjusting the effects of uncontrollable exploratory variables.

Keywords: logistic regression, production rule, grading

References:

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth.

Tian, L., and Tibshirani, R. (2011). *Adaptive index models for marker-based risk stratification.* Biostatistics, **12**, 68–86.

Factors Influencing On Growth Of Garments Industry In Bangladesh

Md. Shahidul Islam and Mohammad Sazzad Mosharrof
Auckland University of Technology

Abstract: If globalization provides the backdrop for drama, then the achievements of the garment industry in Bangladesh are indeed dramatic. The garment industry particularly has played a pioneering role in the development of industrial sector of Bangladesh and has grown rapidly for the last 15 years and now one of the largest garment exporters in the world. The study of our research has examined the successful development process of the Bangladesh garments industry and explored the keys to its success. In point of view we collected some primary and secondary data of garment manufacturers and traders to investigate further the key role and mechanism of technology transfers to operate a garment industry in Bangladesh. After that we apply some statistical models such as random effect model, tobit model and probit model to generate the performance of our variables. Also we use some dummy variables in case of different years for all the models. The result of our statistical models indicate that the high education of manufacturers and enterprise performance are highly significant. The only reason of this close relationship is that manufacturers have to upgrade their skills and they know how continuously in order to survive the intense competition in the world garment market and the high levels of the general human capital of the entrepreneur are needed to manage an increasing number of managers and experts. The result also shows the formal training that the garment entrepreneur has received in a foreign country and the entrepreneur's experience of working at a garments enterprise have small effect on growth of garment industry but not so much high. This is because those garment workers who had acquired skills and know-how but they could not helped smoothly new manufacturers and afford to start trading houses without good marketing and communication skills. But the traders who received formal training abroad have provided higher-valued services for manufacturers and contributed more to the proliferation of manufacturers. We have also found that foreign owned trading houses perform better than indigenous trading houses, which suggests that there still exist skills and know-how to be learned from foreign countries. So technology transfer seems to be a long-term process and its effect also seems last over the long term. Finally the key point of our findings strongly suggest that the performance of manufacturers and traders as well as production technologies are very potential for the high growth of industrial development. It has a great opportunity to earn a lot of foreign currency through developing garment industry and contribute economic development.

Keywords: Garments, Growth of Garment Industry, Performance of Manufacturers and Traders, Statistical Model

Tuesday 12th 16:20 Case Room 3 (260-055)

Comparison Of Exact And Approximate Testing Procedures In Clinical Trials With Multiple Binary Endpoints

Takuma Ishihara and Kouji Yamamoto
Osaka City University

Abstract: In confirmatory clinical trials, the efficacy of a test treatment are sometimes assessed by using multiple primary endpoints. We consider a trial in which the efficacy of a test treatment is confirmed only when it is superior to control for at least one of the endpoints and not clinically inferior for the remaining endpoints. Nakazuru et al. (2014) proposed a testing procedure that is applicable to the above case when endpoints are continuous variables. In this presentation, firstly, we propose a testing procedure in the case that all of the endpoints are binary.

Westfall and Troendle (2008) proposed multivariate permutation tests. Using this methods, we also propose an exact multiple testing procedure.

Finally, we compare an exact and approximate testing procedures proposed above. The performance of the proposed procedures was examined through Monte Carlo simulations.

Keywords: Clinical trial; Multivariate Bernoulli distribution; Non-inferiority; Superiority.

References:

Nakazuru, Y., Sozu, T., Hamada, C. and Yoshimura, I. (2014). A new procedure of one-sided test in clinical trials with multiple endpoints. *Japanese Journal of Biometrics,* **35**, 17-35.

Westfall PH and Troendle JF. (2008). Multiple testing with minimal assumptions. *Biometrical Journal,* **50(5)**, 745-755.

Multiple Function-On-Function Linear Regression With Application To Weather Forecast Calibration

Min-Chia Huang, Xin-Hua Wang, and Lu-Hung Chen
National Chung Hsing University

Abstract:

We suggest a direct approach to estimate the coefficient functions for function-on-function linear regression models. To avoid the risk of discarding useful information for regressions, the approach does not depend on basis representations or dimension reductions. It can accommodate for multiple functional responses and multiple functional predictors on different multidimensional domains, observed on dense or irregular sparse grids. We demonstrate the performances of the approach by simulation studies and a real application on calibrating numerical weather forecasts.

Modelling The Distribution Of Lifetime Using Compound Time-Homogenous Poisson Process

Kien Tran
Victoria University of Wellington

Abstract: Modelling the distribution of lifetime has traditionally been done by constructing a deterministic function for the survival function and/or force of mortality. This paper outlines previous research and presents the author's initial attempts to model the force of mortality and remaining lifetime using time-homogenous compound Poisson processes.

The paper presents two models. In model 1, the force of mortality of an individual is modelled as a random sum of i.i.d random variables (i.e. a compound Poisson process). In model 2, each individual is assumed to have an initial normally distributed innate lifetime, and their remaining life is a shifted compound Poisson process. In other words, we assume that there are random events coming at a constant rate modifying either the force of mortality or remaining lifetime of individuals. Simulations in R are then run to find the optimized parameters and the empirical survival function, force of mortality and distribution of lifetime are then constructed. Finally, these outputs are compared existing models and actual demographic data.

It turns out that for model 1, it is very difficult to model the force of mortality using a time-homogenous compound Poisson process without introducing additional complications such as the inclusion of event times. For model 2, however, if we allow the events to be a Cauchy random variable, then we can model the survival function of New Zealand population much better than several existing well-known specifications such as Weibull.

Keywords: Distribution of lifetime, force of mortality, survival function, time-homogenous compound Poisson process, innate lifetime, R simulation

References:

Khmaladze, E (2013). Statistical methods with application to demography and life insurance. CRC Press.

Weibull, W (1939). A statistical theory of the strength of materials. Generalstabens litografiska anstalts frlag, 1st edition.

Gompertz, B (1825). On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life. Philosophical Transactions of the Royal Society of London, 115, 513-583.

Detecting Change-Points In The Stress-Strength Reliability P(X<Y)

Hang Xu[1], Philip L.H. Yu[1], and Mayer Alvo[2]
[1]Unversity of Hong Kong
[2]University of Ottawa

Abstract: We address the statistical problem of detecting change-points in the stress-strength reliability R=P(X<Y)R=P(X<Y) in a sequence of paired variables (X,Y)(X,Y). Without specifying their underlying distributions, we embed this non-parametric problem into a parametric framework and apply the maximum likelihood method via a dynamic programming approach to determine the locations of the change-points in R. Under some mild conditions, we show the consistency and asymptotic properties of the procedure to locate the change-points. Simulation experiments reveal that in comparison with existing parametric and non-parametric change-point detection methods, our proposed method performs well in detecting both single and multiple change-points in R in terms of the accuracy of the location estimation and the computation time. It offers robust and effective detection capability without the need to specify the exact underling distribution of the variables. Applications to real data demonstrate the usefulness of our proposed methodology for detecting the change-points in the stress-strength reliability R.

Keywords:

Multiple change-points detection; Stress-strength model; Dynamic programming

New Zealand Crime And Victims Survey: Filling The Knowledge Gap

Andrew Butcher and Michael Slyuzberg
NZ Ministry of Justice

Abstract: The key objective of the Ministry of Justice is to ensure that New Zealand has a strong justice system that contributes to a safe and just society. To achieve this objective, the ministry and the wider Justice Sector need to know whether they are focusing their efforts in the right places and really making a difference. It is often difficult because we lack a crucial piece of information: how much crime is actually out there. Administrative data does not provide an answer as only about 30 The New Zealand Crime and Victims Survey (NZCVS) is introduced to fill this knowledge gap. The survey which is currently on the pilot phase was designed to meet the recommendations of Statistics New Zealand and key stakeholders' demand. It will interview about 8,000 of New Zealand residents aged from 15 years old and aims to: provide information about the extent (volumes and prevalence) and nature of crime and victimisation in New Zealand; provide geographical break-down of victimisation; provide extensive victims' demographics; measure how much crime gets reported to Police; understand the experiences of victims; measure crime trends in New Zealand.

The paper summarises the core requirements to NZCVS obtained from extended discussions with key stakeholders and describes key design features to be implemented in order to meet these requirements. These key requirements include, but are not limited to: Measuring the extent and nature of reported and unreported crime across New Zealand; Providing in-depth story-telling of victims' experiences; Providing frequent and timely information to support Investment Approach for Justice and wider decision making; Reducing information gaps by matching the NZCVS with administrative data in Statistics New Zealand's Integrated Data Infrastructure (IDI).

In particular, the paper discusses modular survey design which includes core crime and victimisation questions and revolving modules added annually, stratified random sampling, a new highly automated approach to offence coding through extended screening, measuring harm from being victimised, obtaining respondents' informed consent for data matching, use of survey data for extended analysis and forecasting and other important survey features.

Tuesday 12th 16:40 Case Room 3 (260-055)

Missing Data In Randomised Control Trials: Stepped Multiple Imputation

Rose Sisk and Alain Vandal
Auckland University of Technology

Abstract:

Missing data in Randomised Control Trials is usually unavoidable, but can present considerable issues to analysis in an Intention-to-Treat (ITT) setting. Multiple imputation is often regarded as the most appropriate method of handling missing data when compared with simpler methods such as complete case analysis and mean/mode imputation. However, in practice it can often be tricky to implement when working with large longitudinal datasets. The Sodium Lowering in Dialysate (SOLID) trial is a randomised control trial seeking to improve cardiovascular and other outcomes by lowering the dialysate concentration of sodium of patients on home haemodialysis. The trial contains 99 participants and over 30 primary and secondary outcomes. Missing data from various sources are present at baseline and at follow-up time points. Attempting to multiply impute a large number of outcomes, each measured at up to 4 follow-up times, proved to be a challenging task in this study. Several attempts to obtain sensible imputations were made but many of these failed due to the presence of highly correlated outcomes which were often missing together. This presentation discusses the approach taken to overcome this problem, which involved defining sets of outcomes to impute in various rounds, preventing sets of similar (highly correlated, missing together) outcomes being imputed in the same round. Once a round of imputation was completed, the next set of outcomes to be imputed was matched onto the completed dataset. This process is repeated until the full ITT dataset contains no missing values in any outcomes. We call this “stepped imputation”. Theory from mixed models was also applied to seek measures associated with the missingness mechanism, with the potential to include them in the final model to further reduce any possible bias resulting from missing data. Results from a simulation to test the validity of “stepped imputation” will be presented. In this simulation, an attempt is made to generate data related in a similar way to the outcomes in the SOLID trial. Results from the “gold standard” analysis with no missing data, and the complete case analysis is compared to the stepped imputation method.

Programme And Abstracts For Wednesday 13th Of December

Promoting Your R Package

Hadley Wickham
RStudio

Abstract: Your new statistical or data science tool is much more likely to be used if you provide it in a convenient form, like an R package. But how do people find out that your R package exists? I'll provide a comprehensive overview of the options, including creating excellent documentation (with roxygen2) and vignettes (with rmarkdown), creating a package website (with pkgdown), and promoting your work on social media.

Keywords:

R packages, websites

A Smoothing Filter Modelling Approach For Time Series

Marco Reale[1], Granville Tunnicliffe Wilson[2], and John Haywood[3]
[1]University of Canterbury
[2]Lancaster University
[3]Victoria University of Wellington

Abstract: We introduce different representations of a new model for time series based on repeated application of a filter to the original data. They can represent correlation structure to higher lags with fewer coefficients and they can provide a robust prediction at higher lead times.

Keywords:

Time series, smooting, parsimonious models

Online Learning For Bayesian Nonparametrics: Weakly Conjugate Approximation

Yongdai Kim[1], Kuhwan Jeong[1], Byungyup Kang[2], and Hyoju Chung[2]
[1]Seoul National University
[2]NAVER Corp.

Abstract: We propose a new online learning method for Bayesian nonparametric (BNP) models so called *weakly conjugate approximation* (WCA). We consider classes of BNP priors which are weakly conjugate. Here, 'weakly conjugate prior' means that the resulting posterior can be easily approximated by an efficient MCMC algorithm.

Suppose the whole data set is divided into two groups, say $\mathbf{x} = (\mathbf{x}^{old}, \mathbf{x}^{new})$. Then, the Bayes rule implies $p(\theta \mid \mathbf{x}) \propto p(\mathbf{x}^{new} \mid \theta) p(\theta \mid \mathbf{x}^{old})$, where θ is the parameter. WCA replaces $p(\theta \mid \mathbf{x}^{old})$ with $p^{wk}(\theta \mid \eta)$ where the proxy parameter η is estimated by minimizing the Kullback-Leibler (KL) divergence $\mathbb{E}_{p(\theta \mid \mathbf{x}^{old})} \left\{ \log p(\theta \mid \mathbf{x}^{old}) - \log p^{wk}(\theta \mid \eta) \right\}$. It can be easily approximated when we can generate samples from $p(\theta \mid \mathbf{x}^{old})$. To be more specific, suppose $\theta_1, \ldots, \theta_M$ are samples generated from $p(\theta \mid \mathbf{x}^{old})$. Then, we can estimate η by minimizing $\sum_{j=1}^{M} \left\{ \log p(\theta_j \mid \mathbf{x}^{old}) - \log p^{wk}(\theta_j \mid \eta) \right\} / M$.

To apply WCA for online learning with multiple batches, suppose the whole data $\mathbf{x}$ are divided into multiple small batches as $\mathbf{x} = (\mathbf{x}^{[1]}, \ldots, \mathbf{x}^{[S]})$. A WCA algorithm sequentially approximates $p(\theta \mid \mathbf{x}^{[1]}, \ldots, \mathbf{x}^{[s]})$ by $p^{wk}(\theta \mid \eta_s)$, where eta_s is the proxy parameter minimizing the approximated KL divergence. Since $p^{wk}(\theta \mid \eta)$ is weakly conjugate, we can easily generate samples from $p(\mathbf{x}^{[s]} \mid \theta) p^{wk}(\theta \mid \eta_{s-1})$, and hence easily update η_s.

