www.kss.or.kr

Proceedings of the KSS Winter Conference 2025

2025년도 한국통계학회 동계 학술논문발표회 프로시딩

일 시	2025년 12월 19일(금) ~ 20일(토)
장 소	서울대학교(관악캠퍼스)
주 최	(사)한국통계학회
주 관	(사)한국통계학회, 서울대학교 통계학과, 서울대학교 과학데이터혁신연구소

소지역통계와 경계를 이용한 나만의 지도

SGIS plus 통계지리정보서비스

소지역통계·경계서비스

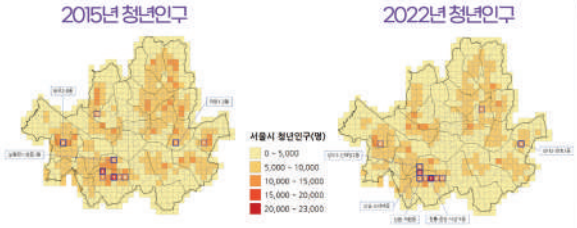
"24년 11월부터 제공"

CSV+SHP

"자료제공 많이 이용해 주세요"

✦ 지도 시각화 분석 사례 ✦

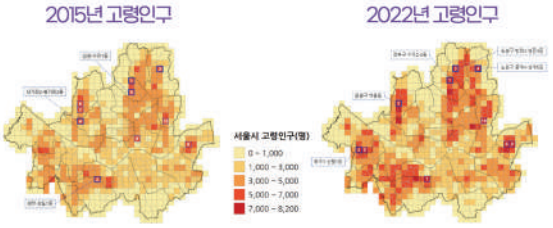
01



서울시 청년인구 격자별 순위 분석

관악구 중심으로 청년인구 순위가 높으며,
인구 분포 변화는 상대적으로 적음

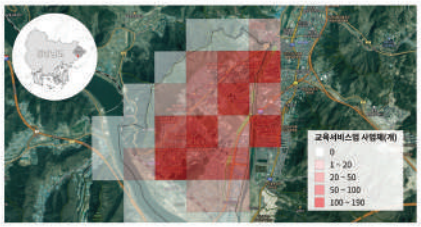
02



서울시 고령인구 격자별 순위 분석

강북구·은평구 등 시 외곽에 고령인구 상위 순위가
분포하며, 고령화 진행확인

03



양산시 물금읍 교육서비스업 격자별 분석

2023년 교육서비스업이 가장 많은
행정동의 격자통계를 배경지도 위에서 시각화

04

R Studio를 활용한 시각화와 자동화

통계분석 도구인 R에 공간데이터 처리와
시각화를 위한 패키지를 설치하여 활용

✦ 이용매뉴얼 · 분석사례집 ✦

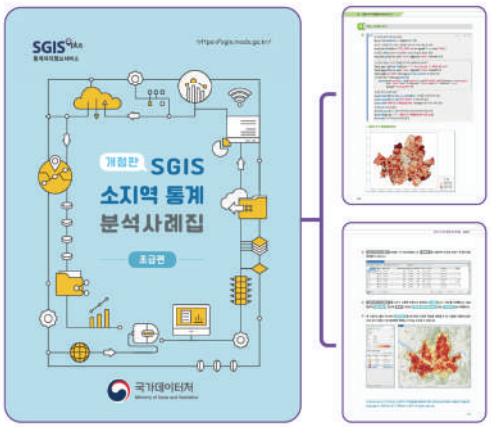
소지역통계 이용매뉴얼

QR코드로 쉽게 확인해보세요!

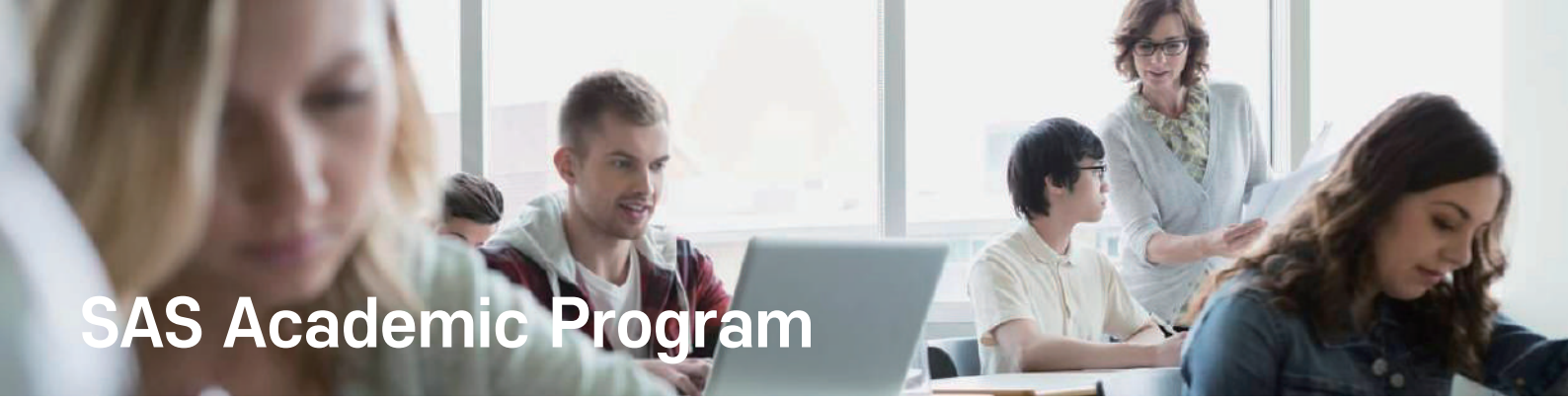
SGIS와 SDC에서 제공하는
통계와 경계에 대한 설명,
홈페이지를 통한 신청방법 등 안내

소지역통계 분석사례집

QR코드로 쉽게 확인해보세요!



대표적인 공간분석 도구 4종
(QGIS, Python, R, ArcGIS)의
기본 사용법을 쉽게 배울 수 있도록 구성



SAS Academic Program

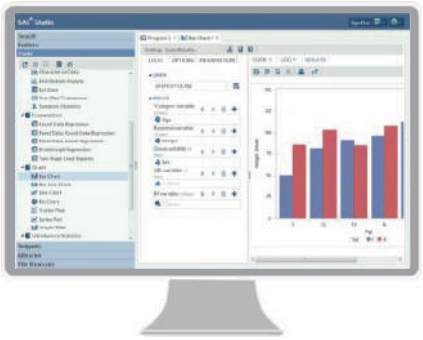
SAS는 분석 및 다양한 스킬에 대한 학습을 통해 분석 전문가로서의 경력을 준비할 수 있도록 교육자와 학생 모두를 위한 자원 및 학습 계획을 지원합니다.

INSTRUCTIONAL MATERIALS | ACADEMIC SOFTWARE | E-LEARNING | CERTIFICATION & BADGING | DISCOUNTS | CAREER RESOURCES

1. SAS Academic 무료 소프트웨어

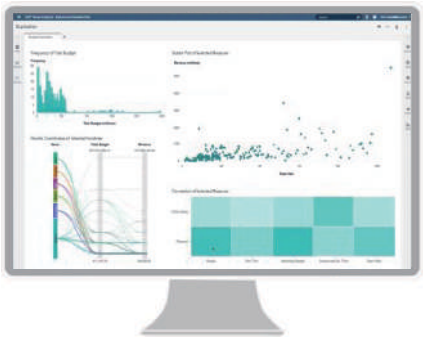
SAS Academic 무료 소프트웨어를 통해 별도의 SAS 설치 없이 인터넷 연결이 가능한 환경에서 SAS를 활용한 학습 및 분석을 진행할 수 있습니다.

SAS OnDemand for Academics

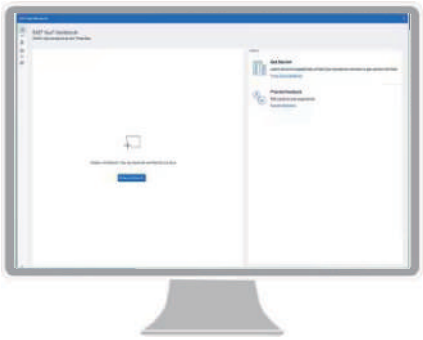


https://welcome.oda.sas.com

SAS® Viya® for Learners



SAS® Viya® Workbench for Learners



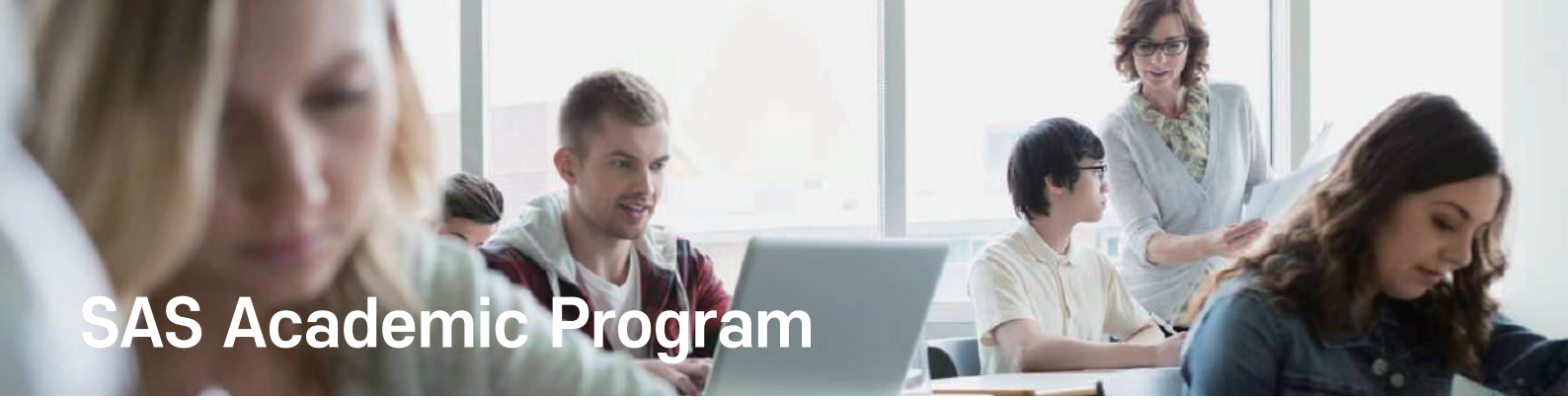
https://www.sas.com/ko_kr/learn/academic-programs/software.html

2. SAS Academic Specialization

학위 연계 프로그램으로 대학 및 대학원은 SAS로 분석 커리큘럼을 설계하고, 학생들의 분석 전문가 경력을 준비할 수 있도록 하는 프로그램입니다.

- ✔ 학위 연계 프로그램
- ✔ 학교별 적절한 참여 수준 선택(Tier 1~3)
- ✔ SAS 지식 및 기술에 대한 심층 교육
- ✔ 무료 SAS 교육 자원 활용

https://www.sas.com/ko_kr/learn/academic-programs/specializations/academic-specializations.html

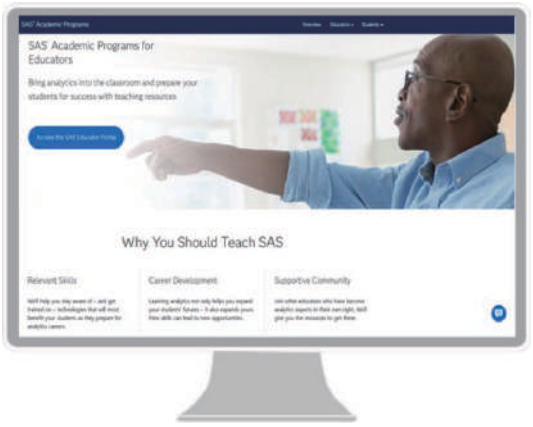


SAS Academic Program

3. 교육/학습 자원

SAS Educator Portal

교육자를 위한 Portal로 SAS 강의를 위한 모든 자원을 무료로 제공합니다.

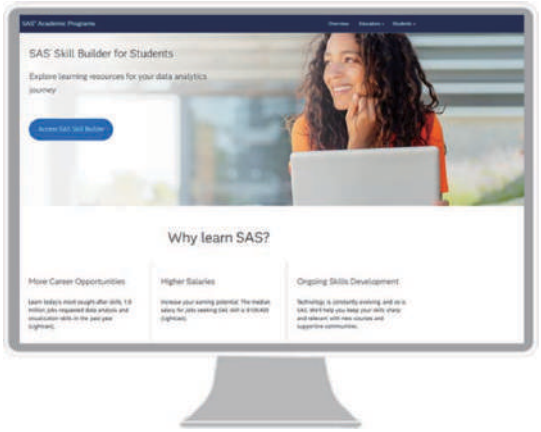


- ✔ 교육 자료
  - 강의 교재
  - 과제 및 프로젝트 자료
- ✔ 이러닝
- ✔ 아카데믹 소프트웨어
- ✔ SAS 자격증 준비 자료
- ✔ 교육자 워크샵 및 강의

https://www.sas.com/ko_kr/learn/academic-programs/educators.html

SAS Skill Builder for Student

학생들을 위한 Portal로 SAS 학습을 위한 모든 자원을 무료로 제공합니다.



- ✔ 학습 자료
  - 튜토리얼
  - SAS Documentations
  - 프로젝트 자료
- ✔ 이러닝
- ✔ 아카데믹 소프트웨어
- ✔ SAS 자격증 준비 자료

https://www.sas.com/ko_kr/learn/academic-programs/students.html

---

SAS Academic 프로그램 및 교육자/학생들을 위한 커뮤니티 안내는 QR 코드를 스캔하여 확인하세요!

SAS Academic 프로그램

Educator Community

Student Community

Proceedings of the KSS Winter Conference 2025

2025년도 한국통계학회
동계 학술논문발표회 프로시딩

일 시 2025년 12월 19일(금) ~ 20일(토)
장 소 서울대학교(관악캠퍼스)
주 최 (사)한국통계학회
주 관 (사)한국통계학회, 서울대학교 통계학과,
서울대학교 과학데이터혁신연구소

본 사업은 기획재정부의 복권기금 및 과학기술정보통신부의 과학기술진흥기금으로
추진되어 사회적 가치 실현과 국가 과학기술 발전에 기여합니다.

2025년도 동계학술논문발표회 초청강연

학회장 초청강연

- 연사 : 안형준(국가데이터처장)
- 제목 : AI 시대, 국가데이터 혁신

집중강연 I

- 연사 : 임성빈(고려대)
- 제목 : Understanding Foundation Model: From Machine Learning to Machine Reasoning

집중강연 II

- 연사 : 모상우(POSTECH)
- 제목 : Foundation Models for Vision, Language, and Robotics

갤럽학술상 수상자 강연

- 연사 : 임요한(서울대)
- 제목 : 고차원 공분산 행렬의 추정과 검정

올해의 대한민국 통계연구자상 수상자 강연

- 연사 : 정성규(서울대)
- 제목 : Generalized Frechet means with random minimizing domains

- 연사 : 이은령(성균관대)
- 제목 : High-dimensional non-sparse additive regression under general multi-dimensional dependency structures

신진 통계학자 학술논문상 수상자 강연

- 연사 : 김형우(국립부경대)
- 제목 : Variable selection in AUC-optimizing classification

목 차

◐ 일반세션 Ⅱ-1 Causal Inference and Complex Data Integration

◐ 일반세션 Ⅱ-2 Statistical Modeling and Applications for High-Dimensional and Structured Data

◐ 학생세션 I-1

◐ 학생세션 I-2

◐ 학생세션 Ⅱ-1

◐ 학생세션 Ⅱ-2

◐ 포스터세션

2025년 동계
학술논문발표회 프로시딩

학회장초청강연

AI 시대, 국가데이터 혁신

안형준[a]

요약: AI의 비약적 발전으로 전 세계는 산업・공공서비스 혁신에 돌입하였고 데이터가 국가 생산성을 높일 새로운 성장동력으로 부상하고 있다. 이런 흐름에 따라 2025년 10월, 통계청은 1990년 개청 이후 35년 만에 국가데이터처로 승격, 공식 출범하였다. 국가데이터처는 데이터를 수집하고 관리하는 역할을 넘어 데이터 혁신을 주도하는 중심 기관으로 거듭나라는 국가적 사명에 부합하여, 국가데이터 거버넌스 구축을 추진한다. 국민이 믿을 수 있고, 쉽게 쓸 수 있는 데이터 제공을 최우선 가치로 삼아 데이터 거버넌스 재정립, AI 활용 데이터 체계 구축, 데이터 허브 기능 강화, AI 대전환(AX) 인프라 구축 등 4개 전략을 마련하였다. 또한 이를 토대로 각 전략별 세부 과제를 추진하여 범정부적 데이터 활용 혁신을 실현하고자 한다. AI 시대, 국가데이터 혁신을 통해 어려운 사회문제 해결을 지원하고, AI가 데이터에 정확하게 접근하고 활용할 수 있는 환경을 조성하는 등 세계 AI 3개 강국 도약을 뒷받침하는 데이터 혁신 국가를 구현할 수 있을 것으로 기대한다.

[a]국가데이터처장

2025년 동계
학술논문발표회 프로시딩

집중강연 I

Understanding Foundation Model: From Machine Learning to Machine Reasoning

임성빈[a]

요약: 본 튜토리얼에서는 최근 인공지능 기술의 패러다임을 주도하는 Foundation Model 을 통계학적 관점에서 바라보고, 기계학습(Machine Learning)을 넘어 기계추론(Machine Reasoning)으로 향하는 최신 연구 개발 방향을 소개한다. 이를 위해 Transformer 구조에서 출발하여 GPT, CLIP 등에서 사용되는 Pretraining 프레임워크 및 In-Context Learning 기법을 통계학적 관점으로 설명하고, 인공지능 모델들의 기계추론 능력을 발전시키기 위한 연구 동향을 다룰 예정이다.

