114 年統計學術研討會會議手冊

會議日期：2025 年 12 月 16 日(星期二)

會議地點：中央研究院

主辦單位：中央研究院統計科學研究所

中國統計學社

合辦單位：行政院主計總處

教育部

主計協進社

中華機率統計學會

會議主席：杜憶萍教授

籌備委員：吳韋瑩教授、吳漢銘教授、李宜真教授、

李燊銘教授、林宗儀教授、孫立憲教授、

張　中教授、張玉媚教授、陳定立教授、

陳春樹教授、陳璿宇教授、溫啟仲教授、

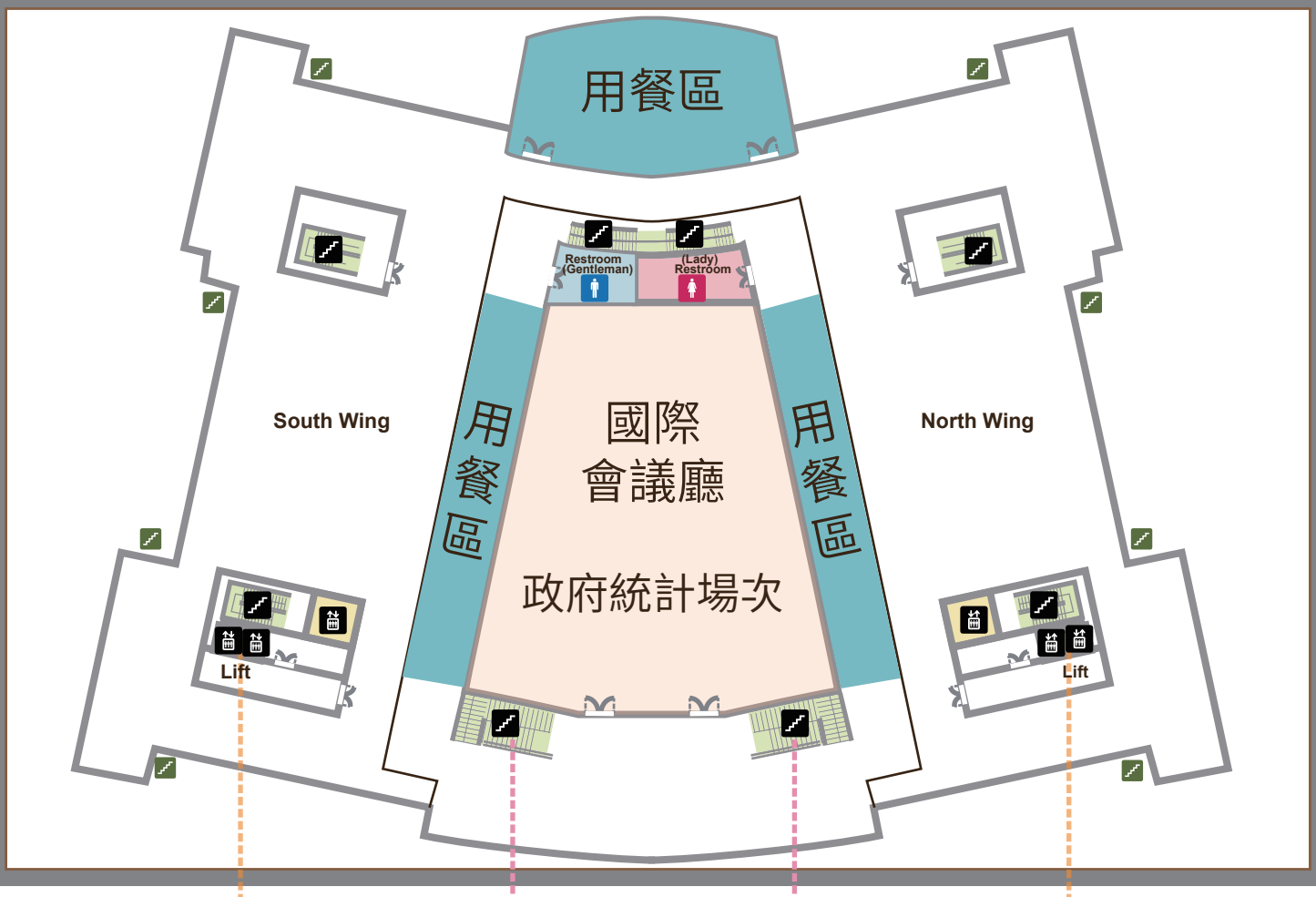
黃士峰教授、黃怡婷教授、黃冠華教授、

蔡政安教授、鄭宗記教授、蘇南誠教授

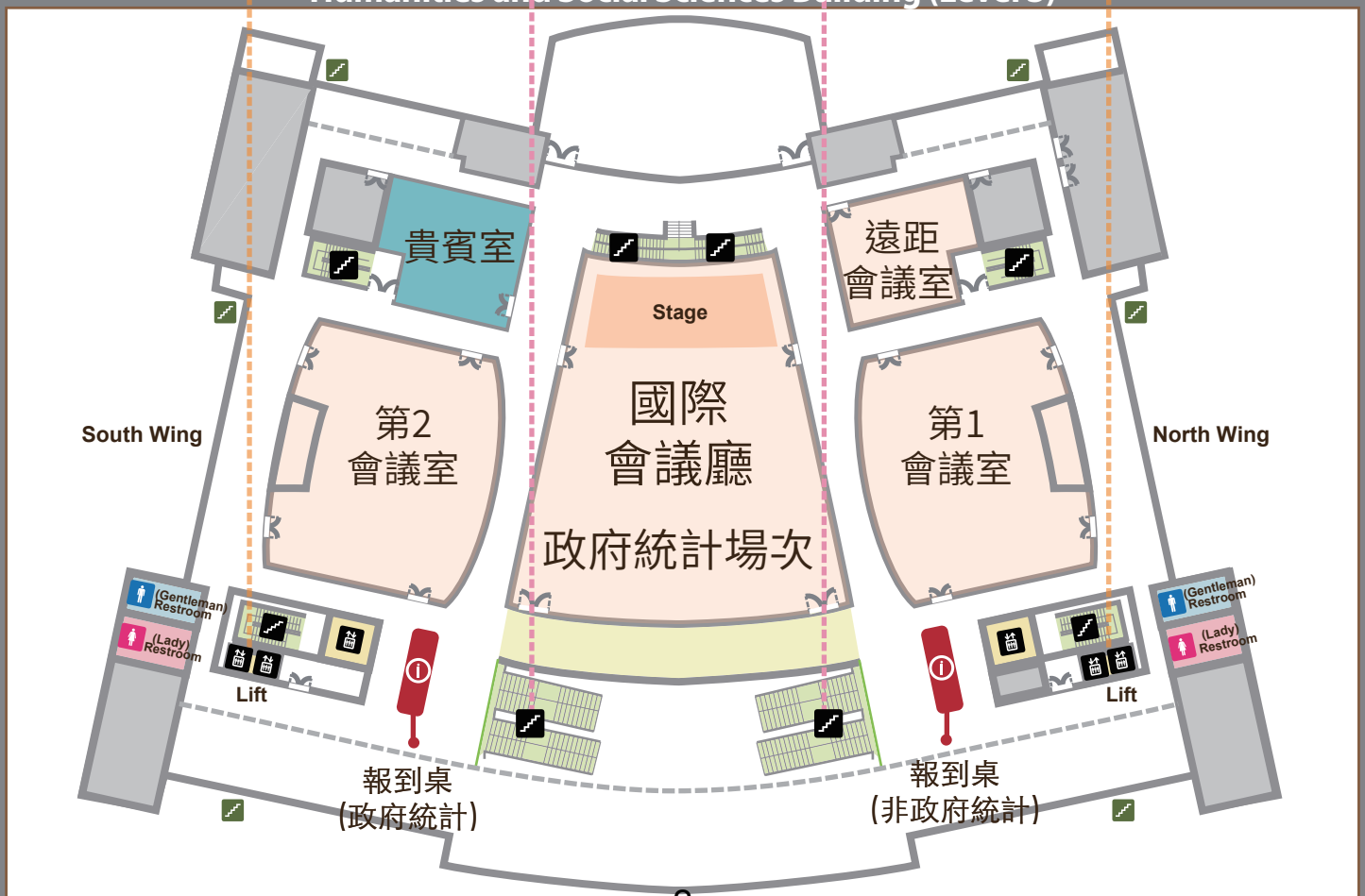
（依姓氏筆畫順序排列）

目錄

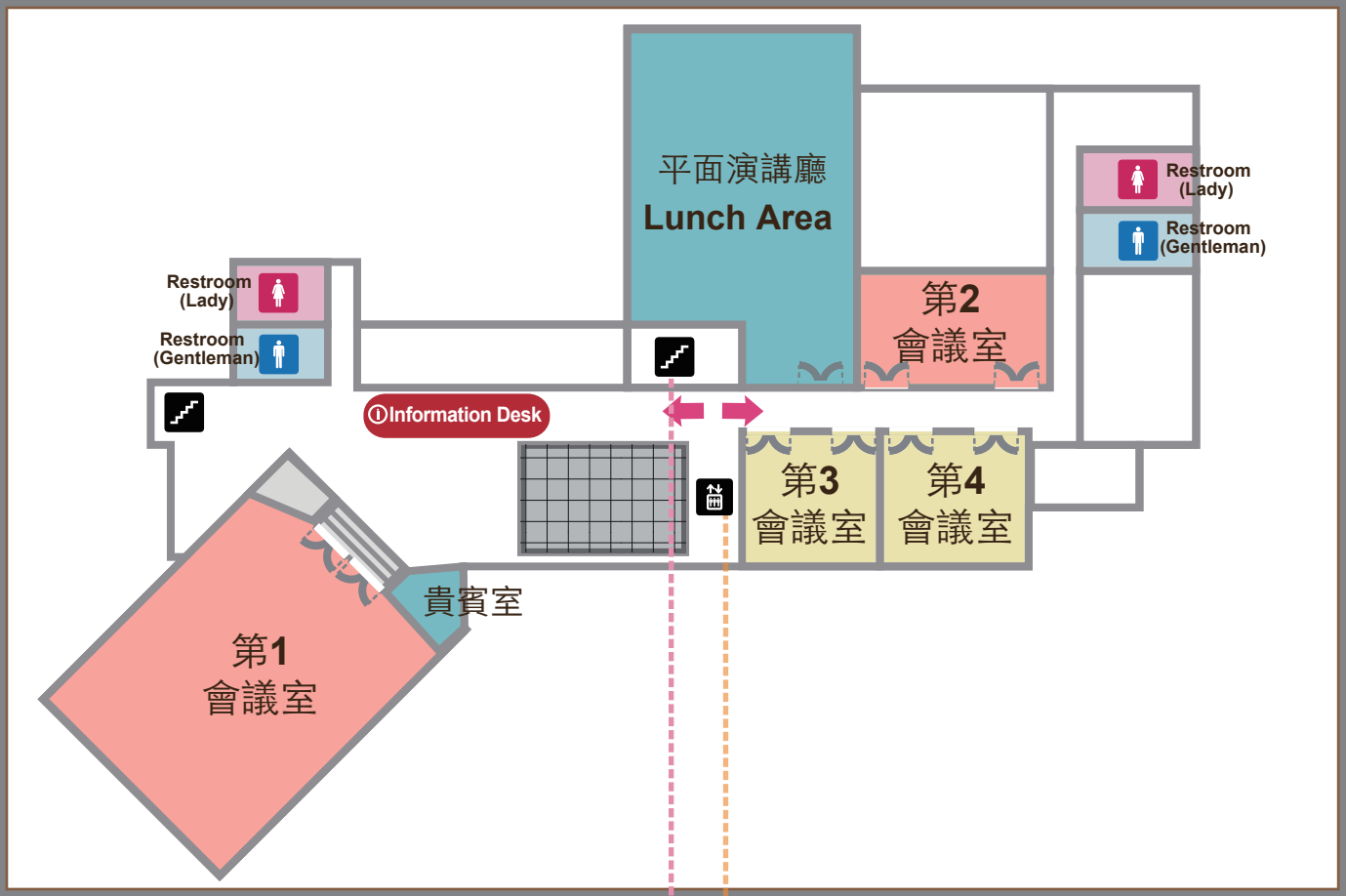
人文社會科學館(人文館) 4樓平面圖
Humanities and Social Sciences Building (Level 4)



人文社會科學館(人文館) 3樓平面圖
Humanities and Social Sciences Building (Level 3)



學術活動中心2樓平面圖
Activity Center (Level 2)



學術活動中心1樓平面圖
Activity Center (Level 1)

中央研究院院區圖

01 院區大門
02 生物醫學科學研究所
03 總務處環安科
04 細胞與個體生物學研究所
04 生物多樣性研究博物館—動物標本館
05 分子生物研究所
06 生物化學研究所 / 生命科學圖書館
08 跨領域科技研究大樓
09 生態時代館、小森林復育區
10 院本部行政大樓（家驊館）
11 黃樓
12 植物暨微生物學研究所
13 資訊科技創新研究中心
13 永續科學中心
14 蔡元培紀念館、Trine & Zen Café
15 統計科學研究所
16 郵局、車廠、員工福利社及萊爾富便利商店
17 生態池
18 基因體研究中心
19 農業科技大樓

20 學術活動中心（四分溪書坊、大禮堂、會議室、住宿、餐廳-微風美饌）
21 中國文哲研究所
22 地球科學研究所
23 綜合體育館
24 人文社會科學館（人文館）
24 人文社會科學聯合圖書館（人文館 1~2 樓）
26 植物分子育種溫室
27 溫室大樓
27 生物多樣性研究中心
28 環境變遷研究中心（環境變遷研究大樓C棟）
28 統計科學研究所（環境變遷研究大樓A棟2~6樓）
28 地球科學研究所（環境變遷研究大樓A棟7~9樓）

30 化學研究所
31 人文社會科學研究中心
31 調查研究專題中心
32 資訊科學研究所
33 物理研究所
33 吳大猷紀念館
34 胡適紀念館
35 近代史研究所
36 歐美研究所
37 歷史語言研究所 / 歷史文物陳列館
38 傅斯年圖書館
39 經濟研究所
40 民族學研究所 / 民族學研究所博物館
41 嶺南美術館（近美大樓）
42 近史所檔案館
43 臺灣考古館

50 中央研究院宿舍群
60 中央研究院國際研究生教學行政大樓（教研大樓）
60 國際研究生學程辦公室及外籍人士服務之行政辦公室（教研大樓 2-3 樓）
60 中央研究院附設幼稚園（教研大樓 1 樓）
61 白樓
61 生物多樣性研究博物館－植物標本館（地下 1 樓）
62 國家生技研究園區
62 生醫轉譯研究中心（國家生技研究園區A棟）