We compare several online learning algorithms by analyzing simulated/real data sets in Dirichlet process mixture models and hierarchical Dirichlet processes topic models. The proposed method shows better accuracy in our experiments.

Keywords:

online learning, weakly conjugate approximation, Dirichlet process mixture model, hierarchical Dirichlet processes

Return to Programme

Wednesday 13th 10:30 Case Room 2 (260-057)

Improving The Production Cycle At Stats NZ With RStudio

Gareth Minshall and Chris Hansen
Stats NZ

Abstract: Stats NZ are looking to move away from the collection and publication of stand-alone surveys to making use of a wide range of data sources and estimation strategies. A key component to enabling this change is to develop the infrastructure which allows analysts to explore, test and use a range of tools which are not traditionally heavily used within National Statistics Offices. One of the tools Stats NZ is looking to make heavier use of is R. This talk will outline the development of internal RStudio and Shiny servers at Stats NZ, and give examples demonstrating the types of innovation RStudio has enabled at Stats NZ to improve the way we produce and disseminate statistics.

Keywords: Shiny, R Markdown, Official Statistics

Acknowledgement:

This work was supported by JSPS KAKENHI Grant Number JP16H02013.

A Max-Type Multivariate Two-Sample Baumgartner Statistic

Hidetoshi Murakami
Tokyo University of Science

Abstract: A multivariate two-sample testing problem is one of the most important topics in nonparametric statistics. Further, a max-type Baumgartner statistic based on the modified Baumgartner statistic (Murakami, 2006) was proposed by Murakami (2012) for testing the equality of two continuous distribution functions. In this paper, a max-type multivariate two-sample Baumgartner statistic is suggested based on the Jurečková and Kalina's ranks of distances (Jurečková and Kalina, 2012). Simulations are used to investigate the power of the suggested statistic for various population distributions. The results indicate that the proposed test statistic is more suitable than various existing statistics for testing a shift in the location, scale and location-scale parameters.

Keywords: Baumgartner statistic, Jurečková & Kalina's ranks of distances, Multivariate two-sample rank test, Power comparison

References:

Jurečková, J. and Kalina, J. (2012). Nonparametric multivariate rank tests and their unbiasedness. *Bernoulli*, **18**, 229–251.

Murakami, H. (2006). A kk-sample rank test based on the modified Baumgartner statistic and its power comparison. *Journal of the Japanese Society of Computational Statistics*, **19**, 1–13.

Murakami, H. (2012). A max-type Baumgartner statistic for the two-sample problem and its power comparison. *Journal of the Japanese Society of Computational Statistics*, **25**, 39–49.

Wednesday 13th 10:30 Case Room 4 (260-009)

Random Search Global Optimization Using Random Forests

Blair Robertson, Chris Price, and Marco Reale
University of Canterbury

Abstract: The purpose of a global optimization algorithm is to efficiently find an objective function’s global minimum. In this talk we consider bound constrained global optimization, where the search is performed in a box, denoted Ω\Omega. The global optimization problem is deceptively simple and it is usually difficult to find the global minimum. One of the difficulties is that there is often no way to verify that a local minimum is indeed the global minimum. If the objective function is convex, the local minimum is also the global minimum. However, many optimization problems are not convex. Of particular interest in this talk are objective functions that lack any special properties such as continuity, smoothness, or a Lipschitz constant.

A random search algorithm for bound constrained global optimization is presented. This algorithm alternates between partition and sampling phases. At each iteration, points sampled from Ω\Omega are classified low or high based on their objective function values. These classified points define training data that is used to partition Ω\Omega into low and high regions using a random forest. The objective function is then evaluated at a number of points drawn from the low region and from Ω\Omega itself. Drawing points from the low region focuses the search in areas where the objective function is known to be low. Sampling Ω\Omega reduces the risk of missing the global minimum and is necessary to establish convergence. The new points are then added to the existing training data and the method repeats.

A preliminary simulation study showed that alternating between random forest partition and sampling phases was an effective strategy for solving a variety of global optimization test problems. The authors are currently refining the method and extending the set of test problems.

Keywords: Bound constrained optimization, classification and regression trees (CART), stochastic optimization

Wednesday 13th 10:50 098 Lecture Theatre (260-098)

gridSVG: Then And Now

Paul Murrell
University of Auckland

Abstract: The **gridSVG** package[@RJ-2014-013] was first developed in 2003 to experiment with features of the SVG format that were not available through a normal R graphics device[@R], such as hyperlinks and animation. A number of different R packages[@rsvgtipsdevice; @cairo; @svglite; @svgannotation] have been developed since then to allow the generation of SVG output from R, but **gridSVG** has remained unique in its focus on generating structured and labelled SVG output. The reason for that was to maximise support for customisation and reuse, particularly unforseen reuse, of the SVG output. Unfortunately, there were two major problems: killer examples of customisation and reuse failed to materialise; and the production of SVG with **gridSVG** was painfully slow. In brief, **gridSVG** was a (sluggish) solution waiting for a problem. This talk charts some of the developments over time that have seen **gridSVG**’s patient wait for relevance ultimately rewarded and its desperate need for speed finally satisfied.

Keywords:

R, statistical graphics, SVG, accessibility

Probabilistic Outlier Detection And Visualization Of Smart Meter Data

Rob Hyndman
Monash University

Abstract: It is always a good idea to plot your data before fitting any models, making any predictions, or drawing any conclusions. But how do you actually plot data on thousands of smart meters, each comprising thousands of observations over time? We cannot simply produce time plots of the demand recorded at each meter, due to the sheer volume of data involved.

I will propose an approach in which each long series of demand data is converted to a single two-dimensional point that can be plotted in a simple scatterplot. In that way, all the meters can be seen in the scatterplot; so outliers can be detected, clustering can be observed, and any other interesting structure can be examined. To illustrate, I will use data collected during a smart metering trial conducted by the Commission for Energy Regulation (CER) in Ireland.

First we estimate the demand percentiles for each half hour of the week, giving us 336 probability distributions per household. Then, we compute the distances between pairs of households using the sum of Jensen–Shannon distances.

From these pairwise distances, we can compute a measure of the "typicality" of a specific household, by seeing how many similar houses are nearby. If there are many households with similar probability distributions, the typicality measure will be high. But if there are few similar households, the typicality measure will be low. This gives us a way of finding anomalies in the data set — they are the smart meters corresponding to the least typical households.

The pairwise distances between households can also be used to create a plot of all households together. Each of the household distributions can be thought of as a vector in KK-dimensional space where K=7×48×99=33,264. To easily visualize these, we need to project them onto a two-dimensional space. I propose using Laplacian eigenmaps which attempt to preserve the smallest distances — so the most similar points in KK-dimensional space are as close as possible in the two-dimensional space.

This way of plotting the data easily allows us to see the anomalies, to identify any clusters of observations in the data, and to examine any other structure that might exist.

Wednesday 13th 10:50 OGGB5 (260-051)

The Joint Models For Nonlinear Longitudinal And Time-To-Event Data Using Penalized Splines: A Bayesian Approach

Thi Thu Huong Pham, Darfiana Nur, and Alan Branford
Flinders University

Abstract: The joint models for longitudinal data and time-to-event data have been introduced to measure the association between longitudinal data and survival time in clinical, epidemiological and educational studies.. The main aim of this talk is to estimate the parameters in the joint models using a Bayesian approach for nonlinear longitudinal data and time-to-event data using penalized splines. To perform this analysis, the joint posterior distribution of hazard rate at baseline, survival and longitudinal coefficient and random effects parameters is first being introduced followed by derivation of the conditional posterior distributions for each of parameter. Based on these target posterior distributions, the samples of parameters are simulated using Metropolis, Metropolis Hastings and Gibbs sampler algorithms. An R program is written to implement the analysis. Finally, the prior sensitivity analysis for the baseline hazard rate and association parameters is performed following by simulations studies and a case study.

Keywords: Bayesian analysis, Joint models, Longitudinal data, MCMC algorithms, Prior sensitivity analysis, Survival data

References:

D. Rizopoulos, D. (2014). The R package JMbayes for fitting joint models for longitudinal and time-to- event data using MCMC. *Journal of Statistical Software,* 72(7):1 – 45.

Brown, E. R., J. G. Ibrahim, J. G., DeGruttola, V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival.*Biometrics,* 61(1):64 – 73.

Wednesday 13th 10:50 Case Room 2 (260-057)

R – A Powerful Analysis Tool To Improve Official Statistics In Romania

Nicoleta Caragea[1,2] and Antoniade Ciprian Alexandru[1,2]
[1]National Institute of Statistics
[2]Ecological University of Bucharest

Abstract: This presentation is focused on how R is used in Romanian official statistics to improve the quality of results provided by different statistical data sources on the base of administrative data. Some benefits for statistical analysis come when it is possible to link administrative records from different registers together, or when they can be linked with censuses or sample surveys. Many of these record linkage or matching methods must be done under statistically conditions, R program being one of the most powerful analysis tool. In Romania, there has been increasing attention in recent years to use R in official statistics, through specialized R courses for statisticians and training on the job sessions. A international conference on R (uRos) is yearly organized to provide a public forum for researchers from academia and institutes of statistics. It is also a continuous work to develop statistics based on Big Data, Romania being part of the ESSnet Big Data Project.

Keywords:

R package, data sources, statistics, matching method, linkage method

Wednesday 13th 10:50 Case Room 3 (260-055)

Simultaneous Test For Mean Vectors And Covariance Matrices In High-Dimensional Settings

Takahiro Nishiyama[1] and Masashi Hyodo[2]
[1]Senshu University
[2]Osaka Prefecture University

Abstract: Let ⍰g1,⍰g2,…,⍰gng\mathbf{X}_{g1}, \mathbf{X}_{g2}, \ldots, \mathbf{X}_{gn_g} be i.i.d. random samples of size ngn_g from a pp-dimensional population Πg\Pi_g (g∈{1,2}g \in \{1, 2\}) with $\mathrm{E}(\mathbf{X}_{gi})={\boldsymbol\mu}_g$ and ${}({}_{gi})=_g$ (i∈{1,…,ng}i \in \{1, \ldots ,n_g\}). In this talk, our primary interest is to test following hypothesis when p>min{n1−1,n2−1}p > \min\{n_1-1, n_2-1 \}: $$\begin{aligned} H_0 : {\boldsymbol\mu}_1 = {\boldsymbol\mu}_1,~ \Sigma_1 = \Sigma_2 \quad \mbox{vs.} \quad H_1 : \mbox{not}~ H_0. \end{aligned}$$

For this problem, we discuss an L2L^2-norm-based test for simultaneous testing of mean vectors and covariance matrices among two non-normal populations. To construct a test procedure, we propose a test statistic based on both unbiased estimator of differences mean vectors proposed by Chen and Qin (2010) and covariance matrices proposed by Li and Chen (2012). Also, we derive an asymptotic distribution of this test statistic and investigate the asymptotic sizes and powers of the proposed test. Finally, we study the finite sample and dimension performance of this test via Monte Carlo simulations.

Keywords: Asymptotic distribution, High-dimensional data analysis, Testing hypothesis

References:

Chen, S.X. and Qin, Y.L. (2010). A two-sample test for high dimensional data with applications to gene-set testing. *Ann. Statist.*, **38**, 808–835.

Li, J and Chen, S.X. (2012). Two sample tests for high-dimensional covariance matrices. *Ann. Statist.*, **40**, 908–940.

Wednesday 13th 10:50 Case Room 4 (260-009)

Dimension Reduction For Classification Of High-Dimensional Data By Stepwise SVM

Elizabeth Chou and Tzu-Wei Ko
National Chengchi University

Abstract: The purpose of this study is to build a simple and intuitive wrapper method, stepwise SVM, for reducing dimension and classification of large p small n datasets. The method employs a suboptimum search procedure to determine the best subset of variables for classification. The proposed method is compared with other dimension reduction methods, such as Pearson product moment correlation coefficient (PCCs), Recursive Feature Elimination based on Random Forest (RF-RFE), and Principal Component Analysis (PCA) by using five gene expression datasets. In this study, we show that stepwise SVM can effectively select the important variables and perform well in prediction. Moreover, the predictions of reduced datasets from stepwise SVM are better than that of the unreduced datasets. Compared with other methods, the performance of stepwise SVM is more stable than PCA and RF-RFE but it is difficult to tell the difference in performance from PCCs. In conclusion, stepwise SVM can effectively eliminate the noise in data and improve the prediction accuracy.

Keywords:

Stepwise SVM, Dimension reduction, Feature selection, High-dimension

Bringing Multimix From Fortran To R

Murray Jorgensen
Auckland University of Technology

Abstract: Multimix is the name for a class of multivariate finite mixture models designed with clustering (*unsupervised learning*) in mind. It is also a name for a program to fit these models, written in Fortran77 by Lyn Hunt as part of her Waikato PhD thesis.

Why convert to R? Although written in the 1990s Multimix is easy to convert to modern GNU Fortran (gfortran) but there are advantages to having an R version available. For users this means a simpler way of reading in the data and describing the form of the model. Also for ongoing development of improvement and modifications of the Multimix models. R's interactive environment provides a more comfortable place for experimentation. Designing the new program. Rather than attempt any sort of translation of the old code, the new R version of Multimix is designed from the beginning as an R program. In my talk I will describe some of the design decisions made and the reasons for them. A particular concern was that the R version be as fast as possible.

How to package up the new program? Two versions of Multimix in R have been developed, a *global* version with many global variables employed, and a *nested* version restricting the scope of variables to the surrounding function. The pluses and minuses of each approach will be described. I am conscious that I may not always have made the best design decisions and comments from others will be welcomed.