[a]고려대

2025년 동계
학술논문발표회 프로시딩

집중강연 II

Foundation Models for Vision, Language, and Robotics

모상우[a]

요약: 최근 인공지능 연구는 개별 과제에 맞춘 모델 설계에서 벗어나, 대규모 데이터와 사전학습을 통해 다양한 문제에 공통적으로 적용할 수 있는 Foundation Model 중심으로 빠르게 발전하고 있다. 본 튜토리얼에서는 시각 표현 학습의 발전을 이끈 DINO, 이미지와 언어를 공동 임베딩 공간에서 정렬시킨 CLIP, 그리고 대규모 언어 모델(LLM)을 통해 언어 이해와 생성 능력을 확장한 연구들을 살펴본다. 이어서 시각과 언어를 통합한 비전-언어 모델(VLM)과, 나아가 로보틱스와 행동 이해로 확장되는 비전-언어-행동 모델(VLA)까지 다루며, 이러한 모델들이 인공지능의 일반화 능력과 실제 응용 범위를 어떻게 넓혀가고 있는지를 설명한다.

[a]POSTECH

2025년 동계
학술논문발표회 프로시딩

한국통계학회 갤럽학술상 수상자 기념 강연

고차원 공분산 행렬의 추정과 검정

임요한[a]

요약: 고차원 공분산행렬 또는 역공분산 행렬의 추정과 검정은 여러 다변량 자료 분석에서 핵심적인 역할을 하며, 그 응용 분야는 금융의 포트폴리오 배분과 위험 관리, 생물정보학에서의 이상발현 유전자 탐색과 유전체 네트웍의 추정, 경제학에서의 인자분석 등 다양한 영역에 걸쳐있다. 본 발표에서는 공분산 행렬과 역공분산 행렬의 추정과 검정과 관련한 발표자의 이전 연구들을, 특히 행렬의 차원이 표본 크기와 함께 발산하거나 표본 크기를 초과할 수 있는 고차원적 상황에 초점을 맞추어 선별적으로 고찰하고, 남아있는 열린 문제들도 논의하고자 한다.

[a]서울대

2025년 동계
학술논문발표회 프로시딩

올해의 대한민국 통계연구자상 수상자 기념 강연

Generalized Frechet Means with Random Minimizing Domains

정성규[a]

Summary: In this talk, I will introduce a novel extension of Frechet means, referred to as generalized Frechet means, as a comprehensive framework for describing the characteristics of random elements. The generalized Frechet mean is defined as the minimizer of a cost function, and the framework encompasses various extensions of Frechet means that have appeared in the literature. The most distinctive feature of the proposed framework is that it allows the domain of minimization for the empirical generalized Frechet means to be random and different from that of its population counterpart. This flexibility broadens the applicability of the Frechet mean framework to various statistical scenarios, including sequential dimension reduction for non-Euclidean data. We establish a strong consistency theorem for generalized Frechet means and demonstrate the utility of the proposed framework by verifying the consistency of principal geodesic analysis on the hypersphere.

[a]서울대

High-dimensional Non-sparse Additive Regression under General Multi-dimensional Dependency

이은령[a]

Summary: In this talk, I will introduce a new framework for estimation and inference in ultra–high-dimensional additive models that does not assume sparsity. The framework accommodates complex predictor dependence by working under general α-mixing conditions and achieves scalability via thresholded operator constructions, while boundary correction improves accuracy near the edges of the covariate domain. A distinctive feature of the proposal is its ability to capture nonsparse effects in truly high-dimensional regimes where traditional sparse methods can fail, thereby broadening the applicability of additive modeling in scientific studies. We establish rigorous theory, including uniform error bounds and computational convergence rates under minimal regularity conditions, and we provide valid inference procedures within the same framework. The practical utility of the method is demonstrated through extensive simulations and real data analyses, where it consistently outperforms existing alternatives in recovering nonsparse structure.

[a]성균관대

2025년 동계
학술논문발표회 프로시딩

신진통계학자 학술논문상 수상자 기념 강연

Variable Selection in AUC-optimizing Classification

김형우[a]

Summary: Optimizing the receiver operating characteristic (ROC) curve is a popular way to evaluate a binary classifier under imbalanced scenarios frequently encountered in practice. A practical approach to constructing a linear binary classifier is presented by simultaneously optimizing the area under the ROC curve (AUC) and selecting informative variables in high dimensions. In particular, the smoothly clipped absolute deviation (SCAD) penalty is employed, and its oracle property is established, which enables the development of a consistent BIC-type information criterion that greatly facilitates the tuning procedure. Both simulated and real data analyses demonstrate the promising performance of the proposed method in terms of AUC optimization and variable selection.

[a]국립부경대

2025년 동계
학술논문발표회 프로시딩

기획세션

Dynamic Functional Connectivity Analysis of Functional MRI Based on Time-varying Partial Correlation[a]

Namgil Lee[b*]

Summary: We propose a time-varying partial correlation as a statistical measure of dynamic functional connectivity (dFC) in the human brain. The proposed measure is constructed based on the copula-based dynamic conditional correlation (DCC) framework, which does not rely on specific distribution assumptions, to estimate time-varying correlations between regions-of-interest (ROIs) in the brain. The proposed time-varying partial correlation provides an effecive approach for inferring sparse dFC structures, being robust to noise distributions and preprocessing procedures in functional MRI data.

Keywords: Dynamic conditional correlation, Generalized AutoRegressive Conditional Heteroscedastic (GARCH), Partial correlation

[a]This work was in part supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (RS-2024-00358572, RS-2024-00336424).

[b]Department of Information Statistics, Kangwon National University, Chuncheon, Gangwon 24341, Republic of Korea. namgil.lee@kangwon.ac.kr

Statistical Inference for Leading Author and Popularity Effects in Collaboration Networks[a]

Hohyun Jung[b*]

Summary: We develop a statistical framework for collaboration networks that incorporates the roles of leading authors and the effects of popularity and latent expertise. In our model, each multi-author paper is initiated by a leading author who selects collaborators based on prior visibility and unobserved ability. This structure captures the asymmetry of co-authorship and quantifies how accumulated popularity influences collaboration patterns. For parameter estimation and latent variable inference, we employ a joint approach combining the Expectation–Maximization algorithm with Gibbs sampling. Simulation studies confirm the validity of the method, and an application to Web of Science data illustrates how the model disentangles authors' latent expertise, referred to as genius levels, from popularity effects. The results provide insights into how leadership roles and recognition dynamics shape scientific collaboration.

Keywords: scientific collaboration, leading author model, popularity effect, collaboration networks

[a]This work was in part supported by the National Research Foundation (NRF) of Korea, under grant RS [2024-00455553].

[b]Assistant Professor, School of Mathematics, Statistics and Data Science, Sungshin Women's University, 34 Da-gil, Bomun-ro, Seongbuk-gu, Seoul, 02844, Korea. hhjung@sungshin.ac.kr

Quantile-Free Uncertainty Quantification in Graph Neural Networks

SoYoung Park[a] · Hwanjun Song[b] · Sungsu Lim[c*]

Summary: Uncertainty quantification (UQ) in graph neural networks (GNNs) is crucial in high-stakes domains but remains a significant challenge. In graph settings, message passing often relies on strong assumptions such as exchangeability, which are rarely satisfied in practice. Moreover, achieving reliable UQ typically requires costly resampling or post-hoc calibration. To address these issues, we introduce Quantile-free Prediction Interval GNN (QpiGNN), a framework that builds on quantile regression (QR) to enable GNN-based UQ by directly optimizing coverage and interval width without requiring quantile inputs or post-processing. QpiGNN employs a dual-head architecture that decouples prediction and uncertainty, and is trained with label-only supervision through a quantile-free joint loss. This design allows efficient training and yields robust prediction intervals, with theoretical guarantees of asymptotic coverage and near-optimal width under mild assumptions. Experiments on 19 synthetic and real-world benchmarks show QpiGNN achieves average 22% higher coverage and 50% narrower intervals than baselines, while ensuring efficiency and robustness to noise and structural shifts.

Keywords: Uncertainty Quantification, Graph Neural Networks

[a]Ph.D. Candidate, Department of Computer Science and Engineering, Chungnam National University, Daejeon 34134, Korea. sypark1452@o.cnu.ac.kr

[b]Assistant Professor, Department of Industrial and Systems Engineering, KAIST, Daejeon 34141, Korea. songhwanjun@kaist.ac.kr

[c]Associate Professor, Department of Computer Science and Engineering, Chungnam National University, Daejeon 34134, Korea. sungsu@cnu.ac.kr

Deep Learning for Small Medical Imaging Datasets: DISH and Glaucoma Case Studies

Dongseok Choi[a*]

Summary: This study explores the feasibility of convolutional neural networks (CNNs) in small medical imaging datasets. We present two applications: spinal X-rays for diagnosing diffuse idiopathic skeletal hyperostosis (DISH), and OCT scans for glaucoma classification. Using a dataset of 116 DISH cases and 262 controls, CNNs trained in R demonstrated high specificity and robust overall accuracy. In the glaucoma study, a multi-modal hybrid model combining CNN-derived NFL thickness and reflectance maps with fully connected networks (FCN) for clinical variables significantly outperformed traditional logistic regression. These results suggest that deep learning, even with limited data, can support accurate classification in medical imaging.

[a]OHSU. choid@ohsu.edu

Sparse Covariate-driven Factorization of High-dimensional Brain Connectivity with Applications to Site Effect Correction

Jun Young Park[a*]

Summary: Large-scale neuroimaging studies often collect data from multiple scanners across different sites, where variations in scanners, scanning procedures, and other conditions across sites can introduce artificial site effects. These effects can bias brain connectivity network, such as functional connectivity (FC). To address this, we propose SLACC (sparse latent covariate-adjusted connectome), a novel method that explicitly parameterizes covariate effects inherent in sparse latent network patterns to disentangle biologically meaningful signals from site-specific heterogeneity. Our model identifies localized site-driven variability within and across brain networks, enabling targeted correction. We develop a penalized Expectation-Maximization (EM) algorithm for parameter estimation, incorporating the Bayesian Information Criterion (BIC) to guide optimization. Extensive simulations validate SLACC's robustness in recovering the true parameters and the underlying connectivity patterns. Applied to the Autism Brain Imaging Data Exchange (ABIDE) dataset, SLACC demonstrates its ability to reduce site effects while preserving biologically relevant associations.

[a]U. Toronto. junjy.park@utoronto.ca

Bayesian Joint Modeling for Hierarchically Structured Medical Data

Seongho Song[a*]

Summary: Joint modeling has been a useful strategy for incorporating latent associations between different types of outcomes simultaneously, often focusing on a longitudinal continuous outcome characterized by a Linear Mixed Effect (LME) sub-model with a binary process, which is commonly specified by a Generalized linear Mixed Model (GLMM) under hierarchically structured medical data. In this talk, we propose a multilevel joint model that encompasses LME and GLMM sub-models through a Bayesian approach. Motivated by the need for timely detection of pulmonary exacerbations and characterization of irregularly observed lung function measurements in people living with cystic fibrosis (CF) receiving care across multiple centers, we apply the model to data arising from the US CF Foundation Patient Registry.

[a]Univ of Cincinnati. seongho.song@uc.edu

Statistical Properties of Initial Sequence Type Variance Estimators for Reversible Markov Chains

Stephen Berg[a*]

Summary: Initial sequence estimators, originally introduced by Geyer (1992), are commonly used to estimate Monte Carlo standard errors in reversible Markov chain Monte Carlo chains. In particular, the initial positive sequence estimator utilizes the property that the sums of adjacent autocovariances are non-negative, summing autocovariances up to the point where this non-negativity condition is violated. While this estimator has been widely used and adapted to different settings, only its asymptotic conservativeness has been shown, while consistency remains an open question. In this talk, we address this gap by investigating the statistical properties of the initial positive sequence estimator. We first study the convergence behavior of its random truncation point. We also propose an alternative initial sequence-type estimator based on a modified truncation rule. For both estimators, we establish consistency and derive bounds on rates of convergence. Finally, through empirical studies using both simulated and real-world data, we validate our theoretical findings and compare the empirical performance of the two initial sequence-type estimators with the standard overlapping batch mean estimator.

[a]Penn State. sqb6128@psu.edu

실무형 효과 검증 체계의 확장: 준실험과 Agent 기반 시뮬레이션 사례

이상현[a]*

요약: 실무에서는 서비스 구조나 캠페인 방식에 따라 무작위 배정이 어려워, A/B 테스트만으로는 정책 효과를 충분히 검증하기 어렵다. 본 발표에서는 이러한 제약 속에서 실험 설계를 확장한 두 가지 사례를 소개한다. 첫 번째는 마케팅 캠페인에 AI 기반 타겟팅을 도입하고, 기존 통계 기반 정책과 비교하여 Difference-in-Differences 방식으로 증분 효과를 검증한 준실험 사례이다. 이를 통해 AI 기반 타겟팅이 기존 정책과 상호보완적으로 작동하며, 저성과 고객군에서도 유의미한 전환 효과를 실험적으로 확인하였다. 두 번째는 추천 시스템의 재현이 어려워 동일 조건의 비교가 불가능한 환경을 반영하여, LLM 기반 고객 Agent 시뮬레이션을 활용해 다양한 추천 정책의 반사실적 효과를 사전적으로 평가한 사례이다. 본 발표는 A/B 테스트의 한계를 넘어, 실무 환경에서도 신뢰도 높은 인과적 효과 검증과 데이터 기반 의사결정을 지원하는 평가 체계의 확립 방향을 제시한다.

주요용어: 인과추론, 준실험, Agent 기반 시뮬레이션, 평가 프레임워크

[a]LG유플러스, AI Scientist

Building an Experiment Culture for Data-Driven Decision-Making in Business

Eliote Yuil Jeong[a]*

Summary: This presentation introduces how a culture of experimentation is being established within global companies through real-world data and analytics practices. Beyond simple A/B testing, experimentation has become a key mechanism for validating intuition and driving data-based decisions across products, content, and marketing operations. We present three representative case studies - (1) game balance adjustment, (2) personalized content recommendation, and (3) lifecycle retargeting message optimization - to demonstrate how experiments are designed, measured, and operationalized. A five-step execution framework is proposed: defining testable questions, structuring experimental splits, automating collaboration pipelines, standardizing success metrics, and archiving failed experiments as learnings. Through this process, experimentation evolves from an analytical tool to a growth-driven organizational mindset, fostering continuous improvement and data-driven innovation in the business context.

Keywords: experiment culture, causal inference, randomized experiment, data-driven decision-making, A/B testing, personalization

[a]The Walt Disney Company

게임 마케팅의 성공을 이끄는 인과추론과 데이터 기반 의사결정 문화 구축 전략[a]

신진수[b]*

요약: 게임 산업은 광고 채널, 인게임 행동 로그, 앱 마켓 등으로부터 방대한 실시간 유저 데이터를 수집하고 있고, 일부 원시자료를 분석용으로 활용하고 있다. 게임 데이터는 유저 획득부터 결제에 이르는 전체 경로를 사용자 수준에서 탐구할 수 있다는 점에서 매력적이나, 채널별 데이터 파편화 및 ATT와 같은 개인정보보호 정책 강화로 인해 마케팅 활동의 순수한 인과적 효과를 식별하는 데 한계가 존재한다. 또한, 관찰 데이터에 내재된 교란 변수(confounding variables)로 인해 변수 간 상관관계를 인과관계로 오인할 수 있다는 점도 연구 수행 단계에서의 주요 도전점이다. 본 발표에서는 이러한 도전과제를 극복하기 위한 통계적 접근법과 그 적용 사례를 소개한다. 먼저 관찰 데이터에 기반하여 마케팅 성과를 분석하는 인과추론 기법을 소개하고, 나아가 ATT 동의율 최적화 및 리타겟팅 캠페인 효과 측정과 같이 통제된 환경에서 가설을 검증하는 온라인 실험의 설계 및 분석 방법을 논의한다. 궁극적으로, 이러한 통계적 방법론을 조직 문화에 내재화하여 '데이터 기반으로 의사결정하는 조직'으로 성장하기 위한 전략적 방향을 제시하고자 한다.

주요용어: 게임 마케팅, 데이터 기반 의사결정, 인과추론, 온라인 실험, 조직문화

[a]본 발표는 기획세션 <비즈니스 현장에서의 임의실험과 인과추론 사례>의 발표입니다.
[b]크래프톤, 시니어 데이터분석가

Bias-corrected Estimation in Causal Mediation Analysis[a]

Jaeho Jeong[b] · Jongho Im[c] · Young Min Kim[d*]

Summary: Causal mediation analysis, based on the counterfactual approach, decomposes the total effect of an exposure into natural direct and indirect effects, with the mediation proportion (MP) quantifying the relative contribution of the mediator. However, the MP, along with its components, the natural indirect effect (NIE) and the natural direct effect (NDE), are functional estimators. In finite samples, their ratio and exponential forms can make them unstable and biased. This paper introduces the problem of transformation-induced bias in regression-based causal mediation analysis and proposes two likelihood-based bias-correction methods. These methods target exponential and ratio functionals, including the NDE, NIE, and MP, and provide closed-form corrections for continuous and binary outcomes and mediators, with or without exposure–mediator interaction. Simulation studies show that ordinary estimators suffer from noticeable finite-sample bias, particularly for the MP and for log-odds scale effects such as binary outcome cases. In contrast, the proposed methods consistently reduce relative bias and mean squared error while preserving large-sample properties. Applications to two real datasets demonstrate that the proposed estimators yield more stable and precise inference, especially for the MP.

Keywords: causal mediation analysis, mediation proportion, transformation-induced bias, bias-correction

[a]This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-00573620, 2025S1A5C3A0201187111, and NRF-2021R1C1C1014407.

[b]Ph.D. Candidate, Department of Statistics, Kyungpook National University, Daegu 41566, Korea. jaeho.jeong@knu.ac.kr

[c]Corresponding author: Associate professor, Department of Statistics and Data Science, Yonsei University, Seoul 03722, Korea. ijh38@yonsei.ac.kr

[d]Corresponding author: Associate professor, Department of Statistics, Kyungpook National University, Daegu 41566, Korea. kymmyself@knu.ac.kr

Causal Inference for Survival Data in Continuous-time under Time-varying Treatment and Confounders

Jooyoung Lee[a*] · Richard Cook[b]

Summary: Estimating causal effects of time-varying treatments in survival analysis traditionally relies on discrete-time approximations of the g-formula, inverse probability weighting, or doubly robust estimators. However, discretization induces bias and inefficiency in continuous-time settings common in electronic health record (EHR) data, where treatment and covariate processes evolve irregularly. We propose a continuous-time doubly-robust estimator for survival and restricted mean survival time (RMST) under dynamic treatment policies. Our approach extends semiparametric efficient influence-function (EIF) theory to continuous time, yielding an orthogonal estimating equation that remains unbiased if either the outcome hazard model or the joint treatment–censoring mechanism is correctly specified, while accommodating time-varying confounders. This frwamework bridges continuous-time causal inference and modern machine learning, enabling orthogonal estimation of policy-specific survival in high-frequency longitudinal health data. Simulation studies demonstrate improved bias and efficiency relative to discrete-time approximations and inverse probability weighting, particularly under strong treatment–confounder feedback.