114 年統計學術研討會議程表

時間	12 月 16 日（星期二）							
08:20 ~ 08:50	報到							
08:50 ~ 09:00	開幕致詞 人文社會科學館國際會議廳							
09:00 ~ 09:40	社員大會							
09:40 ~ 09:50	團體照、休息							
09:50 ~ 10:40	主持人：杜憶萍 大會演講 ： 小笠原 欣幸 台灣政治研究三十年：田野調查與選舉預測							
10:40 ~ 11:00	休息							
11:00 ~ 12:30	Session I							
	人文社會科學館				學術活動中心			
	I-1.國際會議廳	I-2.第一會議室	I-3.第二會議室	I-4.遠距會議室	A-5.第一會議室	A-6.第二會議室	A-7.第三會議室	A-8.第四會議室
	政府統計 1	財務統計	人工智慧浪潮下的思維與辯證	Degradation data analysis and its optimal design	中國統計學社 114 年論文獎	調查分析特別議題	Statistical Inference in Biostatistics	空間統計
	Chair：饒志堅 1. 趙明光 2. 蔡慧瑩 3. 彭貴田 4. 馮漢昌	Chair：黃士峰 1. 林良靖 2. 邱海唐 3. 蔡承翰	Chair：陳璿宇 1. 陳璿宇 2. 李建良 3. 何之行	Chair：李宜真 1. 董弘平 2. 鄭雅珊 3. 曾耀霆	Chair：黃信誠 1. 徐仁瓏 2. 何姵青 3. 黃思緯 4. 廖筱軒 5. 王呈嘉	Chair：李燊銘 1. 謝淑惠 2. Martin Lukusa 3. 王智立	Chair：蔡政安 1. 吳泓熹 2. 蘇士詠 3. 陳彥賓	Chair：吳韋瑩 1. 張志浩 2. 楊洪鼎 3. 黃灝勻
12:30 ~ 13:30	午餐 人文館 4 樓交誼廳、學術活動中心 2F 平面會議室							
13:30 ~ 15:00	Session II							
	I-1.國際會議廳	I-2.第一會議室	I-3.第二會議室	I-4.遠距會議室	A-5.第一會議室	A-6.第二會議室	A-7.第三會議室	A-8.第四會議室
	政府統計 2	象徵性資料分析與應用	深度學習於(醫學)影像分析	實驗設計與因果擴論	臨床試驗統計方法	漫談電信信令資料應用	生物統計	空間資料分析
	Chair：王淑娟 1. 陳泯宏 2. 劉維凱 3. 吳東璋 4. 黃佳婷	Chair：吳漢銘 1. 高翊瑋 2. 高君豪 3. 林良靖 4. 吳漢銘	Chair：黃冠華 1. 彭冠舉 2. 廖柏翔 3. 黃冠華	Chair：蘇南誠 1. 余日彰 2. 陳秉洋 3. 謝瓊如	Chair：張玉媚 1. 許根寧 2. 林惠文 3. 曾議寬	Chair：黃怡婷 1. 黃怡婷 2. 王鴻龍 3. 林柏丞	Chair：張 中 1. 陳瓊梅 2. 黃世豪 3. 黃佳慧	Chair：陳春樹 1. 陳怡如 2. 龔一鴻 3. 曾聖澧
15:00 ~ 15:20	休息							
15:20 ~ 16:50	Session III							
	I-1.國際會議廳	I-2.第一會議室	I-3.第二會議室	I-4.遠距會議室	A-5.第一會議室	A-6.第二會議室	A-7.第三會議室	A-8.第四會議室
	政府統計 3	Recent advances and development in statistical modeling	高維資料與深度學習的統計分析： 慶祝陳素雲教授榮退專場	計算統計	統計及資料科學應用	生物統計		
	Chair：陳惠欣 1. 劉軒宇 2. 游舒淳 3. 張 珣 4. 楊子江	Chair：孫立憲 1. 翁新傑 2. 張育瑋 3. 王俞凱	Chair：楊欣洲 陳素雲老師與多位嘉賓 黃啟瑞 盧鴻興 蕭朱杏 杜憶萍 陳定立 洪弘 葉倚任 王紹宣 張源俊 陳彥賓 森元俊成 謝叔蓉 黃文良	Chair：林宗儀 1. 陳泰賓 2. 王婉倫 3. 黃偉恆	Chair：鄭宗記 1. 何漢葳 2. 周珮婷 3. 鄭宗記	Chair：溫啟仲 1. 黃逸輝 2. 吳裕振 3. 姜 杰 4. 溫啟仲		
17:00 ~ 17:10	晚宴接駁							
17:40 ~ 20:20	晚宴 (By invitation only)							

台灣政治研究三十年：田野調查與選舉預測

小笠原欣幸 Yoshiyuki OGASAWARA
國立清華大學人社學院榮譽講座教授

本演講將分享一位外國學者如何進行台灣選舉研究的歷程。首先簡要介紹在台灣進行選舉田野調查的方式，以及外籍學者在過程中所面臨的優勢與挑戰。接著說明個人過去的選舉預測成果，並探討總統選舉的分析與預測方法論。

民主化以來，台灣政治以選舉為核心運作，選舉也成為社會高度關注的焦點。加上台灣位於印太地區的重要環節，海外學者與媒體對其動向尤為關注。掌握台灣選舉的整體脈絡，並深入分析選舉統計資料，有助於理解台灣政治的發展趨勢。如果其對過去選舉結果的分析是正確的話，應該可以從分析模型來推估未來的投票行為，並且可以以預測結果來修改分析模型。選舉預測的價值不在於「預測本身」，而在於作為理解台灣政治的有效工具，也是檢驗自己分析準確性的最佳工具。

總統大選的時候，台灣選民將投總統票、立委選區票、立委不分區票（政黨票）的三張票。建構三張選票的投票動向模型，是個人進行台灣選舉預測的核心工作。具體而言，透過參考過去的選舉統計資料，民調，觀察各縣市的情況，並在關鍵選區進行田野調查，將把三張票的投票動向分別整理於三個 Excel 表格中，以推估相較於上一次選舉的變動幅度。由於三個 Excel 表格分別依據不同的計算過程與數據來源，因此呈現出不同的預測得票率。對照三個不同的數值，進行不斷地反復對照作業，尋找平衡點。這個平衡點就是得票率的預測。經過這個過程，也可以縮小預測的誤差範圍。

簡言之，選舉預測模型是田野調查與統計分析的綜合成果，兩者缺一不可，必須兼顧。

Keyword: 台灣政治、總統選舉、田野調查、選舉預測

Session I

[I-1]：政府場次 1　　人文館-國際會議廳

Chair：饒志堅（內政部統計處處長）

[I-2]：財務統計　　人文館-第一會議室

Organizer：黃士峰（國立中央大學統計研究所）

Chair：黃士峰（國立中央大學統計研究所）

[I-3]：人工智慧浪潮下的思維與辯證　　人文館-第二會議室

Organizer：陳璿宇（中央研究院統計科學研究所）

Chair：陳璿宇（中央研究院統計科學研究所）

[I-4]：Degradation data analysis and its optimal design 人文館-遠距會議室

Organizer：李宜真（國立成功大學統計學系）

Chair：李宜真（國立成功大學統計學系）

[I-5]：碩士論文獎　　活動中心-第一會議室

Chair：黃信誠（中央研究院統計科學研究所）

[I-6]：調查分析特別議題　　活動中心-第二會議室

Organizer：李燊銘（逢甲大學統計系）

Chair：李燊銘（逢甲大學統計系）

[I-7]：Statistical Inference in Biostatistics　　活動中心-第三會議室

Organizer：蔡政安（國立臺灣大學農藝學系）

Chair：蔡政安（國立臺灣大學農藝學系）

[I-8]：空間統計　　活動中心-第四會議室

Organizer：吳韋瑩（國立東華大學應用數學系）

Chair：吳韋瑩（國立東華大學應用數學系）

11:00 ~ 12:30

BERT 助攻！AI 賦能，大幅提升行業統計分類效能

趙明光

行政院主計總處

摘要

本研究旨在探討利用微調後之 BERT（Bidirectional Encoder Representations from Transformers）模型，協助經濟活動描述之自動分類與編碼，以提升工作效率與精進資料品質。現行多數統計調查之行業問項，仍須倚賴人工判讀，不僅耗時費力，且易受主觀判斷影響，以 110 年工業及服務業普查為例，近 149 萬筆主次要經濟活動資料以人工判讀，須耗費 2 萬 5 千小時（每筆約 1 分鐘估算）。若改以微調後之 BERT 模型自動分類，僅需 19 小時即可完成，大幅節省作業時間，且模型預測準確率達 90％以上。

在模型訓練部份，主要採 110 年普查資料，並導入 ChatGPT 協助資料清洗、內容修補及編碼確認，經剔除重複後，以近 17 萬筆資料進行模型之微調訓練，為提升模型效能與分類準確度，訓練過程採多次迭代方式，並根據每次測試結果持續改進，直至最佳化狀態。

目前相關成果已導入本處自行開發之「行業智能查詢系統」，並初步建置於 eBAS 平台，可協助同仁於執行相關調查（如服務業營運及投資概況調查與工業及服務業普查等）時使用，未來亦可擴展至職類別判讀等專業領域。以發揮大型語言模型在協助處理政府資料上的巨大潛力，為統計作業自動化與智慧化奠定堅實基礎。

關鍵字：BERT 模型、ChatGPT、行業統計分類、人工判讀、準確率、行業智能查詢系統

衛福資料應用新里程

蔡慧瑩

衛生福利部統計處

摘要

人權團體 101 年要求其健保個資退出健保業務以外利用（目的外利用）遭拒，提起訴願、行政訴訟，經敗訴定讞，以適用之個資法等有違憲疑義，聲請釋憲（健保資料庫案）。憲法法庭於 111 年 8 月 12 日公告憲判字第 13 號判決結果，除個資法第 6 條第 1 項但書第 4 款規定合憲外，個資法中欠缺個資保護之獨立監督機制有違憲之虞，而健保資料欠缺明確管理機制及目的外利用欠缺請求停止利用之相關規定則違憲。有關個資保護獨立監督機制部分，個資法已於 112 年 5 月 16 日修正、31 日公布增訂第 1 條之 1，行政院成立「個人資料保護委員會」為個資法之專責獨立監督機關。健保資料管理及目的外利用部分，則由衛福部制定「全民健康保險資料管理條例」專法，以完備法律明確性及個資自主權。本文主要敘述健保資料管理法制化過程及目的外學術利用相關重點。

關鍵字：全民健康保險資料管理條例、個人資料保護法

AI 智慧決策系統－由數據驅動的社會住宅規劃與管理

彭貴田

新北市政府主計處

摘要

社會住宅（以下簡稱社宅）是推動居住正義很重要的一項政策，面對社宅布建區域，若能事先掌握申請熱區，就能提前進行空間規劃與資源配置。因此，本研究運用 Python 程式對影響社宅申請之相關變數，如：人口、租金、房型、生活機能等進行資料清理作業；接續以內政部全國門牌地址定位服務完成生活機能因素（如：醫院、學校、便利商店等）地址轉換部分，並使用地理資訊系統 QGIS 環域分析功能，藉以蒐集社宅周邊相關變數的數量；建立需求預測模型部分，本研究採取數種機器學習演算法（如：隨機森林、XGBoost 等）進行，透過預測結果評估選出最適模型後，導入 SHAP 值來解釋每個變數對模型預測結果的「貢獻值」，像是社宅房型、就業人口、老年人口就是影響社宅申請人的正向變數，以提升模型的透明性與可解釋性，最後以申請人數為目標變數，產製熱區並疊合既有設施圖層，進一步以 QGIS2Web 匯出 HTML「社會住宅視覺化儀表板」，提供決策支援、營運管理與跨局處合作平台，俾利資訊的搜尋。

關鍵字：社會住宅；需求預測；地理資訊系統（QGIS）；環域分析；SHAP；社會住宅視覺化儀表板

教育經費國際比較指標編算之變革

馮漢昌

教育部統計處

摘要

依據 OECD 定義，教育經費指公、私部門用於教育的支出，範圍涵蓋各級政府及家戶，其相關指標用於評估國家挹注教育資源的多寡。

歷來教育經費指標的編製，因會計科目資料重複無法拆分及學校填報資料缺漏等因素，難以完全符合 OECD 定義，為接軌國際，本處進行全面檢討，發展符合 OECD 定義之新編算方式，以會計報告資料取代學校填報資料，運用會計收入面資料，以經費來源區分公部門及私部門教育經費，真實呈現政府於教育經費之投入，除增納「教育衍生的間接費用」，即各級政府編列但未直接使用於學校之教育經費（如教育行政管理費用、非屬特定教育階段之經費等）外，亦計入其他部會（如：國科會、國發會等）挹注之教育經費，進一步完善數據，並引用賦稅負擔率，以顯示我國以有限資源大力挹注教育。

觀察代表性指標「教育經費占國內生產毛額比率」，新編算方式之國民教育（國小、國中、高中）部分自 2.0%上升至 3.0%，高等教育部分自 1.2%增至 1.6%，其中高等教育公部門占比自 0.5%增至 1.1%，與 OECD 國家平均一致。

關鍵字：教育經費、國際比較

LIMOS – LightGBM Interval Merton's One-period-portfolio Selection

Liang-Ching Lin,*[1] Hao-Chien Huang[1] and Sz-Wei Charng[1]
[1]Department of Statistics and Institute of Data Science, National Cheng Kung University, Tainan, Taiwan.

Abstract

The modern portfolio theory can assist us in allocating wealth to risky and risk-free assets reasonably well by using some statistical methods. In this study, we will focus on evolving Merton's portfolio problem. Instead of the conventional parameter estimations based on only the closing prices, we include the opening, high, low, and closing prices to enlarge the database as much as possible to make the parameter estimations much more accurate. Furthermore, the covariances are estimated using the bivariate symbolic interval-valued variables under a copula function. In addition, we use LightGBM to predict the transaction directions in which the prices and many statistics are included in the features. In real data analysis, we demonstrate the usefulness of combining the aforementioned methods by showing the portfolio profits of selecting 10 stocks in 2018 and 2019. The results particularly show the superiority of the proposed strategy over the conventional method: the profits can be transformed from negative profit to positive profit and have around 60% to 95% annually.

Keywords: LightGBM; Merton's portfolio problem; Symbolic interval-valued estimation.