Keywords:

multivariate finite mixture models, clustering, package, global, local

Specification Of GARCH Model Under Asymmetric Error Innovations

Oyebimpe Adeniji, Olarenwaju Shittu, and Kazeeem Adepoju
University of Ibadan

Abstract: An empirical analysis of the mean return and conditional variance of Nigeria Stock Exchange (NSE) index is performed using various error innovations in GARCH models. Conventional GARCH model which assumed normal error term failed to capture volatility clustering, leptokurtosis and leverage effect as a result of zero skewness and kurtosis respectively. We re-modify error distributions of GARCH (p,q) model inference using some thick-tailed distributions. Method of Quasi-Maximum Likelihood Estimation (MLE) was used in parameter estimation. The robust model that explained the NSE index is determined by loglikelihood and model selection Criteria. Our result shows that GARCH model with fat-tailed densities improves overall estimation for measuring conditional variance. The GARCH model using Beta-Skewed-t distribution is the most successful for forecasting NSE index.

Keywords:

GARCH, Nigeria stock index, Maximum Lilkelihood Estimation (MLE), Beta Skewed -t distributions

Wednesday 13th 11:10 OGGB5 (260-051)

Performance Of Bayesian Credible Interval For Binomial Proportion Using Logit Transformation

Toru Ogura[1] and Takemi Yanagimoto[2]
[1]Mie University Hospital
[2]Institute of Statistical Mathematics

Abstract: The confidence or the credible interval of the binomial proportion pp is one of most widely employed statistical analysis methods, and a variety of methods have been proposed. The Bayesian credible interval attracts recent researches' attentions. One of the promising methods is the highest posterior density (HPD) interval, which implies the shortest possible interval enclosing 100(1−α)100(1-\alpha)% of the probability density function. The HPD interval is often used because it is narrow compared to other credible intervals. However, the HPD interval has some drawbacks when the binomial proportion is a small. To dissolve them, we calculate first a credible interval by the HPD interval of the logit transformed parameter, θ=log{p/(1−p)}\theta=\log\{p/(1-p)\}, instead of pp. Note that θ\theta and pp are the canonical and the mean parameters of the binomial distribution in the exponential family, respectively. Writing the HPD interval of θ\theta as (θl,θu)(\theta_{l}, \theta_{u}), we define the proposed credible interval of pp as (pl,pu)=(eθl/(1+eθl),eθu/(1+eθu))(p_{l}, p_{u})= \big(e^{\theta_{l}} / (1+e^{\theta_{l}}), \, e^{\theta_{u}}/(1+e^{\theta_{u}}) \big). It is explored in depth, and numerical comparison studies are conducted to confirm its favorable performance, especially when the observed number is small, such as 0 or 1. Practical datasets are analyzed to examine the potential usefulness for applications in medical fields.

Keywords: Bayesian credible interval, binomial proportion, highest posterior density interval, logit transformation, zero count

References:

Newcombe, R.G. (2012). *Confidence Intervals for Proportions and Related Measures of Effect Size*. Florida: Chapman and Hall/CRC.

Wednesday 13th 11:10 Case Room 2 (260-057)

Statistical Disclosure Control With R: Traditional Methods And Synthetic Data

Matthias Templ
Zurich University of Applied Sciences

Abstract: The demand for and volume of data from surveys, registers or other sources containing sensible information on persons or enterprises have increased significantly over the last several years. At the same time, privacy protection principles and regulations have imposed restrictions on the access and use of individual data. Proper and secure microdata dissemination calls for the application of statistical disclosure control methods to the data before release. Traditional approaches to (micro)data anonymization, including data perturbation methods, disclosure risk methods, data utility and methods for simulating synthetic data have been made available in R. After introducing the audience to the R packages sdcMicro and simPop, the presentation will focus on new developments and research for generating close-to-reality synthetic data sets using specific model-based approaches. The resulting data can work as a proxy of real-world data and they are useful for training purposes, agent-based and/or microsimulation experiments, remote execution as well as they can be provided as public-use files. The strength and weakness of the methods are highlighted and an (brief) application to the Euorpean Statistics of Income and Living Condition Survey is given.

Keywords: Statistical Disclosure Control, Anonymization, Disclosure Risk, Synthetic Data

References:

Templ, M. (2017). *Statistical Disclosure Control for Microdata. Methods and Applications in R*, Springer International Publishing. doi:10.1007/978-3-319-50272-4

Templ, M., Kowarik, A., Meindl, B. (2015). Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. *Journal of Statistical Software*, 67(4), 1-36. doi:10.18637/jss.v067.i04

Templ, M., Kowarik, A., Meindl, B., Dupriez, O. (2017). Simulation of Synthetic Complex Data: The R Package simPop. *Journal of Statistical Software*, 79(10), 1-38. doi:10.18637/jss.v079.i10

Wednesday 13th 11:10 Case Room 3 (260-055)

High Dimensional Asymptotics For The Naive Canonical Correlation Coefficient

Mitsuru Tamatani[1] and Kanta Naito[2]
[1]Doshisha University
[2]Shimane University

Abstract: In this talk we investigate the asymptotic behavior of the estimated naive canonical correlation coefficient under the normality assumption and High Dimension Low Sample Size (HDLSS) settings. In general, canonical correlation matrix is associated with canonical correlation analysis which is useful in studying the relationship between two sets of variables. However, in HDLSS settings, the within-class sample covariance matrix Σ^\hat{\Sigma} is singular, because the rank of Σ^\hat{\Sigma} is much less than the number of dimension. To avoid the singularity of Σ^\hat{\Sigma} in HDLSS settings, we utilize the naive canonical correlation matrix with replacing sample covariance matrix by its diagonal part only. We derive the asymptotic normality of the estimated naive canonical correlation coefficient, and compare the results of our numerical studies to the theoretical asymptotic results.

Keywords: High dimension low sample size, Naive canonical correlation coefficient, Asymptotic normality

References:

Tamatani, M., Koch, I. and Naito, K. (2012). *Journal of Multivariate Analysis*, **111**, 350–367.

Srivastava, M. S. (2011). *Journal of Multivariate Analysis*, **102**, 1190–1103.

Fan, J. and Fan, Y. (2008). *The Annals of Statistics*, **36**, 2605–2637.

Deep Learning High-Dimensional Covariance Matrices

Philip Yu and Yaohua Tang
University of Hong Kong

Abstract: Modeling and forecasting covariance matrices of asset returns play a crucial role in finance. The availability of high frequency intraday data enables the modeling of the realized covariance matrix directly. However, most models in the literature depend on strong structural assumptions and they also suffer from the curse of dimensionality. To solve the problem, we propose a deep learning model which treats each realized covariance matrix as an image. The network structure is designed with simplicity in mind, and yet provides superior accuracy compared with several advanced statistical methods. The model could handle both low-dimensional and high-dimensional realized covariance matrices.

Keywords: Deep learning, Realized covariance matrix, Convolutional neural network

References:

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 86, 2278–2324.

Shen, K., Yao, J. and Li, W. K.(2015). Forecasting High-Dimensional Realized Volatility Matrices Using A Factor Model. *ArXiv e-prints*.

Tao, M., Wang, Y., Yao, Q. and Zou, J. (2011). Large volatility matrix inference via combining low-frequency and high-frequency approaches. *Journal of the American Statistical Association*, 106, 1025–1040.

Wednesday 13th 11:30 098 Lecture Theatre (260-098)

R In Industry – Application On Pipe Renewal Planning

Glenn Thomas
Harmonic Analytics

Abstract: R has become an increasingly used tool in industry to practically help councils and organisations with their asset management challenges. We will demonstrate some of the practical tools Harmonic Analytics has developed using R to assist in asset management.

One specific example demonstrated will be recent work for a New Zealand council that was experiencing challenges in long term planning around its three waters infrastructure. In particular, challenges stem from the limited information about pipe condition. Using past work order history as proxy for pipe failures, we present a tool that uses a pipe break model to inform replacement strategies. The developed tool allows users to generate and compare both data driven and engineering based scenarios through a variety of lenses, ranging from annual replacement length to service level outcomes. A number of visualisations are available to support comparisons. Data driven scenarios are driven from a variety of perspectives, such as traditional age based replacement, probability of failure and minimising the expected number of pipe breaks across the network.

This kind of work is an exciting step forward, as councils show interest in collaboration and pooling data to improve accuracy.

Empirical Comparison Of Some Algorithms For Automatic Univariate ARMA Modeling Using RcmdrPlugin.SPSS

Dedi Rosadi
Universitas Gadjah Mada

Abstract: In some application of time series modeling, it is necessary to obtain forecast of various types of data automatically and possibly, in real-time. For instances, to forecast large number of univariate series every day, or to do a real-time processing of the satellite data. Various automatic algorithms for modeling ARMA models are available in the literature, where here we will discuss three methods in particular. One of the method is based on a combination between the best exponential smoothing model to obtain the forecast, together with state-space approach of the underlying model to obtain the prediction interval (see Hyndman, 2007). The second method, which is more advanced method, is based on X-13-ARIMA-SEATS, the seasonal adjustment software by the US Census Bureau (see Sax , 2015). From our previous study in Rosadi (2016), we found that these methods are perform relatively well for SARIMA data. Unfortunately, these approaches do not working well for many of ARMA data. Therefore in paper we extend the study by considering an automatic modeling method based on genetic algorithm approach (see Abo-Hammour, et.al., 2012). These approaches are implemented in our R-GUI package RcmdrPlugin.Econometrics which now already integrated in our new and more comprehensive R-GUI package, namely RcmdrPlugin.SPSS. We provide application of the methods and the tool. From some empirical studies, we found that for ARMA data, the method based on genetic algorithm performs better than the other approaches.

Keywords: Automatic ARMA modeling, genetic algorithm, exponential smoothing, X-13-ARIMA, R-GUI

References:

Abo-Hammour, Z. E. S., Alsmadi, O. M., Al-Smadi, A. M., Zaqout, M. I., & Saraireh, M. S. (2012). ARMA model order and parameter estimation using genetic algorithms. *Mathematical and Computer Modelling of Dynamical Systems*, **18(2)**, 201–221.

Hyndman, R. J. (2007). forecast: Forecasting functions for time series, R package version 1.05. URL: http://www.robhyndman.info/Rlibrary/forecast/.

Sax, C. (2015). Introduction to seasonal: R interface to X-13ARIMA-SEATS, https://cran.r-project.org/web/packages/seasonal/vignettes/seas.pdf.

Rosadi, D. (2016). Automatic ARIMA Modeling using RcmdrPlugin.SPSS, Presented in *COMPSTAT 2016*, Oviedo, Spain, 23-26 August 2016.

Bayesian Optimum Warranty Length Under Type-II Unified Hybrid Censoring Scheme

Tanmay Sen[1], Biswabrata Pradhan[2], Yogesh Mani Tripathi[1], and Ritwik Bhattacharya[3]
[1]Indian Institute of Technology Patna
[2]Indian Statistical Institute Kolkata
[3]Centro de Investigacionen Matematicas

Abstract: This work considers determination of optimum warranty length under Type-II unified hybrid censoring scheme. Consumers are willing to purchase a highly reliable product with certain cost constraint. To assure the product reliability and also to remain profitable, the manufacturer provides warranties on product lifetime. Moreover, censoredlifetime data are available in practice, to assess the reliability of the product. Therefore, determination of an appropriate warranty length based on censored lifetime data is an important issue to the manufacturer. It is assumed that the lifetime follows a lognormal distribution. We consider a combine free replacement and pro-rata warranty policy (FRW/PRW). The life test is conducted under Type-II unified hybrid censoring scheme. The warranty length is obtained by maximizing an expected utility function.The expectation is taken with respect to the posterior predictive model for time to failure given the available data obtained under Type-II unified hybrid censoring scheme. A real data set is analyzed to illustrate the proposed methodology. We propose a non-linear prorate warranty policy and compare them with linear warranty policy. It is observed that non-linear prorate warranty policy give larger warranty length with maximum profit

Keywords:

Lognormal distribution, FRW/PRW policies, Optimum warranty length, MH algorithm

Imputation Of The 2016 Economic Census For Business Activity In Japan

Kazumi Wada[1], Hiroe Tsubaki[2], Yukako Toko[1], and Hidemine Sekino[3]
[1]National Statistics Center
[2]Institute of Statistical Mathematics
[3]The Statistics Bureau

Abstract: R has been used in the field of official statistics in Japan for over ten years. This presentation takes up the case of the 2016 Economic Census for Business Activity. The Census aims to identify the structure of establishments and enterprises in all industries on a national and regional level, and to obtain basic information to conduct various statistical surveys by investigating the economic activity of these establishments and enterprises. The major corporate accounting items, such as sales, expenses and salaries, surveyed by the census require imputation to avoid bias. Although ratio imputation is a leading candidate, it is well known that the ratio estimator is very sensitive to outliers; therefore, we need to take appropriate measures for this problem.

Ratio imputation is a special case of regression imputation; however, the conventional ratio estimator has a heteroscedastic error term, which is the obstacle of robustification by means of M-estimation. New robust ratio estimators are developed by segregating the homoscedastic error term with no relation to the auxiliary variable from the original error. The computation of the estimators are made by modifying iterative reweighted least squares (IRLS) algorithm, since it is easy to calculate and fast to converge. The proposed robustified ratio estimator broadens the conventional definition of the ratio estimator with regards to the variance of the error term in addition to effectively alleviating the influence of outliers. The application of the robust estimator is expected to contribute to the accuracy of the Census results.

An random number simulation to confirm the characteristics of these estimators, deciding imputation domains by CART (classification and regression tree), model selection and preparing necessary rates by domain for the census data processing are conducted within the R programming environment.

Keywords: GNU R, Outlier, Iteratively reweighted least squares, Ratio estimator, Official statistics

Acknowledgement:

This work was supported by JSPS KAKENHI Grant Number JP16H02013.

Applying Active Learning Procedure To Drug Consumption Data

Yuan-Chin Chang
Academia Sinica

Abstract: We apply the method of active learning to build a binary classification model for drug consumption data. Due to the nature of active learning, subject selection is an major issue is its learning process. There are many kinds of subject selection schemes proposed in the literature. The subject recruiting procedure may also depend on its learning target criterion such as accuracy, area under ROC curve and so on. Moreover, in practical active learning scenarios, the label information of samples can only be revealed as they are recruited into training data set, and we will pay the domain experts to label these selected sample. Therefore, to consider the labelling cost, how/when to stop an active learning procedure is always an important and challenging problem in active learning. In this talk, we propose an active learning procedure targeting at area under an ROC curve, and based on the idea of robustness, we then used a modified influential index to locate the most informative samples, sequentially, such that the learning procedure can achieve the target efficiently. We then apply our procedure to drug consumption data sets.