Keywords: Continuous-time causal inference, g-computation, doubly robust estimation, time-varying treatment, survival analysis.

[a]Department of Applied Statistics, Chung-Ang University, 84, Heukseok-ro, Dongjak-gu, Seoul (06974), Korea. jooylee@cau.ac.kr

[b]Department of Statistics and Actuarial Science, University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada. rjcook@uwaterloo.ca

Developing Statistical Methods for Selection-biased Self-matched Data[a]

Seungjae Lee[b*] · Hyunman Sim[c] · Woojoo Lee[c,d]

Summary: Self-matched study designs, such as those comparing pre- and post-exposure periods within the same individuals, are well suited for assessing the effects of transient interventions on acute outcomes, and thus offer the major advantage of eliminating both measured and unmeasured time-invariant confounding. However, sample selection bias may arise if the study includes more individuals who were exposed to the intervention and subsequently developed a particular disease. To address this issue, we propose novel methods for analyzing selection-biased self-matched data with binary and count outcomes. The proposed methods explicitly account for sample selection bias through a selection ratio model and allow valid inference using a conditional likelihood approach. We illustrate our proposed methods by analyzing data from the humidifier disinfectant victim cohort to evaluate the effect of humidifier disinfectant exposure on respiratory diseases.

Keywords: conditional likelihood approach, humidifier disinfectant, sample selection bias, selection-biased self-matched data, selection ratio model

[a]This work was supported by the Humidifier Disinfectant Health Center for Epidemiological Evaluation and by the National Institute of Environment Research (NIER), funded by the Ministry of Environment (MOE) of the Republic of Korea (Grant No. NIER-2025-04-03-001).

[b]Department of Applied Statistics, Kyonggi University, Korea
[c]Institute of Health and Environment, Seoul National University, Korea
[d]Department of Public Health Sciences, Graduate School of Public Health, Seoul National University, Korea

Conditional Mean Dimension Reduction for Tensor Time Series

Chung Eun Lee[a*]

Summary: The dimension reduction problem for a stationary tensor time series is addressed. The goal is to remove linear combinations of the tensor time series that are mean independent of the past, without imposing any parametric models or distributional assumptions. To achieve this goal, a new metric called cumulative tensor martingale difference divergence is introduced and its theoretical properties are studied. Unlike existing methods, the proposed approach achieves dimension reduction by estimating a distinctive subspace that can fully retain the conditional mean information. By focusing on the conditional mean, the proposed dimension reduction method is potentially more accurate in prediction. The effectiveness of the proposed method is illustrated by extensive simulations and two real-world data applications.

[a]Baruch College, CUNY. chungeun.lee@baruch.cuny.edu

Heterogeneous Treatment Effects under Network Interference: A Nonparametric Approach Based on Node Connectivity

Heejong Bong[a*]

Summary: In network settings, interference between units makes causal inference more challenging as outcomes may depend on the treatments received by others in the network. Typical estimands in network settings focus on treatment effects aggregated across individuals in the population. We propose a framework for estimating node-wise counterfactual means, allowing for more granular insights into the impact of network structure on treatment effect heterogeneity. We develop a doubly robust and non-parametric estimation procedure, KECENI (Kernel Estimator of Causal Effect under Network Interference), which offers consistency and asymptotic normality under network dependence. The utility of this method is demonstrated through an application to microfinance data, revealing the node-wise impact of network characteristics on treatment effects.

[a]Purdue. bong0@purdue.edu

A Semi-parametric Global Bandit Framework for Flexible and Scalable Sequential Decision-making

Hyebin Song[a]*

Summary: Sequential decision-making problems are central to applications like personalized medicine and adaptive clinical trials. The multi-armed bandit (MAB) framework is a powerful tool for these settings, but real-world scenarios present key challenges: feedback often arrives in batches, patient covariates must be incorporated, and outcomes from related treatments (e.g., different drug dosages) are not independent. In this talk, I will introduce a novel semi-parametric framework for batched bandits that directly addresses these complexities. Our approach leverages the single-index regression (SIR) model to capture the relationship between arm rewards via a shared parameter, striking a balance between interpretability and model flexibility. We propose a new algorithm, Batched single-Index Dynamic binning and Successive arm elimination (BIDS), which employs a successive arm elimination strategy guided by the single-index direction. We establish theoretical regret bounds for two settings: one with a known pilot direction and another where the direction is learned from data. Critically, when a pilot direction is available with sufficient accuracy, our approach achieves the minimax-optimal regret rate for non-parametric batched bandits (with dimension d=1), effectively circumventing the curse of dimensionality. Finally, I will showcase results from simulations and real-world data applications that highlight the practical advantages of BIDS over competing non-parametric bandit algorithms.

[a]Penn State. hps5320@psu.edu

Leveraging External Individualized Prediction Models in Bayesian Survival Analysis

Mi-Ok Kim[a*]

Summary: Individualized risk prediction algorithms, such as the Prostate Cancer Risk Assessment tool, are increasingly used to predict cancer relapse or progression. Since these algorithms are typically trained on large datasets, effectively integrating their outputs can enhance the efficiency of analyzing individual studies. In this research, we consider Cox regression analysis for right-censored time-to-event outcomes, incorporating external information provided by large-scale prediction models. We adopt a Bayesian inference in estimating the baseline hazard at each distinct time point. External information is integrated through the Kullback–Leibler (KL) divergence, leading to informative priors for Bayesian analysis. The performance of the proposed model is demonstrated through simulation studies and an application to data from a prostate cancer clinical trial.

[a]UCSF. miok.kim@ucsf.edu

2025년 동계
학술논문발표회 프로시딩

일반논문

Statistical Properties of Deep Heaviside Networks[a]

Insung Kong*[b] · Juntong Chen[c] · Sophie Langer[d] · Johannes Schmidt-Hieber[f]

Summary: Motivated by biological neurons, the first models of artificial neurons employed the Heaviside function (unit step function) as activation function. Because of the biological plausibility and lower energy consumption, it is still of interest to investigate the specific approximation theoretic and statistical properties of deep Heaviside networks (DHNs). We show that DHNs have limited expressiveness but that this can be overcome by including either skip connections or neurons with linear activation. We provide upper and lower bounds for the VC dimensions and approximation rates of these network classes. As an application, we derive statistical convergence rates for DHN fits in the nonparametric regression model.

Keywords: deep neural networks, biological neural networks, Heaviside activation function, statistical learning theory

[a]This work is supported by ERC grant A2B (grant agreement No. 101124751).
[b]Department of Applied Mathematics, University of Twente, Drienerlolaan 5, Enschede, 7522 NB, Overijssel, Netherlands
[c]Department of Probability and Statistics, Xiamen University, 422 Siming S Rd, Siming District, Xiamen, Fujian, China
[d]Faculty of Mathematics, Ruhr University Bochum, Universitätsstraße 150, Bochum, 44801, Nordrhein-Westfalen, Germany
[f]Department of Applied Mathematics, University of Twente, Drienerlolaan 5, Enschede, 7522 NB, Overijssel, Netherlands

Semi-supervised Learning of Noisy Mixture of Experts Models[a]

Oh-Ran Kwon[b*] · Gourab Mukherjee[c] · Jacob Bien[d]

Summary: The mixture of experts (MoE) model is a versatile framework for predictive modeling that has gained renewed interest in the age of large language models. A collection of predictive "experts" is learned along with a "gating function" that controls how much influence each expert is given when a prediction is made. This structure allows relatively simple models to excel in complex, heterogeneous data settings. In many contemporary settings, unlabeled data are widely available while labeled data are difficult to obtain. Semi-supervised learning methods seek to leverage the unlabeled data. We propose a novel method for semi-supervised learning of MoE models. We start from a semi-supervised MoE model that was developed by oceanographers. This model makes the strong assumption that the latent clustering structure in unlabeled data maps directly to the influence that the gating function should give each expert in the supervised task. We relax this assumption, imagining a noisy connection between the two, and propose an algorithm based on least trimmed squares, which succeeds even in the presence of misaligned data. Our theoretical analysis characterizes the conditions under which our approach yields estimators with a parametric rate of convergence. Simulated and real data examples demonstrate the method's efficacy.

Keywords: transfer learning, multi-view data, least trimmed squares, mixture of Gaussians

[a]This work was supported by a grant by the Simons Collaboration on Computational Biogeochemical Modeling of Marine Ecosystems/CBIOMES (Grant ID: 549939 to JB). The authors acknowledge the Center for Advanced Research Computing (CARC) at the University of Southern California for providing computing resources that have contributed to the research results reported within this publication. We thank Elizabeth C. Atwood, Bror Jonsson, and Thomas Jackson for many useful conversations in understanding chlorophyll prediction models used in oceanography and Matias Salibian Barrera for a helpful email about robust estimators.

[b]Assistant Professor, Department of Statistics, The Ohio State University

[c]Associate Professor, Department of Data Sciences and Operations, Univeristy of Southern California

[d]Professor, Department of Data Sciences and Operations, Univeristy of Southern California

동형암호 기반 프라이버시 보존 벡터 유사도 검색: RAG 시스템에서의 Inversion Attack 차단

이가람[a]*

요약: 대규모 언어모델(LLM)의 검색증강생성(Retrieval-Augmented Generation, RAG)은 응답의 최신성과 신뢰성을 높이기 위해 최신정보 반영 및 출처기반 응답을 가능하게 하는 핵심 기술로 주목받고 있다. 그러나 최근 연구(Morris et al., EMNLP 2023)는 RAG 파이프라인의 벡터DB에 저장된 임베딩 벡터로부터 복원 공격(inversion attack)을 통해 원문 텍스트의 87% 이상을 높은 정확도로 복원할 수 있음을 입증하였다. 고차원 벡터 공간상의 점으로 표현되는 임베딩 벡터는 의미적 유사도(semantic similarity) 검색을 가능하게 하는 한편, 원본 데이터의 정보를 풍부하게 함축하여 복원 공격에 취약하다는 근본적 한계를 지닌다는 것을 의미한다. 이는 의료기록, 기업기밀, 개인정보 등 민감 데이터를 다루는 RAG 응용에서 개인정보 유출 및 산업 보안 위협 등 심각한 보안 위협을 초래한다. 본 연구에서는 RAG 파이프라인에 동형암호 (Homomorphic Encryption, HE) 기반의 Encrypted Vector Similarity Search를 활용하여, 벡터 수준에서 암호화된 상태를 유지하면서도 RAG의 검색 정확도와 효율성을 유지할 수 있음을 실험적으로 검증하였다. GenTok AI 및 Wikipedia 데이터셋(N=250K)을 활용한 실험에서, CKKS(Cheon-Kim-Kim-Song) 동형암호 스킴을 활용하여 암호화된 상태에서 코사인 유사도 계산을 수행함으로써, 서버가 벡터DB에 직접 접근하지 않고도 Top-k 검색을 가능하게 한다. 동형암호 기반 RAG는 평문 대비 1.93배 증가에 그친 반면, nDCG@10, Recall@10이 0.99이상의 높은 검색 성능을 유지하였다. 또한 암호문 상태에서 전체 질의 응답 시간은 쿼리당 1초 내외(검색+생성 기준)로 실시간 서비스에 적합한 수준임을 확인하였다. 이 연구는 민감 데이터 기반 LLM 어플리케이션에 있어 프라이버시 보호와 성능 간의 균형을 실현하는 실질적 해결책을 제시하며, GDPR·HIPAA 등 강력한 프라이버시 규제가 적용되는 의료·금융·국방 분야에서 고성능 AI 도입을 위한 핵심 기반 기술로 작용할 것으로 기대한다.

주요용어: 검색증강생성 (Retrieval-Augmented Generation, RAG), 동형암호, 임베딩 복원 공격, 프라이버시 보존 검색

[a](08791) 서울특별시 관악구 관악로 98 삼정빌딩 3층 크립토랩 사업개발실 공공사업팀 연구원.
garamlee@cryptolab.co.kr

Spatial Causal Inference with Difference-in-differences: The Impact of the Confirmation of GTX-A Dongtan Station on Apartment Prices in Hwaseong

Mijeong Kim[a*]

Summary: This study applies a Difference-in-Differences (DID) model that accounts for both spatial dependence and spillover effects to address violations of the Stable Unit Treatment Value Assumption (SUTVA) in spatial causal inference. While conventional DID models estimate direct effects under the parallel trends assumption, spatial dependence and spillovers in spatial data challenge this assumption. Delgado and Florax (2015) proposed a DID model incorporating spillovers but not spatial dependence, and Bardaka et al. (2018) and Chagas et al. (2016) combined DID with Spatial Error and Spatial Lag models, respectively, to address spatial dependence. However, both studies had limitations in statistical inference of spatial causal effects. This study reviews these methodologies and aims to improve the inference for both models. Based on this framework, we analyze the impact of the confirmation of the GTX-A Dongtan station on apartment prices in Hwaseong, separating the effects into direct and indirect components.

Keywords: Difference-in-Differences, Spatial causal inference, Spatial dependence model, Spatial spillover effect

[a]Associate Professor, Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760 Republic of Korea. m.kim@ewha.ac.kr

Overlap Weights for Binary Outcomes: a Performance Assessment[a]

Seo Young Park[b*] · Jaeil Ahn[c] · Jae Hoon Lee[d] · Jaewoo Kwon[e] · Hana Lee[f]

Summary: Inverse probability weighting (IPW) is a widely used method to estimate the causal effect of treatment from observational data. However, it can be unstable when extreme propensity score (PS) values lead to very large weights. Overlap weights (OW), which emphasize subjects in areas of covariate overlap, reduce the influence of extreme PS without excluding participants. While OW method has shown strong performance in simulations with continuous outcomes, its utility in binary outcome settings—common in health research—has not been thoroughly evaluated. We conducted simulation studies to evaluate the performance of OW in comparison to other PS weighting methods including IPW, trimmed IPW, and matching weights, in settings with extreme PS values and a binary outcome. The performance of the PS weighting methods was further illustrated through an application to real world data from a study on pancreatic ductal adenocarcinoma. In simulation studies, IPW's performance deteriorated markedly as the overlap in the covariate distribution decreased. In contrast, OW achieved exact covariate balance and consistently showed the highest efficiency among all methods evaluated. In the application to real-world data characterized by low treatment prevalence and substantial covariate imbalance, OW also outperformed the other methods in terms of both standard error and covariate balance. These findings suggest superior performance of OW in terms of covariate balance and estimation efficiency in settings with extreme PS and a binary outcome.

Keywords: Overlap weights, Propensity score, Inverse probability weighting

[a]This work was supported by 2024 Oversees Training Fund from Korea National Open University.
[b]Department of Statistics and Data Science, Korea National Open University, 86(Dongsuong-dong), Daehak-ro, Jongno-gu, Seoul 03087
[c]Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University
[d]Department of Surgery, Asan Medical Center, University of Ulsan College of Medicine, Division of Hepato-Biliary and Pancreatic Surgery
[e]Department of Surgery, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine
[f]Center for Drug Evaluation and Research, United States Food and Drug Administration

Estimating Survivor Attributable Fraction in the Presence of Truncation by Death[a]

Hyunman Sim[b,c*] · Hongsoo Kim[c,d] · Woojoo Lee[b,d]

Summary: As a super-aged society is approached, it is important to quantify the contribution of treatment or exposure for efficient resource allocation. Although the population attributable fraction (PAF) has been widely used in epidemiology, its use is challenging in gerontology because some older people die before their outcome assessment, making the outcome ill-defined. This problem is known as truncation by death. Nevertheless, if the PAF were estimated by restricting the entire population to those who were alive at the outcome assessment time, the resulting estimate suffers from the sample selection bias. To address this issue, we define the survivor AF, derive an identification formula, and develop a multiply robust estimator, which remains consistent if at least two models of the three models-the outcome model, the propensity score model, and the principal strata model-are correctly specified. Simulation studies demonstrate the finite-sample performance of the developed estimators.

Keywords: attributable fraction, principal stratification, truncation by death, causal inference, gerontology

[a]This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government(MSIT) (No. RS-2025-02304335, Precision Multimorbidity Prediction and Prevention Technology Using AI Digital Biomarkers for Intrinsic Capacity).

[b]Institute of Health and Environment, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul, 08826. smisk77@gmail.com

[c]AI Institute, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul, 08826

[d]Graduate School of Public Health, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul, 08826. lwj221@gmail.com

Weight Calibration in the Joint Modelling of Medical Cost and Mortality

Seong Hoon Yoon[a,b*], Alain Vandal[a], Claudia Rivera-Rodriguez[a]

Summary: Joint modelling of longitudinal and time-to-event data is a method that recognises the dependency between the two data types, and combines the two outcomes into a single model, which leads to more efficient estimates. These models are applicable when individuals are followed over a period of time, generally to monitor the progression of a disease or a medical condition, and also when longitudinal covariates related to the time-to-event variable are available. However, some longitudinal datasets (e.g. medical cost) are usually obtained using a complex sampling design rather than simple random sampling, which needs to be recognised in the statistical analysis. To address this, we combine survey calibration with standard joint modelling. The proposed method is applied to the survival and cost data on patients with diagnosed dementia in New Zealand.