High-impact variable selection for clustering in multinomial logistic regression

Hai-Tang Chiou[1], Chun-Wei Dong[2], Chun-Ying Lin[2], and Shih-Feng Huang[2]

[1] Department of Mathematics, National Chung Cheng University
[2] Graduate Institute of Statistics, National Central University

Abstract

This study investigates high-impact variable selection for identifying clusters in high-dimensional multinomial logistic regression. For datasets with known class labels, variables are initially selected using an aggressive variant of the Chebyshev greedy algorithm, and then refined through a high-dimensional information criterion. For unlabeled data, pseudo-labels are generated via clustering, and the variable selection is applied iteratively, updating both labels and selected variables at each step. Simulation studies demonstrate that the proposed methods effectively identify variables relevant for distinguishing underlying classes.

Keyword: Chebyshev greedy algorithm, High-dimensional information criterion, High-dimensional variable selection

Order Placement Strategy Based on Trend Forecast of Bulk Shipping Hire Rate Methods

Cheng-Han Chua [1] and Meihui Guo[2]

[1] Department of Statistics, Tunghai University
[2] Department of Applied Mathematics, National Sun Yat-sen University

Abstract

In this talk, we introduce a data-driven methodology to optimize chartering strategy within the steel industry, with a specific focus on Taiwan's bulk carrier imports from Australia. A high charter rate means the steel industry needs to incur higher costs to rent the bulk carrier. To reduce these costs, we propose a strategy that aims to place orders when the charter rate is relatively low, and consists of two steps: (1) predicting the trend of charter rates for the upcoming month using an ARX model, and (2) determining the optimal time to charter a vessel based on the predicted trend from step (1). Through our research, we identify several key factors that influence charter rates, including fuel prices, demand fluctuations, and geopolitical events. Our findings demonstrate a substantial cost reduction ratio (CRR) of 0.10, representing a 10% cost reduction, leading to significant savings compared to alternative strategies. This study provides valuable insights to enhance supply chain management and improve the efficiency of raw material imports in the steel industry.

Keywords: Maritime Transport, Bulker Charter Rate, Forecasting, Order Placement, Strategy Design.

人工智慧浪潮下的思維與辯證：以精準健康與醫療為例

陳璿宇
中央研究院統計科學研究所

Abstract

人工智慧（AI）在近年來展現出突破性的發展，無論是深度學習模型還是生成式 AI，皆已逐步融入我們的生活與專業領域，帶來前所未有的便利與可能性。然而，AI 的基礎來自於資料庫，若資料涵蓋範圍不足或存在偏差，將直接影響模型的準確性與應用價值。此外，使用者是否具備足夠的 AI 基礎知識與問題解決思維，也決定了 AI 能否真正發揮輔助與創新功能。在 AI 工具唾手可得的今日，研究人員與實務工作者面臨的挑戰，不僅是學會使用 AI，更在於如何發展出真正具創新性與科學價值的研究成果。本次演講將聚焦於 AI 在精準健康與醫療領域中的應用角色，並深入探討其所需的知識基礎、創意思維，以及由此衍生的倫理、偏誤與跨領域合作等議題。期望透過本次分享，激發與會者對於 AI 應用的深層思考，並培養面對 AI 浪潮所需的批判性思維與跨域整合能力，共同邁向更智慧且人本的未來醫療。

Keyword: 人工智慧、資料庫、生成式語言模型、精準醫療

Beyond AI-Risk-Classifications

李建良

中央研究院法律學研究所

Abstract

本座談探討 AI 風險治理的法律問題，擬跳脫 AI 風險分級的思考框架，以歐盟 AI Act 第 5 條禁止規定為基礎，討論有哪些 AI 的運用不被容許、應加以禁止？理由何在？在台灣有無必要？法制上是否可行？

Keyword: AI、EU-AI Act、風險治理、法律管制。

AI 時代下的資料治理

何之行

中央研究院歐美所

摘要

隨著人工智慧技術的迅速發展，資料已成為推動創新與決策的重要資產。然而，AI 對於資料的高度依賴亦帶來前所未有的治理挑戰，包括個人資料保護的邊界、資料取用的倫理爭議以及生成內容所引發的幻覺風險。如何在促進技術進步與維護個人權益之間取得平衡，已成為當前資料治理的核心議題。本演講旨在探討 AI 時代下資料治理的規範與挑戰。內容將聚焦於：生成式 AI 所帶來個資保護之邊界與難題、AI 模型訓練過程中的資料偏誤與透明度問題，以及資料治理框架的重構。藉由跨領域的對話與實例分析，盼能勾勒出一兼具創新與課責的資料治理藍圖，為建構可信賴的 AI 生態系奠定基礎。

關鍵字: 生成式 AI、資料治理、偏見、透明度、課責

Optimal designs for gamma degradation tests

Hung-Ping Tung[1], Yu-Wen Chen[1]

Department of Industrial Engineering and Management, National Yang Ming Chiao Tung University

ABSTRACT

This study analytically investigates the optimal design of gamma degradation tests, including the number of test units, the number of inspections, and inspection times. We first derive optimal designs with periodic inspection times under various scenarios. Unlike previous studies that typically rely on numerical methods or fix certain design parameters, our approach provides an analytical framework to determine optimal designs. In addition, the results are directly applicable to destructive degradation tests when number of inspection is one. The investigation is then extended to designs with aperiodic inspection times, a topic that has not been thoroughly explored in the existing literature. Interestingly, we show that designs with periodic inspection times are the least efficient. We then derive the optimal aperiodic inspection times and the corresponding optimal designs under two cost constraints.

Keywords: Reliability; Degradation tests; Gamma process; Inspection time; Optimal design.

Optimal Test Planning for Heterogeneous Wiener Processes

Ya-Shan Cheng[1], Chien-Yu Peng[2]

[1] Graduate Institute of Statistics, National Central University
[2] Institute of Statistical Science, Academia Sinica

Abstract

Degradation models based on heterogeneous Wiener processes are commonly used to assess information on the lifetime of highly reliable products. An optimal test plan given limited resources is generally obtained using numerical methods for heterogeneous Wiener processes. However, numerical searches for optimal test plans have the disadvantage of being time-consuming and may provide unclear explanations for the findings. To overcome these difficulties, we derive an explicit expression for decision variables (such as the termination time, number of measurements, and sample size) of *D*- and *V*-optimal test plans with cost constraints. The theoretical results not only ensure that the optimal test plan is found, but also provide clear insights into the decision variables affected by model parameters and experimental costs. Some numerical examples are presented to support the efficiency and applicability of the optimal test plans.

Keyword: First passage time, Intraclass correlation coefficient, Invariance, Mean-time-to-failure, Random coefficients

Statistical Inference for the Kumaraswamy Parameters Using Progressively Censored Data

Yao-Ting Tseng

Department of Statistics and Data Science, Tamkang University, Tamsui 251301, Taiwan

Abstract

In this paper, we delve into the process of making inferences about the parameters of the Kumaraswamy distribution using progressively type-II censored data. Our approach involves employing the method of maximum likelihood to derive point estimates for the model parameters. We establish the existence and uniqueness of these maximum likelihood estimators. Additionally, we present pivotal quantities that enable the construction of exact confidence intervals and joint confidence regions for the model parameters. To assess the performances of our proposed estimation techniques, we conduct comprehensive simulation studies. In conclusion, we apply the introduced estimation methods to analyze and discuss the results obtained from a real dataset, providing practical insights into their performance.

基於統計架構之冗餘感知自適應層剪枝方法：應用於影像分類與深偽檢測

徐仁瓏

成功大學統計學系碩士班

摘要

本研究提出一種統計式自適應層剪枝方法，以有效移除神經網路中的冗餘結構並提升模型壓縮效率。相較於需多次試驗與超參數調整的傳統剪枝技術，本方法運用曼–惠特尼 U 檢定（Mann–Whitney U test）與科摩哥洛夫–史密諾夫檢定（Kolmogorov–Smirnov test），直接分析預訓練模型各殘差區塊之特徵分佈差異，自動判定冗餘層並予以移除，無需手動設定或額外資料，降低剪枝成本並提升可解釋性。本方法廣泛驗證於影像分類與深偽影像檢測，包括 CIFAR-10、CIFAR-100、ImageNet、FaceForensics++ 與 Celeb-DF 資料集，並應用於 ResNet 與 EfficientNetV2 之多種架構。結果顯示，本方法可依據任務難度與模型深度自適應調整剪枝比例，在多數設定下同時提升準確率與壓縮效益，展現良好穩定性與泛化能力。此外，本研究亦比較推論延遲、剪枝搜尋時間，並進行性能與壓縮間的權衡分析；同時透過 Grad-CAM 視覺化檢視模型注意力，全面驗證本方法於壓縮效率、運算成本與特徵學習上的優勢。

關鍵字：層剪枝、統計檢定、影像分類、深偽影像檢測

多變數門檻結構下門檻自迴歸模型之估計方法研究

何姵青

清華大學統計與數據科學研究所

摘要

門檻自迴歸（Threshold Autoregressive, TAR）模型在非線性時間序列建模中，是能夠捕捉序列區制轉換（regime-switching）的重要模型架構之一。然而，傳統門檻自迴歸模型多假設門檻變數為已知或單一變數，無法處理由多個變數共同驅動的複雜動態結構。

為克服此一限制，本篇論文提出一套廣義的門檻自迴歸建模框架，允許門檻變數為多個變數的線性組合。此估計問題被重新表述為一個不可微分的模型選取問題，核心目標在於估計主導潛在區制結構的門檻方向（即線性組合的係數向量）。

本篇論文進一步提出一套三步驟的整合估計程序，稱為 BO-TAR（Bayesian Optimization for Threshold Autoregressive Modeling），結合貝氏最佳化（Bayesian Optimization, BO）與正交貪婪演算法（Group Orthogonal Greedy Algorithm, GOGA）。BO 用於對門檻方向進行全域搜尋，而 GOGA 則負責估計對應的區制分割與各區制下的自迴歸結構。模型選取則透過高維赤池資訊準則（High-Dimensional Akaike Information Criterion, HDAIC）加以評估，進而達成門檻方向、區制結構與自迴歸結構的整合估計，並兼顧計算效率。

透過多組模擬實驗，本研究驗證所提方法在各種情境下皆具良好估計表現。在實證應用中，BO-TAR 成功辨識出傳統自迴歸模型難以捕捉的非線性規則、非對稱且具區制依賴的動態特徵。此結果凸顯 BO-TAR 在處理具非線性與時間變異特性之實際時間序列資料上的實用價值。

關鍵字：貝氏最佳化、模型選取、非線性時間序列、區制切換、門檻自迴歸模型、門檻估計

單位元高斯過程最佳分割的次層搜尋法

黃思緯

清華大學統計與數據科學研究所

摘要

高斯過程為一種靈活且具統計基礎的半參數局部平滑化方法。但在資料量有限時，其表現高度依賴核函數的設計。對比於文獻中常見的相乘型核函數，加法型核函數雖能捕捉變數間各種交互作用，但兩者分別有彈性不足與易過度配適的問題。為此，Unital Gaussian Process （UGP） 透過對變數集合進行分割，建構出混有相乘及加法結構特性的彈性核函數，以提升模型對複雜資料之適應能力。然而，UGP 的效能高度依賴所挑選的分割，而尋找最適分割的過程，會面臨分割數量隨變數個數增加而呈階乘級上升，與不同分割的 UGP 模型間缺乏包含關係等挑戰。本文基於分割間「較細」與「較粗」關係所建立之階層式結構，提出一套最適分割搜尋法。其以最細分割為起點，利用對次層的分割進行分類，從而有效搜尋最適分割。透過模擬研究，我們發現此方法不僅能大量節省窮舉搜尋所帶來的高昂計算成本，準確搜尋出最適分割，同時亦可維持優異的預測效能。

關鍵字：相加型核、階層式結構、核函數、模型選擇、類神經網路、相乘型核

改進時間空間跨欄模式及其應用

廖筱軒

逢甲大學統計與精算碩士班

摘要

This study introduces a modified spatial hurdle integer-valued generalized autoregressive conditional heteroscedasticity （INGARCH） model that enhances prediction for spatiotemporal count data, particularly under over-dispersion and zero inflation. To ensure parsimony, the model builds on existing spatial hurdle INGARCH frameworks and incorporates an empirical Bayes approach to estimate the probability of positive counts, offering greater flexibility in modeling zero values. We apply Bayesian inference with an adaptive Markov Chain Monte Carlo（MCMC）algorithm to efficiently estimate the model parameters. We assess model performance using the Deviance Information Criterion （DIC） and predictive accuracy metrics, including mean squared error （MSE）, mean absolute error （MAE）, and mean absolute scaled error （MASE）. Simulation studies and real applications to weekly dengue fever case data from four provinces in Thailand and four cantons in Costa Rica demonstrate the model's effectiveness. Seasonal patterns are incorporated using either climate covariates or Fourier terms to better capture in-sample dynamics. The improved model captures spatiotemporal dependencies more accurately and handles excess zeros more robustly than traditional spatial hurdle and zero-inflated INGARCH models. These results underscore the model's utility in advancing the analysis of spatiotemporal disease patterns.

關鍵字：Spatial INGARCH model, Hurdle model, Markov chain Monte Carlo method, Count time series, Generalized Poisson distribution, Dengue fever

多重有序中介因子之機制分析
—平均因果效應與族群介入效應

王呈嘉

陽明交通大學統計學研究所

摘要

過去二十年間，因果效應的機制研究中，中介效應與交互作用是核心的研究議題。學者 VanderWeele 提出的單一中介因子四路徑拆解法，則是為中介效應與交互作用的研究奠定基礎的理論框架。然而，當其他學者試圖將單一中介因子拓展至多重中介因子的時候，卻面臨諸多挑戰，因為因果效應的路徑數量隨著中介因子數目增加呈現超越指數的增長。依據本研究室 Liao 提出的雙有序中介因子二十八路徑拆解法，便呈現此問題的複雜性，雖然當初二十八路徑拆解法並未完整的引入反事實模型的定義、辨識假設、結果，但也為雙有序中介因子的機制分析提供了研究的方向。因此，本文將回顧本研究室 Liao 提出的雙有序中介因子二十八路徑拆解法，利用反事實模型定義此二十八個路徑特定效應，並且確立辨識假設以及呈現辨識結果，並與學者 Miles 提出的「中介效應準則」進行比較，仔細的討論在雙有序中介因子二十八路徑拆解法是否滿足過去主流學界對於中介效應的合理性。

此外，本文也進一步將雙有序中介因子二十八路徑拆解法拓展至族群介入效應（population intervention effect, PIE）的尺度。族群介入效應和族群可歸因分率（population attributable fraction, PAF），在數學公式上有著相近的關係，上述的 PIE、PAF 的尺度都是公共衛生研究評估移除暴露因子對於整體族群的重要指標。而族群介入效應的尺度下，則會發展成雙有序中介因子三十三路徑拆解法，同樣的呈現其反事實模型以及辨識假設、結果，並且比較過往學者提出的 PIE、PAF 拆解法，期望為公共衛生研究的統計指標提供完整的理論詮釋。

關鍵字：平均因果效應、族群介入效應、中介效應、交互作用

Detecting Social Desirability Bias in Online Surveys: Experimental Evidence from the Cheating Detection Triangular Model

Shu-Hui Hsieh
Research Center for Humanities and Social Sciences, Academia Sinica

Abstract

This study examines the effectiveness of the Cheating Detection Triangular Randomized Response Model (CDTRM) in reducing social desirability bias in online surveys on sensitive topics. Using a between-subjects experimental design, participants were randomly assigned to either a CDTRM group or a direct questioning (DQ) group, in which respondents were instructed to answer directly. DQ often leads to misreporting or nonresponse when addressing sensitive issues are involved. Drawing on data from over 5,000 respondents in a national Taiwanese web panel, the study compares estimated cheating rates and sensitive trait prevalence across conditions. The results demonstrate CDTRM's ability to detect impression management behaviors and produce more accurate prevalence estimates.