Keywords: ROC curve, area under curve, active learning, influential index

References:

Calders, T. and Jaroszewicz, S. (2007). Efficient auc optimization for classification. In *Knowledge Discovery in Databases: PKDD 2007*, pages 42–53. Springer.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. , 69(346):383–393.

R For Everything

Jared Lander
Lander Analytics

Abstract: Everyone knows I love R. So much that I never want to leave the friendly environs of R and RStudio. Want to download a file? Use download.file. Want to create a directory? Use dir.create. Sending an email? gmailr. Using Git? git2r. Building this slideshow? rmarkdown. Writing a book? knitr. Let's take a look at everyday activities that can be done in R.

Keywords: R, RMarkdown, knitr, email, football, git, download, data, plotting, modeling, logistic regression

References:

Lander, J. (2017). *R for Everyone, Second Edition.* New York: Addison-Wesley.

R Package For New Two-Stage Methods In Forecasting Time Series With Multiple Seasonality

Anupama Lakshmanan and Shubhabrata Das
Indian Institute of Management Bangalore

Abstract: Complex multiple seasonality is an important emerging challenge in time series forecasting. We propose a framework that segregates the task into two stages. In the first stage, the time series is aggregated at the low frequency level (such as daily or weekly) and suitable methods such as regression, ARIMA or TBATS, are used to fit this lower frequency data. In the second stage, additive or multiplicative seasonality at the higher frequency levels may be estimated using classical, or function-based methods. Finally, the estimates from the two stages are combined.

In this work, we build a package for implementing the above two-stage framework for modeling time series with multiple levels of seasonality within R. This would make it convenient to execute and possibly lead to more practitioners and academicians adopting it. The package would allow the user to decide the specific methods to be used in the two stages and also the separation between high and low frequency. Errors are calculated for both model and validation period, which may be selected by the user and model selection choices based on different criterion will be facilitated. Forecast combination may also be integrated with the developed routine. The schematics will be presented along with demonstration of the package in several real data sets.

Keywords:

Additive seasonality, ARIMA, forecast combination, high frequency, low frequency, multiplicative seasonality, polynomial seasonality, regression, TBATS, trigonometric seasonality

Analysis Of Official Microdata Using Secure Statistical Computation System

Kiyomi Shirakawa[1,3], Koji Chida[2], Satoshi Takahashi[2], Satoshi Tanaka[2], Ryo Kikuchi[2], and Dai Ikarashi[2]
[1]National Statistics Center
[2]NTT
[3]Hitotsubashi University

Abstract: We introduce some important functions on a secure computation system and empirically evaluate them using the statistical computing software R. The secure computation is a cryptographic technology that enables us to operate data while keeping the data encrypted. Due to the remarkable aspect, we can construct a secure on-line analytical system to protect against unauthorized access, computer virus and internal fraud. Moreover, the function of secure computation has a benefit for privacy.

So far, we developed a secure computation system that runs R as a front-end application. In this research, we focus on the analysis of official microdata using our secure computation system. By employing the R script language to secure computation, we can potentially make new functions for the analysis of official microdata on our secure computation system. We show some examples of functions on the system using the R script language. A demonstration experiment to verify the practicality and scalability of the system in the field of official statistics is also in our scope.

Keywords:

Secure Computation, Security, Privacy, Big Data, Official Statistics, R

Presenting Flexi, A Statistical Program For Fitting Variance Models

Martin Upsdell
AgResearch

Abstract:

Flexi is a statistical program designed to fit variance based models. In this talk I will explore the advantages and disadvantages of the variance based model compared to the more commonly adopted mean based approach. Several examples will be given where the properties of variance based models provide a clearer understanding of the data. To illustrate the differences in the approach to the data I will compare Television and Progressive Graphics File methods of transferring a picture. The Television builds up the global picture from individual pixels describing a local area of the picture, whereas the Progressive Graphics File proceeds from the global value of the median colour of the whole picture to the local value of each individual pixel by successive refinements. This gives a coarse blocky picture at the start which refines into a detailed picture at the end. Mean based models are like television pictures whereas variance based models are like Progressive Graphics File pictures. The advantages and disadvantages of the two methods will be discussed.

Space And Circular Time Log Gaussian Cox Processes With Application To Crime Event Data

Alan Gelfand
Duke University

Abstract: We view the locations and times of a collection of crime events as a space-time point pattern modeled as either a nonhomogeneous Poisson process or a more general log Gaussian Cox process. We need to specify a space-time intensity. Viewing time as circular, necessitates a valid separable and nonseparable covariance functions over a bounded spatial region crossed with circular time. Additionally, crimes are classified by crime type and each crime event is marked by day of the year which we convert to day of the week.

We present marked point pattern models to accommodate such data. Our specifications take the form of hierarchical models which we fit within a Bayesian framework. We consider model comparison between the nonhomogeneous Poisson process and the log Gaussian Cox process as well as separable vs. nonseparable covariance specifications. Our motivating dataset is a collection of crime events for the city of San Francisco during the year 2012.

Wednesday 13th 14:10 098 Lecture Theatre (260-098)

Cluster-Wise Regression Models Combined By A Quasi-Linear Function

Kenichi Hayashi[1], Katsuhiro Omae[2], and Shinto Eguchi[3]
[1]Keio University
[2]The Graduate University for Advanced Studies
[3]Institute of Statistical Mathematics

Abstract: Suppose that there are multiple heterogeneous subgroups in a dataset. In the "Big data" era, this would be a natural assumption for many fields of application such as medicine, biology, marketing, psychology, etc. Then, conventional linear regression models result in not only poor prediction performance but also misleading interpretation of analyses. In this study, we propose an extension of cluster-wise regression models ϕ−1(∑k=1Kpk(x⃗)ϕ(β⃗k⊤x⃗))\phi^{-1}\left(\sum_{k=1}^Kp_k(\vec{x})\phi(\vec{\beta}_k^\top\vec{x})\right), where ϕ\phi is a strictly increasing function, x⃗∈ℝd\vec{x}\in\mathbb{R}^d, β⃗k\vec{\beta}_k is a regression coefficient for kkth cluster and pk(x⃗)p_k(\vec{x}) is a non-negative function satisfying ∑k=1Kpk(x⃗)=1\sum_{k=1}^Kp_k(\vec{x})=1 for any x⃗\vec{x}. We show that the proposed model has flexibility in clustering and "averaging" multiple regressors and hence includes the existing methods such as Späth (1981), DeSarbo et al. (1989) as special cases.

Keywords: Cluster-wise regression, Generalized linear models, Population heterogeneity

References:

DeSarbo, W.S., Oliver, R.L., and Rangaswamy, A. (1989). A simulated annealing methodology for clusterwise linear regression. *Psychometrika*, **54**, 707–736.

Späth, H. (1979). Algorithm 39: Clusterwise linear regression. *Computing*, **22**, 367–373.

Wednesday 13th 14:10 Case Room 1 (260-005)

Hierarchical Structural Component Analysis Of Gene-Environment Interactions

Sungkyoung Choi[1], Seungyeoun Lee[2], and Taesung Park[3]
[1]Yonsei University
[2]Sejong University
[3]Seoul National University

Abstract: Gene-environment interactions (GEI) are known to be one possible avenue for addressing the missing heritability problem in genome-wide association studies (GWAS). Although many statistical methods have been proposed for identifying and analyzing GEI, most of these consider interactions between a single genetic variants such as single nucleotide polymorphism (SNPs) by the environment. In this study, we proposed a new statistical method for gene-based GEI analysis, Hierarchical structural CoMponent analysis of Gene-Environment Interaction (HisCoM-GEI). HisCoM-GEI is based on generalized structured component analysis, and can consider hierarchical structural relationships among SNPs in a gene. HisCoM-GEI can effectively aggregate all possible pairwise SNP-Environment interactions into a latent variable by imposing a ridge penalty, from which it then performs GEI analysis. Furthermore, HisCoM-GEI can evaluate both gene-level and SNP-level analyses. We applied the HisCoM-GEI to the cohort data of the Korea Associated Resource (KARE) consortium to identify GEIs between genes and alcohol intake on the blood pressure traits.

Keywords:

Gene-environment interaction, SNP, gene, GWAS

Wavelet-Based Power Transformation Of Non-Gaussian Long Memory Time Series

Kyungduk Ko[1] and Chul Eung Kim[2]
[1]Boise State University
[2]Yonsei University

Abstract: We consider a power transformation through the well-known Box-cox transformation to induce normality from non-Gaussian long memory processes and propose a Bayesian method to simultaneously estimate the transformation parameter and long memory parameter. To ease computational burdens due to the dense variance-covariance matrix of long memory time series, we base our statistical inference on the wavelet domain rather than the original data domain. For a joint estimation of the parameters of interest, posterior estimations are carried out via Markov chain Monte Carlo (MCMC). An application to German stock return data is presented.

Keywords: Box-Cox transformation, Discrete wavelet transform, Long memory, MCMC, Normality

References:

Dahlhaus, R. (1990). Efficient location and regression estimation for long range dependent regression models. *Annuals of Statistics*, 23, 1029–1047.

Ko, K. and Lee, J. (2008). Confidence intervals for long memory regressions. *Statistics and Probability Letters*, 78, 1894–1902.

Lee, J. and Ko, K. (2007). One-way analysis of variance with long memory errors and its application to stock return data. *Applied Stochastic Models in Business and Industry*, **23**, 493–502.

Wednesday 13th 14:10 Case Room 2 (260-057)

Cross Covariance Estimation For Integration Of Multi-Omics Data

Johan Lim[1], Hiromi Koh[2], and Hyungwon Choi[2]
[1]Seoul National University
[2]National University of Singapore

Abstract: In integrative analysis of multiple types of -omics data, it is often of interest to infer associations between two different types of molecules. The prevailing analysis methods depend on ensemble of brute-force pairwise univariate association tests between the two types, best exemplified by expression quantitative loci (eQTL) mapping. In a nutshell, this problem can be generally framed as a sparse cross-covariance matrix. In this work, we propose a two-stage estimator of cross covariance matrix $\mathbf{\Sigma}_{\mathbf{XY}}$ between pp-vector $\mathbf{X}$ and qq-vector $\mathbf{Y}$, assuming that the two variables have regulatory relationships and that we know a group structure in the variables in $\mathbf{X}$. We first decompose the covariance matrix of $\mathbf{X}$, $\mathbf{\Sigma_{XX}}$, into systematic covariance consistent with the functional group information (G) $\mathbf{\Sigma}_{\mathbf{XX}}{}^{G}$ and the residual covariance not explained by the group information ���(R){\mathbf{\Sigma_{XX}}}^{(R)}. Following this decomposition, we estimate the cross covariance matrix by multi-response group lasso, yielding ���=����(p×q)=(���(G)+���(R))�(p×q){\mathbf{\Sigma_{XY}}}= {\mathbf{\Sigma_{XX}}}{\mathbf{B}}_{(p \times q)} = ({\mathbf{\Sigma_{XX}}}^{(G)} + {\mathbf{\Sigma_{XX}}}^{(R)}) {\mathbf{B}}_{(p \times q)}. As a result of this decomposition, ���{\mathbf{\Sigma_{XY}}} can also be expressed as the sum of a systematic term and a residual term, breaking down the cross-covariance into a fraction attributable to pathway-level regulation and the rest. We applied the method to epigenetic regulation analysis of mRNA expression by DNA methylation in the The Cancer Genome Atlas invasive breast cancer cohort.

Keywords: Cross covariance matrix, data integration.

References:

Simon, N., Friedman, J. and Hastie, T. (2013). *A Blockwise Descent Algorithm for Group-penalized Multiresponse and Multinomial Regression*

Koboldt,D.C. and others. (2012). *Nature*, **490**, 61-70.

Wednesday 13th 14:10 Case Room 3 (260-055)

Relationships Between Linguistic Characteristics And The Use Of Māori Loanwords In New Zealand English.

Steven Miller and Andreea Calude
University of Waikato

Abstract: We present the initial results from a project looking at the linguistic and socio-linguistic characteristics that affect the prevalence of Māori loanwords in the use of New Zealand English, and describe the paths we see this research taking in the next few years.

Loanwords are words that originate in one language (the donor language) and enter into, and are productively used within another language (the host language). For our initial research, we were particularly interested in the use of Māori loanwords in spoken New Zealand English, as found within the Wellington Corpus of Spoken New Zealand English.

We used generalised linear mixed effects models to determine if there were significant relationships between the linguistic characteristics of the loanwords used / words replaced, demographic features of the speakers, and the ethnicity of the audiences.

We found that linguistic characteristics of the loanwords and their English counterparts affect the probability of using the loanword for both Pākehā and Māori speakers, there was a difference in the probability of using a loanword between the sexes for Māori speakers only, and Māori speakers moderated the use of loanwords in conversations depending on the ethnicity of their audience.

We will briefly describe the next phase of the research that will use network modelling to characterise the use of Māori loanwords in written media.

Keywords:

Linguistics, loanwords, Māori, GLMM

Wednesday 13th 14:10 Case Room 4 (260-009)

Transfer Regression And Predictive Distributions

Shigetoshi Hosaka[1] and Jinfang Wang[2]
[1]Hosaka Clinic of Internal Medicine
[2]Chiba University

Abstract: We introduce the *transfer regression*, a method for constructing prior distributions for parameters defining generalized linear models (GLM). The transfer regressions are based frequency tables, usually obtained by categorizing the continuous variables. So obtained prior information are transferred to the parameters defining the second stage GLM based on detailed data. at the second stage based on more detailed data. We illustrate these ideas by showing how to compute posterior predictive probabilities of contracting diabetes based on HbA1c data obtained from comprehensive medical examinations.