Keywords: Joint Modelling, Longitudinal Data, Survival Analysis, Weight Calibration

[a]Department of Statistics, The University of Auckland
[b]School of Computing and Mathematical Sciences, The University of Waikato

Tobit INAR Models for Count Time Series with Negative Autocorrelation[a]

Hee-Young Kim[b*] · Christian H. Weiss[c] · Fukang Zhu[d]

Summary: Thinning-operator-based integer-valued autoregressive moving-average (INARMA) models are quite popular for stationary count time series, the two main cases of which are the purely autoregressive INAR and purely moving-average INMA model. Like the ordinary ARMA models, the INARMA models have a linear conditional mean such that the classical Yule–Walker equations for the autocorrelation function (acf) hold. But a known drawback is that the dependence parameters have to be non-negative such that also the resulting acf can only take non-negative values. Modeling count time series with negative acf values in a simple construction is a long-standing open problem that has not been satisfactorily resolved so far. To address this problem, we propose simple and flexible frameworks based on the Tobit modeling approach: Tobit INAR model. Stochastic properties, approximate linearity of the conditional mean, maximum likelihood and approximate estimation for the model parameters, and related simulations for both kinds of models are given. Two real-world data examples about a chemical process and beer sales are analyzed in detail, and it is shown that the proposed model outperforms existing ones. Extension of the Tobit INAR model to zero-inflated counts as well as to bounded counts are also discussed.

Keywords: count time series, INARMA model, linear model, negative autocorrelation, Tobit model

[a]This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07045707, NRF-2021R1F1A1048309).
[b]Division of Big Data Science, Korea University, Sejong, South Korea. starlike.kim@gmail.com
[c]Department of Mathematics and Statistics, Helmut Schmidt University, Hamburg, Germany
[d]School of Mathematics, Jilin University, Changchun, China.

Functional Protein Biomarkers Based on Distributions of Single-Cell Expression Levels: From Linear to Nonlinear Quantile Index Predictors[a]

Misung Yi[b*] · Tingting Zhan[c] · Inna Chervoneva[c]

Summary: Background: Protein biomarkers of cancer progression and response to therapy are increasingly important for personalized medicine. Current histocytometry methods enable single-cell quantification of biomolecules in tumor tissue sections through multiplex immunohistochemistr. Advanced quantitative pathology platforms provide distributions of cellular signal intensity (CSI) levels across entire cell populations. However, this rich cell-by-cell biomarker information is typically reduced to a single mean or converted into a simple proportion of biomarker-positive cells, failing to exploit intra-tumor heterogeneity that may provide important prognostic information. Methods: We developed a comprehensive framework using distributions of functional single-cell protein expression levels as cancer biomarkers. The quantile index (QI) biomarker is defined as a weighted average of CSI distribution quantiles in individual tumors, where the weight for each quantile is determined by fitting a functional regression model for a clinical outcome. We extended this to nonlinear QI (nlQI) biomarkers allowing the association between the CSI quantile function and outcome to vary nonlinearly. An algorithm was developed for selecting optimal cutoffs for dichotomizing cell signal intensity distribution quantiles as predictors of continuous, categorical, or survival outcomes. For many functional proteins, single-cell expressions vary independently of spatial localization, and incorporation of spatial information may not affect prognostic value. Results: Linear and nonlinear QI biomarkers based on single-cell expressions of ER, Ki67, TS, CyclinD3, PCNA, PD-L2, and PR were derived and evaluated as predictors of progression-free survival (PFS) or high mitotic index in a large breast cancer dataset. The QI biomarkers demonstrated improved prognostic value compared with standard mean signal intensity predictors. Performance was validated using an independent external cohort. For proteins significantly associated with PFS, optimal quantile biomarkers yielded either larger or similar effect sizes as compared to mean signal intensity biomarkers. Simulation studies demonstrated that nlQI biomarkers yield higher predictive power than linear QI biomarkers when between-tissue variability in CSI distributions is substantial. Conclusions: The proposed approach can be applied to any cell-level expressions of proteins or nucleic acids from immunohistochemistry or other single-cell technologies. R packages Qindex and hyper.gam implementing these methods are freely available on CRAN, featuring user-friendly interfaces and visual tools for exploring integrand surfaces. These tools address the need for biomarkers accounting for heterogeneous protein expression levels in tissues.

Keywords: Quantile index, Single-cell imaging, Multiplex immunofluorescence, Distribution quantiles, Protein biomarker, Functional regression

[a]This work was supported by the National Institutes of Health, U.S. Department of Health and Human Services grants R01CA222847 (I.C., T.Z., and H.R.) and R01CA253977 (H.R. and I.C.). Generation of the underlying data was supported by a Komen Promise grant KG091116 awarded to a team of investigators led by H.R. and the late Edith P. Mitchell.

[b]Dankook University. 152, Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do, 16890, Korea. misung.yi@dankook.ac.kr
[c]Division of Biostatistics & Bioinformatics, Department of Pharmacology, Physiology & Cancer Biology, Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, PA 19107, United States

Probabilistic Data Augmentation for Vibrational Spectroscopic Analysis under Physical-Property Variability[a]

Seung Jee Yang[b*] · Haeseong Jeong[c] · Jomjai Peerapattana[d] · Hoeil Chung[c]

Summary: In high-dimensional vibrational spectroscopy data, uncontrolled spectral variations caused by physical differences among samples—such as variations in measurement conditions and preparation—can distort predictor distributions and degrade regression accuracy. Conventional alignment-based approaches, including domain-invariant partial least squares (di-PLS) and domain-adversarial neural networks (DANN), attempt to address this issue by aligning feature spaces but rely on restrictive assumptions, such as similar response distributions or labeled target data, that rarely hold in practice. We present a probabilistic data augmentation framework that enhances robustness by expanding the training distribution toward unobserved yet physically plausible regions of the predictor–response space. The method combines neighborhood-based geometric modeling with stochastic data generation guided by local covariance in a latent domain. Synthetic samples are drawn around interpolated anchor points using multivariate Gaussian sampling, capturing realistic local variability in both predictors and responses. This nonparametric framework models local dependencies and uncertainty in regression without target-domain information. Experiments demonstrate stabilized predictive performance under experimentally induced variability, providing a practical and interpretable approach for robust modeling of heterogeneous chemical data.

Keywords: Data augmentation, Distribution shift, Spectroscopic data, Chemometrics

[a]This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (2020R1A6A1A06046728, RS-2025-25438303).
[b]Research Institute for Convergence of Basic Sciences, Hanyang University, Seoul, Korea
[c]Department of Chemistry, Hanyang University, Seoul, Korea
[d]Department of Pharmaceutical Technology, Faculty of Pharmaceutical Sciences, Khon Kaen University, Khon Kaen, 40002, Thailand

2025년 동계
학술논문발표회 프로시딩

학생논문

Eigenstructure Inference for High-dimensional Covariance with Generalized Shrinkage Inverse-Wishart Prior[a]

Seongmin Kim[b*] · Kwangmin Lee[c] · Sewon Park[d] · Jaeyong Lee[e]

Summary: In multivariate statistics, estimating the covariance matrix is essential for understanding the interdependence among variables. In high-dimensional settings, where the number of covariates increases with the sample size, it is well known that the eigenstructure of the sample covariance matrix is inconsistent. The inverse-Wishart prior, a standard choice for covariance estimation in Bayesian inference, also suffers from posterior inconsistency. To address the issue of eigenvalue dispersion in high-dimensional settings, the shrinkage inverse-Wishart (SIW) prior has recently been proposed. Despite its conceptual appeal and empirical success, the asymptotic justification for the SIW prior has remained limited. In this paper, we propose a generalized shrinkage inverse-Wishart (gSIW) prior for high-dimensional covariance modeling. By extending the SIW framework, the gSIW prior accommodates a broader class of prior distributions and facilitates the derivation of theoretical properties under specific parameter choices. In particular, under the spiked covariance assumption, we establish the asymptotic behavior of the posterior distribution for both eigenvalues and eigenvectors by directly evaluating the posterior expectations for two sets of parameter choices. This direct evaluation provides insights into the large-sample behavior of the posterior that cannot be obtained through general posterior asymptotic theorems. Finally, simulation studies illustrate that the proposed prior provides accurate estimation of the eigenstructure, particularly for spiked eigenvalues, achieving narrower credible intervals and higher coverage probabilities compared to existing methods. For spiked eigenvectors, the performance is generally comparable to that of competing approaches, including the sample covariance.

Keywords: covariance, eigenstructure, shrinkage inverse-Wishart prior, posterior convergence rate, spiked covariance model

[a]This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. NRF-2023R1A2C1003050).

[b]Department of Statistics, Seoul National University. zlatjdals@snu.ac.kr
[c]Department of Big Data Convergence, Chonnam National University. klee564@jnu.ac.kr
[d]Department of Statistics, Sookmyung Women's University. swpark0413@sookmyung.ac.kr
[e]Department of Statistics, Seoul National University. leejyc@gmail.com

Penalized Maximum Likelihood Estimation for Latent Class Analysis

Jimin Park[a*] · Byungtae Seo[a]

Summary: Latent Class Analysis (LCA) is a statistical model that assumes categorical data arise from unobservable latent classes. The model allows the observed variables to have different distributions across these latent classes. However, if some variables share the same distribution across all latent classes, they are considered irrelevant for clustering. We propose a new method that simultaneously estimates parameters and selects relevant variables for clustering using a penalized maximum likelihood estimation framework. We also present results from simulation studies that demonstrate the performance of our method, as well as a real data analysis that illustrates its practical applicability.

Keywords: Latent Class Analysis, Variable Selection, Penalized Estimation

[a]Department of Statistics, Sungkyunkwan University, 88 Samgaksan-ro, Gangbuk-gu, Seoul, Korea

James-Stein Estimation of Spiked Eigenvectors under the Generalized Spiked Population Model

Giheon Seong[a*] · Seungki Hong[b] · Sungkyu Jung[c]

Summary: In high-dimensional Principal Component Analysis (PCA), sample eigenvectors are known to be biased. Goldberg et al. (2025) proposed a James–Stein shrinkage framework to address this in the high dimension, low sample size (HDLSS) regime where the sample size n is fixed and the dimension $p \to \infty$, which leverages external information modeled as a "target subspace." This work provides a generalization of this framework to the random matrix theory (RMT) regime where $n, p \to \infty$ with their ratio converging to a positive constant. We introduce a new estimator for the asymptotically optimal shrinkage parameter that is consistent in both the RMT and original HDLSS regimes. We prove our proposed method outperforms the sample eigenvector when the target subspace is informative and, crucially, that it is harmless even when the target subspace is non-informative. Our work provides a practical method for enhancing eigenvector estimation, as it requires no a priori knowledge of the signal strength or the quality of the target subspace.

Keywords: PCA, Random Matrix Theory, James–Stein Estimator

[a]Ph.D Candidate. (Presenter) Department of Statistics, Seoul National University, 25-401, 1, Gwanak-ro, Gwanak-gu, Seoul, Korea. heon1998@snu.ac.kr
[b]Meritz Fire & Marine Insurance Co., Ltd, 382, Gangnam-daero, Gangnam-gu, Seoul, Korea. skgaboja@gmail.com
[c]Professor. Department of Statistics, Seoul National University, 25-436, 1, Gwanak-ro, Gwanak-gu, Seoul, Korea. sungkyu@snu.ac.kr

A Comparative Study of the Two-stage Meta-analytic Method and Bayesian Hierarchical Models for Small-area Estimation for the Ozone-mortality Association

Junhwan Lee[a]*, Yeonseung Chung[a]

Summary: In environmental epidemiology, the two-stage meta-analytic method has been widely applied to estimate the association between ozone exposure and mortality across multiple areas. However, this approach faces limitations in small-area estimation, as the relatively few mortality cases in small regions often lead to unstable effect estimates. To address these challenges, Bayesian hierarchical models—both spatial and non-spatial—have been proposed as alternatives. Yet, existing comparisons of these methods have been largely restricted to empirical analyses rather than simulation-based evaluations. In this study, we conducted simulation experiments to systematically compare the two-stage method with spatial and non-spatial Bayesian hierarchical models in the context of small-area estimation of ozone–mortality associations. Two simulation settings were designed, with and without spatial correlation in the effect parameter. Across all scenarios and sample sizes, the two-stage method consistently demonstrated the weakest performance. The non-spatial Bayesian model generally outperformed the others when no spatial dependence was present, while the spatial Bayesian model showed higher performance under spatially correlated settings. These findings provide strong evidence supporting Bayesian approaches as effective alternatives to the two-stage method for small-area estimation of the short-term ozone–mortality association.

Keywords: simulation study, two-stage method, Bayesian hierarchical model, spatial Bayesian hierarchical model, model comparison, ozone, mortality, small-area estimation

[a]Department of Mathematical Sciences, Korea Advanced Institute of Science and Technology

Distributed Reduced-Rank Regression for Large-scale Data

Hongshin Choi[a*] · Seyoung Park[b]

Summary: Modern data analysis often involves processing large-scale, high-dimensional data distributed across multiple machines. In such settings, low-rank regression suffers from limitations in both statistical estimation and computational efficiency. Although convex nuclear-norm regularization offers computational convenience, it introduces bias for large singular values. Moreover, existing distributed optimization methods typically rely on global Lipschitz constants for step-size selection, which require iterative synchronization across nodes and lead to severe communication bottlenecks. To overcome these challenges, we propose a communication-efficient adaptive nuclear normalization method. The approach applies a first-order approximation of the nonconvex Smoothly Clipped Absolute Deviation (SCAD) penalty, transforming the problem into an iterative sequence of weighted nuclear-norm minimizations. In addition, it eliminates global aggregation at each iteration by replacing globally determined Lipschitz-based step sizes with locally computable adaptive updates using the Barzilai–Borwein rule. Extensive experiments demonstrate that the proposed method improves both prediction accuracy and estimation quality while matching or outperforming existing distributed nuclear-norm baselines, often with substantially fewer communication rounds and reduced transmission cost.

Keywords: Distributed estimation, Adaptive Nuclear Norm, SCAD, Low-rank regression

[a]Department of Statistics and Data Science, Yonsei University
[b]Corresponding author: Department of Statistics and Data Science, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea. ishspsy@yonsei.ac.kr

Randomized QLP Decomposition for Third-order Tensors with Unitary Transform

Youngwook Kwon[a*] · Hee-Seok Oh[a]

Summary: Recently, randomized algorithms have gained considerable attention as efficient techniques for dimension reduction in large-scale data across various scientific fields. In this talk, we introduce a randomized algorithm for third-order tensor decomposition based on the tensor-tensor product (t-product) using a unitary transform. Our approach is motivated by randomized tensor approximation methods that depend on random projections of each frontal slice of a tensor. However, these methods still incur significant computational costs when applying SVD or column-pivoted QR decomposition to the slices. To improve the efficiency of randomized algorithms, we propose a randomized tensor QLP decomposition (rt-QLP) without pivoting for third-order tensors, extending the matrix-based QLP to the tensor setting in the transformed domain. Deterministic and probabilistic error bounds are derived by combining properties of the t-product with existing error analysis results of matrix QLP. The effectiveness and efficiency of the proposed method are demonstrated through extensive numerical experiments on tasks such as data compression, image completion, and facial recognition.

Keywords: Randomized algorithm, Tensor decomposition, Transformed t-product, Randomized tensor QLP decomposition

[a]Department of Statistics, Seoul National University, Seoul 08826, Republic of Korea

KSP: Kolmogorov-Smirnov Metric-based Post-Hoc Calibration for Survival Analysis[a]

Jeongho Park[b*] · Daheen Kim · Cheoljun Kim · Hyungbin Park · Sangwook Kang · Gwangsu Kim

Summary: We propose a new calibration method for survival models based on the Kolmogorov–Smirnov (KS) metric. Existing approaches—including conformal prediction, D-calibration, and Kaplan–Meier (KM)-based methods—often rely on heuristic binning or additional nonparametric estimators, which undermine their adaptability to continuous-time settings and complex model outputs. To address these limitations, we introduce a streamlined *KS metric-based post-processing* framework (KSP) that calibrates survival predictions without relying on discretization or KM estimation. This design enhances flexibility and broad applicability. We conduct extensive experiments on diverse real-world datasets using a variety of survival models. Empirical results demonstrate that our method consistently improves calibration performance over existing methods while maintaining high predictive accuracy. We also provide a theoretical analysis of the KS metric and discuss extensions to in-processing settings.

Keywords: Survival analysis, calibration, post-processing, deep learning

[a]This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) [RS-2023-00218377], Global Learning & Academic research institution for Master's·PhD students, and Postdocs (LAMP) Program of the NRF grant funded by the Ministry of Education [RS-2024-00443714].

[b](Shinchon-dong, Yonsei University) 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea. wjdgh4325@yonsei.ac.kr

Insurance Ratemaking with Endogenous Deductibles[a]

Dong Ha Lee[b*] · Himchan Jeong[c] · Peng Shi[d]

Summary: This project examines the insurance ratemaking with endogenous deductibles. In non-life insurance, the deductible controls the frequency and severity of reported claims, so its endogeneity has been argued in terms of policyholder‘s behaviour and asymmetric information to insurers. In this context, we consider two scenarios: without underreporting and with underreporting. Underreporting refers to the case where a loss event is not reported to the insurer if the loss amount is below the deductible. These two scenarios provide a solid procedure to detect endogeneity of the deductible depending on the presence of underreporting. For empirical analysis, a dataset from the Wisconsin Local Government Property Insurance Fund (LGPIF) is chosen. The endogeneity of the deductible is investigated by comparing the results in each model assuming its exogeneity. For the joint modeling for a given deductible, a pair copula is adopted to capture the dependence from the endogeneity of the deductible. The results of joint modeling improved the goodness-of-fit than modeling with exogenous deductibles. Notably, severity joint modeling gives a consistent outcome regardless of underreporting or not.