Keywords: Cheating Detection, Social Desirability Bias, Indirect Questioning, Sensitive Behavior Prevalence, Online Survey Methods

Adjusting Self-Protective and Non-Response Behaviours in Sensitive Prevalence Estimation by a Two-Stage Multilevel Randomised Response Technique

Martin T. Lukusa[1], Pier Francesco Perri[2] , Shu-Hui Hsieh[3]

[1] Department of Statistics, Feng Chia University, Taichung, Taiwan

[2] *Department of Economics, Statistics and Finance 'Giovanni Anania', University of Calabria, Arcavacata di Rende, Italy*

[3] *Research Center for Humanities and Social Sciences, Academia Sinica, Taipei, Taiwan*

Corresponding to: Research Center for Humanities and Social Sciences, Academia Sinica, Taipei, Taiwan

Abstract

When using the customary direct questioning approach to collect sensitive information in a survey, some respondents may have no problem disclosing their true status or opinion, while others may be reluctant to reveal them. The problem of response bias is likely to arise as some responses tend to be more socially desirable than truthful. Hence, the indirect questioning approach can be used to guarantee privacy protection and reduce the influence of social desirability bias. This paper introduces a new technique which com-bines, both the direct and indirect questioning approaches and allows Bayesian estimates of the prevalence of multisensitive attributes to be obtained by taking into account the estimated proportion of honest respondents under direct questioning. Our proposal stems from a real survey in Taiwan and is illustrated with two motivating examples concerning voting behaviours and sexual identity. The empirical analysis reinforces that the proposed two-stage multilevel method is satisfactory in mitigating the effects of respondents' self-protective behaviour, and produces results that are more reliable than those based on the traditional direct questioning approach and better than a previous version of the multi-level randomised response method that ignores the presence of cheating behaviours in the direct questioning stage.

Keyword: Bayesian estimation; cheating behaviours; honest responses; indirect questioning techniques; sexual identity; social desirability bias; voting behaviour.

大專校院畢業生流向調查視覺化系統與資料採礦技術應用之實例分享

王智立

銘傳大學應用統計與資料科學系教授

中文摘要

大專校院畢業生流向調查視覺化系統是一套整合資料收集、分析與視覺呈現的工具，目的在於協助學校了解畢業生的就業、升學與其他流向趨勢。視覺化系統可以包含幾種功能，例如:1.互動式儀表板：可依條件（如學校、科系、畢業年度）呈現圖表、地圖與時間軸呈現畢業生流向。2.地理分布分析：顯示畢業生在國內外各地的就業或升學分布，支援地圖視覺化。3.趨勢追蹤與統計摘要：分析歷年流向變化提供各類流向的比例、平均薪資、產業別分布等關鍵指標。

依據資料收集架構可採用橫斷面資料分析（Cross-sectional Analysis），分析某一特定時間點（如某年度）畢業生的流向分布、比較不同學院、學系、性別、地區的畢業生在同一時間的就業或升學情況。亦可以採用縱斷面資料分析（Longitudinal Analysis），追蹤同一群體（如某屆畢業生）在不同時間點的流向變化、觀察畢業生職涯發展、轉職趨勢、薪資成長...等。

近年來，資料數據快速取得，資料量越來越大，由許多文獻可知資料採礦技術在各種產業之應用已經相當廣泛。如何應用資料採礦技術，而挖掘出有用的資訊，本研究探討的主要方向。本研究運用大專校院就業職能平台(University Career and Competency Assessment Network；UCAN) 職涯興趣量表測評結果中 16 類與畢業一年和畢業三年就業調查資料為例，以個案大學部學生在校時所測驗的 UCAN 職涯興趣資料，並勾稽該批大學生分別在畢業後一年、三年之畢業生流向調查資料庫，採資料採礦技術建構出適當的模型。研究結果如下：以決策樹

(Decision Tree)與貝氏分類法(Bayesian classifier)資料採礦技術建模，結果發現於UCAN 平台 16 類職涯結果中畢業後一年和畢業後三年實際就業類型之決策樹所建構出的模型準確率較高。在 Holland code 6 類理論中，以資料採礦建模方式建模預測，結果發現 Holland code 6 類結果中貝氏分類法建構的模型準確率較高。

關鍵詞：資料採礦、決策樹、貝氏分類法、橫斷面資料、縱斷面資料、

Statistical framework for analyzing heterogeneous response patterns within microbial treatments for agriculture

Szu-Ting Tan[1], Wen-Po Chuang[1], Steven Hung-Hsi Wu[1]

[1]Department of Agronomy, National Taiwan University, Taiwan

Abstract

Agricultural biologicals are microbial products that can increase crop yield and reduce chemical usage, offering sustainable solutions to food security. However, discovering biological products has several unique challenges. Unlike chemical products, microorganisms form complex three-way interactions with plants and environments. This often results in heterogeneous responses, due to some microbes are inactive or only partially active. Basic statistical methods might not be able to identify promising candidates with heterogeneous responses. A mixture model framework was developed to address heterogeneous treatment response and model it as two components: active and inactive microbes. The model estimates the proportion of active microorganisms and the effect size of microorganisms. Simulation studies demonstrated that the model can identify treatments with only a moderate proportion of active microbes, where these treatments are often missed by basic statistical approaches. Experimental validation using tomato experiments confirmed the robustness and accuracy of this model. This model provides a more efficient screening approach for potential biological products compared to conventional methods. By accounting for heterogeneous biological behaviours and responses, it accelerates the discovery of effective products for sustainable agriculture.

Keyword: Agricultural biologicals, microbial product, heterogeneous responses, mixture model, simulation results.

Estimating cohort-based life expectancy using the age-period-cohort model: Application to all-cause mortality in Taiwan

曾莛勛[1]，蘇士詠[2]

[1] 國立臺灣大學統計碩士學位學程

Abstract

The cohort-based life expectancy is intuitive for application and interpretation but it is difficult to be observed. The age-cohort model was used to estimate the cohort-based life expectancy but the methodology is concerned. We aimed to provide a better estimation of cohort-based life expectancy by using the age-period-cohort model, complete life table and single-year age/period aggregated data to improve the limitations of the previous study. The population-level data of all-cause mortality in Taiwan was used for demonstration and application. This study found the similar results of the age, period and cohort effects between men and women in Taiwan; a J-shaped curve of age effect, a weak period effect, and an overall declining cohort effect. For cohort-based life expectancy, the obvious sex difference and similar temporal change were found; the cohort-based life expectancy for both genders are increased from 1891 to 2022 birth year, and the cohort-based life expectancy in women are all higher than those in men. The cohort-based life expectancy in 2022 being lower than the previous 13 years (between 2009 and 2021) for men and the previous 10 years (between 2012 and 2021) for women. We used the all-cause

mortality data in Taiwan for evaluating the population health from the different perspective of cohort-based life expectancy. This study also found the influence of COVID-19 outbreak on cohort-based life expectancy which implied that such pandemic had diminished the population health including the youngest generation.

Keyword: Life expectancy; mortality; complete life table; age-period-cohort analysis

Bioimage Classification in Sport-Cognition Applications

Yan-Bin Chen (陳彥賓)

Master Program in Statistics, National Taiwan University

Abstract

Bioimage classification provides valuable insights into the connection between sport and cognition. This study explores methods for analyzing sport-related bioimages, such as EEG signals, to identify cognitive states linked to athletic activity. Using statistical techniques and deep learning models, we extract features that distinguish levels of attention and motor coordination. Results indicate that combining advanced machine learning with tailored preprocessing improves classification accuracy across tasks. These findings suggest that bioimage classification can support personalized training, enhance cognitive performance, and deepen understanding of the interaction between athletic activity and cognitive processes.

Keyword: bioimage, machine learning

Threshold Boundary Logistic Regression for Binary Data

Chih-Hao Chang[1], Shih-Feng Huang[2]

[1] National Chengchi University [2]National Central University

Abstract

This talk introduces Threshold Boundary Logistic Regression (TBLR) for binary outcomes, which links covariates to both the logistic component and a linear or nonlinear threshold that partitions the feature space for distinct logistic models. We develop an iterative two-stage sample-splitting estimator that recasts the non-differentiable likelihood into an optimization: threshold parameters are obtained by minimizing a weighted classification loss, while logistic parameters are updated by likelihood maximization. Under suitable conditions, we establish consistency, oracle-optimal convergence rates, and asymptotic normality. Computation uses both Mixed-Integer Programming (MIP) and Weighted SVM (WSVM) as solvers: for linear boundaries we warm-start MIP with a WSVM solution—improving estimation at extra runtime—whereas nonlinear boundaries are solved by WSVM only. Simulations and real-data analyses illustrate finite-sample behavior and feasibility in nonlinear regimes. We further outline an extension to count responses—Threshold Boundary Poisson Regression (TBPR)—adapting the two-stage scheme to the Poisson likelihood and log link, and we will present preliminary empirical analyses that demonstrate applicability and the modeling workflow for count data.

Keyword: Maximum likelihood (ML), mixed integer programming (MIP), two-stage sample-splitting, weighted support vector machine (WSVM).

A GEE-Based Hurdle Framework for Zero-Modified Spatiotemporal Count Data with Environmental Applications

Yang, Hong-Ding

Department of Applied Mathematics, National Chiayi University

Abstract

We propose a generalized estimating equation (GEE) framework for analyzing zero-modified spatiotemporal count data, where excessive or reduced zeros coexist with correlation across space and time. A hurdle model allows us to separately model the probability of zeros and the mean of positive counts, capturing flexible covariate effects. In the estimation procedure, we update the working correlation matrix via a semiparametric approach and employ a block jackknife for reliable variance estimation. Simulation studies demonstrate that the proposed approach provides accurate slope estimation and satisfactory coverage rate of interval estimation under both zero-inflated and zero-deflated scenarios. We further illustrate the method using extreme rainfall data from Northern Taiwan, where temperature and precipitation are shown to significantly influence the occurrence of extremely heavy rain. This combined methodological and applied contribution points out the practical value of GEE-based inference in complex environmental settings. (This is a joint work with Dr. Shen, Chung-Wei and Dr. Chen, Chun-Shu.)

Keyword: hurdle model; spatiotemporal correlation; generalized estimating equations; block jackknife; environmental application

DeepKriging on the Sphere: Basis Adaptation for Spherical Data

Hao-Yun Huang

Department of Applied Mathematics, National Dong Hwa University

Abstract

In this talk, we present a spherical extension of the DeepKriging framework designed for large-scale spatial prediction on global domains. Many existing deep spatial models rely on Euclidean distances, which can distort spatial relationships when applied to spherical data such as global climate or remote sensing measurements. In contrast, classical geostatistical approaches that correctly handle spherical geometry are often computationally infeasible for modern large datasets. To address these challenges, we develop a spherical multi-resolution DeepKriging model that integrates thin-plate spline basis functions defined directly on the sphere with a neural network architecture. The proposed approach captures both large-scale smooth variation and fine-scale spatial detail while maintaining geometric consistency on $\mathbb{S}^2$. We further incorporate Huber's loss into the training procedure to improve robustness against outliers. Simulation studies and real-world applications to global sea surface temperature and precipitation data demonstrate that the proposed framework achieves accurate, stable, and scalable prediction, offering a flexible and robust tool for modeling complex spatial processes on the sphere.

Keywords: DeepKriging, kriging, multi-resolution spline basis functions, thin-plate spline functions on the sphere, non-Gaussian spatial process

Session II

[II-1]：政府場次 2　　人文館-國際會議廳

Chair：王淑娟（農業部統計處處長）

[II-2]：象徵性資料分析與應用　　人文館-第一會議室

Organizer：吳漢銘（國立政治大學統計學系）

Chair：吳漢銘（國立政治大學統計學系）

[II-3]：深度學習於(醫學)影像分析　　人文館-第二會議室

Organizer：黃冠華（國立陽明交通大學統計學研究所）

Chair：黃冠華（國立陽明交通大學統計學研究所）

[II-4]：實驗設計與因果擴論　　人文館-遠距會議室

Organizer：蘇南誠（國立臺北大學統計系）

Chair：蘇南誠（國立臺北大學統計系）

[II-5]：臨床試驗統計方法　　活動中心-第一會議室

Organizer：張玉媚（東海大學統計學系）

Chair：張玉媚（東海大學統計學系）

[II-6]：漫談電信信令資料應用　　活動中心-第二會議室

Organizer：黃怡婷（國立臺北大學統計系）

Chair：黃怡婷（國立臺北大學統計系）

[II-7]：生物統計　　活動中心-第三會議室

Organizer：張 中（國立中山大學應用數學系）

Chair：張 中（國立中山大學應用數學系）

[II-8]：空間資料分析　　活動中心-第四會議室

Organizer：陳春樹（國立中央大學統計研究所）

Chair：陳春樹（國立中央大學統計研究所）

13:30 ~ 15:00

臺南市畜牧產業狀況

陳泯宏

臺南市政府主計處

摘要

畜牧業是農業體系中不可或缺的一環，除提供肉品、乳品與蛋品等主要動物性蛋白來源，亦影響民生物價與糧食安全。臺南市位處南部平原，具備良好農業資源與飼養條件，長期發展出多元畜禽產業，逐漸形成地方特色的產業聚落。為掌握臺南市畜牧業現況，本文利用畜牧業相關資料進行分析，探討產業聚集情形及競爭優劣。