Keywords: Bayesian generalized linear models, Markov chain Monte Carlo methods, posterior predictive distributions

References:

Andrew D. Martin, Kevin M. Quinn and Jong Hee Park (2011). MCMCpack: Markov Chain Monte Carlo in R, *Journal of Statistical Software*, **42**, 1–21.

Kass, R. E. and Wasserman, L. (1996). The Selection of Prior Distributions by Formal Rules, *Journal of the American Statistical Association*, **91**, 1343–1370.

Wang, J. and Hosaka, S. (2017). Cell regression and reference prior, Symposium on "*Statistical Modelling and Computational Algorithms*", at Nagoya University, Febuary 18–19, 2017.

Wednesday 13th 14:30 098 Lecture Theatre (260-098)

An Overview Of The Correspondence Analysis Family

Eric Beh
University of Newcastle

Abstract: Correspondence analysis (CA) is well known to be a member of the family of multivariate analysis techniques and is concerned with the visualisation of the association between two or more categorical variables. The classic texts of Greenacre (1984) and Lebart, Morineau and Warwick (1984), for example, provide an excellent technical, practical and historical account of development of CA up to that period. What is less well known is that the literature on CA extends well beyond the traditional approaches that can be found in many multivariate texts and often there are disciplines that redefine the way in which it is performed. For example, the various fields of ecology have successfully germinated variants including *canonical correspondence analysis* and *detrended correspondence analysis*. However the scope, and literature, of CA is not confined to these examples. Beh and Lombardo (2014, Section 1.6.3) and provide a comprehensive list of members of the "family" which, now, number about 50 members. I shall provide an overview of some of the popular, and not-so-popular, members of the CA family.

Keywords: Correspondence analysis, Multiple CA, Family of analyses

References:

Beh, E. J. and Lombardo, R. (2014). *Correspondence Analysis: Theory, Practice and New Strategies*. Chichester: Wiley.

Greenacre, M. J. (1984), *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Lebart, L., Morineau, A. and Warwick, K. M. (1984). *Multivariate Descriptive Statistical Analysis*. New York: John Wiley & Sons.

Wednesday 13th 14:30 Case Room 1 (260-005)

Testing For Genetic Associations In Arbitrarily Structured Populations

Minsun Song
Sookmyung Women's University

Abstract: We present a new statistical test of association between a trait and genetic markers, which we theoretically and practically prove to be robust to arbitrarily complex population structure. The statistical test involves a set of parameters that can be directly estimated from large-scale genotyping data, such as those measured in genome-wide associations studies. We also derive a new set of methodologies, called a genotype-conditional association test, shown to provide accurate association tests in populations with complex structures, manifested in both the genetic and non-genetic contributions to the trait. We demonstrate the proposed method on a large simulation study and on the real data. Our proposed framework provides a substantially different approach to the problem from existing methods.

Keyword:

Genome-wide association studies, Latent variable, Population structure

Threshold Determination For The Meteorological Data Quality Control In Korea

Yung-Seop Lee[1], Hee-Kyung Kim[1], and Myungjin Hyun[2]
[1]Dongguk University
[2]KMA National Climate Data Center

Abstract: The raw meteorological data need to be cleaned since they are from the diverse sources such as ASOS(Automated Synoptic Observing System) and AWS(Automatic Weather Station). The meteorological data in South Korea is observed from about 100 ASOS and 500 AWS. In order to produce the high qualified meteorological data, several data quality control algorithms are applied. In this study, cluster analysis for almost 600 meteorological sites is applied depending on their climatic characteristics. After clustering, we propose the several threshold algorithms in the given cluster. The proposed threshold values for data quality control algorithms will be adequate to Korea climate condition by cluster and month. Thresholds of QC algorithms, which are step test, persistence test and climate range test, are determined. Through these algorithms and threshold, the qualified meteorological data can be produced for the improved forecast accuracy.

Keywords:

meteorological data quality control, threshold values, cluster analysis, step test, persistence test, climate range test.

Wednesday 13th 14:30 Case Room 2 (260-057)

Regularized Noise-Reduction Methodology For High-Dimensional Data

Kazuyoshi Yata and Makoto Aoshima
University of Tsukuba

Abstract: In this talk, we consider principal component analysis (PCA) methods in high-dimensional settings. We first consider asymptotic properties of the conventional estimator of eigenvalues. We show that the estimator is affected by the high-dimensional noise structure directly, so that it becomes inconsistent. In order to overcome such difficulties in a high-dimensional situation, Yata and Aoshima (2012) developed a new PCA method called the noise-reduction (NR) methodology. We show that the NR method can enjoy consistency properties not only for eigenvalues but also for PC directions in high-dimensional settings. The estimator of the PC directions by the NR method has a consistency property in terms of an inner product. However, it does not hold a consistency property in terms of the Euclid norm. With the help of a thresholding method, we modify the estimator and propose a regularized NR method. We show that it holds the consistency property of the Euclid norm. Finally, we check the performance of the new NR method by using microarray data sets.

Keywords: eigenstructure, large pp small nn, PCA, spiked model

References:

Yata, K. and Aoshima. M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of Multivariate Analysis*, **105**, 193–215.

Computation Of Influence Functions For Robust Statistics

Maheswaran Rohan
Auckland University of Technology

Abstract: Robust statistics are often computed when outliers are present. One of the diagnostics tools for assessing the robustness of estimation is the influence function, which measures the impact on a statistic of adding new data to or removing existing data from the data set. It is also useful for computing the standard error of the statistic.

The computation of influence function for closed form estimates is relatively easy in comparison to that for non-closed form estimates. However, robust statistics are often not in closed form and are computed using iterative algorithms. Obtaining the analytical form of the empirical influence functions of robust statistics for multiple parameters is rare in the current literature and not easy.

In this talk, I use matrix algebra including matrix derivation to show how influence functions for robust statistics can be obtained analytically, particularly in M-estimators with multiple of parameter vectors.

Keywords:

Keywords M-estimators, One-step influence function, Jacobian matrix

Adaptive Model Averaging In High-Dimensional Linear Regression

Tzu-Chang Forrest Cheng[1], Wei-Cheng Hsiao[2], and Ching-Kang Ing[2]
[1]National Central University
[2]National Tsing Hua University

Abstract: This paper aims to propose a data-adaptive model averaging estimation in the high-dimensional framework. To this end, We first consider the orthogonal greedy algorithm (OGA) proposed by Ing and Lai (2011) to construct a set of nested models. The high-dimensional model averaging criteria (HDMMA) suggested by Ing (2016) is considered upon the OGA nested models, while the penalty term is unknown and needed to be estimated. We then use the cross-validation to select the optimal penalty. This method of penalty selection is shown to be optimally adaptive to a wide class of data generating processes. Furthermore, the resultant HDMMA estimator based on the selected penalty is shown to be asymptotic rate efficient. Finally, numerical studies in this paper are expected to shed some light on the choice of data splitting ratio for the cross-validation.

Keywords: Adaptive penalty, Cross-validation, High dimension, Model averaging, Greedy algorithm

References:

Ing, C.-K. and Lai, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statist. Sinica*, **21**, 1473–1513.

Ing, C.-K. (2016). Model averaging in high-dimensional regressions. *Unpublished Technical Report*.

Wednesday 13th 14:50 098 Lecture Theatre (260-098)

Model-Based Clustering For Multivariate Categorical Data With Dimension Reduction

Michio Yamamoto
Okayama University

Abstract: A novel model-based clustering procedure for multivariate categorical data is proposed. The proposed model assumes that each response probability has a low-dimensional representation of the cluster structure, which is constructed by weights for categorical variables and scores for cluster representatives. For the visualization of the cluster structure, we define low-dimensional scores for individuals as convex combinations of scores for cluster representatives, which may be interpretable in a similar manner to the archetypal analysis developed by Cutler and Breiman (1994). Because the proposed model has the so-called rotational indeterminacy, it is needed to conduct rotation methods after parameter estimation to obtain interpretable results. Instead of this two-step approach, we develop a penalized likelihood procedure that imposes a sparsity-inducing penalty on the weights for categorical variables. To optimize the proposed penalized likelihood criterion, we develop an expectation-maximization (EM) algorithm with gradient projection and coordinate descent. It is shown that there is trade-off relation between the convergence rate of the algorithm and the cluster recovery.

Keywords: clustering, categorical data, dimension reduction, EM algorithm, sparse estimation

References:

Cutler, A., Breiman, L. (1994). Archetypal analysis. *Technometrics*, **36**, 338–347.

Yamamoto, M., Hayashi, K. (2015). Clustering of multivariate binary data with dimension reduction via L1L_{1}-regularized likelihood maximization. *Pattern Recognition*, **48**, 3959–3968.

Wednesday 13th 14:50 Case Room 1 (260-005)

Phylogenetic Tree-Based Microbiome Association Test

Sungho Won
Seoul National University

Abstract: Microbial metagenomics data has large inter-subject variation and operational taxonomic units (OTU) for each species are usually very sparse. Because of these problems, non-parametric approaches such as Mann-Whitney U test and Wilcoxon rank-sum test have been utilized. However these approaches suffer from low statistical powers for association analyses and thus investigation on efficient statistical analyses is necessary. Main goal in my thesis is to propose phylogenetic Tree-based Microbiome Association Test (TMAT) for association analyses between microbiome abundances of each OTU and disease phenotype. Phylogenetic tree reveals similarity between different OTUs, and thus was used to provide TMAT. TMAT calculates score test statistics for each node and test statistics for all nodes are combined into a single statistics by minimum p-value or Fisher's combing p-value method. TMAT was compared with existing methods with extensive simulations. Simulation studies show that TMAT preserves the nominal type-1 error and its statistical powers were usually much better than existing methods for considered scenarios. Furthermore it was applied to atopic diseases and found that community profiles of Enterococcus is associated.

Keywords:

NGS; phylogenetic treel Microbiome Association Test

Fitting Additive Hazards Model Using Calibrated Weights For Case-Cohort Data

Hyuntae Kyung and Sangwook Kang
Yonsei University

Abstract: A case-cohort design is an efficient study design for analyzing failure time data by reducing the cost and effort of conducting a large cohort study. Estimation of regression coefficients is typically done through a weighted estimating equation approach whose weight is the inverse of the sampling probabilities. Several techniques to enhance the efficiency by estimating weights or calibrating weights based on auxiliary variables have been developed for Cox models. In this paper, we propose to extend these methodologies to semiparametric additive hazards models. The proposed estimators are demonstrated to be more efficient, via simulation studies, than the usual Horvitz-Thompson type estimator. We illustrate a use of the proposed estimators by using the National Wilms Tumor Study data.

Keywords:

Estimating Equations, Semiparametric Model, Survey Sampling, Survival Analysis

Selecting The Number Of Principal Components

Yunjin Choi
National University of Singapore

Abstract: Principal Component Analysis (PCA) is one of the most popular methods in multivariate data analysis, which can be applied to covariance matrices. Despite the popularity of the method, there is no widely adopted standard approach to select the number of principal components to retain. To address this issue, we propose a novel method utilizing the hypothesis testing framework and test whether the currently selected principal components capture all the statistically significant signals in the given data set. While existing hypothesis testing approaches do not enjoy the exact type 1 error property and lose power under some scenarios, the proposed method provides an exact type 1 error control along with decent size of power in detecting signals. Central to our work is the post-selection inference framework which facilitates valid inference after data-driven model selection; the proposed hypothesis testing method provides exact type 1 error controls by conditioning on the selection event which leads to the inference. We also introduce a possible extension of the proposed method for high-dimensional data.

Keywords:

Principal component analysis, post-selection inference, hypothesis testing

Wednesday 13th 14:50 Case Room 3 (260-055)

Rolling Survival Extrapolation Algorithm For Estimating Life Years Lost Among Subjects Exposed To Long-Term Air Pollution

Jing-Shiang Hwang and Tsuey-Hwa Hu
Academia Sinica

Abstract: Measure of expected years of life lost (EYLL) of a cohort of subjects living with specific conditions would be useful for quantifying and thereby comparing the societal burden of different conditions. One promising approach of estimating EYLL is based on relative survival between the index cohort and an age- and sex-matched reference population generated from vital statistics to extrapolate survival function of the index cohort. The EYLL is then estimated by computing the area between the survival curve of the reference population and extrapolated survival curve of the index cohort. In this talk, we will introduce our newly developed method called rolling survival extrapolation algorithm which consists of two major stages. First, we apply logit transformation to the relative survival so that the transformed curve beyond follow-up would approximate to a nearly straight line. Second, similar to the rolling forecast process for predicting the future over a set period of time, we take advantage of the accurate short-term extrapolation of restricted cubic splines models to guide the transformed relative survival forward step-by-step using the model updated data. There are some studies that provide general evidence for long-term associations of air pollution with hospital admissions and death of various causes. We have found no published epidemiological studies on the effects of long-term air pollution exposure and life years lost. With the proposed method, we estimate EYLL from long-term exposure to air pollution among Taiwanese adult population at rural township and city district levels. The results show that elderly people living in a worse local air pollution for decades long had more expected years of life lost after adjusting social economic status.

Keywords:

Life expectancy, expected years of life lost, air pollution health effects, particulate matter

Wednesday 13th 14:50 Case Room 4 (260-009)

Enhancing The Flexibility Of Regression Modeling By Liquid Association

Ker-Chau Li
Academia Sinica

Abstract: Multivariate regression aims at the study of the relationship between one set of input variables X and one set of output variables Y. Challenges occur when no parametric model is known and yet the number of variables is large. To overcome the difficulties, dimension reduction methods under the inverse regression viewpoint have been investigated by many authors. Liquid association (LA) depicts the change in the covariation of two variables X and Y as a third variable Z varies. In this talk, I will describe a framework to illustrate how the LA methodology can help increase the modeling flexibility of multivariate regression in analyzing complex data.