Keywords: Dependence modeling, Frequency-severity, Asymmetric Information

[a]This work is was in part supported by the Korea Life Insurance Association Social Contribution Committee.
[b]Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Dr W, Burnaby, BC, V5A 1S6 Canada. dongha_lee@sfu.ca
[c]Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Dr W, Burnaby, BC, V5A 1S6 Canada. himchan_jeong@sfu.ca
[d]Risk and Insurance Department, Wisconsin School of Business, 5191D Grainger Hall, 974 University Ave, Madison, WI 53597, USA. pshi@bus.wisc.edu

Robust Bayesian Estimation in Conditionally Heteroscedastic Time Series Models[a]

Jeongho Lee[b*] · Junmo Song[c]

Summary: Outliers can seriously distort statistical inference by inducing excessive sensitivity in the likelihood function, thereby compromising the reliability of Bayesian estimation. To address this issue, we develop a robust Bayesian estimation method for conditionally heteroscedastic time series models by extending the density power divergence (DPD) framework to the Bayesian setting. The resulting DPD-based posterior distribution, controlled by a tuning parameter, achieves a smooth balance between efficiency and robustness. We establish the asymptotic properties of the proposed estimator. In particular, the DPD-based posterior is shown to satisfy a Bernstein-von Mises type theorem, converging to a normal distribution centered at the minimum DPD estimator. Furthermore, the corresponding Bayes estimator, defined as the expected DPD-based posterior estimator (EDPE), is asymptotically equivalent to the minimum DPD estimator. Monte Carlo simulations based on GARCH models confirm that the proposed method performs well under both uncontaminated and contaminated data, maintaining robustness where the ordinary Bayesian estimator becomes severely biased. An empirical application to Bitcoin return data further demonstrates the practical advantages of the proposed robust Bayesian framework for financial time series analysis.

Keywords: Conditionally heteroscedastic time series model, Robust Bayesian estimation, Density power divergence, GARCH model

[a]This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2025-20122969).

[b]Graduate student, Department of Statistics, Kyungpook National University, Daegu 41566, Korea. gho0327@knu.ac.kr

[c]Corresponding author: Professor, Department of Statistics, Kyungpook National University, Daegu 41566, Korea. jsong@knu.ac.kr

Dynamic Effect Analysis of Housing Price

Seo In Jang[a*]

Summary: This study quantitatively examines the inter-regional return spillover structure from both static and dynamic perspectives, utilizing weekly housing price data from Seoul's 25 districts over the period 2010–2020. Methodologically, we employ vector autoregression (VAR) and generalized forecast error variance decomposition (GFEVD) to compute directional spillover indices (TO, FROM, NET) and the total connectedness index (TCI). Additionally, we derive the stationary distribution (π) from the transition matrix and separately interpret Viral Centrality (VC). The analysis reveals that in the static network, Songpa-gu exhibits strong net transmitter characteristics, while Gangnam-gu displays relatively net receiver properties. Yeongdeungpo-gu exhibits high steady-state outflows and moderate Viral Centrality (VC), suggesting its role as a persistent source in the steady state and a potential dampener for surrounding regional housing prices. The TCI from the 52–week rolling window remains generally stable, but significant spikes (sharp increases) are identified in late July 2012, early 2014, mid–March 2015, early August 2017, and mid-March 2020. By integrating stationary distribution and Viral Centrality (VC) with the FEVD–based Spillover Index, this study provides a refined explanation of the differential roles in the inter-regional transmission mechanism.

Keywords: Spillover Effect, Housing Price, Network Analysis, Transition Matrix

[a]Department of Statistics, Kyungpook National University

Memorize Early, Then Query: Inlier-memorization-guided Active Outlier Detection

Minseo Kang[a*] · Seunghwan Park[b] · Dongha Kim[c]

Summary: Outlier detection (OD) aims to identify abnormal instances, known as outliers or anomalies, by learning typical patterns of normal data, or inliers. Performing OD under an unsupervised regime–without any information about anomalous instances in the training data–is challenging. A recently observed phenomenon, known as the inlier-memorization (IM) effect, where deep generative models (DGMs) tend to memorize inlier patterns during early training, provides a promising signal for distinguishing outliers. However, existing unsupervised approaches that rely solely on the IM effect still struggle when inliers and outliers are not well-separated or when outliers form dense clusters. To address these limitations, we incorporate active learning to selectively acquire informative labels, and propose IMBoost, a novel framework that explicitly reinforces the IM effect to improve outlier detection. Our method consists of two stages: 1) a warm-up phase that induces and promotes the IM effect, and 2) a polarization phase in which actively queried samples are used to maximize the discrepancy between inlier and outlier scores. In particular, we propose a novel query strategy and tailored loss function in the polarization phase to effectively identify informative samples and fully leverage the limited labeling budget. We provide a theoretical analysis showing that the IMBoost consistently decreases inlier risk while increasing outlier risk throughout training, thereby amplifying their separation. Extensive experiments on diverse benchmark datasets demonstrate that IMBoost not only significantly outperforms state-of-the-art active OD methods but also requires substantially less computational cost.

Keywords: Outlier Detection, Deep Generative Models, Active Learning, IM Effect

[a]Master's student, Department of Statistics, Sungshin Women's University, (02844) 2, Bomun-ro 34 Da-gil, Seongbuk-gu, Seoul, Korea. krystalms0306@gmail.com

[b]Associate Professor, Department of Information Statistics, Kangwon National University, (24341) 1, Gangwon-daehak-gil, Chuncheon-si, Gangwon, Korea. stat.shpark@kangwon.ac.kr

[c]Assistant professor, School of Mathematics, Statistics and Data Science, Sungshin Women's University, (02844) 2, Bomun-ro 34 Da-gil, Seongbuk-gu, Seoul, Korea. dongha0718@gmail.com

Scalable and Efficient Multiple Imputation for Case-Cohort Studies via Influence Function-Based Supersampling[a]

Jooho Kim[b*] · Yei Eun Shin[b,c]

Summary: Two-phase sampling designs have been widely adopted in epidemiological studies to reduce costs when measuring certain biomarkers is prohibitively expensive. Under these designs, investigators commonly relate survival outcomes to risk factors using the Cox proportional hazards model. To fully utilize covariates collected in phase 1, multiple imputation methods have been developed to impute missing covariates for individuals not included in the phase 2 sample. However, performing multiple imputation on large-scale cohorts can be computationally intensive or even infeasible. To address this issue, Borgan et al. (2023) proposed a random supersampling (RSS) approach that randomly selects a subset of cohort members for imputation, albeit at the cost of reduced efficiency. In this study, we propose an influence function-based supersampling (ISS) approach with weight calibration. The method achieves efficiency comparable to imputing the entire cohort, even with a small supersample, while substantially reducing computational burden. We further demonstrate that the proposed method is particularly advantageous when estimating hazard ratios for high-dimensional expensive biomarkers. Extensive simulation studies are conducted, and a real data application using the National Institutes of Health-American Association of Retired Persons (NIH-AARP) Diet and Health Study is provided to illustrate the effectiveness of the proposed method.

Keywords: Influence function, Multiple imputation, Supersampling, Weight calibration

[a]This research was supported by the National Research Foundation of Korea(NRF) Grant funded by the Korea government(MSIT) (RS-2023-00211561) and the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (RS-2024-00462238).

[b]Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, South Korea
[c]College of Liberal Studies, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, South Korea. shin.y@snu.ac.kr

A New Algorithm for the Maximum Likelihood Estimator in Erlang Mixtures[a]

KyeongA Yang[b*] · Byungtae Seo[c]

Summary: The Erlang mixture model is widely used to model insurance losses and other aggregated risks. However, finding the maximum likelihood estimate (MLE) of its parameters remains challenging because the shape parameters are integer-valued. This discreteness complicates the use of standard expectation–maximization (EM) algorithms and often leads to convergence toward local optima. Although several alternative algorithms have been proposed, they tend to be computationally intensive and unstable across different initializations. In this study, we propose a novel algorithm that employs a gradient function to efficiently guide the exploration of the discrete parameter space. The proposed method improves both stability and computational efficiency compared to existing approaches. Through numerical experiments, we demonstrate that our algorithm consistently achieves higher or comparable log-likelihood values with reduced variability, offering a practical and reliable solution for MLE in Erlang mixture models.

Keywords: Erlang mixtures, Gradient function, EM algorithm, Maximum likelihood estimation.

[a]This work was in part supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (No. RS-2025-02216235) and (No. RS-2022-NR069313).

[b]Department of Statistics, Sungkyunkwan University, 25-2 Sungkyunkwan street, Jongno-Gu, Seoul 03063, Korea. yka2524@gmail.com

[c]Department of Statistics, Sungkyunkwan University, 25-2 Sungkyunkwan street, Jongno-Gu, Seoul 03063, Korea. seobt@skku.edu

Bayesian Piecewise Shape-restricted Regression for Estimating Minimum Mortality Temperature Range in Temperature-mortality Studies

Seunghyun Eom[a*] · Yeonseung Chung[b]

Summary: Numerous epidemiological studies have explored the relationship between temperature and mortality, often interpreting the minimum mortality temperature (MMT) as an indicator of population adaptation to local thermal environments. Traditionally, this relationship has been modeled using U- or J-shaped curves, with the MMT defined as the temperature at which mortality risk is minimized. However, growing evidence indicates that temperature-mortality curves often exhibit a flat-bottom shape, suggesting that a range of temperatures—rather than a single point—may be associated with the lowest mortality risk. In such cases, it is more appropriate and realistic to define the MMT as a range rather than a unique value. In this study, we propose a Bayesian piecewise shape-restricted regression model that explicitly defines the MMT as a range. The model imposes monotonicity constraints on both sides of the MMT range, ensuring that temperatures beyond this range are associated with higher mortality risk. Posterior inference is performed using a Sequential Monte Carlo sampler with a Hamiltonian Monte Carlo kernel. We conducted simulation studies to evaluate the performance of the proposed approach and applied it to real-world data to identify minimum mortality temperature ranges in a temperature–mortality association study.

Keywords: Bayesian shape-restricted regression, temperature–mortality association, minimum mortality temperature (MMT), Sequential Monte Carlo (SMC) sampler

[a]PhD Candidate, Graduate School of Data Science, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea
[b]Associate Professor, Department of Mathematical Sciences, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

A Missing Value Imputation Method for High-dimensional Tabular Data Using DeepInsight and Image Inpainting[a]

Jeseok Lee[b*] · Byungwon Kim[c]

Summary: While the increasing scale and complexity of modern data is a key characteristic of data science, it has exacerbated the severity of the missing value problem. As the number of features (dimensionality) increases, conventional simple statistical imputation methods distort the complex distribution of the data, while existing methods, such as multivariate imputation algorithms, face challenges of excessive computational cost. To overcome these limitations, this study proposes a novel hybrid approach that applies the image inpainting technique from computer vision to tabular data. The process begins by employing the DeepInsight algorithm to transform the tabular data into image data, where highly correlated features are grouped into the same pixel locations. Following this transformation, a two-stage imputation strategy that leverages the new data structure is performed. For pixels containing a mix of missing and non-missing values, imputation is primarily performed using information from the non-missing values within that same pixel group. Conversely, when a pixel consists entirely of missing values, it is treated as a 'lost region' of the image, and an AI-based inpainting model is applied to restore its value. Finally, the image with all values imputed is transformed back into tabular data. The proposed method is compared with existing methods using data from The Cancer Genome Atlas (TCGA). By converging technologies from heterogeneous domains, this research presents a new direction for addressing the missing value problem in high-dimensional data. While the increasing scale and complexity of modern data is a key characteristic of data science, it has exacerbated the severity of the missing value problem. As the number of features (dimensionality) increases, conventional simple statistical imputation methods distort the complex distribution of the data, while existing methods, such as multivariate imputation algorithms, face challenges of excessive computational cost. To overcome these limitations, this study proposes a novel hybrid approach that applies the image inpainting technique from computer vision to tabular data. The process begins by employing the DeepInsight algorithm to transform the tabular data into image data, where highly correlated features are grouped into the same pixel locations. Following this transformation, a two-stage imputation strategy that leverages the new data structure is performed. For pixels containing a mix of missing and non-missing values, imputation is primarily performed using information from the non-missing values within that same pixel group. Conversely, when a pixel consists entirely of missing values, it is treated as a 'lost region' of the image, and an AI-based inpainting model is applied to restore its value. Finally, the image with all values imputed is transformed back into tabular data. The proposed method is compared with existing methods using data from The Cancer Genome Atlas (TCGA). By converging technologies from heterogeneous domains, this research presents a new direction for addressing the missing value problem in high-dimensional data.

Keywords: Missing Value Imputation, High-Dimensional Data, DeepInsight, Image Inpainting

[a]This work was in part supported by the Korea Research Foundations, Korea, under grant RS-2023- 00213626.
[b]Ph.D student, Department of Statistics, Kyungpook National University, Daegu 41566, Korea. dlakakwns@knu.ac.kr
[c]Corresponding author:Associate Professor, Department of Statistics, Kyungpook National University, Daegu 41566, Korea. byungwonkim@knu.ac.kr

A Statistical Framework for Assessing Synthetic Tabular Data Quality

Jimin Kim[a*] · Jongwoo Song[b]

Summary: Synthetic tabular data has emerged as a practical solution for safeguarding privacy, augmenting limited datasets, and mitigating class imbalance. Despite significant progress in generative modeling, the evaluation of synthetic data remains an open challenge. A central requirement is to determine whether synthetic data are statistically indistinguishable from real data while preserving structural relationships among variables. TabEvalIdx is introduced as a unified index that consolidates multiple statistical perspectives into a single bounded score. The framework jointly considers marginal distributions and inter-variable dependencies, enabling interpretable and consistent comparisons across datasets and models. Experiments on benchmark datasets demonstrate its ability to provide stable, discriminative, and practically meaningful evaluations.

Keywords: Tabular Data Generation, Performance Metric, Statistical Similarity

[a]Ewha Womans University, Seoul 03760, Korea. jmpape21@gmail.com
[b]Ewha Womans University, Seoul 03760, Korea. josong@ewha.ac.kr

Scaling Up ROC-optimizing Support Vector Machines[a]

Gimun Bae[b*] · Seung Jun Shin[c]

Summary: The ROC-SVM, introduced by Rakotomamonjy (2004), directly maximizes the area under the ROC curve (AUC) and is widely used for imbalanced classification. A major drawback, however, is its computational cost: training requires evaluating all O(n^2) positive–negative pairs, which quickly becomes infeasible as the sample size grows. To reduce this burden, we propose a scalable algorithm based on incomplete U-statistics, which reduces the complexity from O(n^2) to O(n). We further extend the approach to nonlinear settings through a low-rank approximation for efficient kernel matrix computation, while preserving statistical accuracy. As a result, training time decreases from hundreds of seconds to just a few seconds without significant loss in predictive performance, making the model feasible for large-scale imbalanced datasets.

Keywords: Large-scale learning, Imbalanced binary classification, ROC curve, etc

[a]This work was in part supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT), grant number 2023R1A2C1006587.
[b]Master's student, Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea. bgd5517@korea.ac.kr
[c]Professor, Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea. sjshin@korea.ac.kr

Online Gradient Descent와 Thompson Sampling을 통한 효율적 Heavy-tailed 선형 밴딧 알고리즘[a]

선우영민[b*] · 김지수[b,c]

요약: 본 논문은 두꺼운 꼬리 (heavy-tailed) 보상을 가지는 확률적 선형 밴딧 (Stochastic Linear Bandits) 문제를 해결하기 위한 새로운 효율적 알고리즘을 제안한다. 지금까지 두꺼운 꼬리 보상을 취급하는 여러 선형 밴딧 알고리즘들이 제안되어 왔으며, 특히 online mirror descent (OMD) 업데이트를 활용한 계산 효율적인 알고리즘도 제안된 바 있다 [Wang, Jing, et al. (2025)][d]. 그러나 해당 알고리즘은 역행렬 연산으로 인해 각 스텝에서 $O(d^2)$의 계산 복잡도가 요구되며, 여기서 d는 입력 feature의 차원이다. 이러한 계산 비용을 줄이기 위해, 본 논문은 online gradient descent (OGD) 업데이트를 채택하여 각 스텝에서 $O(d)$의 계산 복잡도로 작동하는 새로운 효율적인 두꺼운 꼬리 보상을 취급하는 선형 밴딧 알고리즘을 제안한다. 제안된 알고리즘은 OGD 업데이트를 통해 누적된 경험들을 이용하면서, Thompson Sampling (TS) 기반의 확률적 탐색을 결합하여 탐험과 이용간의 균형을 달성한다. 우리는 특히 두꺼운 꼬리 보상을 야기하는 노이즈의 분포가 대칭적(symmetric)일 경우에 feature 분포에 관한 고유값에 대한 가정 아래에서 제안된 알고리즘의 regret의 상한이 $\tilde{O}(\sqrt{T})$를 달성함을 보이며, 여기서 T는 총 라운드이다. 이는 현재까지 제안된 두꺼운 꼬리 보상을 취급하는 선형 밴딧 알고리즘들 중 가장 낮은 regret의 상한에 해당한다. 마지막으로 우리는 시뮬레이션 실험을 통해 개발된 알고리즘의 성능을 검증한다.

주요용어: 확률적 선형 밴딧, 두꺼운 꼬리 보상, online gradient descent

[a]이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. RS-2020-II201336, 인공지능대학원지원(울산과학기술원); No.2022-0-00469, 자율 드론 실용화를 위한 목적지향 강화학습 핵심기술 개발).

[b]울산과학기술원 인공지능대학원

[c]울산과학기술원 산업공학과

[d]Wang, Jing, et al. "Heavy-tailed linear bandits: Huber regression with one-pass update." arXiv preprint arXiv: 2503.00419 (2025).

Locally Optimal Private Sampling

Hrad Ghoukasian[a*] · Bonwoo Lee[b*] · Shahab Asoodeh[c]

Summary: We study the problem of sampling from a distribution under local differential privacy (LDP). Given a private distribution P, the goal is to generate a single sample from a distribution that remains close to P in f-divergence while satisfying the constraints of LDP. This task captures the fundamental challenge of producing realistic-looking data under strong privacy guarantees. While prior work by Park et al. (NeurIPS'24) focuses on global minimax-optimality across a class of distributions, we take a local perspective. Specifically, we examine the minimax error in a neighborhood around a fixed distribution P_0, and characterize its exact value, which depends on both P_0 and the privacy level. Our main result shows that the local minimax error is determined by the global minimax error when the distribution class is restricted to a neighborhood around P_0. To establish this, we (1) extend previous work from pure LDP to the more general functional LDP framework, and (2) prove that the globally optimal functional LDP sampler yields the optimal local sampler when constrained to distributions near P_0. Building on this, we also derive a simple closed-form expression for the locally minimax-optimal samplers which does not depend on the choice of f-divergence. We further argue that this local framework naturally models private sampling with public data, where the public data distribution is represented by P_0. In this setting, we empirically compare our locally optimal sampler to existing global methods, and demonstrate that it consistently outperforms global minimax samplers.