本文首以畜牧業生產層面進行分析，討論的變數包含畜牧物種的飼養場數及在養數量，並針對臺南市在養數量於全國排名較為突出的物種進一步分析，探究歷年變化與飼養行政區分布狀況。接著進入畜牧業供給層面，包含各種畜產品的產量及產值，以洞悉臺南市畜產品產值結構及各種畜產品產值於全國排名狀況。此外，本文亦計算畜產品的區位商數（Location Quotient），以瞭解臺南市畜產品產業集中程度及競爭優劣狀況。最後利用 K-Means 演算法，以 104 年至 113 年畜產品區位商數，將畜產品分為 5 群，不同群的畜產品間存有不同競爭優劣勢情形。

依據分析結果，臺南市在多種畜牧物種或畜產品均處於全國前段的地位，惟不同畜產品產業集中或競爭優劣狀況有所差異，本文分析之結果與目前臺南市政府相關單位推動政策方向一致，期臺南市畜牧業能夠在市府的輔導下持續成長，維持高度競爭力及永續發展。

關鍵字：畜牧業、區位商數、K-Means 分群演算法

調查於「產銷平衡」政策之應用－以大蒜為例

劉維凱

農業部農糧署

摘要

大蒜為臺灣重要敏感作物之一，產區高度集中（雲林種植面積占全國 95%），產量與價格波動對農民收益及市場穩定影響甚鉅。藉由農糧署相關指定統計調查案例為基礎，探討統計調查在產銷平衡決策上的支援角色。

回顧全球大蒜產業現況，依 FAO 年報顯示，中國生產量占全球約四分之三，臺灣大蒜市場需求約 6.2 萬公噸，長期仰賴國產與進口並行之供應結構，進口占比約四分之一，主要自阿根廷與埃及進口。

在調查方法上，考量大蒜種植期與採收期集中，勞力需求高，產業面臨勞動力短缺與產銷失衡之雙重風險，故大蒜種植面積調查，使用科技工具 APP 依地籍地號逐筆現地調查，再輔以農業部農業試驗所衛星判釋交叉驗證，以確保調查的真實性與準確性；產量調查方面，規劃配置樣本農戶，並由各公所選定「具代表性」之農戶進行試割作業，依試割結果並綜合考量轄內生產情況，如實查報單位面積產量。相關調查結果則於每年 1 至 5 月發布於農政與農情期刊，以更即時提供決策參用。

量價分析顯示，近四年國產量維持約 5 萬公噸，產地價格於 110 年因疫情及國際行情大幅波動，惟近三年已趨於平穩。依大蒜生產成本調查，113 年每公頃總成本約 52.5 萬元，加計自給人工費、地租及資本利息後，農家每公頃可獲利 29.3 萬元，顯示其仍具經濟效益；另再運用生產成本資料設算「監控價格」，建置預警機制。

結論指出，統計調查能為決策單位提供合理種植規模與市場預警依據，協助避免因價格高漲而盲目擴種，造成供過於求並損及收益。未來將持續整合地面調查與遙測技術，並結合即時市場數據，以強化我國農業產銷調節能力。

關鍵字：大蒜、農業統計、產量預測、衛星判釋、產銷平衡、監控價格

結合資料庫查詢技術與決策樹模型
分析公費肺癌篩檢結果的重要特徵

吳東璋

高雄市政府主計處

摘要

癌症仍是國人主要死因，其中以肺癌致死率最高，且存活率因期別差異極大，凸顯早期篩檢的重要性。衛福部自 111 年起推動公費肺癌篩檢，針對高風險族群提供每 2 年 1 次的胸部低劑量電腦斷層攝影檢查，期望降低死亡率並提升健康水準。本文利用約 2 年半的篩檢資料，分析篩檢者特徵與結果，找出罹癌與非罹癌者及不前後期別患者的關鍵特徵，進而為政府制定更有效的肺癌防治政策提供建議。

透過資料庫交叉分析篩檢資料得知，相較於重度吸菸者，具有家族史者的肺癌罹患風險更高，且在女性中表現尤為明顯；男性多與吸菸相關，而女性則多與家族史相關，特別是若家族史來自女性親屬時，罹癌偵測率最高。整體而言，不論性別，家族史皆是比重度吸菸更具影響力的肺癌風險因子。

透過決策樹分析確診肺癌者特徵，「吸菸史包年」是影響肺癌診斷期別的首要因素，決策樹的第一個分歧點落在 40 包年，超過此門檻者後期肺癌風險明顯提高；而在 40 包年以下或無吸菸史者中，教育程度成為關鍵差異，教育程度較低者會因健康知識與檢查意識不足，較易延誤診斷。而家族史會受教育程度調節其差異：在教育程度較低者，是否有父母罹癌造成的差異較大；而教育程度較高者，家族史是否來自父母則影響相對較小。

關鍵字：公費、肺癌篩檢、重度吸菸、家族史、決策樹

嘉義市 YouBike 微笑單車使用情形探討

黃佳婷

嘉義市政府主計處

摘要

嘉義市為因應低碳城市與永續交通發展政策，積極推廣 YouBike 微笑單車作為市民短程移動與綠色運具的選擇。本文透過 112 年度營運報告與投保者租借紀錄，探討嘉義市 YouBike 之使用情形，藉由統計方法之運用以瞭解市民騎乘行為與使用趨勢。研究結果顯示，全年租借總量逾 175 萬車次，租借量與降雨呈現顯著負相關，使用高峰主要集中於上下班與上下課時段，且熱門站點以嘉義火車站周邊為核心。進一步運用 K-means 分群演算法分析各站點租借特性，將站點劃分為高頻、中頻與低頻三類，呈現不同地區與使用時段的差異。路線分析則指出，平日時段使用目的地以學校與辦公區為主，而假日則轉向觀光景點。另採用 PrefixSpan 演算法進行三站以上之序列分析，顯示「嘉義車站門戶－圓環噴水池商圈」及「嘉義大學校區內循環」為具代表性的騎乘模式。綜合上述，本研究建議未來可強化高頻站點之車輛調度，以確保供需平衡，同時加強低頻站點推廣，並持續優化整體路網配置與安全設施，以提升服務品質並促進市民對綠色交通工具之使用意願，進而落實低碳城市的發展目標。

關鍵字：YouBike、微笑單車、共享單車、K-means、PrefixSpan

應用象徵性資料分析方法於電子病歷應用之探析

高翊瑋 [1,2]

[1] 銘傳大學應用統計與資料科學學系 [2] 銘傳大學金融科技應用學系

Abstract

隨著醫療大數據的快速累積，如何有效分析電子病歷中的複雜資料結構成為重要議題。象徵性資料分析（Symbolic Data Analysis, SDA）提供了處理複雜資料型態的新視角，其在醫療領域的應用潛力值得深入探討。

本研究以某醫院的電子病歷資料庫為例，探索 SDA 方法在電子病歷分析的應用可能性。研究過程中，我們檢視了醫療資料的特性、評估了現有 SDA 方法的適用性，並使用 R 語言相關套件進行初步的實作嘗試。

本演講將分享研究探索的歷程與初步發現，包括資料處理的實務經驗、方法應用的觀察、以及過程中識別的研究方向。同時將討論 SDA 在醫療資料分析所面臨的機會與挑戰，並提出未來可能的發展建議。期待透過本次分享，與與會學者交流討論，共同推進此跨領域研究的發展。

Keyword: 象徵性資料分析(Symbolic Data Analysis), 電子病歷(Electronic Health Records), 醫療大數據(Medical Big Data)

Matrix Visualization of Categorical Big Data

高君豪

淡江大學統計與資料科學學系

Abstract

近年來，資訊科技的迅速發展導致需要處理和分析的資料量大幅增加。涵蓋計算技術和統計分析方法的巨量資料分析已成為一個關鍵的研究趨勢。在巨量資料的深入分析中，資料視覺化和探索性資料分析(EDA)將扮演重要角色。然而，值得注意的是，當前巨量資料分析的研究主要集中在連續型資料上，對類別資料的分析則相對獲得較少關注。在象徵型資料分析(Symbolic Data Analysis, SDA)的範疇內，類別模態多值型(Categorical Modal Multi-Valued type)的概念為類別巨量資料分析提供了另一種途徑。基於已知群組或分群方法的結果，類別巨量資料轉換為類別模態多值型象徵型資料，再利用 SDA 的方法進行分析。

本研究針對使用矩陣視覺化方式對類別巨量資料進行視覺化的三個主要挑戰進行相關研究並提出解決方案：(1)計算和排序關係矩陣的計算能力限制。(2)有效利用色彩空間來表示類別矩陣資料。(3)在有限螢幕空間的限制下有效呈現類別巨量資料矩陣圖。我們提出使用廣義相關圖(GAP)結合象徵型資料分析來進行矩陣視覺化和群集分析。通過將類別巨量資料轉換為模態多值型象徵資料，並採用廣義相關圖中的關係矩陣、排序和矩陣視覺化方法，建構一個能夠有效視覺化類別巨量資料的新 EDA 工具：cBigGAP。旨在克服類別巨量資料的關係矩陣計算、排序和呈現等相關的挑戰。

Keyword: Symbolic data analysis, Modal multi-valued data, Matrix visualization, Generalized association plots, Categorical big data, Dimension reduction

Symbolic Interval-Valued Regression Model

Liang-Ching Lin

Department of Statistics and Institute of Data Science, National Cheng Kung University, Tainan, Taiwan

Abstract

From micro data to macro data, symbolic data analysis can provide statistical inferences in the macro data while preserving as much information as possible from the micro data. In this study, we focus on the symbolic interval-valued regression model. The microdata are reorganized into intervals by using the largest and smallest order statistics. Then, we develop innovative symbolic interval-valued regression models to construct the relationship between two or more intervals. By defining the negative sign to the intervals, we can always keep the natural order that the higher value of the dependent variable will be larger than the lower value of the dependent variable, even when the parameters are negative. Firstly, we develop the simple linear symbolic interval-valued regression model and derive the corresponding maximum likelihood estimators (MLEs). In addition, the Fisher information matrix of the MLEs is shown, and the asymptotic normality is given. Next, we extend the aforementioned model to a multiple linear symbolic interval-valued regression model, and the corresponding maximum likelihood estimators are also provided. Monte Carlo simulations and real data analysis confirm the validity of the proposed method.

Keywords: asynchronous observations, order statistics, regression, symbolic data analysis.

Sliced Inverse Interval Estimation for Sufficient Dimension Reduction: a Symbolic Data Analysis Approach

吳漢銘

國立政治大學統計學系

Abstract

Sufficient Dimension Reduction (SDR) has become a central topic in statistics and machine learning, providing powerful tools for reducing the dimensionality of covariates in a nonparametric manner. This paper proposes a novel SDR method, Sliced Inverse Interval Estimation (SIIE), which draws inspiration from Symbolic Data Analysis (SDA) to improve the estimation of Effective Dimension Reduction (EDR) directions. Existing SDR approaches based on inverse conditional moments typically rely on single summary statistics, such as the mean, median, kth moment, variance, mean difference, or quantiles, thus neglecting much of the structural information contained within slices of the data. To address this limitation, SIIE introduces a new kernel matrix for central subspace estimation that incorporates interval information, yielding a richer representation of slice structures. We establish two theorems: under mild conditions, the directions estimated from inverse intervals span the EDR subspace, and the central subspace estimator is root-n consistent. The effectiveness of SIIE is demonstrated through simulation studies and real data analyses, where it is compared with existing inverse conditional moments–based SDR methods. The results show that SIIE, through its novel use of inverse interval estimation, more effectively captures the conditional distributional features of the response given high-dimensional covariates, thereby advancing both the theory and application of SDR.。

Keyword: Central subspace; Dimension reduction; Interval-valued data; Sliced inverse regression; Symbolic data analysis

Adaptive Forward-Backward Splitting for Multi-Layer Convolutional Dictionary Learning

Guan-Ju Peng

Institute of Data Science and Information Computing,

National Chung Hsing University

Abstract

Multi-layer convolutional sparse coding (ML-CSC) is a powerful technique for obtaining parsimonious representations of signals in machine learning applications. However, achieving maximum parsimony without sacrificing data fidelity requires learning a dictionary from data, resulting in the non-convex non-smooth optimization problem of multi-layer convolutional dictionary learning (ML-CDL). Existing ML-CDL solvers derived from traditional dictionary learning employ the convex $\ell 1$ function as the sparsity-inducing function to promote convergence; nevertheless, analyzing the convergence of candidate alternate optimization schemes is still challenging. In this paper, we reformulate the ML-CDL problem by replacing the convolution with the Hadamard product to reduce computational complexity and, to achieve this without destroying the consistency between the variables in the time and frequency domains, additional constraints are introduced to the reformulated problem. We also modify a forward-backward splitting algorithm to solve the reformulated ML-CDL problem using simultaneous coefficients and dictionary elements updates across all layers. In contrast, these variables are updated alternately in the previous approaches. The proposed algorithm includes a parameter adaptation scheme to improve optimization performance and ensure convergence. In numerical simulations, the proposed method outperforms existing methods by achieving convergence in fewer iterations and generating a smaller final functional value. The high quality of recovered images demonstrates the applicability of multi-layer sparse representation to image reconstruction.

Keyword: multi-layer convolutional sparse representation, dictionary learning, forward-backward splitting

基於少量資料微調的醫學影像語意分割模型應用與評估：以腸胃穿孔疾病腹腔游離氣體判讀為例

Fine-tuning Medical Image Segmentation Models with Limited Data: A Use Case on Abdominal Free Air Detection

廖柏翔[1, 2], 黃冠華[2]

[1] 臺北榮民總醫院急診部 [2] 國立陽明交通大學統計學研究所

Abstract

腸胃穿孔為致命急重症，在腹部電腦斷層中呈現氣腹（pneumoperitoneum），即有從腸胃道破出游離空氣 （free air）出現在腹膜中，如何有效率標註病灶游離空氣區域為一極具挑戰性的任務，不僅因其在臨床上或是影像上的出現頻率低，且形態變化多端且易和正常腸胃道空氣混肴。目前在影像切割任務中表現優異的 foundation model（如 SAM-Med3D）則為已用現有公開大型醫學影像資料庫進行訓練可以準確切割正常解剖構造或已有公開標註資料之結構，然而這些模型並未能分辨游離空氣等異常影像。

本研究嘗試以少量由放射科醫師人工標註之氣腹 CT 影像作為起始資料，進行 SAM-Med3D 模型微調（fine-tuning）。接著將更新後的模型應用於未標註資料，產出預測遮罩（segmentation mask），並由專家進行修正。此反覆進行之人機協作流程可持續提升模型表現，降低標註負擔。整體策略仿效 TotalSegmentator 的 iterative training 架構，強調自舊資料生成新預測，再由人為校正回饋模型，逐步強化模型辨識能力，設計一套半自動互動式的標註流程，以最小的人力成本建立高品質氣腹 segmentation 資料集，進而提升模型對稀有病灶的辨識能力。

本研究期望達成以下三項目標：(1) 加速氣腹標註流程；(2) 發揮大型 foundation model 的通用性，檢測少量樣本分批進入模型時對於預測的效果改善幅度；(3) 實驗設計比較此模型減少人工標註時間並計算與放射科專家人工標注之遮罩（segmentation mask）重疊面積評估（Dice score 等 segmentation metrics）。

Keyword: free air, foundation model, fine-tuning, segmentation

Deep Learning for Paired Chest X-Ray Image Analysis: Classification and Generation

戴煜偉[1], 黃冠華[1]

[1]國立陽明交通大學統計學研究所

Abstract

Longitudinal analysis of chest X-ray (CXR) images is crucial for monitoring disease progression and treatment response, but manual comparison of these images is challenging. Deep learning offers promising avenues for automating this process. This thesis investigates the application of deep learning to paired CXR images on two tasks: (1) multi-class classification of treatment responses and (2) predictive generation of post-treatment images. For classification, we proposed a Siamese network with a novel correlation-based feature fusion module. For predictive generation, a conditional generative adversarial network (cGAN) based on the Pix2PixHD framework was implemented. Experimental results showed that the Siamese network effectively categorized treatment outcomes and the cGAN model can produce anatomically coherent outputs.