Keywords:

Sliced inverse regression, liquid association

Clusterwise Low-Rank Correlation Analysis Based On Majorization

Kensuke Tanioka[1], Satoru Hiwa[2], Tomoyuki Hiroyasu[2], and Hiroshi Yadohisa[2]
[1]Wakayama Medical University
[2]Doshisha University

Abstract: Given correlation matrices between variables of subjects and these classes of subjects, it is important to get the distinctive local networks for each class. For example, in fMRI data analysis, such the situation is observed. In concretely, each correlation matrix between regions of interests for his/her brain is observed, and each information of class is get through the experiment. In this presentation, to achieve the purpose, we proposed simultaneous analysis for both clustering of variables and low-rank approximation of correlation matrices corresponding to each class. For the estimation, we adopt the majorization algorithm based on Pietersz and Groenen (2004) and Simon and Abell (2010). Through the proposed method, we can get the distinctive sparse correlation matrices corresponding to classes, while we have to determine the number of clusters.

Keywords: sparse estimation, clustering variables, ALS

References:

Pietersz, R., and Groenen, J.F (2004). *Rank Reduction of Correlation Matrices by Majorization.* Quant.Finance, **4**: 649–662.

Simon, D., and Abell, J. (2010). *Majorization Algorithm for Constrained Correlation Matrix Approximation*, Linear Algebra and its Apprications,, **432**, 1152-1164.

Bayesian Analysis For Fitting Zero-Inflated Count Data With Data Augmentation

Beomseuk Hwang[1] and Zhen Chen[2]
[1]Chung-Ang University
[2]National Institutes of Health

Abstract: Count data with excess zeros are common in epidemiological studies. Zero-inflated Poisson (ZIP) model or zero-inflated negative binomial (ZINB) model can be usually used in these cases. From Bayesian perspective, however, the ZIP and ZINB models are not straightforward to fit, usually requiring manual tunings in the Markov chain Monte Carlo algorithm. We consider the auxiliary mixture sampling through several data augmentations that effectively transform the non-linear and non-Gaussian problem in zero-inflated regression model into a set of linear and Gaussian one. The auxiliary mixture sampling results in tuning-free algorithms in MCMC. We demonstrate how the auxiliary mixture sampling can be applied to an epidemiological case study.

Keywords:

Auxiliary mixture sampling, ZIP model, ZINB model, Markov chain Monte Carlo

Wednesday 13th 15:10 Case Room 2 (260-057)

Towards A Sparse, Scalable, And Stably Positive Definite (Inverse) Covariance Estimator

Joong-Ho Won
Seoul National University

Abstract: High dimensional covariance estimation and graphical models is a contemporary topic in statistics and machine learning having widespread applications. The problem is notoriously difficult in high dimensions as the traditional estimate is not even positive definite. An important line of research in this regard is to shrink the extreme spectrum of the covariance matrix estimators. A separate line of research in the literature has considered sparse inverse covariance estimation which in turn gives rise to graphical models. In practice, however, a sparse covariance or inverse covariance matrix which is simultaneously well-conditioned and at the same time computationally tractable is desired. There has been little research at the confluence of these three topics. In this paper we consider imposing a condition number constraint to various types of losses used in covariance and inverse covariance matrix estimation. This extends the approach by Won, Lim, Kim, and Rajaratnam (2013) on multivariate Gaussian log likelihood. When the loss function can be decomposed as a sum of an orthogonally invariant function of the estimate and its inner product with a function of the sample covariance matrix, we show that a solution path algorithm can be derived, involving a series of ordinary differential equations. The path algorithm is at- tractive because it provides the entire family of estimates for all possible values of the condition number bound, at the same computational cost of a single estimate with a fixed upper bound. An important finding is that the proximal operator for the condition number constraint, which turns out to be very useful in regularizing loss functions that are not orthogonally invariant and may yield non-positive-definite estimates, can be efficiently computed by this path algorithm. As a concrete illustration of its practical importance, we develop an operator-splitting algorithm that imposes a guarantee of well-conditioning as well as positive definiteness to recently proposed convex pseudo-likelihood based graphical model selection methods (Zhang and Zou, 2014; Khare, Oh, and Rajaratnam, 2015).

This is a joint work with Sang-Yun Oh (UC Santa Barbara) and Bala Rajaratnam (UC Davis).

Wednesday 13th 15:10 Case Room 3 (260-055)

Tick-By-Tick Effect On The Inference Of Ultra-High Frequency Data

Zhi Liu
University of Macau

Abstract:

In the inference of ultra-high frequency data, the existing methods are challenged by the tick-by-tick effect, namely, the transactions recorded simultaneously at a time point. In this talk, we will discuss how it influences the estimation of integrated volatility matrix. The effect of microstructure noise and jumps will be discussed as well. We propose to use a double-averaging procedure to deal with these issues. The related asymptotic distribution of the proposed estimator is established. Even interestingly, the double-averaging estimator achieves the "oracle"" property, that is, the asymptotic efficiency is the same as that of the case that the exact trading time of transactions are fully observed. Simulation studies support the theoretical results. The estimation procedure is illustrated through a real data analysis.

High Mortality Predictions With Lines Or Curves Fitted To Over-Dispersed Exposure-Mortality Data

John Maindonald
Statistics Research Associates

Abstract:

Two types of models are considered and compared. The first is a generalized linear model with logit link, with quasibinomial error, and with weighting function that is designed to give reduced weights at the two ends of the scale, relative to mortalities of around 50%. The other approach is to apply a logit transform, and then use a linear model. The logit transformed data appears, for the data that motivated this paper, consistent with the usual linear model variance homogeneity assumptions. For use of a generalized linear model, the standard linear model diagnostics require modifica- tion so that high mortality points do not unduly distort the smooth that is standardly shown for the scale-location plot. A further issue is that 100% mortality points appear to distort a response that, for mortalities less than 100%, is close to linear on the scale of the link function.

Lattice Polytope Samplers

Martin Hazelton
Massey University

Abstract: Statistical inverse problems occur when we wish to learn about some random process that is observed only indirectly. Inference in such situations typically involves sampling possible values for the latent variables of interest conditional on the indirect observations. This talk is concerned with inverse problems for count data, for which the latent variables are constrained to lie on the integer lattice within a convex polytope (a bounded multidimensional polyhedron). An illustrative example arises in transport engineering where we observe vehicle counts entering or leaving each zone of the network, then want to sample possible interzonal patterns of traffic flow consistent with those entry/exit counts. Other problems with this structure arise when conducting exact inference for contingency tables, and when analysing capture-recapture data in ecology.

In principle such sampling can be conducted using Markov chain Monte Carlo methods through a random walk on the lattice polytope, but it is challenging to design algorithms for doing so that are both computationally efficient and have guaranteed theoretical properties. The seminal work of Diaconis and Sturmfels (1998) on Markov bases addresses some of the theoretical issues, but has significant practical limitations. In this talk I shall discuss some preliminary findings based on a more geometric approach to sampler design.

Keywords: lattice bases, Markov bases, MCMC, statistical linear inverse problem

References:

Diaconis, P., and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics* **26**, 363-397.

Statistical Modelling And Analysis Of Cosmic Microwave Background Data

Andriy Olenko
La Trobe University

Abstract: Analysis of the Cosmic Microwave Background (CMB) radiation is a remarkable research area in cosmology whose results won two Nobel prizes in physics in 1978 and 2006 for the discovery of the CMB radiation and its anisotropy. Spurred on by a wealth of satellite data, intensive investigations in the past few years have resulted in many relevant physical and mathematical formalisms to describe and characterise CMB radiation. At the same time, these investigations have raised a number of challenges, theoretical and practical. Studies of deviations from isotropy and Gaussianity, the two fundamental assumptions of cosmological models of the early Universe, form the core of recent experimental and theoretical research in cosmology.

Recent results on modelling CMB evolution and approximation of corresponding random fields will be discussed. Some new approaches to test Gaussianity using multifractality will be illustrated using CMB data. Finally, a new R package for CMB data will be presented.

The presentation is based on joint research with Vo Anh (QUT), N.Leonenko (Cardiff university), P.Broadbridge, D. Fryer, Yu.G. Wang (La Trobe University). This research was supported under the Australian Research Council's Discovery Project DP160101366.

Keywords: random fields, spatial statistics, cosmic microwave background data, R package

References:

Anh, Vo, Broadbridge, P., Olenko, A., Wang Yu.G. On approximation for fractional stochastic partial differential equations on the sphere. Submitted.

Measure Of Departure From Marginal Average Point-Symmetry For Two-Way Contingency Tables With Ordered Categories

Kiyotaka Iki
Tokyo University of Science

Abstract: For the analysis of two-way contingency tables with ordered categories, Yamamoto, Tahata, Suzuki and Tomizawa (2011) considered a measure to represent the degree of departure from marginal point-symmetry. The maximum value of the measure cannot distinguish two kinds of marginal complete asymmetry with respect to the midpoint. The present article proposes a measure which can distinguish two kinds of marginal asymmetry with respect to the midpoint. It also gives large-sample confidence interval for the proposed measure.

Keywords: Asymmetry, marginal proportional point-symmetry, marginal point-symmetry, measure, model, ordered category

References:

Tomizawa, S. (1985). *Biometrical Journal*, **27**, 895–905.

Wall, K.D. and Lienert, G.A. (1976). *Biometrical Journal*, **18**, 259–264.

Yamamoto, K., Tahata, K., Suzuki, M. and Tomizawa, S. (2011). *Statistica*, **71**, 367–380.

Wednesday 13th 16:00 Case Room 2 (260-057)

Sparse Estimates From Dense Precision Matrix Posteriors

Beatrix Jones and Amir Bashir
Massey University

Abstract: A variety of computationally efficient Bayesian models for the covariance matrix of a multivariate Gaussian distribution are available. However, all produce a relatively dense estimate of the precision matrix, and are therefore unsatisfactory when one wishes to use the precision matrix to consider the conditional independence structure of the data. This talk considers the posterior of model fit for these covariance models. We then undertake post-processing of the Bayes point estimate for the precision matrix to produce a sparse model whose expected fit lies within the upper 95% of the posterior over fits. Extensions to finding sparse differences between inverse covariance matrices are also considered. We illustrate our findings with moderate dimensional data examples from metabolomics.

Keywords:

Gaussian graphical models, precision matrices, Bayesian models, metabolomics

Wednesday 13th 16:00 Case Room 3 (260-055)

Dimension Reduction Strategies For Modeling Bi-Clustered High Dimensional Data

Michael Van Supranes and Joseph Ryan Lansangan
University of the Philippines Diliman

Abstract:

A three-stage framework is developed for fitting a mixture of regressions for high dimensional data. The method combines a hierarchical agglomerative grouping algorithm, regression-based clustering, and a sequential, group-wise sparse estimation called Layered Elastic Net Selection (LENS). A simulation study is used to compare the method with LASSO-type and PC-based strategies in terms of predictive accuracy, selection optimality, and clustering accuracy. All simulation scenarios are high dimensional (n<<p), wherein the best subset of predictors may or may not vary among clusters. When the group of most important predictors varies among regression components, the combination of ordinary least squares (OLS) and LENS for mixture of regressions outperforms LASSO-type and PC-based strategies in terms of prediction and clustering accuracy. Based on simulation, the method (termed as MixLENS) results to optimal variable selection, and applying OLS on selected variables results to better prediction and clusters. OLS-MixLENS may result to a more interpretable model that is as predictive as a full model (e.g. Mixture of PC Regressions). In general, MixLENS is likely to select an optimal small subset of predictors for modeling.

Investigating Methods To Produce Price Indexes From Big Data

Alan Bentley, Mathew Stansfield, and Sam Olivecrona
Stats NZ

Abstract: We will present our latest findings on a promising method to use big data for enhancing and improving our current data collections and price measurement. We apply a fixed effects regression approach including using the Fixed Effects Window Splice (FEWS) technique to produce unrevised price indexes at near real-time. We will give examples using daily web scraped food data and administrative rent data.

Stats NZ recently signed up to purchase a trial supply of daily web-scraped online price data from PriceStats, the commercial counterpart of MIT's Billion Prices Project. This data captures, in real-time, online prices for a wide range of different NZ retailers. Preliminary research on measuring rent price change from government administrative data held by MBIE will also be used to illustrate our approach to the art of using *big data*. This data source highlights the opportunities and obstacles of coverage, timing and quality adjustment.

References:

Krsinich, F (2016). The FEWS index: Fixed effects with a window splice. *Journal of Official Statistics*, Vol. 32, No. 2, 2016, pp. 375--404

Computing Entropies With Nested Sampling

Brendon Brewer
University of Auckland

Abstract: The Nested Sampling algorithm, invented in the mid-2000s by John Skilling, represented a major advance in Bayesian computation. Whereas Markov Chain Monte Carlo (MCMC) methods are usually effective for sampling posterior distributions, Nested Sampling also calculates the marginal likelihood integral used for model comparison, which is a computationally demanding task. However, there are other kinds of integrals that we might want to compute. Specifically, the entropy, relative entropy, and mutual information, which quantify uncertainty and relevance, are all integrals whose form is inconvenient in most practical applications. I will present my technique, based on Nested Sampling, for estimating these quantities for probability distributions that are only accessible via MCMC sampling. This includes posterior distributions, marginal distributions, and distributions of derived quantities. I will present an example from experimental design, where one wants to optimise the relevance of the data for inference of a parameter.

Keywords: Bayesian inference, Nested Sampling, Markov Chain Monte Carlo, Information theory

References:

Brewer, B. J. (2017). *Computing Entropies with Nested Sampling.* Entropy, 19, 422.

Skilling, J. (2006). *Nested Sampling for General Bayesian Computation.* Bayesian analysis, 1(4), 833-859.

Spline-Based Drift Models For High Temperature Operating Life Tests

Vera Hofer and Thomas Nowak
University of Graz

Abstract: Since the proper operation of semiconductor devices is of crucial importance for the reliability of a vast range of products, issues concerning quality control are of central relevance to manufacturers. This quality control task is concerned with high temperature operating life tests, where devices are exposed to high temperatures, pressures or humidity, which causes the devices to age artificially fast.