Keywords: private sampling, local minimax, local differential privacy, functional LDP

[a]Department of Computing and Software, McMaster University. ghoukash@mcmaster.ca
[b]E2 3220, Department of Mathematical Science, Korea Advanced Institute of Science & Technology. righthim@kaist.ac.kr
[c]Department of Computing and Software. McMaster University. asoodeh@mcmaster.ca

Inference on Gaussian mixture models with dependent labels[a]

Seunghyun Lee[b*] · Rajarshi Mukherjee[c] · Sumit Mukherjee[d]

Summary: Gaussian mixture models are widely used to model data generated from multiple latent sources. Despite their popularity, most theoretical research assumes that the labels are either independent and identically distributed, or follow a Markov chain. It remains unclear how the fundamental limits of estimation change under more complex dependence. This presentation addresses this question for the spherical two-component Gaussian mixture model. We first show that for labels with an arbitrary dependence, a naive estimator based on the misspecified likelihood is $\sqrt{n}$-consistent. Additionally, under labels that follow an Ising model, we establish the information-theoretic limitations for estimation, and discover an interesting phase transition as dependence becomes stronger. The Ising model is a popular quadratic interaction model that incorporates dependence via a network. When the dependence is smaller than a threshold, the optimal estimator and its limiting variance exactly match the independent case, for a wide class of Ising models. On the other hand, under stronger dependence, estimation becomes easier and the naive estimator is no longer optimal. Hence, we propose an alternative estimator based on the amortized variational approximation of the likelihood, and argue its optimality under a specific Ising model.

Keywords: Gaussian mixture model, hidden Markov random field, Ising model, phase transition

[a]This work is supported by NSF awards CAREER-8529216-01 and DMS-2113414.

[b](Presenting author) Department of Statistics, Columbia University, 1255 Amsterdam Avenue New York, NY 10027 USA. sl4963@columbia.edu, sm3949@columbia.edu

[c]Department of Biostatistics, Harvard University, 677 Huntington Avenue, Boston, MA 02115 USA. ram521@mail.harvard.edu

[d]Department of Statistics, Columbia University, 1255 Amsterdam Avenue New York, NY 10027 USA. sm3949@columbia.edu

2025년 동계
학술논문발표회 프로시딩

포스터세션

종단 오믹스 자료 발현 분석을 위한 R 패키지 개발[a]

강하주[b*] · 박선철[c] · Nguyen Phuoc Long[d] · 박성오[e]

요약: 종단 오믹스(omics) 자료는 동일한 생물학적 샘플에서 여러 시간 지점에 걸쳐 수천 개의 분자 특성을 반복 측정하는 데이터이다. 그러나 기존 분석 패키지인 maSigPro는 오차의 상관구조를 독립으로 가정하여 시간적 종속성을 반영하지 못하고, 개체간 이질성을 고려하지 않아 통계적 추론의 왜곡과 동적 발현 변화 해석에 제약을 준다. 이를 보완하기 위해 본 연구에서는 선형 혼합효과모형(linear mixed-effects model)을 활용하여 시간 상관 구조와 개체 간 이질성을 명시적으로 반영하는 새로운 분석 패키지를 제안한다. 제안된 패키지는 데이터 전처리, 유전자 수준의 모형 적합 및 통계적 검정, 시간적 발현 패턴 시각화, 시간 추세 기반 클러스터링으로 구성된다. 실제 유전체(genomics) 자료에 적용한 결과, 기존의 독립 가정 기반 접근법보다 시간 변동 양상 탐지 성능이 향상되었으며, 발현 패턴을 효율적으로 요약할 수 있었다. 본 연구는 종단 오믹스 자료 발현 분석에서 시간적 종속성을 반영하는 실용적이고 통계적으로 정립된 분석 틀을 제시하며, 유전체를 포함한 다양한 오믹스 연구로의 확장 가능성을 보여준다.

주요용어: 오믹스 자료, R 패키지, 종단 자료 분석, 선형 혼합 효과 모형

[a]This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00339064).

[b]Department of Applied Statistics, College of Natural Sciences, Hanyang University, 222-1 Wangsimni-ro, Seongdong-gu, Seoul 04763, Republic of Korea. haju0903@gmail.com

[c]Department of Mathematics; Research Institute for Natural Sciences, College of Natural Sciences, Hanyang University, 222 Wangsimni-ro, Seongdong-gu, Seoul 04763, Republic of Korea. pscstat@hanyang.ac.kr

[d]Graduate Institute of Biomedical Sciences and Molecular Medicine Research Center, College of Medicine, Chang Gung University, 259 Wenhua 1st Rd, Guishan District, Taoyuan City 333, Taiwan. bsngphuoclong@gmail.com

[e]School of Mathematics, Statistics and Data Science, Sungshin Women's University, 2 Bomun-ro 34da-gil, Seongbuk-gu, Seoul 02844, Republic of Korea. spark6@sungshin.ac.kr

Multivariate T Mixture of Experts

고병준[a*] · 서병태[b]

요약: Mixture of Experts는 데이터 내 잠재 그룹에 따라 공변량과 반응 변수의 관계가 달라지는 이질성을 모델링하는 데에 효과적인 프레임워크이다. 그러나 기존 연구들은 주로 일변량 반응변수에 국한되어 있으며, 전문가 네트워크의 정규분포 가정으로 인해 이상치나 두꺼운 꼬리를 가진 분포에 취약하여 군집 할당 및 회귀 계수 추정의 정확도가 저하되는 한계를 가진다. 본 연구에서는 이러한 한계를 극복하고자, Experts에 다변량 t-분포를 적용한 강건한 다변량 전문가 혼합 모형(TMoE)을 제안한다. 제안 모형은 t-EIGEN 패밀리를 기반으로 공분산 행렬에다양한 구조적 제약을 가해 추정해야 할 모수의 수를 줄이고, 효율적이고 안정적인 추정을 가능하게 한다. 또한 EM 알고리즘을 통해 모수를 안정적으로 추정하는 방법을 제시하였다. 모의실험과 실증 분석 결과, 제안한 다변량 TMoE 모형은 이상치가 포함된 환경에서 자유도 모수를 조절하여 이상치의 영향을 효과적으로 완화하였으며, 기존 정규분포 기반 모형보다 우수한 예측 및 군집화 성능을 보임을 입증하였다.

주요용어: Mixture of Experts, EM알고리즘, t-EIGEN 패밀리, 강건성

[a](03063) 서울시 종로구 성균관로 25-2, 성균관대학교 통계학과. qudwns0583@skku.edu
[b](03063) 서울시 종로구 성균관로 25-2, 성균관대학교 통계학과.

Site Selection for Public Bike Stations in Seoul Using Spatially Clustered Regression with Bus Ridership Patterns[a]

Junyoung Ko[b*] · Seung Jee Yang[c] · Jaehong Jeong[d]

Summary: Efficiently selecting sites for public bike stations in urban mobility systems requires a thorough understanding of spatial heterogeneity and its relationship with nearby public transit flows. This study extends the Spatially Clustered Regression (SCR) model by incorporating bus ridership time series patterns to jointly capture spatial and temporal coherence in local transit dynamics. Consequently, the relationship between bike-sharing demand and adjacent bus usage becomes clearer. However, achieving strong spatial and temporal coherence often conflicts with maintaining high regression performance. To address this trade-off, we adopt a multi-objective optimization framework that simultaneously optimizes clustering quality and regression accuracy. The resulting set of Pareto-optimal models balances spatio-temporal clustering consistency with predictive performance. These Pareto solutions provide valuable insights into the spatio-temporal clustering of bike-sharing demand and its interactions with public transportation systems, facilitating the identification of potential locations for new bike stations in Seoul.

Keywords: Spatially clustered regression, Multi-objective optimization, Bike sharing, Site selection

[a]This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00345167).
[b]Department of Applied Statistics, Hanyang University, Seoul, 04763, South Korea
[c]Research Institute for Convergence of Basic Sciences, Hanyang University, Seoul, Korea
[d]Department of Mathematics, Hanyang University, Seoul, 04763, South Korea

Median-Polish Kriging on a Horseshoe Domain with Wing Isolation[a]

Gulim Alkenova[b*] · Hyoung Moon Kim[c]

Summary: Spatial interpolation over irregular or non-convex domains, such as a horseshoe-shaped region, poses significant challenges for conventional kriging due to domain topology and boundary discontinuities. In this study, we implement Median Polish Kriging (MPK) to achieve robust spatial prediction across a non-convex horseshoe domain. Spatial prediction on irregular, nonconvex domains poses challenges for classical interpolation methods due to disconnected regions and artificial smoothing across voids. We present an implementation of Median Polish Kriging (MPK) tailored to a complex horseshoe-shaped domain using a wing isolation strategy to prevent cross-domain leakage. MPK separates large-scale trends and small-scale variation by first applying Tukey's median polish to a gridded representation of the data, then kriging the residuals. To avoid cross-wing leakage, the horseshoe polygon was split horizontally into top and bottom wings. MPK was performed independently on each wing using synthetic data, and the resulting rasters were aligned and mosaicked to reconstruct the full domain. The split –fit–stitch procedure produced smooth, realistic fields within each wing while preserving the open gap between them. This method provides a simple, geometry-aware framework for robust kriging on complex, nonconvex spatial domains.

Keywords: Median Polish Kriging, Spatial Interpolation, Non-convex Geometry

[a]This work was in part supported by the National Research Foundation of Korea (NRF), grant funded by the Korean government (MSIT)(RS-2024-00357199).

[b]Data Science, Konkuk University, Seoul, Korea

[c]Corresponding author: Applied Statistics, Konkuk University, Seoul, Korea. hmk966a@gmail.com

Joint Bayesian Additive Regression Trees for Prediction and Causal Inference

Nayeon Kim[a*] · Minjin Ha[b]

Summary: In real-world clinical data, it is common to partition patients into subgroups according to demographic and genetic factors, and heterogeneity in predictive relationships is frequently observed in this process. Meanwhile, some shared structure may exist across subgroups. Therefore, this study aims to preserve group-specific differences while sharing common structure by extending JointBART to prediction and causal inference to implement partial pooling across groups. Specifically, we fit a separate BART ensemble to each group and jointly estimate variable-selection probabilities under a coupled prior, so that group heterogeneity is preserved while common components are selectively shared. In simulations, we evaluated performance against competing models across diverse regression structures (linear, nonlinear, and mixed), non-Gaussian errors, and settings without information sharing. JointBART exhibited consistent superiority in variable-selection capability and predictive accuracy across all scenarios. In the real-data application, we conducted causal-inference analysis to identify synthetic lethality (SL) pairs using cancer cell line data, thereby further confirming the applicability of JointBART to causal inference and its interpretability.

Keywords: JointBART, Causal inference, Subgroup analysis, Partial pooling

[a]Master degree, Department of Biostatistics and Computing, Yonsei University Graduate School. skdus0704@yuhs.ac
[b]Associate professor, Biohealth Data Science, Graduate school of Transdisciplinary Health Science, Yonsei University. mjha@yuhs.ac

SVDD-Based Charts with DTW Kernel for Time-Series Anomaly Detection

DongGeun Kim[a*] · SangWoo Kim[b] · SooHyun Ahn[c]

Summary: The Support Vector Data Description (SVDD)-based control chart, such as OC-SVM and D^2 charts, have been widely used for process monitoring and anomaly detection by describing the boundary of in-control data. However, conventional RBF kernel-based approaches often fail to handle temporal misalignment in time-series data, which limits their ability to detect process shifts accurately. To address this limitation, this study proposes SVDD-based OC-SVM and D^2 charts incorporating a Dynamic Time Warping (DTW) kernel, which better measures temporal similarity between sequences. The proposed charts capture nonlinear and temporally distorted patterns, improving robustness for time-series anomaly detection. Applications to benchmark datasets, including ECG5000 and Ford A, demonstrate the potential of DTW-based methods for reliable process monitoring.

Keywords: Dynamic Time Warping (DTW), Support Vector Data Description (SVDD), D^2 Control Charts, Time-Series Anomaly Detection

[a]B.S.Graduate, Department of Mathematics, Ajou University
[b]Undergraduate Student, Department of Mathematics, Ajou University
[c]Professor, Department of Mathematics, Ajou University

시계열 데이터의 노이즈 제거 방법 비교 분석[a]

김동혁[b*] · 이종민[c]

요약: 본 연구는 경계와 이상점이 많은 시계열 신호에서 이동평균의 과도한 평활화 문제와 이동중앙값의 계산비용·확장성 한계를 보완할 대안을 탐색하기 위해, 로버스트 방법과의 비교와 데이터 유형별 선택 지침 확립을 목적으로 한다. 이를 위해 이미지 분야에서 제안된 M-smoother 아이디어와 Huber M-estimation의 IRLS(Iteratively Reweighted Least Squares)절차를 창 기반으로 적용한 Moving Huber M-estimation을 정의하고 이동평균·이동중앙값·단순지수평활법과 함께 비교하였다. 모의 실험은 경계가 많은 신호와 부드러운 선형신호를 구성한 뒤, 가우시안·라플라스·가우시안 혼합·코시 등 네가지 노이즈를 추가하여 L_1, L_2, L_∞ 지표로 성능을 평가하였고, 외삽 가능성 평가는 실제 데이터(2021년 GameStop 일별 고가, 도로 교통 센서 점유율, 트위터 'GOOG' 언급량)에 적용해 확인하였다. 그 결과, 신호 구조와 노이즈 특성에 따라 최적 방법이 달라짐을 확인했다. 신호에 경계가 존재하며 노이즈 수준이 낮을 때 M-estimation 기반 방법들이 우수한 성능을 보여줬고, 신호가 경계 없이 부드럽게 변하는 경우에는 단순지수평활법이 성능이 가장 좋았다. 또한 코시 분포와 비슷한 환경에서는 이동중앙값이 뛰어난 성능을 보여주었다. 결론적으로 단일 방법의 보편적 우위는 없으며, 신호의 유형을 고려한 선택이 필요하고, 이동평균과 이동중앙값의 장단점을 보완하는 실용적 대안으로서 창 기반 Moving Huber M-estimation의 적용 가능성을 제안한다.

주요용어: 시계열, 디노이징, 로버스트 통계, M-estimation 등

[a]이 논문은 2025 부산대학교 대학원혁신실의 지원을 받아 진행된 논문입니다.
[b](46540) 부산광역시 금정구 부산대학로63번길 2, 부산대학교, 통계학과, 대학원생. yhsb1849@naver.com
[c](46540) 부산광역시 금정구 부산대학로63번길 2, 부산대학교, 통계학과, 교수. jongminlee9218@gmail.com

Lightweight Statistical Detection of Imperceptible Poisoning with Downscaling and Latent Shift Analysis[a]

Donghyeon Kim[b*] · Hye-Young Jung[c]

Summary: This study investigates the statistical characteristics of imperceptible poisoning—visually undetectable perturbations that distort model representations—and proposes a lightweight detection method without additional learning. A Variational Autoencoder (VAE) encoder is employed in combination with image downscaling to construct a latent distribution of clean data, where downscaling serves to attenuate high-frequency noise and amplify distributional differences between clean and poisoned samples. Latent shift values derived from the VAE representations are statistically analyzed, and classification based on fitted distribution intersections demonstrates that simple statistical comparison can effectively detect imperceptible poisoning without model-dependent procedures. The proposed approach offers a practical and statistically grounded preprocessing strategy for large-scale dataset purification.

Keywords: Imperceptible Poisoning, Latent Shift, Statistical Detection, High-Frequency Components

[a]his work was in part supported by the Korea Research Foundations, Korea, under grant KRF-2022R1F1074939.
[b](15588) Department of Mathematical Data Science, College of Computing, Hanyang University, 55 Hanyangdaehak-ro, Ansan, Republic of Korea. tails2002@hanyang.ac.kr
[c](15588) Department of Mathematical Data Science, College of Computing, Hanyang University, 55 Hanyangdaehak-ro, Ansan, Republic of Korea. hyjunglove@hanyang.ac.kr

Machine Learning for Gait Data: Improved Classification of Berg Balance Scale Using Multivariate Functional PCA Combined with the Square-root Velocity Framework

Minseok Kim[a*], Jieun Cho[b], Hogene Kim[c], Jooyoung Lee[d]

Summary: Post-stroke gait asymmetry is a major contributor to impaired balance and increased fall risk. Conventional analyses using discrete gait parameters or standard principal component analysis (PCA) often fail to capture the continuous, multivariate, and nature of gait motion. This study introduces a framework integrating Multivariate Functional Principal Component Analysis (MFPCA) with the Square-Root Velocity Framework (SRVF) to separate amplitude and phase variability in post-stroke gait trajectories. Kinematic data from 43 stroke participants were processed using PCA, MFPCA, and MFPCA combined with SRVF. The Synthetic Minority Oversampling Technique (SMOTE) was applied to address class imbalance between high- and low-BBS groups. The MFPCA-SRVF model with a linear SVM on the paretic-side achieved the highest classification performance (Accuracy = 0.889; F1 = 0.833; AUC = 0.903), outperforming both PCA and unaligned MFPCA. SRVF alignment corrected for phase distortions, revealing biomechanical differences such as reduced ankle push-off, exaggerated hip rotation in low-BBS participants. The framework enhances both classification accuracy and clinical interpretability, enabling quantitative assessment of motor deficits and personalized rehabilitation planning for stroke survivors.