Keyword: Chest X-rays, Deep learning, Longitudinal analysis, Siamese network, Conditional generative adversarial networks

An Unsupervised Adaptive Approach for Causal Inference in Competing Risks Data Using Separable Effects

廖柏竣[1], 黃彥棕[2], and 余日彰[1]

[1]國立台北大學

[2]中央研究院

Abstract

This study investigates causal effects when the primary time-to-event outcome is unobserved due to truncation caused by competing risks, wherein a competing event precludes its observation. Employing the separable effects framework, we disentangle the direct effect of exposure on the primary outcome from the indirect effect mediated through the competing event. Our goal is to develop a robust, unsupervised method that accommodates such data structures while retaining flexibility and estimation efficiency, thereby addressing biases due to dependent censoring. We propose a method based on the generalized transformation model, which adapts to a wide range of data distributions without requiring excessive parameterization. This unsupervised approach incorporates confounders as covariates for proper adjustment and is computationally efficient, making it suitable for moderate sample sizes. We theoretically establish that the proposed estimator is asymptotically consistent and weakly converges to a Gaussian process, thereby ensuring valid statistical inference. Simulation studies demonstrate that the proposed method outperforms existing nonparametric approaches by mitigating model misspecification, even when the true model lies outside the transformation class. Moreover, it efficiently adjusts for confounding without incurring heavy computational costs. Finally, we apply the method to assess the causal effect of hepatitis B virus (HBV) infection on the incidence of liver cancer as the first primary cancer, accounting for competing risks using data from the REVEAL study. The results suggest that HBV infection directly increases the incidence of liver cancer.

Keywords: Causal inference, Separable effect, Competing risks

Swarm-Based Search Procedure for Finding Optimal Multi-Stage Designs for Phase II Clinical Trials

Ping-Yang Chen

Department of Statistics, National Taipei University

Abstract

Multi-stage Phase II clinical trials offer advantages over single-stage designs by enabling interim analyses that can accurately inform early termination of the trial if there is evidence that the treatment is likely to be ineffective or effective. However, identifying optimal designs for multi-stage trials poses considerable computational challenges. In addition to having to optimize many integer-valued variables, there are multiple constraints, including order constraints. Traditional exhaustive search methods lack scalability and quickly become computationally infeasible when the number of stages is three or more. To overcome this challenge, we utilize a spherical coordinate system and reformulate the design problem as a continuous optimization task. The new formulation enables us to efficiently use Particle Swarm Optimization (PSO) to extend Simon's celebrated two-stage Phase II designs to three or more stages. Specifically, we show that our proposed search procedure not only reproduces the two-stage designs and certain three-stage designs found in the literature but also able to achieve the results more efficiently than traditional exhaustive search methods. We provide R codes for reproducing the optimal designs in this paper, which can be easily customized to generate tailor-made optimal designs for specific user needs.

Keyword: Adaptive Design, Maximum Sample Size, Minimax Design, Nature-Inspired Metaheuristics, Particle Swarm Optimization

Biomarkers and Disease Progression in Precision Medicine: Causal Insights from a Large-Scale Taiwanese Cohort

Ai-Ru Hsieh[1]

[1] Department of Statistics, National Taipei University, New Taipei City, 237, Taiwan

Abstract

Mendelian Randomization (MR) is widely applied to strengthen causal inference. However, standard two-sample MR may be biased by correlated exposures, temporal variation, and pleiotropy. Extensions such as multivariable MR (MVMR) and time-varying MR (TVMR) address these challenges. We evaluated their methodological performance and clinical relevance by investigating the causal role of glycated hemoglobin (HbA1c) in microvascular complications. Using data from the Taiwan Biobank, we used HbA1c as the exposure, while microvascular biomarkers were the outcomes. Causal estimates were compared across two-sample MR, MVMR, and TVMR. This study demonstrates that advanced MR frameworks can jointly address methodological biases and generate clinically meaningful insights. Our findings provide evidence that long-term glycemic exposure contributes causally to microvascular outcomes. Moreover, we also point to the need for integrating methodological advances with clinical application.

Keyword: Mendelian Randomization; Taiwan Biobank; HbA1c

Statistical Inference for Seamless Phase II/III Clinical Trials

Ken-Ning Hsu[1]

[1] PharmaEssentia Corp.

Abstract

Seamless phase II/III trials have been advocated following the US Food and Drug Administration's Critical Path Initiative lunched in 2004. Statistical methods have been proposed to address the treatment selection (phase II portion) and final test (phase III portion) issues. We compare two popular methods of Shun et al. (2008) and Friede et al. (2011) in terms of the familywise error rate and power as a function of interim analysis information time and the correlation between the short-term surrogate endpoint, on which the treatment selection is based, and the long-term primary endpoint, on which the final test is focused.

Keyword: Familywise error rate, exact distribution, treatment selection, combination of p-values, select the winner design

Two-Stage Calibration for GAMLSS: Distributional Regression with Validation Data under Unmeasured Confounding

Hui-Wen Lin, PhD

Department of Mathematics, Soochow University, Taipei, Taiwan

Abstract

This paper introduces a Two-Stage Calibration (TSC) procedure for Generalized Additive Models for Location, Scale, and Shape (GAMLSS), aimed at correcting unmeasured confounding in large observational studies. By leveraging a small validation subsample where the confounder is measured, the TSC method calibrates the full conditional distribution, including location, scale, skewness, and kurtosis, to achieve unbiased estimation while maintaining efficiency. Theoretical results establish the consistency and asymptotic normality of the TSC-GAMLSS estimators, and a nested nonparametric bootstrap is proposed for valid inference. A real data application to the National Health and Nutrition Examination Survey (NHANES) data demonstrates the practical utility of TSC-GAMLSS in estimating systolic blood pressure distributions while adjusting for body mass index (BMI) as a confounder.

Keywords: TSC-GAMLSS, unmeasured confounding, observational studies, bootstrap.

Semiparametric methods for the Interval-censored c-index

曾議寬

國立中央大學統計研究所

Abstract

The concordance index (c-index) is a widely used metric in medical research to assess the discriminatory ability of predictive models in distinguishing between low-risk and high-risk patients. It also evaluates the performance of models in predicting the probability of experiencing an event of interest. While numerous methods exist for estimating the c-index in right-censored survival data, research on c-index estimation for interval-censored survival data remains limited. To address this gap, we propose the development of semiparametric methods for estimating the c-index in interval-censored survival data. Our approach will integrate semiparametric hazard regression models, including the Cox proportional hazards model, the accelerated failure time (AFT) model, and transformation models. We will investigate the large-sample properties of the proposed estimation methods to ensure their theoretical validity. To evaluate their practical performance, we will conduct comprehensive simulation studies. Finally, we will apply our proposed methodologies to real-world data from the Taiwan AIDS cohort study to demonstrate their effectiveness in medical research applications.

Keyword: AIDS, Concordance index, Interval censor

電信資料視角下雙北居民平日日常生活圈的研究

黃怡婷[1*]、謝意盛[1]、王薇淳[1]、王鴻龍[1]、詹士樑[2]、林柏丞[2]

[1] 國立臺北大學統計學系

[2] 國立臺北大學不動產與環境學系

Abstract

智慧科技的發展，讓生活便利，並順帶留下寶貴的時空資訊，若能有效整合這些資料，可以提供或輔助過往需大量仰賴人工或問卷取得的資訊，如，人流或交通流量熱點、人流移動或通勤模式等。本研究利用去識別雙北整合某月份平日的電信資料，討論彙整時空電信資料為人口居住與移動特徵的方式，藉以瞭解雙北居民平日日常生活圈的規模與平日移動的範圍，並輔以國土地利用現況調查與開放街圖 (Open Street Map) 爬取興趣點資料，整合成網格的特徵資料，並利用空間統計模型分析生活圈規模與國土地利用現況調查與興趣點資料的關聯。研究結果可了解雙北生活圈樣貌與影響生活圈的相關因素。

Keyword: 電信信令、空間統計模型、國土地利用現況調查、開放街圖

探討整合信令資料推估日夜人流起訖矩陣之統計模型

黃怡婷[1]、嚴潔翎[1]、詹士樑[2]、林柏丞[2]、王鴻龍[1]*

[1]國立臺北大學統計學系

[2]國立臺北大學不動產與城鄉學系

起訖矩陣 （Origin-Destination Matrix, OD 矩陣）為常用彙整人口流動的重要方法之一，但傳統 OD 矩陣建制的資料收集需要仰賴大量的人力。近年因行動裝置普及，該裝置可紀錄用戶行動數據，讓彙整人口移動的方式有機會便容易。礙於個資法，一般是無法取得詳細用戶行動軌跡，但可取得彙整夜日縣市級別的 OD 矩陣。然而，從縣市級別的 OD 資訊無法應用到都會人口移動與交通動線資訊，本研究探討運用觀光人流的重力模型 （Gravity Model）來估算行政區級別的 OD 矩陣的可行性。

將台灣區分為六大生活圈，依生活圈建構行政區或鄉鎮級別的 OD 矩陣，本研究將常用的重力模型轉換為對數線性模型，加入區域的產業型態、土地利用等鄉鎮層級的興趣點（Point of Interest, POI）資訊來增加模型的解釋能力。另外，加入零膨脹模型來解決行政區或鄉鎮級別的 OD 矩陣存在過多零的情況，以及空間自迴歸模型（Spatial Autoregressive Model, SAR）及條件自迴歸模型（Conditional Autoregressive Model, CAR）來配適空間相依性。並以赤池信息準則 (Akaike Information Criterion, AIC)、均方根誤差 (Root Mean Square Error, RMSE) 與平均絕對誤差 (Mean Absolute Error, MAE) 作為模型篩選的評估指標。

研究結果顯示，除原有重力模型所考慮的變數，起訖點人數與距離，納入 POI（興趣點）可顯著提高的模型的準確性。在平假日的比較中，模型在假日時的表現更加穩定，而在人口密集地區，模型的預測誤差較高。此外，本研究亦探討空間相依性對跨區移動推估表現的影響，考慮空間結構後，在跨區移動的準確性顯著提高，採用以迄點構建的空間權重矩陣效果最為明顯，且推估值正確性高，說明納入地理資訊及空間結構都能有效改善模型表現。

關鍵詞：起訖矩陣、信令資料、卜瓦松迴歸、零膨脹卜瓦松迴歸、空間統計模型

利用人流時序特徵分析都市商業紋理之空間異質性

林柏丞[1], 徐貫瑩[2]

[1]國立臺北大學不動產與城鄉環境學系 助理教授
[2]國立臺北大學不動產與城鄉環境學系 碩士

摘要

隨著行動電話的普及，電信信令資料能夠在大範圍且精細的時空尺度上提供人口移動資訊。其中，經由時空網格加總的行動電話用戶數是常見的電信信令資料形式，從時間維度觀察，每個地理網格內的用戶數會隨時間變化形成時間序列，透過時間序列特徵的萃取，如人潮高峰時間與持續時間等，可有效簡化時間序列的複雜性，並保留人流時空規律的特徵。都市人口移動的高度時空規律，反映了人與環境之間的互動關係，而都市商業紋理則在此互動過程中扮演關鍵角色。興趣點（Point of Interest, POI）是反映人們造訪特定地點的意圖與偏好，亦是進行商業紋理分析的重要資料基礎。本研究以臺北都會區為研究範圍，透過時序特徵萃取建立描述人流聚集現象的指標，包括人流峰值(MAX)、人流聚集頻率(CAM)、首波高峰時段(FLMax)、高峰傾斜型態(Sk)、高人流持續時間(LSAM)與低人流持續時間(LSBM)，進一步應用多尺度地理加權迴歸(Multiscale Geographically Weighted Regression, MGWR)分析不同 POI 類別與人流時間序列特徵之間的空間關聯，分析結果顯示，不同 POI 類型對人流時間序列特徵的空間影響存在顯著差異。整體而言，本研究的空間分析架構，不僅提供都市空間行為的定量解釋途徑，也提供在無須取得完整時間序列情況下進行人流分析的可行途徑。

關鍵詞：電信信令、時間序列特徵、空間異質性

A Quantile Cure Model with Fully or Partially Functional Covariate Effects

Chyong-Mei Chen

Institute of Public Health, School of Medicine, National Yang Ming Chiao Tung University

Abstract

The quantile regression has several attractive features, such as its ability to allow covariate effects to vary at different quantile levels and to handle heteroscedasticity in data easily, which make it a viable alternative when analyzing data with continuous outcomes in recent years. In particular, it has been used in modeling survival data with and without a cured fraction. In this work, we propose novel estimating equation approaches to estimate a mixture cure model where the latency survival time is modeled by a quantile regression. The proposed estimation methods enjoy a double robustness in the sense that a misspecification in one of the two parts in the mixture cure model will not affect the estimation in the other part. The methods do not require the global log-linear assumption in the quantile regression, and they allow mixed effects of functional and constant effects in the regression when the log-linear assumption is hold in an interval of quantile levels. We established the asymptotic properties of the proposed estimators. Our simulation studies demonstrated the double robustness and the efficiency gains in the proposed estimators. An application of the proposed model and methods to data from a lung cancer study revealed new and interesting findings that were not reported in a previous analysis of the data. This is joint work with Prof. Yingwei Peng.