Based on measurements of a random sample of devices, the aim of this work is to compute tolerance limits, such that all subsequent measurements during the stress test stay within their predefined specification limits with a given high probability. These tolerance limits can then be used by automated test equipment for the quality control of devices directly from the production line without their prior exposure to stress test conditions.

In this study, we model the drift behavior of electrical parameters using linear and cubic hermite splines, which are assumed to resemble the true, yet unobserved drift behavior. These spline models allow for the computation of probabilities for an electrical parameter to stay or leave its specification limits at a given point in time. While a very restrictive choice of the tolerance limits might achieve a high level of reliability, the resulting yield loss might get unnecessarily high. Therefore, we formulate an optimization problem that maximizes the probability for initial measurements to be within the tolerance limits (in order to minimize the resulting yield loss) and where the reliability requirement is formulated as a constraint. A derivative-free search algorithm is proposed for this optimization problem, which is then used to test the performance and validity of the model.

Acknowledgment

This work was supported by the ECSEL Joint Undertaking under grant agreement No. 662133 - PowerBase. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and Austria, Belgium, Germany, Italy, Netherlands, Norway, Slovakia, Spain and United Kingdom.

Keywords: quality control, tolerance limits, splines, reliability engineering

A New Approach To Distribution Free Tests In Contingency Tables

Thuong Nguyen
Victoria University of Wellington

Abstract: We will discuss in this talk a new construction for a class of distribution free goodness of fit tests for the classical problem: testing independence in contingency tables. The point is that this problem has been stayed with only one asymptotically distribution free goodness of fit test for a long time, the chi-square test. We will show that our class of new distribution free goodness of fit tests is very wide and discuss the cases where the new tests perform better than the conventional chi-square test.

Keywords: Contingency tables, distribution free, goodness of fit tests

References:

Khmaladze, E., (2013). Note on distribution free testing for discrete distribution, *Annals of Statistics*, **41**, 2979–2993

Khmaladze, E., (2016). Unitary transformations, empirical processes and distribution free testing, *Bernoulli*, **22**, 563–588

Nguyen, T.T.M., (2017). A new approach to distribution free tests in contingency tables, *Metrika*, **80**, 153–170

A Bayesian Inference For Time Series Via Copula-Based Markov Chain Models

Li-Hsien Sun
National Central University

Abstract:

We study the non-standardized Student's *t*-distribution for fitting serially correlated observations where serial dependence is described by the copula-based Markov chain. Due to the computational difficulty of obtaining maximum likelihood estimates, alternatively, we develop Bayes inference using the empirical Bayes method through the resampling procedure. We provide the simulations to examine the performance and also analyze the stock price data in empirical studies for illustration.

Wednesday 13th 16:20 Case Room 3 (260-055)

Modified Gene Shaving Algorithm - A Dimension REduction And Clustering Method

Donna Mae Santos[1], Erniel Barrios[2], and Joseph Ryan Lansangan[2]
[1]Quirino State University
[2]University of the Philippines Diliman

Abstract: High dimensional data exist in digital images, financial time series and gene expression microarrays. Dealing with high dimensionality has become a challenge, where the difficulty lies on how to visualize and explore the high dimensional function or data set. Gene shaving is a statistical method which is based on Principal Component Analysis (PCA) that has proven its worth in visualization and exploration of microarray data. In this paper, the gene shaving algorithm was modified using PCA and Sparse Principal Component Analysis (SPCA), and the modified algorithms were explored in terms dimension reduction and clustering of variables in a more general (not necessarily microarray) high dimensional data setting.The proposed algorithms were evaluated through a simulation study. Simulation results suggest that the modified algorithms identify a singly cluster of variables that may already best explain the variations in the entire data and/or that already are the most informative. Also, the algorithms may produce overlapping clusters, whose variables in the succeeding clusters (other than the first cluster) are those that may provide information not inherent to the first cluster. The modified algorithms are thus potential and useful for exploration and identification of a group of variables worth for further investigation, as well as clustering/groups of variables for understanding variable structures/relationships.

Keywords:

High dimensional data, Cluster of Variables, Gene Shaving, Principal Component Analysis, Sparse Principal Component Analysis, Dimension Reduction

Wednesday 13th 16:20 Case Room 1 (260-005)

The Potential Of Web Scraping

Mathew Stansfield and Sam Olivecrona
Stats NZ

Abstract:

As part of Stats NZ's initiative to explore alternative methods of data collection for price indexes we are currently investigating web scraping. In the short term we are looking to substitute manual online collection, in the medium term we are looking to produce a real-time *Digital Food Price Index*. We will demonstrate a basic technique of web scraping through the R package rvest. We will discuss some of the challenges and limitations of web scraping more sophisticated websites, and using web scraped data in the production of official price indexes. Some challenging questions in this area are regarding the reliability, relevance, and ethics of using internet prices. Will web scraping make us too reliant on external data sources? Do online prices reflect what consumers pay? Is the publicly available data truly public, or does it belong to the company?

A Simple Method To Construct Confidence Bands In Functional Linear Regression

Masaaki Imaizumi[1] and Kengo Kato[2]
[1]Institute of Statistical Mathematics
[2]Unversity of Tokyo

Abstract:

This paper develops a simple method to construct confidence bands, centered at a principal component analysis (PCA) based estimator, for the slope function in a functional linear regression model with a scalar response variable and a functional predictor variable. The PCA- based estimator is a series estimator with estimated basis functions, and so construction of valid confidence bands for it is a non-trivial challenge. We propose a confidence band that aims at covering the slope function at “most” of points with a prespecified probability (level), and prove its asymptotic validity under suitable regularity conditions. Importantly, this is the first paper that derives confidence bands having theoretical justifications for the PCA-based estimator. We also propose a practical method to choose the cut-off level used in PCA-based estimation, and conduct numerical studies to verify the finite sample performance of the proposed confidence band. Finally, we apply our methodology to spectrometric data, and discuss extensions of our methodology to cases where additional vector-valued regressors are present.

Wednesday 13th 16:40 OGGB5 (260-051)

Separation Of Symmetry For Square Contingency Tables With Ordinal Categories

Kouji Tahata
Tokyo University of Science

Abstract: Symmetry and asymmetry models are used to analyze a *square* contingency table with ordinal categories. Caussinus (1966) pointed out that the symmetry model, which indicates the structure of symmetry for cell probabilities, could be separated into the structure of symmetry for odds-ratios and that of symmetry for marginal distributions. This result provides the reason for poor fit of the symmetry model when it occurs for a real dataset. Also, other separations of the symmetry model have been given. For example, Kateri and Agresti (2007), and Saigusa et al. (2015). In this paper, we consider the separation of symmetry by using the generalized asymmetry models. A theorem which the likelihood ratio statistic for testing goodness of fit of the symmetry model is asymptotically equivalent to the sum of those for testing the generalized asymmetry model and the moment equality model under some conditions is given. A simulation study is presented.

Keywords: ff-divergence, moment equality, orthogonality, quasi-symmetry

References:

Caussinus, H. (1966). Contribution à l'analyse statistique des tableaux de corrélation. *Ann. Fac. Sci. Univ. Toulouse* **29**, 77–182.

Kateri, M. and Agresti, A. (2007). A class of ordinal quasi-symmetry models for square contingency tables. *Statist. Probab. Lett.* **77**, 598–603.

Saigusa, Y., Tahata, K. and Tomizawa, S. (2015). Orthogonal decomposition of symmetry model using the ordinal quasi-symmetry model based on ff-divergence for square contingency tables. *Statist. Probab. Lett.* **101**, 33–37.

Wednesday 13th 16:40 Case Room 3 (260-055)

Testing For Presence Of Clustering Effect In Multilevel Model With High-Dimensional Predictors

Frances Claire San Juan[1], Erniel Barrios[2], and Joseph Ryan Lansangan[2]
[1]Cirrolytix Research Services
[2]University of the Philippines Diliman

Abstract:

As big data become more accessible with the boom of data analyzing software, creating value through analytics has grown in demand. Dealing with large data sets in anomaly detection problems, accurate tagging of anomalies is oftentimes lacking and expensive. Unsupervised learning via clustering analysis can be performed to derive labelled data, but used alone, is prone to high false alarm rates. We propose a nonparametric procedure to test presence of clustering effect in a multilevel model with a large set of predictors. Model estimation is done through principal component regression (PCR) and two-way analysis-of-variance (ANOVA), embedded in a backfitting algorithm. Hypothesis test is based on sieve bootstrap. A simulation study showed that the test is effective in detecting high clustering effects, and is optimal when sample size exceeds the number of predictors. The test can be a useful support tool to help address limitations of existing cluster-based methods in anomaly detection.

Towards A Big Data CPI For New Zealand

Alan Bentley
Stats NZ

Abstract: In our digital age, what's the best way to measure inflation? We find an abundance of new data sources, yet these are *found data* in the sense that measuring CPI inflation is a secondary use of the data. Coverage and access become key issues. Automated, scalable, price measurement methods are a must. In this session we take stock and explore the path to a *big data* Consumers Price Index (CPI) for New Zealand.

We discuss the drivers for change and the benefits to a big data approach, reflecting on our early adoption of model-based approaches to price measurement, such as using a hedonic model for second-hand cars, and retail transaction---scanner data---for consumer electronics products (including TVs, computers, digital cameras) in the Consumers Price Index.

We look towards the future by considering the opportunities that are currently in front of us. Notably, we recently signed up to the supply of daily web-scraped online price data from PriceStats, the commercial counterpart of MIT's Billion Prices Project.

Keywords:

Big data, multilateral price indexes, model-based inflation measurement

Programme And Abstracts For Thursday 14th Of December

Keynote: Thursday 14th 9:10 098 Lecture Theatre (260-098)

ALTREP: Alternate Representations Of Basic R Objects

Luke Tierney
University of Iowa

Abstract:

The ALTREP provide framework provides infrastructure to support for alternate representations of basic R objects. Some examples include R vectors with data in memory-mapped files, compact representation of arithmetic sequences, deferred computations, and adding meta-data to objects. This talk will outline the framework, present some examples of its use, and describe the current state of incorporating the framework into the R distribution.

Penalized Vector Generalized Additive Models

Thomas Yee[1], Chanatda Somchit[2], and Chris Wild[1]
[1]University of Auckland
[2]University of Phayao

Abstract: Over the last two decades generalized additive models (GAMs) have become an indispensible tool for modern data analysis and regression. First-generation GAMs as developed by Hastie and Tibshirani are based on backfitting (e.g., the gam R package). Second-generation GAMs have automatic smoothing parameter selection (e.g., the mgcv package by Simon Wood) and are based on, e.g., P-splines. Until recently, these two implementations were largely confined to the exponential family. However, since the 1990s, the vector generalized linear and additive model (VGLM/VGAM) classes were developed by Yee and coworkers, and these are a much larger class of models. First-generation VGAMs were based on vector splines and vector backfitting. This talk will describe 2nd-generation VGAMs using O-splines and P-splines. We illustrate them by examples, to show that automatic smoothing parameter selection based on optimizing a predictive quantity such as generalized cross validation can be very useful. The speaker's VGAM package implementation will be described.

Keywords: Automatic smoothing parameter selection, O-splines, P-splines, Vector generalized additive models, VGAM R package

References:

Yee, T.W. (2015). *Vector Generalized Linear and Additive Models: With an Implementation in R*. New York, USA: Springer.

A Package For Multiple Precision Floating-Point Computation On R

Ei-Ji Nakama[1] and Junji Nakano[2]
[1]COM-ONE Ltd.
[2]Institute of Statistical Mathematics

Abstract: As recent requirements for numerical computation performed by R become larger and more complicated, errors from floating-point arithmetic become problematic. In R, double precision floating-point arithmetic is usually performed, but it may not be adequate or precise for some situations. To avoid and detect errors of double precision floating-point arithmetic, multiple precision arithmetic is useful. Several multiple precision arithmetic packages exist on R, but their abilities are limited. Therefore we provide another multiple precision arithmetic package Rmpenv, which can handle multiple precision arithmetic for real and complex numbers, matrix product and inversion, etc. We also provide a syntactic sugar to make easy the multiple precision computation on R. We utilize a free and open source MPACK library for multiple precision arithmetic and linear algebra computation.

Keywords:

Double precision, floating-point arithmetic, MPACK

Dissimilarities Between Groups Of Data

Nobuo Shimizu[1], Junji Nakano[1], and Yoshikazu Yamamoto[2]
[1]Institute of Statistical Mathematics
[2]Tokushima Bunri University

Abstract: We often have “big data” expressed by both continuous real variables and categorical variables. When their sizes are huge, it is almost impossible to see and check each individual data. Then we divide them into small number of groups which have clear domain meanings. We express each group by using information up to second order moments. For example, means, variances and covariances are used to summarize many continuous real variables, and a Burt matrix which consists of contingency tables by pairs of categorical variables are used to summarize many categorical variables. We call such a set of descriptive statistics “aggregated symbolic data (ASD)”.

We here propose dissimilarities between two ASDs by utilizing pseudo-likelihood ratio test statistic and chi-squared test statistic. Former one is theoretically derived and the latter one is heuristically given. We adopt two dissimilarities for clustering districts in Tokyo by ASD derived from huge real estate data.

Keywords:

Aggregated symbolic data, Chi-squared test statistic, clustering, pseudo-likelihood ratio test statistic

Comparison Of Tests Of Mean Difference In Longitudinal Data Based On Block Resampling Methods

Hirohito Sakurai and Masaaki Taguri
National Center for University Entrance Examinations

Abstract: Let us consider a two-sample problem in longitudinal data, and discuss comparison of tests of mean difference using block resampling methods. The testing methods are based on moving block bootstrap (MBB), circular block bootstrap (CBB) and stationary bootstrap (SB). These block resampling techniques are used to approximate the null distributions of the following four types of test statistics: sum of absolute values of difference between two mean sequences (T1T_1), sum of squares of difference between two mean sequences (T2T_2), area-difference between two mean curves (T3T_3), and difference of kernel estimators based on two mean sequences (SnS_n). Our testing algorithm generates blocks of observations in each sample similar to MBB, CBB or SB, and draws resamples *with replacement* or *without replacement* from the mixed blocks which are generated by two samples. In the context of block resampling, a resample is usually generated *with replacement* from blocks of observations, however our discussion also includes block resampling *without replacement* similar to permutation analogy for MBB, CBB and SB, with T1T_1, T2T_2, T3T_3 and SnS_n, respectively. Monte Carlo simulations are carried out to examine the empirical level and power of the testing methods.