Keywords: Stroke, Gait analysis, Multivariate Functional PCA (MFPCA), Square-Root Velocity Framework (SRVF), Functional data analysis, Machine learning

[a]Department of Data Science and Applied Statistics, Chung-Ang University, 84, Heukseok-ro, Dongjak-gu, Seoul (06974), Korea. ipad39@cau.ac.kr
[b]Translational Research Centre on Rehabilitation Robots, National Rehabilitation Centre, Ministry of Health & Welfare, 58, Samgaksan-ro, Gangbuk-gu, Seoul (01022), Korea. wldms880226@hanmail.net
[c]Translational Research Centre on Rehabilitation Robots, National Rehabilitation Centre, Ministry of Health & Welfare, 58, Samgaksan-ro, Gangbuk-gu, Seoul (01022), Korea. hogenekim@gmail.com
[d]Department of Applied Statistics, Chung-Ang University, 84, Heukseok-ro, Dongjak-gu, Seoul (06974), Korea. jooylee@cau.ac.kr

Longitudinal Boosting with Mixed Effects[a]

Minjeong Kim[b*] · Jaejik Kim[c]

Summary: Longitudinal data, which measures the same subjects repeatedly over time, is commonly used in medical and social sciences. Machine learning models for correlated data, such as mixed effects random forest and mixed effects gradient boosting, have been proposed as the demand for prediction models grows. However, existing mixed effect-based tree models assume residuals are independent, implying that random effects account for all within-subject correlations. This assumption may degrade predictive performance in the presence of within-subject serial correlation or heteroscedasticity, which are inherent characteristics of longitudinal data. To address this limitation, we propose a longitudinal boosting model with mixed effects that incorporates subject-specific covariance matrices into the gradient boosting algorithm. The proposed model enables flexible prediction by explicitly modeling the error covariance matrices through subject-specific and time covariates. The performance of the model is verified through extensive simulations by applying various missing data mechanisms and random effect sizes in both linear and nonlinear structures. We further demonstrate the practical applicability of the proposed model via performance evaluation on real-world datasets.

Keywords: Longitudinal data, regression tree, gradient boosting, mixed effect

[a]This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Korea government (No. RS-2024-00463500, RS-2024-00407300, RS-2025-16067563).
[b]Graduate Student, Department of Statistics, Sungkyunkwan University
[c]Professor, Department of Statistics, Sungkyunkwan University. jaejik@skku.edu

Adaptive Spike-and-Slab Priors for Bayesian Transfer Learning in Linear Regression

Minju Km[a*] · Kyoungjae Lee[a]

Summary: This study proposes an improved prior specification for Bayesian transfer learning with the conditional spike-and-slab prior, aimed at improving inference accuracy and computational efficiency. conditional spike-and-slab prior estimates target regression coefficients while determining similarities at the covariate-level across multiple source datasets. For each covariate, a latent informative indicator decides whether to transfer information, and source coefficients follow a conditional spike-and-slab prior centered at the target coefficient. In the original prior, sparsity and shrinkage are treated independently, which can lead to excessive information loss in dense settings. To address this, we adopt the Adaptive Correlated Spike-and-Slab (ACSS) prior to transfer learning, which adaptively links sparsity and shrinkage. It applies stronger shrinkage in dense models and weaker shrinkage when only a few covariates are relevant. By integrating the ACSS prior into conditional spike-and-slab prior covariate selection during transfer becomes more effective. Gibbs sampling is used for inference, jointly updating indicators, coefficients, and prior parameters. The model adapts to various transfer structures without extensive hyperparameter tuning, leading to improved prediction accuracy and computational efficiency, and simulation studies confirm its effectiveness in both dense and sparse settings.

Keywords: Bayesian inference, Transfer learning, High-dimensional data, Spike-and-slab prior

[a]Department of Statistics, Sungkyunkwan University, 25-2 Sungkyunkwan-ro Jongno-gu, Seoul 03063, Korea

A DTW-Based Kernel Extension of D-SVM Charts for Robust Process Monitoring

Sangwoo Kim[a*] · Donggeun Kim[b] · Soohyun Ahn[c]

Summary: The distance-based support vector machine (D-SVM) chart is a nonparametric monitoring method designed to detect process shifts by quantifying the SVM-derived distance (score) between in-control reference data and real-time observations. It has demonstrated strong performance for high-dimensional and non-normal data. However, the conventional D-SVM chart employs an RBF kernel based on Euclidean distance, which limits its effectiveness when time-series data exhibit temporal misalignment. To overcome this limitation, this study proposes a DTW-based D-SVM chart, in which the Dynamic Time Warping (DTW) distance is incorporated into the RBF kernel. The proposed chart effectively captures temporal distortions among time-series samples, maintaining the desired ARL_0 under in-control conditions and achieving a low ARL_1 when the process shifts out of control, as demonstrated in simulation results. A simulation study using the ECG5000 dataset was conducted to evaluate the proposed approach, and the results confirm that the DTW-based D-SVM chart remains robust under temporal misalignment conditions.

Keywords: Anomaly Detection, Dynamic Time Warping, Process Monitoring, Average Run Length

[a]Undergraduate Student, Department of Mathematics, Ajou University
[b]B.S. Graduate, Department of Mathematics, Ajou University
[c]Professor, Department of Mathematics, Ajou University

A Bayesian Nonparametric Method for Confounder Selection with Mixed Covariates[a]

Seokho Kim[b*] · Chanmin Kim[b]

Summary: Accurate estimation of causal effects requires appropriate adjustment for confounding variables. Bayesian tree-based models such as Bayesian Additive Regression Trees (BART) and its Dirichlet extension have shown promise for flexible nonparametric modeling and variable selection. However, these models assume continuous covariates and exhibit a bias against selecting binary or categorical covariates with fewer unique values, as such variables are less frequently chosen for tree splits. This leads to suboptimal confounder control and biased causal estimates. To overcome this issue, we propose a Weighted BART model for confounder selection. The proposed method employs a two-stage procedure: (1) potential confounders are first identified through multivariate BART using both treatment and outcome variables as responses, and (2) the resulting variable inclusion probabilities are integrated into a weighted Dirichlet prior used as a common prior between the exposure and outcome models. By incorporating these adaptive weights, the proposed model increases the probability of selecting binary and categorical variables as confounders while mitigating the bias due to the inclusion of many non-confounders. The proposed method provides a Bayesian nonparametric approach for high-dimensional causal inference and confounder selection, especially with mixed-type covariate environments common in public health.

Keywords: Bayesian Additive Regression Trees, Causal Inference, Mixed-type covariates

[a]This work is supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (RS-2025-00554477, RS-2024-00407300).

[b]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-1, Seonggyungwan-ro, Jongno-gu, Seoul, Republic of Korea. chanmin.kim@skku.edu

Raw-Data Driven Functional Data Analysis with Multi Adaptive Functional Neural Networks for Ergonomic Risk Classification Using Facial and Bio-signals Time-series Data

Suyeon Kim[a*] · Afrooz Shaker[b] · Seyed Shayan Darabi[b] · Eunsik Kim[c] · Kyongwon Kim[d]

Summary: Ergonomic risk classification during manual lifting tasks is crucial for the prevention of workplace injuries. This study addresses the challenge of classifying lifting task risk levels (low, medium, and high risk, labeled as 0, 1, and 2) using multi-modal time-series data comprising raw facial landmarks and bio-signals (electrocardiography [ECG] and electrodermal activity [EDA]). Classifying such data presents inherent challenges due to multi-source information, temporal dynamics, and class imbalance. To overcome these challenges, this paper proposes a Multi Adaptive Functional Neural Network (Multi-AdaFNN), a novel method that integrates functional data analysis with deep learning techniques. The proposed model introduces a novel adaptive basis layer composed of micro-networks tailored to each individual time-series feature, enabling end-to-end learning of discriminative temporal patterns directly from raw data. The Multi-AdaFNN approach was evaluated across five distinct dataset configurations: (1) facial landmarks only, (2) bio-signals only, (3) full fusion of all available features, (4) a reduced-dimensionality set of 12 selected facial landmark trajectories, and (5) the same reduced set combined with bio-signals. Performance was rigorously assessed using 100 independent stratified splits (70% training and 30% testing) and optimized via a weighted cross-entropy loss function to manage class imbalance effectively. Results demonstrated that the integrated approach, fusing facial landmarks and bio-signals, achieved the highest classification accuracy and robustness. Furthermore, the adaptive basis functions revealed specific phases within lifting tasks critical for risk prediction. These findings underscore the efficacy and transparency of the Multi-AdaFNN framework for multi-modal ergonomic risk assessment, highlighting its potential for real-time monitoring and proactive injury prevention in industrial environments.

Keywords: Functional Data Analysis, Neural Network, Classification

[a]Department of Statistics, Ewha Womans University
[b]University of windsor
[c]Department of Mechanical, Automotive, and Materials Engineering, University of Windsor
[d]Department of Applied Statistics, Department of Statistics and Data Science, Yonsei University

A Design-Based Matching Framework for Staggered Adoption with Time-Varying Confounding

Suehyun Kim[a*] · Dahai Jung[b] · Kwonsang Lee[c]

Summary: Causal inference in longitudinal datasets has long been challenging due to dynamic treatment adoption and confounding by time-varying covariates. Prior
work either fails to account for heterogeneity across treatment adoption cohorts and treatment timings or relies on modeling assumptions. In this work, we develop a novel design-based framework for inference on group- and time-specific treatment effects in panel data with staggered treatment adoption. We establish identification results for causal effects under this structure and introduce corresponding estimators, together with a block bootstrap procedure for estimating the covariance matrix and testing the homogeneity of group-time treatment effects. To implement the framework in practice, we propose the Reverse-Time Nested Matching algorithm, which constructs matched strata by pairing units from different adoption cohorts in a way that ensures comparability of covariate histories at each treatment time. Applying the algorithm to the Netflix-IPTV dataset, we find that while Netflix subscription does not significantly affect total IPTV viewing time, it does negatively affect VoD usage. We also provide statistical evidence that the causal effects of Netflix subscription may vary even within the same treatment cohort or across the same outcome and event times.

Keywords: Causal inference; Design-based inference; Nested structure; Matching algorithm; Panel data

[a]Department of Statistics, Seoul National University
[b]Department of Statistics, Sungkyunkwan University
[c]Corresponding Author: Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, 08826, Republic of Korea. kwonsanglee@snu.ac.kr

A Novel Closed-form Asymptotically Efficient Estimator for the Gumbel Distribution[a]

Seung-Hwan Kim[b*] · Hyoung-Moon Kim[c]

Summary: In extreme value theory, the Gumbel distribution plays a central role due to its analytical tractability and broad applicability. Conventional estimators, such as the method of moments estimator and the maximum likelihood estimator, have notable drawbacks: method of moments estimator often suffers from low accuracy, while maximum likelihood estimator, although asymptotically efficient, is computationally intensive. To address these issues, we propose a new closed-form asymptotically efficient estimator that achieves high accuracy comparable to maximum likelihood estimator but with substantially reduced computational cost. Through rigorous theoretical development, we establish the asymptotic efficiency of a new closed-form asymptotically efficient estimator. Simulation studies under multiple scenarios demonstrate that a new closed-form asymptotically efficient estimator achieves performance nearly indistinguishable from maximum likelihood estimator, while requiring significantly less computation time, particularly for large samples. Furthermore, the empirical application to daily COVID-19 mortality to case ratio data highlights the suitability of the Gumbel distribution for modeling health-related extreme data. These findings suggest that a new closed-form asymptotically efficient estimator not only provides a computationally efficient and statistically reliable alternative to existing methods but also expands the practical relevance of the Gumbel distribution to both natural and health data domains.

Keywords: Asymptotically efficient estimator, Extreme value theory, Gumbel distribution, Extreme value health data

[a]This work was in part supported by the National Research Foundation of Korea (NRF), grant funded by the Korean government (MSIT)(RS-2024-00357199).
[b]Seoul, Korea. Applied Statistics, Konkuk University
[c]Corresponding author: Seoul, Korea, Applied Statistics, Konkuk University. hmk966a@gmail.com

Enhancing Gene Set Enrichment Analysis Using Mirror Statistics for Highly Correlated Genomic Data[a]

Yuna Kim[b*] · Hokeun Sun[c]

Summary: In the analysis of high-dimensional genomic data, analyzing single-genes has limitations due to the multiple testing problem and restricted biological interpretation. To address this, Gene Set Enrichment Analysis (GSEA), which focuses on common functions or pathways, is widely used. Fast GSEA (FGSEA) is particularly advantageous as it can rapidly and accurately estimate even extremely small p-values. However, FGSEA inherently relies on moderated t-statistics, which fails to account for gene correlations. This can lead to an overestimation of significance and an increased number of false positives when genes within the same pathway are highly correlated. To overcome this limitation, this study proposes a novel approach based on Multiple Data Splitting (MDS) using mirror statistics. Mirror statistics were calculated by applying various regularization methods, including Ridge, Lasso, Elastic Net, and Debiased Lasso, and then integrated with FGSEA. Our simulation studies and breast cancer data analysis results showed that the proposed approach generally improved the power compared to conventional statistics. In particular, the Ridge based mirror statistic effectively controlled the False Discovery Rate (FDR) under strong correlation conditions. Therefore, this study highlights the critical importance of selecting appropriate statistics in GSEA and presents a new strategy for controlling false positives and enhancing the reliability of biological interpretation in high-dimensional genomic data analysis.

Keywords: Gene set enrichment analysis, Mirror statistics, FDR control, High-dimensional data

[a]This work was supported by the National Research Foundations of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-16065086).
[b]Master's Student, Department of Statistics, Pusan National University, Busan 46241, Korea. yunaa1026@pusan.ac.kr
[c]Professor, Department of Statistics, Pusan National University, Busan 46241, Korea. hsun@pusan.ac.kr

Variational Autoencoder-based Out-of-Distribution Generation with Latent Mixing[a]

Hojin Lee[b] · Jinwoo Jung[c] · Seoyeon Hwang[d] · Yoon-Yeong Kim[e*]

Summary: Out-of-distribution (OOD) detection is a critical task for reliable deep neural networks, yet the absence of explicit OOD samples during training often limits their robustness. To address this, we propose a Variational Autoencoder (VAE)-based virtual outlier generation framework that synthesizes pseudo-OOD samples from in-distribuiton (ID) data in a structured and controllable manner. Our VAE model employs two encoders and a shared decoder; each encoder extracts ID-relevant and ID-irrelevant features, while a single decoder reconstructs two distinct outputs of the original ID image and a noise-like variant from each features respectively. The VAE is trained with regularization losses for each feature along with reconstruction losses. To ensure a proper disentanglement between two features, we introduce a total correlation–based disentanglement loss, which is implemented through an adversarial discriminator that minimizes the dependency between the two latent spaces. Finally, the two features are mixed to reconstruct the original image as a safeguard to prevent latent collapse. Moreover, we design a novel scheduling strategy for the mixing coefficient, allowing the model to generate diverse virtual OOD samples covering from far-OOD to near-OOD. Experimental results demonstrate that our approach outperforms existing baselines, while qualitative analyses with t-SNE visualizations confirm the effective disentanglement of latent space.

Keywords: Out-of-Distribution Detection, Virtual Outlier Generation, Variational Autoencoder, Feature Disentanglement

[a]This work was supported by the National Research Foundation of Korea (NRF) gran funded by the Korea government (MSIT) (202503042002).

[b]Department of Statistics, University of Seoul, 163 Seoulsiripdae-ro, Dongdaemun-gu, Seoul 02504, Korea. lhj23@uos.ac.kr

[c]Department of Statistics, University of Seoul, 163 Seoulsiripdae-ro, Dongdaemun-gu, Seoul 02504, Korea. wjdwlsuuu@uos.ac.kr

[d]Department of Economics and Data Science, University of Seoul, 163, Seoulsiripdae-ro, Dongdaemun-gu, Seoul 02504, Korea. seoyeon0518@uos.ac.kr

[e]Department of Statistics, University of Seoul, 163 Seoulsiripdae-ro, Dongdaemun-gu, Seoul 02504, Korea. yykim@uos.ac.kr

Modified Log-rank Test Statistic in the Presence of Dependent Censoring

Unchong Kim[a*] · Chung Mo Nam[b]

Summary: The log-rank test, the standard two-sample procedure for survival data, assumes independent right-censoring. In clinical studies, however, loss to follow-up often correlates with patients' event risk, violating this assumption and inflating Type I error while reducing power. We propose a copula-based modification of the log-rank statistic that directly models dependence between event and censoring times. For each censored subject, conditional on the observed censoring time, we estimate the probability of event occurrence in each subsequent interval and redistribute censored mass accordingly, yielding revised pseudo event counts and risk sets. Simulations spanning Kendall's τ, copula families (Clayton, Gumbel, Frank), and between-group censoring imbalance—and allowing misspecification of the copula family, τ, and margins—show that the proposed test maintains nominal size and markedly reduces inflation relative to the standard test, especially under unequal censoring and strong dependence. Performance is robust to moderate τ misspecification (≈±0.2) and to margin misspecification (exponentially generated data analyzed under Weibull margins). Under the null, Kaplan–Meier tends to overestimate and Copula–Graphic to underestimate the survival curve, whereas our estimator closely matches the truth. By correcting expected event counts and risk sets under dependent censoring without covariates, the method improves size control and interpretability and offers a practical alternative when event–censoring dependence and censoring imbalance are anticipated.

Keywords: Log-rank test, Dependent censoring, Copula, Type I error

[a]Doctoral degree, Department of Biostatistics and Computing, Yonsei University Graduate School. silvergun@yuhs.ac
[b]Professor, Division of Biostatistics, Department of Preventive Medicine, Yonsei University College of Medicine. CMNAM@yuhs.ac

Interpreting the Dynamic Effects of Wearable-Derived Physiological Signals on Glycemic Variability in Pre-diabetes Using Distributed Lag Non-linear Mixed Models

Ji Eun Kim[a*] · Jae Keun Yoo[b]

Summary: Prediabetes management is constrained by the invasiveness and cost of Continuous Glucose Monitoring (CGM). While research on glucose prediction using wearable devices has been active, there is a lack of studies clarifying the complex relationships between physiological signals and blood glucose. This study aims to interpret the dynamic, non-linear, and lagged effects of wearable-derived physiological signals and dietary information on the glucose response in prediabetic patients. Data collected from 12 prediabetic patients over 8-10 days—including CGM, wearable signals (EDA, IBI, HRV, etc.), and dietary logs—were analyzed using a Distributed Lag Non-linear Mixed Model (DLNMM). A cross-basis function was used to estimate the combined effects according to predictor intensity and time lag. The analysis revealed three key findings: First, a high degree of inter-individual heterogeneity in glucose response was confirmed (ICC=0.93), with this variability being particularly pronounced in Autonomic Nervous System (ANS) related variables. Second, ANS and stress indicators (HRV, EDA) were found to have a greater cumulative impact on glycemic variability than dietary factors. Third, a clear physiological feedback loop was discovered, where a drop in glucose preceded the EDA stress response, which was subsequently followed by a rise in glucose. This study demonstrates, using DLNM, that autonomic nervous system responses are more crucial than dietary factors in glucose management and that their impact is highly individualized. This highlights the necessity of personalized intervention strategies, including stress management.