Keyword: Estimating equation; Inverse probability censoring weight; Mixture cure model; Quantile regression model

Conditional Independence Testing for General Sufficient Dimension Reduction Methods

Hsin-wen Chang[1], Kerby Shedden[2], Shih-Hao Huang[3]

[1] Academia Sinica [2]University of Michigan [3]National Central University

Abstract

We study conditional independence testing within the sufficient dimension reduction (SDR) framework. The goal is to assess whether selected predictors contribute to explaining the response after controlling for the others, with SDR alleviating the curse of dimensionality and preserving modeling flexibility. We propose a novel procedure that performs conditional independence testing by combining appropriate residualization with SDR dimension testing. The procedure is adaptable to a broad class of SDR methods, allowing the direct application of existing dimension tests. Simulations show our procedure achieves empirical performance comparable or superior to that of existing methods in several settings.

Keywords: Conditional independence test, Coordinate test, Dimension test, Residualization, Sufficient dimension reduction

Improving unbiasedness of the proportional hazards model estimator with Cox and Snell's bias approximation and jackknife resampling

Chia-Hui Huang[1], Ruby Chiu-Hsing Weng[1]

[1] Department of Statistics, National Chengchi University, Taiwan

Abstract

Bias approximation has played an important role in maximum likelihood estimation method, and numerous bias calculation techniques have been proposed under different contexts. For the semiparametric proportional hazards model, which is the standard regression method to study the time-to-event data, the existing work applied the bias formula and derived the approximate bias of Cox's estimator based on the partial likelihood function. In this work, we instead use the joint likelihood function and utilize the counting process approach to develop approximate bias of Cox's estimator. Explicit expressions for the higher order partial derivatives are derived, which facilitate the bias calculation techniques. We also incorporate jackknife resampling method and propose a Jackknife-Cox-Snell method that processes the biasedness of Cox's estimator through two steps, where the first step aims for removing the analytical terms derived from Cox and Snell's formula and the second step reduces the residual bias term then. A comprehensive simulation study is performed to evaluate the usefulness of the proposed bias-corrected method and the results show that our proposed estimator outperforms the existing corrected estimators. We also apply the proposed method to two sets of survival data for comparison and illustration. ◦

Keyword: Bias calculation, Counting process, Cox and Snell's formula, Jackknife, Nonparametric maximum likelihood estimator, Proportional hazards model

Geographically weighted Poisson-Tweedie model for count data

陳怡如[1], 李沂瑾[1]
[1] 政治大學統計學系

Abstract

Geographically weighted regression (GWR) is a widely used localized modeling technique for exploring spatial heterogeneity in data relationships. Owing to its intuitive specification and interpretability, recent years have witnessed active developments in extending the GWR framework to accommodate count response variables. However, most existing approaches either rely on restrictive distributional assumptions (e.g., Poisson, negative binomial) that may not capture the full range of dispersion and distributional complexities, or adopt two-part mixture formulations (e.g., zero-inflated models) that can complicate estimation and interpretation. In this study, we propose the geographically weighted Poisson-Tweedie model (GWPTM), which integrates the Poisson–Tweedie family within the GWR framework to provide a unified and flexible tool for spatial count data analysis. By specifying variance as a power function of the mean, GWPTM accommodates a broad range of mean–variance relationships and embeds a class of count distributions under a single framework, including Poisson and negative binomial distributions as special cases along with other members of the Poisson–Tweedie family. This formulation enables the model to naturally account for features such as equi-, over-, and underdispersion, zero inflation, asymmetry, and tail heaviness, while allowing both regression coefficients and distributional parameters to vary across space. We develop estimation procedures, draw statistical inference for local parameters, and discuss practical modeling considerations. Simulation experiments are conducted to examine and validate the model performance. Finally, we apply the proposed method to analyze dengue data from Taiwan, demonstrating its practical value through empirical illustration.

Keyword: Geographically weighted regression, Spatial heterogeneity, Count data analysis, Poisson-Tweedie model, Overdispersion, Zero inflation

A Flexible Scan Statistic for Spatial Cluster Detection: A Dengue Fever Case Study in Southern Taiwan

Yi-Hung Kung

Department of Statistics and Information Science, Fu Jen Catholic University, Taiwan

Abstract

Dengue fever remains a major mosquito-borne viral disease, with over 50–100 million infections annually and a disproportionate burden in the Asia-Pacific region. Taiwan illustrates both the successes and challenges of control: stringent quarantine and vector management kept incidence low from 2003 to 2014, yet global resurgence led to large epidemics in 2014–2015 and again in 2023. These recurring outbreaks underscore the urgent need for improved surveillance and cluster detection methods.

Geographical cluster analysis has become central to identifying hotspots, but existing approaches often fail to adjust for spatial autocorrelation and environmental covariates, producing biased or unstable results. To overcome these limitations, we introduce SEMSCAN, a framework embedding scan statistics within a spatial error model. This integration allows simultaneous estimation of covariate effects and spatial clustering, while directly accounting for residual spatial dependence. SEMSCAN improves robustness, reduces computational burden, and enhances the validity of hotspot detection in heterogeneous risk environments. This framework offers a practical and rigorous tool for strengthening dengue surveillance and guiding targeted vector control strategies.

Keyword: Spatial Epidemiology, Spatial Scan Statistics, Disease Hotspot Detection, Spatial Autocorrelation

Assessing Predictive Performance in Spatial Count Data: A Comparison of Advanced Modeling Approaches

曾聖澧[1], 張家輔[2], 施嘉翰[2]

[1] 國立中興大學統計學研究所 [2]國立中山大學應用數學系

Abstract

Spatial count data, common in fields like ecology and epidemiology, often exhibit heterogeneity and spatial dependence, challenging traditional assumptions.

This study conducts a comparative analysis of spatial count data models, focusing on predictive performance under varying data characteristics and parameter settings. Seven modeling approaches are evaluated: Gaussian Copula Negative Binomial (GC-NB) models using GHK and GQT methods, spatial Poisson and negative binomial random fields, Gaussian misspecification models, and Gaussian random fields with different likelihood methods. Through simulations and real datasets, predictive accuracy is assessed using RMSE metrics, considering factors such as overdispersion, spatial correlation, and range parameters.

Results show GC-NB models consistently outperform others in accuracy and robustness, even when assumptions are violated. Other models struggle with high overdispersion or sensitivity to data structure. Real data analysis supports the effectiveness of GC-NB models in capturing spatial clustering and high-density areas, offering valuable guidance for spatial count data modeling.

Keyword:

Spatial Count Data, Predictive Performance, Gaussian Copula, Overdispersion

Session III

[III-1]：政府場次 3　　人文館-國際會議廳

Chair：陳惠欣（交通部統計處處長）

[III -2]：Recent advances and development in statistical modeling

人文館-第一會議室

Organizer：孫立憲（國立中央大學統計研究所）

Chair：孫立憲（國立中央大學統計研究所）

[III -3]：高維資料與深度學習的統計分析：慶祝陳素雲教授榮退專場　　人文館-第二會議室

Organizer：陳定立（中央研究院統計科學研究所）

Chair：楊欣洲（中央研究院統計科學研究所）

[III -4]：計算統計　　人文館-遠距會議室

Organizer：林宗儀（東海大學 統計學系）

Chair：林宗儀（東海大學 統計學系）

[III -5]：統計及資料科學應用　　活動中心-第一會議室

Organizer：鄭宗記（淡江大學統計與資料科學學系）

Chair：鄭宗記（淡江大學統計與資料科學學系）

[III -6]：生物統計　　活動中心-第二會議室

Organizer：温啟仲（淡江大學應用數學與數據科學學系）

Chair：温啟仲（淡江大學應用數學與數據科學學系）

15:20 ~ 16:50

AI 浪潮驅動我國製造業逆勢翻紅的關鍵產品分析

劉軒宇

經濟部統計處

摘要

自 111 年底 OpenAI 推出 ChatGPT3.5 後，生成式 AI 的應用開始逐漸發酵，帶動高效能運算需求不斷攀升，我國因擁有領先全球的半導體先進製程技術及較完整的產業供應鏈，成為 AI 浪潮下的關鍵生產重地。

對此，本文試圖藉由觀察我國製造業各產品國內生產價值的變動趨勢，離析此波 AI 商機中表現亮眼的關鍵產品，並探究 AI 浪潮對我國製造業及出口結構之影響，最後輔以國際比較及簡述本土供應鏈強化競爭力之作為。

關鍵字：AI、高效能運算、半導體

育嬰留停新趨勢
—促進男性公教人員育嬰留職停薪申請之性別分析

游舒淳

臺中市政府主計處

摘要

我國在推動性別主流化工作方面已取得顯著成就，性別平等表現位居先進國家之列。然而，在低生育率、兩性勞參率差距等結構性議題上仍有待改善。「請領育嬰留職停薪津貼男女性比例」是衡量家庭育兒責任分擔與職場性別平等的關鍵性別統計指標。本研究旨在透過跨部門的性別統計數據分析，探討育嬰留職停薪（育嬰留停）申請在公教人員與勞工群體間的性別差異及其背後的主因。

統計分析結果顯示，儘管總體育嬰留停申請件數在少子化趨勢下仍穩定成長，且男性申請比例已由早期不足 5%大幅提升至 112 年的 25.6%，呈現育兒責任分擔轉變的趨勢。然而，數據亦揭露一項關鍵的性別落差：公教保險男性申請比例與就業保險（勞工）男性申請比例之間存在約 10 個百分點的顯著落差。此差異主因在於公教保險的育嬰留停津貼計算基礎僅為本俸，導致其所得替代率（約 40%）遠低於勞工（約 80%），致使公教家庭，特別是男性公教人員，面臨更高的經濟成本，抑制了其申請育嬰留停的意願。

為促進公務職場的性別平等，本研究運用性別影響評估提出政策建議，核心在於透過經濟誘因來消除公教群體的性別落差。建議採行「促進公保給付平等」方案：將公保育嬰留停津貼的計算基準擴大，加發等同本俸 80%的津貼補助，使公教人員與勞工的所得替代率趨於一致，以大幅提高男性公教人員申請意願。同時，亦建議併行「激勵男性育嬰參與」方案，透過親職教育課程與實務育嬰公假，強化準爸爸員工在育兒過程中的實質參與度。本研究的發現與建議，可為政府在推動更具性別敏感度及可近性的育兒支持政策上，提供重要的統計實證基礎與政策方向。

關鍵字：育嬰留職停薪、性別統計、公教保險、就業保險、所得替代率

結合交通事故大數據及空間資訊
—探討桃園市交通事故特性及熱點

張　珣

桃園市政府主計處

摘要

為確保行人和車輛能夠在更安全的環境中共存，降低交通事故風險，本文利用本府警察局 108 年至 113 年桃園市道路交通事故資料，觀察事故之肇事原因大多為車輛駕駛人因素，且主要使用運具為機車，次為汽車，考量不同年齡層之駕駛事故有其差異性，爰以羅吉斯迴歸模型分析駕駛於各年齡層死傷程度之重要特徵，同時分析造成行人交通事故之死傷風險，結果顯示汽機車駕駛於半夜及凌晨時段死傷風險較高，機車駕駛於交叉路口內及汽車駕駛於路肩、路緣死傷風險較高，另機車路上翻車、摔倒有高死傷風險；行人於凌晨時段、行人穿越道死傷風險較高，又 65 歲以上行人死傷風險較高。

為呈現本市交通事故的空間分布，經由前述分析結果，挑選汽機車駕駛及行人於交通事故中較高死傷風險之變數類別，計有 12 項，以 QGIS 繪製熱點及熱區地圖，結果顯示高死傷風險路段多位於桃園火車站及中壢火車站附近道路，其中以桃園區境內的中正路、春日路、三民路三段及中山路較常發生死傷情形。本文建議加強取締用路人未依規的行為，持續鼓勵民眾搭乘大眾運輸，強化駕駛交通安全教育宣導，積極推動人本交通及行人教育宣導，以降低交通事故風險，保障民眾生命安全。

關鍵字：交通事故、駕駛、行人、熱點

企業動態統計下的性別創業差異

楊子江

財政部統計處

摘要

本研究以財稅資料為基礎，導入 OECD 與歐盟統計局提出之企業動態統計（Business Demography Statistics）概念，串接營業稅、綜所稅、勞保及戶政等跨域資料，建構企業出生與存續指標，以占創業家數 6 成、性別資訊可信度較高之獨資組織為標的，採羅吉斯迴歸模型分析創業者特徵與行業選擇的關聯性，並應用 Kaplan-Meier 存活曲線及 XGBoost Cox 模型，探討影響企業存續的關鍵因素。研究發現，近年女性創業貢獻日益顯著，112 年占比達 4 成 5，且我國獨資企業女性創業占比在國際位階持續進步。受延遲婚育及低薪經濟壓力影響，創業者呈年輕化，未婚創業者家數 5 年內增 8 成，兼職創業比例亦增至近 3 成。此外，首次創業且具家庭創業經驗者占整體 4 成，凸顯家庭支援於創業的重要性。創業行業選擇方面，兩性皆高度集中，女性居冠者為電子購物及郵購業，男性為餐館，惟部分行業存在單一性別極化現象，凸顯女性創業偏向興趣與兼顧家庭考量，男性則著重專業與技能延伸。而年輕、未婚者創業則多以興趣導向，離婚者注重現金流與彈性。創業存活分析方面，年銷售額低於百萬元、離婚者及年輕創業者（24 歲以下）的企業存活率較低，中年創業者（35 歲至 44 歲）、有偶者及首次創業者則相對較高。影響企業存活的重要變數依序為企業年銷售額、創業者年齡及婚姻狀態。

關鍵字：企業動態統計、創業存續率、性別創業差異、獨資組織、財稅資料

Precise—Then Even More Precise: Revisiting Discrete Option Pricing

Hsin-Chieh Wong[1]（翁新傑）

[1]Department of Statistics & Fintech and Green Finance Center,
National Taipei University, New Taipei City, Taiwan

Abstract

The continuity correction approximation developed by Broadie, Glasserman, and Kou (1997, 1999) (BGK) for discretely monitored barrier and lookback options remains one of the most widely used analytical results. However, despite a generation of subsequent research, no sharper or higher-order refinement of the continuity correction has been established for these options. This paper presents a refined asymptotic expansion of the BGK approximation for path-dependent option prices, using the smooth-fit principle. The resulting higher-order correction continues to shift the barrier level, but the adjustment factors are no longer determined solely by the underlying volatility. Instead, the new formula incorporates the drift and introduces interpretable economic terms that shed further light on the structure of the correction. Numerical examples are also provided to illustrate and validate the theoretical improvements.

Keyword: Overshoot, first-passage time, Laplace transform, boundary crossing.