Keywords: moving block bootstrap, circular block bootstrap, stationary bootstrap, with/without replacement, empirical level/power

References:

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. New York: Springer.

SSREM: A Summary-Statistics-Based Random Effect Model To Estimating Heritability, Co-Heritability And Effect Sizes In GWAS Data Analysis

Jin Liu[1] and Can Yang[2]
[1]Duke-NUS Medical School
[2]Hong Kong University of Science and Technology

Abstract: Most existing methods for GWAS data analysis require individual-level genotype data as their input. However, it is often not easy to get access to individual-level data, due to many practical issues, such as privacy protection and disagreement on data-sharing among multiple research groups. In this talk, we introduce SSREM, a Summary-Statistics-based approach to estimating heritability, co-heritability and effect sizes in GWAS data analysis. This is achieved by Bayesian analysis with the standard random-effect prior and a summary-statistics-based likelihood function. We have implemented a parallel Gibbs sampling strategy, which allows us to handle genome-wide-scale datasets. Our analysis results suggest that summary-statistics-based analysis can achieve comparable performance to individual-level data analysis.

Keywords:

Summary statistics; Genome-wide association study; Probabilistic model; Gibbs sampling; Heritability; Co-heritability

Thursday 14th 10:50 098 Lecture Theatre (260-098)

Consistency Of Linear Mixed-Effects Model Selection With Inconsistent Covariance Parameter Estimators

Chihhao Chang
National University of Kaohsiung

Abstract: For linear mixed-effects models with data collected within one cluster, the maximum likelihood estimators of covariance parameters cannot be estimated consistently. Hence the asymptotic behaviors of likelihood-based information criteria, such as Akaike's information criterion (AIC) are rarely discussed in literature. In the contrast, the number of the clusters is generally assumed going to infinity with the sample size to guarantee the consistency of the covariance parameter estimators and thereby guarantees the consistency of the model selection procedures. In this talk, under some mild conditions, we establish asymptotic theorems for ML estimators of covariance parameters when the number of clusters is fixed. Furthermore, the asymptotic behaviors of the generalized information criterion, which includes AIC as special cases, are well studied in our research.

References:

Fan, Y. and Li, R. (2012). Variable selection in linear mixed effects models. In: *The Annals of Statistics*, **40**, 2043 - 2068.

Jiang, J., Rao, J. S. Gu, Z. and Nguyen, T. (2008). Fence methods for mixed model selection. In: *The Annals of Statistics*, **36**, 1669-1692.

Müller, S., Scealy, J. L. and Welsh, A. H. (2013). Model Selection in Linear Mixed Models. In: *Statistical Science*, **28**, 135-167.

Sun, Y., Zhang, W. and Tong, H. (2007). Estimation of the covariance matrix of random effects in longitudinal studies. In: *The Annals of Statistics*, **35**, 2795-2814.

Thursday 14th 10:50 OGGB4 (260-073)

An Incomplete-Data Fisher Scoring With An Acceleration Method

Keiji Takai
Kansai University

Abstract: Incomplete data complicate conventional statistical analyses because the analyses presume complete data are always available. The primary problem is the complication of the parameter estimation. The parameter estimation is based on the observed-data log-likelihood function that consists of the sum of the logarithm of the marginalized likelihood with respect to the missing values, and thus the log-likelihood function becomes complicated to handle. The EM algorithm was proposed to make it easy to handle the log-likelihood function. However, the EM algorithm still has some problems that are often criticized (McLachlan and Krishnan, 2002); namely, slow convergence and unavailability of the standard error.

In my talk, I propose an incomplete-data Fisher scoring (IFS) method with an acceleration method to overcome these problems. The IFS method takes a Newton-Raphson type iteration, but it produces exactly the identical sequence or an approximate sequence to the sequence produced by the EM algorithm. The notable feature of the IFS is that the IFS can accelerate itself by adjusting its steplength and can produce the standard error with the functions used only for the acceleration. The convergence rate is faster than the EM algorithm. In the talk, I provide the convergence theorem and practical examples.

Keywords: Incomplete data, EM algorithm, Fisher scoring, acceleration method

References:

Barnett, J.A., Payne, R.W. and Yarrow, D. (1990). *Yeasts: Characteristics and identification: Second Edition.* Cambridge: Cambridge University Press.

McLachlan, G., and Krishnan, T. (2002). The EM algorithm and extensions, 2nd Edition. Wiley.

(ed.) Barnett, V., Payne, R. and Steiner, R. (1995). *Agricultural Sustainability: Economic, Environmental and Statistical Considerations*. Chichester: Wiley.

Payne, R.W. (1997). *Algorithm AS314 Inversion of matrices Statistics*, **46**, 295–298.

Payne, R.W. and Welham, S.J. (1990). A comparison of algorithms for combination of information in generally balanced designs. In: *COMPSTAT90 Proceedings in Computational Statistics*, 297–302. Heidelberg: Physica-Verlag.

Interactive Visualization Of Aggregated Symbolic Data

Yoshikazu Yamamoto[1], Junji Nakano[2], and Nobuo Shimizu[2]
[1]Tokushima Bunri University
[2]Institute of Statistical Mathematics

Abstract: When we have new "big data", the first step may be to visualize them. For visualizing continuous multivariate data, interactive parallel coordinate plot is known to be appropriate. However, the number of data is huge and some variables are categorical, a simple parallel coordinate plot is not available. We propose to divide big data into rather small groups and summarize them as aggregated symbolic data (ASD), and visualize them by triangular arranged parallel coordinate plots.

We have developed a statistical graphics software for this purpose. Our software equips interactive operations such as selection and linked highlighting, and is written by Java, R, and big data processing technologies such as Apache Hadoop and Apache Spark.

Aggregated symbolic data is a set of descriptive statistics calculated by up to second order moments of variables in each group. We also propose further summarization of ASD to describe characteristics of each variable and a pair of variables for visualizing the difference among ASDs. Real example data are visualized by our software and interpreted intuitively.

Keywords:

Apache Hadoop, Apache Spark, Parallel coordinate plot, Symbolic data analysis

Thursday 14th 10:50 Case Room 2 (260-057)

Analysis Of Spatial Data With A Gaussian Mixture Markov Random Field Model

Wataru Sakamoto
Okayama University

Abstract: In spatial data, detecting regions with higher relative risk is of primary interest. A latent Markov random field model with Gaussian mixture component is introduced, in which the probit or the logit of the mixture weight for each location follows a Gaussian Markov random field such as an intrinsic auto-regressive model (Besag *et al.*, 1991). A mixture model with spatially correlated weights was proposed by Fernández and Green (2002), and our modeling with Gaussian mixture Markov random field can be extended to the cases of involving covariate and random effects. Parameters are estimated by a Bayesian approach, and the posterior mean of the mixture weight for each location, which varies smoothly, gives meaningful interpretation for spatial structure. Our computation was conducted with R Stan package, in which the Hamiltonian Monte Carlo method is implemented. Some applications to disease mapping data are illustrated.

Keywords: Bayesian modeling, spatial cluster detection, spatial correlation

References:

Fernández, C. and Green, P. J. (2002). Modelling spatially correlated data via mixtures: a Bayesian approach. *J. Roy. Statist. Soc. B*, **64**, 805–826.

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.*, **43**, 1–59.

Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications.* Chapman and Hall.

Thursday 14th 10:50 Case Room 3 (260-055)

Forward Selection In Regression Models Based On Robust Estimation

Shan Luo[1] and Zehua Chen[2]
[1]Shanghai Jiao Tong University
[2]National University of Singapore

Abstract: For the purpose of feature selection in ultra-high dimensional regression models, it is required that a sequence of candidate models and a criterion to select the "best" model from them are available. Under different scenarios, various methods have been proposed to achieve these two goals. Intuitively, it is straightforward to choose appropriate loss and penalty functions in a regularization method to accommodate specific characteristics of the data. However, the computation could be expensive for certain cases. From recent studies, we can see that sequential method is promising to produce good candidate models for ultra-high dimensional data. Moreover, it can be easily extended to complex models other than the linear regression model. In this paper, we propose a new feature selection method based on robust estimation.

Keywords:

Feature selection, robust estimation, sequential method

Selecting Generalised Linear Models Under Inequality Constraints

Daniel Gerhard
University of Canterbury

Abstract: Model selection by information criteria can be used to identify a single best model or for inference based on weighted support from a set of competing models, incorporating model selection uncertainty into parameter estimates and estimates of precision. Anraku (1999) proposed a modified version of the well known Akaike information criterion, selecting models in the one-way analysis of variance models when the population means are subject to monotone trends. A generalization of this order-restricted information criterion was proposed by Kuiper et al. (2011), allowing a restriction of population means by a mixture of linear equality and inequality constraints.

An extension to this approach is presented, applying the generalized order-restricted information criterion to model selection from a set of generalized linear models. The class of models can comprise linear equality or inequality constraints of population parameters assuming a distribution of the exponential family for the response. The methodology is illustrated using the open source environment R with the add-on package goric.

Keywords: Model selection, Order-restriction, GLM

References:

Anraku, K. (1999). An information criterion for parameters under a simple order restriction. *Biometrika*, **86**, 141–152.

Kuiper, R.M., Hoijtink, H., Silvapulle, M.J. (2011). An Akaike-type information criterion for model selection under inequality constraints. *Biometrika*, **98**, 495–501.

Thursday 14th 11:10 OGGB4 (260-073)

Improvement Of Computation For Nonlinear Multivariate Methods

Masahiro Kuroda[1], Yuichi Mori[1], and Masaya Iizuka[2]
[1]Okayama University of Science
[2]Okayama University

Abstract: Nonlinear multivariate methods (NL-MM) using optimal scaling as a quantification technique can analyze any data including quantitative and qualitative variables. The alternating least squares (ALS) algorithm is the most popular iterative algorithm in NL-MM. While the algorithm has a stable convergence property, it requires many iterations and a large computational cost, especially for a large data set involving many qualitative variables, because its convergence is linear. It is therefore important to improve the speed of computation when NL-MM with the ALS algorithm is applied. Kuroda and his co-workers tried to accelerate the convergence of the ALS algorithm using the vector ε\varepsilon (vε\varepsilon) accelerator. In this talk, the vε\varepsilon acceleration for the ALS algorithm is implemented in NL-MM, e.g., nonlinear principal component analysis and nonlinear factor analysis, and the performances are demonstrated in numerical experiments.

Keywords: Alternating least squares algorithm, Optimal scaling, Acceleration of convergence

References:

Gifi, A. (1990). *Nonlinear multivariate analysis*. Wiley.

Kuroda, M., Mori, Y., Iizuka, M. and Sakakihara, M. (2011). Acceleration of the alternating least squares algorithm for principal components analysis. *Computational Statistics and Data Analysis*, **55**, 143–153.

Mori, Y., Kuroda, M. and Makino, N. (2016). *Nonlinear principal component analysis and its Applications*. JSS Research Series in Statistics, Springer.

Thursday 14th 11:10 Case Room 3 (260-055)

Feature Selection In High-Dimensional Models With Complex Block Structures

Zehua Chen[1] and Shan Luo[2]
[1]National University of Singapore
[2]Shanghai Jiao Tong University

Abstract: We consider feature selection in multivariate regression models where the response variables as well as the covariates are high-dimensional and both have intrinsic group structures. The models arise naturally in many biology studies for detecting associations between multiple traits and multiple features where the traits and features are embedded in biological functioning groups such as genes or pathways. We propose a sequential procedure for selecting the feature groups based on a correlation principle. At each step of the procedure, the response groups are fitted to already selected feature groups and the residuals are obtained for the response groups, then, the feature group which has the highest correlation with the residuals of any response group is selected next. The correlation measure is the trace of the sample canonical correlation matrix between two vectors. The EBIC is used as the stopping rule of the procedure. This procedure possesses the property of selection consistency. Compared with a group penalization approach, our method is more accurate and demands much less computation.

Keywords: Canonical correlation, correlation principle, grouped data, simultaneous feature selection, selection consistency

References:

Luo, S., and Chen, Z. (2017). *Sequential group feature selection by correlation principle in sparse high-dimensional models with complex block structures*. Manuscript, submitted.

Li, Y., Nan, B. and Zhu, J. (2015). *Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. Biometrics* **71(2)**, 354–363.

Thursday 14th 11:30 098 Lecture Theatre (260-098)

Statistical Generalized Derivative Applied To The Profile Likelihood Estimation In A Mixture Of Semiparametric Models

Yuichi Hirose and Ivy Liu
Victoria University of Wellington

Abstract: There is a difficulty in finding an estimate of variance of the profile likelihood estimator in the joint model of longitudinal and survival data. We solve the difficulty by introducing the "statistical generalized derivative". The derivative is used to show the asymptotic normality of the estimator without assuming the second derivative of the density function in the model exists.

Keywords: Efficiency, Efficient information bound, Efficient score, Implicitly defined function, Profile likelihood, Semi-parametric model, Joint model, EM algorithm, Mixture model

References:

Hsieh, F., Tseng, Y.K. and Wang, J.L. (2006). *Joint modeling of survival and longitudinal data: likelihood approach revisited.* Biometrics **62**, 1037–1043.

Hirose, Y. (2016). *On differentiability of implicitly defined function in semi-parametric profile likelihood estimation.* Bernoulli **22** 589–614.

PREEDALIKIT ET AL. (2016). *Joint modeling of survival and longitudinal ordered data using a semiparametric approach.* Australian & New Zealand Journal of Statistics **58**, 153–172.