Bayesian Item Response Theory Models with Local Dependence

張育瑋[1], 劉孟展[2]

[12] 國立政治大學統計學系

Abstract

Based on response on several items by each of subjects, Item Response Theory (IRT) models aim to make inference on latent factors, item discrimination, item difficulties, and so on. Local independence is a model assumption required to form the likelihood function in classical and neat IRT models. However, the assumption has been found to be violated in many real applications, such as testlet items, mixture models, and time limit tests. Therefore, local dependence (LD) modelings gain much attention in recent year. In the current study, we focus on the LD model in Chen, Li, Liu, and Ying (2018), where they incorporated the Ising model (1925) into IRT models to capture LD among items and proposed some frequentist estimation procedure. In order to further allow those LD parameters to borrow information from each others, we propose some Bayesian inference for the LD model. The main challenge of statistical inference is the intractable posterior distributions. We solve the issue using the pseudo likelihood methods. Simulation studies are conducted to demonstrate the validation of the proposed estimation procedure under various conditions.

Keyword: Bayesian estimation; Ising models; Item Response Theory models; local dependence; pseudo likelihood.

Change point estimation for Gaussian time series data with copula-based Markov chain models

Li-Hsien Sun[1], Yu-Kai Wang[1], Lien-Hsi Liu[1], Takeshi Emura[2], and Chi-Yang Chiu[3]

[1]Graduate Institute of Statistics, National Central University, Taoyuan City 320317, Taiwan

[2]Institute of Statistical Mathematics, Tokyo, Japan

[3]Division of Biostatistics, Department of Preventive Medicine, University of Tennessee Health Science Center, Memphis, TN 38163, USA

Abstract

This paper proposes a method for change-point estimation, focusing on detecting structural shifts within time series data. Traditional maximum likelihood estimation (MLE) methods assume either independence or linear dependence via autoregressive models. To address this limitation, the paper introduces copula-based Markov chain models, offering more flexible dependence modeling. These models treat a Gaussian time series as a Markov chain and utilize copula functions to handle serial dependence. The profile MLE procedure is then employed to estimate the changepoint and other model parameters, with the Newton-Raphson algorithm facilitating numerical calculations for the estimators. The proposed approach is evaluated through simulations and real stock return data, considering two distinct periods: the 2008 financial crisis and the COVID-19 pandemic in 2020.

Keywords: Change point, Copula, Serial dependence, Sequential analysis, Gaussian time series, Markov chain

Integrating Multi-Distribution Initialization for Robust CNN Models in Medical Image Recognition

Tai-Been Chen[1], Yung-Hui Huang[2]

[1]Department of Radiological Technology Faculty of Medical Technology, Teikyo University, Tokyo, Japan

[2] Department of Medical Imaging and Radiological Science, I-Shou University, Kaohsiung, Taiwan

Abstract

The stability and accuracy of convolutional neural networks (CNNs) for medical image recognition are strongly influenced by weight initialization strategies. Conventional approaches such as He and Xavier initialization often yield inconsistent results across diverse imaging modalities, particularly when data distributions deviate from Gaussian assumptions. To address this limitation, this study introduces a *multi-distribution initialization* strategy that integrates diverse statistical distributions—including normal, uniform, beta, gamma, and exponential—to generate representative starting weights. By combining and rescaling samples from multiple distributions, the method provides a more versatile initialization scheme that improves convergence stability and generalization.

Experimental evaluation was conducted using five Kaggle datasets: breast ultrasound (benign, malignant, normal), brain MRI (glioma, meningioma, pituitary, healthy; tumor vs. no-tumor), chest X-ray (COVID-19, pneumonia, normal), and kidney CT (cyst, stone, tumor, normal). Models were trained under three optimizers (Adam, SGDM, RMSProp) and assessed with classification accuracy, Cohen's Kappa, and runtime.

Results demonstrate that multi-distribution initialization achieves robust performance across modalities, offering stable accuracy and significantly reduced training time in tasks such as chest X-ray and kidney CT classification. Compared to He and Xavier methods, which showed higher variance across datasets, the proposed strategy consistently balanced precision and efficiency.

This research establishes multi-distribution initialization as a practical approach for robust CNN training in medical image recognition, with strong potential for deployment in both clinical and computationally constrained environments.

Keywords: CNN, weight initialization, multi-distribution strategy, medical image recognition, generalization, robustness.

Multivariate Contaminated Normal Linear Mixed Models with Censored and Missing Outcomes Applied to Alzheimer's Disease Data

王婉倫[1,*], 林宗儀[2]

[1]國立成功大學 統計學系
[2]國立中興大學 統計學研究所

Abstract

The paper proposes a robust approach to jointly modeling multiple repeated clinical measures with intricate features. More specifically, we aim to expand the scope of the multivariate linear mixed model by using the multivariate contaminated normal distribution. The proposed model, called the MCNLMM-CM, is designed to handle minor outliers effectively while simultaneously accommodating censored measurements and intermittent missing responses. An expectation conditional maximization either (ECME) algorithm is developed to estimate the parameters of the proposed model in situations involving missing at random responses. We also provide techniques for approximating the asymptotic standard errors of model parameters, recovering censored data, imputing missing values, and identifying outliers. The proposed methodology is inspired by and applied to data from the Alzheimer's Disease Neuroimaging Initiative cohort study, which involves longitudinal clinical measurements of patients with mild cognitive impairment. A simulation study is conducted to evaluate the finite-sample properties of the parameter estimators.

Keyword: Censored data; EM algorithm; Longitudinal data; Missing data; Outliers

Nonparametric Profile Monitoring with Error-Prone Auxiliary Variables

Wei-Heng Huang

Department of Statistics, National Taipei University

Abstract

Profile monitoring has been one of the important topics in statistical process control. Unlike the usual settings that monitor a variable, profile monitoring aims to incorporate an auxiliary variable to characterize the primarily interested variable. To efficiently adopt an auxiliary variable to connect the main variable under the unknown relationship, nonparametric regression is a possible strategy. However, in applications, variables are possibly subject to measurement errors, which may lead to unreliable results if ignored in the analysis. In this paper, we investigate nonparametric profile monitoring when an auxiliary variable is contaminated with measurement error. We consider the regression calibration correction on the Nadaraya-Watson estimator. To monitor the variable and to detect the out-of-control samples, we adopt the asymptotic normality of the corrected Nadaraya-Watson estimator to derive the confidence band and treat it as the control limits. A series of numerical settings are considered to assess the performance of the proposed method, and the results verify the importance of taking measurement error effects into account.

Keywords ： Nadaraya-Watson regression; nonparametric regression; regression calibration; statistical process control.

Exploring Arbitrariness in Sentencing for Murder in Taiwan, 2013-2024

Han-Wei Ho [1], Keng-Wei Fan [2], Kuan-Ting Chen [3]

[1] Institutum Iurisprudentiae, Academia Sinica, Taiwan; [2] Department of Law, National Dong Hwa University, Taiwan; [3] Cornell Law School, U.S.A.

Abstract

Preventing arbitrary punishment, especially the death penalty, is a fundamental human rights goal and a key objective for judiciaries worldwide. The idea of a non-arbitrary decision, inspired by the concept of *Rule of Law*, is meant to be informed by reasonable, relevant factors (no implicit bias), and be consistent and unaffected by the noise to a certain degree (robustly predictable). This paper aims to rigorously examine whether court decisions are arbitrary for intentional murder, the most common crime eligible for the death penalty in Taiwan. We first follow common empirical studies that focus on the former, legally irrelevant factors — demographic disparities — with regressions. Then, to evaluate potential arbitrariness holistically by applying the principle of "treat like cases alike", we use UMAP, a dimension reduction algorithm, to characterize case similarity given plenty of relevant features, with outcomes marked. The two outcomes received could be considered significantly different if the symbols imposed on UMAP to represent different sentences are in the immediate proximity. Our results not only provide prima facie reason to suspect potential arbitrariness in the criminal judgments, but also offer a new direction for both the empirical discourse on sentencing practices and the contentious issue of the death penalty.

Keywords: arbitrary(-ness), death penalty, sentencing, treat like cases alike, dimension reduction (DR), (intentional) murder

Bridging Deep Learning and Statistical Dimension Reduction: An SNN–PCA Framework for Efficient Learning

周珮婷

國立政治大學統計學系

Abstract

Deep learning models offer strong predictive capability but are often criticized for high computational cost and limited interpretability, particularly in data analysis. This study introduces a statistically grounded framework that integrates Siamese Neural Networks (SNNs) with Principal Component Analysis (PCA) to enhance representation learning and model efficiency. The SNN is employed as a nonlinear feature extractor that maps high-dimensional inputs into a latent space with improved class separability. These learned representations are then incorporated into classical statistical learning models to strengthen anomaly detection and classification performance. PCA is further applied to SNN-derived features and network weight structures to quantify contribution, reduce dimensionality, and prune redundant neurons. Empirical studies demonstrate that the SNN–PCA framework improves F1-score and balanced accuracy while substantially reducing network complexity. This work illustrates how statistical dimension reduction can be synergistically combined with deep learning to deliver interpretable, computationally efficient, and high-performing models for structured data applications.

Keyword: Siamese Neural Networks (SNN), Principal Component Analysis (PCA), Dimensionality Reduction, Classification, Network Pruning

Dynamic clustering township level data by the mixture of longitudinal factor analyzer approach

黃欣卉[1], 鄭宗記[2]

[1]國立政治大學統計學系 [2]國立政治大學統計學系

Abstract

Previous studies have demonstrated that counties, towns, districts, and cities (collectively referred to as townships) can be grouped into several clusters, with each cluster exhibiting similar patterns across regional and demographic indicators. Such classifications help policymakers formulate region-specific development strategies. However, the development trajectories of these townships are characterized by substantial heterogeneity and temporal dynamics. Traditional clustering approaches are limited in their ability to capture both the complex structure of latent classes and the latent factors that evolve over time in longitudinal data. The present study applies a mixture of longitudinal factor analyzer approach to township-level data spanning 2011 to 2022 in Taiwan. This method uncovers the diverse developmental trajectories of townships and identifies distinct developmental types, along with the underlying latent factors that drive them, such as socioeconomic vitality and aging pressure. The resulting cluster classifications extend beyond conventional measures of urbanization to encompass the economic, social, and cultural characteristics unique to each area.

Keyword: longitudinal clustering, longitudinal factor analysis, mixture of longitudinal factor analyzer, township level data

The estimation of the AFT model with measurement errors–the Buckley James approach

Yih-huei Huang

Department of Applied Mathematics and Data Science, Tamkang University

Abstract

The Accelerated Failure Time (AFT) model is a widely used framework in survival analysis, providing an intuitive interpretation of the effects of covariates on log-lifetime. This paper addresses the estimation problem when covariates are subject to measurement error. We begin by correcting the bias in the estimating function using the traditional approach, assuming no censoring, and then extend the method by computing its conditional expectation given the censoring indicator, in line with the Buckley–James estimation framework. However, the computation requires estimating the distribution of the adjusted lifetime. Since measurement error induces dependence between the adjusted lifetime and censoring time, we propose employing Beran's estimator to address this complication. Simulation results demonstrate that the proposed method outperforms conventional regression calibration and remains consistent under completely random censoring.

Bayesian Analysis of Mixed Interval Censored Data Under Bernstein Proportional Hazards Cure Model

吳裕振 (Yuh-Jenn Wu)
中原大學應用數學系
yuhjenn@cycu.edu.tw

摘要

本論文主要是提出一個貝氏統計方法來分析具有治癒子群的混合區間設限存活資料，並探討該群的治癒率(不含未發病)。雖然已經有很多專家學者研究如何分析具有治癒子群的右設資料，研究具有治癒子群之區間設限存活資料還相對少。近年來也有一些方法被提出來討論如何分析具有治癒子群之區間設限存活資料。比如 Liu (2012)利用非混合比例勝算比治癒模型來分析現狀數據；Li, etal. (2019)及 Yang, etal. (2021)則在混合區間設限資料研究方面，提出更廣義於比例勝算比和比例風險比治癒模型的非混合型聚類治癒模型。雖然這三種模型都是具有生物學上的涵義，他們都使用最大概似估計法來處理問題。對於這類問題，在 Chen(2023)使用 Yang, etal. (2021)提出的模型下，但他使用貝氏統計方法來分析混合區間設限存活資料,而且和其他學者提出使用韋伯分佈或階梯函數函數建立潛在因子活化時間的累積分佈函數等方法不同，他使用的是伯氏多項式多項式來建立，但他在模擬研究方面，計算覆蓋率並不是很精確，也沒有證明貝氏估計的一致性。本論文我們探討比Chen(2023)所提出的較簡單的情形，即比例風險比治癒模型的貝氏統計方法來分析,實作馬可夫鏈蒙特卡羅方法來計算貝氏估計,而且證明了貝氏估計的一致性,在模擬研究方面，估計和計算覆蓋率都有不錯的表現。

關鍵詞： 貝氏統計,比例風險比治癒模型,混合區間設限資料,伯氏多項式

This is a joint work with Li-Hsueh Cheng, Wei-Fan Chen, Ching-Ti Liu and Li-Chu Chien.

A modified two one-sided test for reference-scaled average bioequivalence

姜杰 [1]

[1] 淡江大學應用數學與數據科學系

Abstract

The US FDA has released the first revision of the draft guidance “Statistical Approaches to Establishing Bioequivalence” at the end of 2023. In this revision, the reference-scaled average bioequivalence (RSABE) is recommended for in-vitro permeation tests and assessments of narrow therapeutic index drugs. The corresponding test is suggested to be based on an upper confidence bound by Howe method for the linear combination of squared mean difference and variance(s). The statistical chrematistics for the confidence bound is difficult to derive and, hence, the sample size determination is often based on simulation. In this study, we proposed a modified two one-sided test (MTSOT) alternatively with corresponding sample size determination. Simulation shows that the proposed MTOST can control the type I error rate and provide sufficient level of power. Moreover, it is powerful than the Howe method.

Keyword: reference-scaled average bioequivalence; sample size determination; two one-sided test

Adaptive Lasso for Randomized Response Survival Data

Chi-Chung Wen

Institute of Public Health, National Yang Ming Chao Tung University

Abstract

Surveys on sensitive topics frequently yield current status survival data that are subject to response bias. The randomized response technique (RRT) offers a principled approach to mitigating such bias by enhancing respondent privacy. In this work, we present an adaptive lasso procedure for variable selection in current status survival data collected under RRT. We establish the oracle property with valid standard error inference and accommodate the latent RRT structure through a modified shooting algorithm within an EM framework. Simulations and an application to survey data on extramarital sex demonstrate the method's validity and practical relevance.

Keyword: Current status data; Oracle property; Privacy-preserving; Shooting algorithm, Variable selection