Keywords: Distributed Lag Non-linear Model, Mixed-Effects Model, Glycemic variability, Wearable Devices, etc

[a]Seoul 03760, Republic of Korea, Department of Statistics, Ewha Womans University. happyjieun@ewha.ac.kr
[b]Corresponding Author: Professor, Seoul 03760, Republic of Korea, Department of Statistics, Ewha Womans University. peter.yoo@ewha.ac.kr.

Mixed-Effect Neural Processes for Multi-Annotator Semi-Supervised Medical Image Segmentation[a]

Chanyeong Kim[b*] · Heejin Kim[c] · Weonyoung Joo[d]

Summary: Medical image segmentation in real-world scenarios differs from typical image segmentation tasks in several key aspects. First, gathering the definitive ground-truth labels is challenging; instead, diverse ground-truth-like labels from multiple experts are assigned to a single image, depending on varying opinions. Moreover, due to high labeling costs and patient privacy concerns, labeled data are often scarce, making fully supervised settings infeasible. To address these issues, this work proposes a novel mixed-effect neural processes framework, focusing on the semi-supervised medical image segmentation problem in a multi-annotator setting. Specifically, since uncertainty in the multi-annotator semi-supervised setting arises from both image ambiguity and individual annotator variability, we employ two neural process modules to model each type of uncertainty as a mixed-effect model. Experiments on benchmark medical image datasets, such as RIGA and QUBIQ, empirically demonstrate the superiority of the proposed method.

Keywords: Neural Processes, Mixed Effect, Multi-Annotator Semi-Supervised Learning, Medical Image Segmentation

[a]This work was supported by (1) the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00217022); (2) the Basic Science Research Program (Priority Research Institute) through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1 A6A1A10039823); and (3) the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00334477).

[b]Department of Statistics, EWHA Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. chanyeong@ewha.ac.kr

[c]Department of Statistics, EWHA Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. kim.heejin@ewha.ac.kr

[d]Corresponding Author: 5Department of Statistics, EWHA Womans University, 2, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. weonyoungjoo@ewha.ac.kr

exMRP: An R package for Explainable Multilevel Regression and Post-stratification Models based on Machine Learning with SHAP[a]

Chae-Yun Kim[b*] · Eun-Kyung Lee[b]

Summary: This study introduces exMRP, an R package that enhances explainability and performance in Multilevel Regression and Post-stratification (MRP) models. MRP is a standard approach to estimate subnational public opinion. However, traditional MRP models face several challenges: difficulties in choosing the appropriate variables, limited predictive performance, and poor interpretability. To address these limitations, exMRP incorporated various machine learning techniques—such as Best Subset Selection, Gradient Boosting Machines (GBM), Principal Component Analysis (PCA), Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machines (SVM), Random Forests, and Ensemble Bayesian Model Averaging (EBMA)—into the MRP framework. Furthermore, by introducing SHapley Additive exPlanations (SHAP) values, we improved model explainability, allowing comparisons of the influence of predictive variables across different hierarchical levels. Ultimately, we developed the R package that implements this entire pipeline, offering data preprocessing, model training with hyperparameter tuning, SHAP value calculation, and post-stratification. This package provides a practical tool that makes it easy to apply subnational predictive models with high accuracy and explainability.

Keywords: R Package, Multilevel Regression and Post-stratification, Shapley value, Machine Learning

[a]This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2023-00272083).

[b]Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemum-gu, Seoul, 03760, South Korea

Fast and Asymptotically Efficient Estimation in the Weighted Exponential Family[a]

Hyeonwoo Kim[b*] · JungJae Choi[b] · Hyoung-Moon Kim[c]

Summary: In this paper, we propose a new efficient estimator for the Weighted Exponential Family and its underlying components. These components form a flexible class of distributions within the standard Exponential Family, characterized by a positive support and a generator function. For these models, maximum likelihood estimators are often not available in closed-form and must be obtained through numerical optimization. We therefore develop asymptotically efficient estimator for these distributions by applying the classical theorem of Lehmann and Casella. Monte Carlo simulations demonstrate that the proposed estimator performs nearly identically to numerically computed maximum likelihood estimator, while generally outperforming previously proposed closed-form estimators.

Keywords: weighted exponential Family, asymptotic efficiency, closed-form estimator

[a]This work was in part supported by the National Research Foundation of Korea (NRF), grant funded by the Korean government (MSIT)(RS-2024-00357199).

[b]Seoul, Korea. Applied Statistics, Konkuk University

[c]Corresponding author: Seoul, Korea, Applied Statistics, Konkuk University. hmk966a@gmail.com

CCTV 영상을 이용한 교통 사고 탐지 시스템[a]

김현주[b*] · 곽일엽[c]

요약: 교통의 편의성이 증대함과 동시에 교통사고 발생 빈도도 증가하면서 이에 대한 대응과 탐지의 중요성이 부각되고 있습니다. 본 연구에서는 정형 및 비정형 영상 데이터에서도 성능을 향상시키는 데 초점을 맞춰 사전 학습된 영상 시각 트랜스포머 모델을 이용한 교통사고 분류 방법을 제안합니다. 기존의 교통 사고 분류 모델들은 이미지, 혹은 영상 중 일부 프레임만 사용하는 데에 반해, 본 모델은 영상 자체를 이용하므로 분류에 있어 시공간적인 특징을 반영하고 있습니다. 제안된 모델은 이미지 처리의 최첨단 아키텍처인 비전 트랜스포머(ViT)를 기반으로 하며, 동영상 이해에 순수 트랜스포머 기반 모델을 성공적으로 적용한 첫 번째 사례 중 하나입니다. 본 논문에서는 다양한 분류를 가진 Kinetics400 데이터로 사전 학습된 모델을 사용하였습니다. 벤치마크 데이터세트인 CADP와 UTIC에서 제공하는 실제 국내 CCTV 영상들을 사용하여 실험한 결과, 우리의 접근 방식이 우수한 분류 성능을 달성한다는 것을 보였습니다. 또한 UCF-Crime Dataset을 통한 일반화 성능 평가 역시 좋은 성능을 보였습니다. 본 연구는 교통 사고 분류 연구에 있어 시공간적인 특징을 반영함으로서 여러 환경에서 발생하는 사고들에 대한 인식률을 높이는 데에 기여합니다.

주요용어: Vision Transformer, Car Crash Detection, Deep Learning, Video Data.

[a]본 논문은 한국대학교의 지원을 받아서 연구된 것이며, 연구과제번호는 2024a12345-7입니다.
[b](06974) 서울특별시 동작구 흑석로 84, 중앙대학교 일반대학원. khyeonju19@gmail.com
[c](06974) 서울특별시 동작구 흑석로 84, 중앙대학교 일반대학원. ilyoup.kwak@gmail.com

시계열 블랙박스 모형 기반 SCFI 예측의 해석 및 시각화[a]

김현태[b*] · 김재직[c]

요약: 상하이 컨테이너 운임지수(Shanghai Containerized Freight Index; SCFI)는 해운 물류 운임을 결정하는 기준으로 사용되는 해운업계에서 매우 중요한 지표로 자리하고 있다. 코로나19 팬데믹 이후 SCFI의 변동성이 급격히 증가하면서 해운 산업의 불확실성 관리를 위한 정교한 예측 모형의 필요성이 대두되었고, 이에 최근 시계열 딥러닝과 같은 블랙박스 모형들이 높은 예측 정확도를 제공하게 되었다. 그러나, 블랙박스 모형의 전형적인 특성인 해석 가능성의 한계는 실무 현장에서의 활용을 제약하는 핵심적인 문제로 지적되고 있다. 이에 본 연구에서는 시차를 고려하는 모든 시계열 블랙박스 모형을 해석하기 위한 설명 가능한 인공지능(eXplainable AI; XAI) 방법론을 적용하여, SCFI 블랙박스 모형의 투명성과 신뢰성을 향상하는 것을 목표로 한다. 이러한 SCFI의 시계열 블랙박스 모형은 운임 계약이 최소 3개월 이전에 이루지는 해운업계의 특성에 따라 보통 미래 13주까지의 SCFI를 예측하는 것을 목표로 한다. 따라서 모형의 예측값이 13차원의 벡터이므로, 본 연구에서는 이 예측 결과의 해석을 위해 벡터 SHAP을 도입하고 이를 효율적으로 시각화하는데 집중한다. 본 연구는 SCFI 블랙박스 예측 모형에 대한 해석적 프레임워크를 제시함으로써, 해운 산업의 데이터 기반 의사결정을 지원하고 AI 기술의 실무 적용성을 높이는 데 기여할 것으로 기대된다.

주요용어: 상하이 컨테이너 운임지수, 시계열 블랙박스 모델, 변수 선택, 벡터 SHAP

[a]이 논문은 정부(과학기술정보통신부) 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. RS-2024-00407300, RS-2025-16067563).
[b](03063) 서울시 종로구 성균관로 25-2, 성균관대학교 통계학과, 석사과정
[c](03063) 서울시 종로구 성균관로 25-2, 성균관대학교 통계학과, 교수. jaejik@skku.edu.

A Functional Principal Component-based Approach to Group Comparisons: Developing the Functional Measurement Invariance (FMI) Framework[a]

Hyeri Kim[b*] · Youngseuk Cho[c]

Summary: In the growing field of integrated Functional Data Analysis (FDA) and Structural Equation Modeling (SEM), the absence of a framework for testing measurement invariance has remained a key methodological challenge. This study addresses this gap by proposing and validating a 4-step Functional Measurement Invariance (FMI) framework based on Functional Principal Component Analysis (fPCA) and permutation tests. By sequentially evaluating configural, metric, and scalar invariance prior to latent mean comparison, the framework allows researchers to clearly distinguish true group differences from measurement bias. A Monte Carlo simulation confirmed the framework's high accuracy in discriminating between these two scenarios. This research provides a robust methodological foundation for ensuring statistical validity in group comparisons across diverse functional data, such as time-series and repeated measures, and contributes an R package that provides a solid analytical foundation for future FSEM studies.

Keywords: Functional Data Analysis, Functional Measurement Invariance, Functional Principal Component Analysis, Group Comparison

[a]This research was supported by research funds from the 2025 Pusan National University University Innovation Support Project.

[b]Graduate student, Department of Statistics, Pusan National University, Busan 46241, Korea. gpfl3616@naver.com

[c]Professor, Department of Statistics, Pusan National University, Busan 46241, Korea. choys@pusan.ac.kr

Mirror Statistics-guided Variable Selection for Enhanced Clustering in High-dimensional Data

Yurim No[a*] · Hoyoung Park[b]

Summary: We propose a novel framework for unsupervised learning that integrates mirror statistics_based variable selection with downstream clustering. The mirror statistic is designed to quantify the structural relevance of each variable by comparing perturbed clustering solutions, enabling effective separation of informative (signal) variables from irrelevant (noise) ones in high-dimensional settings. We formally define the mirror statistic and demonstrate its utility in identifying features that are stable and influential across clustering perturbations. Variables with high mirror scores are retained, and clustering is performed on the reduced feature space. Through extensive simulations and real data applications, we show that this approach improves cluster recovery, reduces noise-induces distortions, and enhances interpretability. The proposed method is particularly effective when true clusters ar e driven by a sparse subset of covariates.

Keywords: Clustering stability, High-dimensional data, Mirror statistics, Perturbation-based analysis, Signal-noise separation, Unsupervised learning, Variable selection

[a]Department of Statistics, Sookmyung Women's University. noyurim5126@sookmyung.ac.kr
[b]Department of Statistics, Sookmyung Women's University. hyparks@sookmyung.ac.kr

Estimating Structural Shifts in Graph Domain Adaptation via Pairwise Likelihood Maximization[a]

Huiyun Noh[b*] · Wooseok Ha[c]

Summary: Graph domain adaptation (GDA) emerges as an important problem in graph machine learning when the distribution of the source graph data used for training is different from that of the target graph data used for testing. While much of the prior work on GDA has focused on the idea of aligning node representations across source and target domains, recent studies show that such approaches can be suboptimal in the presence of conditional structure shift (CSS), where the distribution of graph edges conditioned on labels changes across domains. In this work, we develop a unified framework to solve CSS and show that existing GDA methods for CSS arise as special cases of our framework. This framework further allows us to develop a new method, Pairwise-Likelihood maximization for graph Structure alignment (PLSA), which uses rich information from pairwise nodes and edges to improve the estimation of target connection probabilities. We establish conditions under which our method is identifiable and introduce a simple edge reweighting scheme based on importance weights to align the source and target graphs. Theoretically, under the contextual stochastic block model (CSBM), we derive finite-sample guarantees using recent results in matrix concentration inequalities for U-statistics. We complement our theoretical results with empirical studies that demonstrate the effectiveness of our method.

Keywords: Graph domain adaptation, graph structure shift, distribution matching, U-statistics

[a]This work was partly supported by the National Research Foundation (NRF) of Korea grant funded by the Korea government (MSIT) (RS-2025-24523569).

[b]Integrated M.S.&PhD, Program, Department of Mathematical Sciences, 291, Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea. huiyun.noh@kaist.ac.kr

[c]Assistant Professor, Department of Mathematical Sciences, 291, Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea. haywse@kaist.ac.kr

Projective Resampling with Outer Product Gradient for Multivariate Response Dimension Reduction

Hoyeon Ryu[a*] · Kyongwon Kim[b]

Summary: In high-dimensional data analysis, Sufficient Dimension Reduction (SDR) is a powerful statistical method for efficiently extracting key information from complex data structures. However, when involving multivariate response variables, existing SDR techniques struggle to accurately estimate the central subspace due to the curse of dimensionality and complex dependency structures. To address this, various methods combining Projective Resampling (PR) with existing SDR techniques (SIR, SAVE, DR, etc.) have been proposed. However, these inverse regression-based approaches have limitations, requiring strong model assumptions such as Linear Conditional Mean (LCM) and Constant Conditional Variance (CCV). To address these limitations, this study proposes the PR-OPG (Projective Resampling–Outer Product of Gradient) method. OPG (Outer Product of Gradient) is a forward regression-based nonparametric dimension reduction method that estimates the gradient of the conditional mean function of the response variable in the predictor space via local linear regression, then reconstructs the central mean subspace by averaging the outer products of these gradients. This method guarantees effective dimension reduction without assumptions like LCM or CCV, provided the support of the predictor variables is convex and the conditional expectation function changes sufficiently smoothly (i.e., is differentiable). Therefore, OPG is applicable under weaker assumptions and performs stable, robust estimation based on mean structures even for high-dimensional data with nonlinear structures or heteroskedasticity. This study theoretically establishes the consistency and convergence rate of PR-OPG. Through simulations and real data analysis, it demonstrates that PR-OPG achieves more accurate estimation of the true structural dimension, higher explanatory power, and lower RMSE compared to other multivariate response SDR methods such as PR-SIR, PR-SAVE, and PR-DR.

Keywords: Sufficient Dimension Reduction (SDR), Projective Resampling (PR), Outer Product of Gradient (OPG), Multivariate Response, High Dimensional Data Analysis

[a]Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemum-gu, Seoul, 03760, South Korea. rhy214@ewhain.net

[b]Department of Statistics and Data Science, Department of Applied Statistics, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, 03722, South Korea

머신러닝을 이용한 아파트 미분양률 예측: 경기도를 중심으로

명진소[a*] · 유재근[b]

요약: 본 연구의 목적은 머신러닝 기법을 활용하여 경기도 미분양 아파트의 미분양률을 예측하고, 이에 영향을 미치는 요인을 살펴보는 것이다. 주택 미분양 사태는 우리나라에서 반복적으로 발생해 왔으며, 이에 대한 다양한 연구가 지속적으로 이루어져 왔다. 기존 연구들은 분양 아파트의 초기계약률, 계약성패여부, 미분양주택수, 미분양여부 등을 종속 변수로 설정하고, 다중회귀모형, 로지스틱회귀모형, 시계열모형으로 이를 분석해왔다. 그러나, 이러한 전통적인 통계모형은 분석 결과의 타당성을 확보하기 위해 여러 가정을 충족해야 하며, 복잡한 데이터 구조나 변수 간의 비선형 관계를 효과적으로 파악할 수 없다는 한계가 있다. 이에 본 연구에서는 트리 기반 머신러닝 알고리즘인 Random Forest, XGBoost, LightGBM을 사용하여 미분양률을 예측하고자 한다. 분석에는 경기도에서 제공하는 2015년부터 2024년까지의 미분양 아파트 현황 데이터를 사용하였다. 성능지표로는 RMSE와 R^2를 이용하였으며 그리드 서치를 통해 모델의 최적 하이퍼파라미터를 탐색하였다. 모델 튜닝 이후 각 알고리즘의 변수 중요도를 비교하고, Partial Dependence Plot(PDP)을 활용하여 주요 독립변수들과 미분양률 간의 관계를 해석하였다.

주요용어: 미분양률, 미분양 아파트, 트리기반 머신러닝

[a](03760) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과, 석사과정. somyung@ewhain.net
[b](03760) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과, 교수. peter.yoo@ewha.ac.kr

사전학습 언어모델 기반 낚시성 기사 탐지 및 대안 제목 생성에 관한 비교 연구[a]

문수연[b*] · 한상태[c]

요약: 낚시성 기사는 기사 본문과 달리 기사 제목을 과장하거나 자극적으로 구성하여 독자들의 호기심을 유발한다. 하지만 이는 독자들이 언론을 신뢰하지 않는 결과를 초래하므로 본 연구는 낚시성 기사를 탐지하고, 신뢰할수 있는 제목을 생성하는 모델을 구축하고자 한다. 낚시성 기사 탐지를 위해 사전 학습된 BERT(Bidirectional Encoder Representations from Transformers)와 Sentence-BERT를 기반으로 기사 제목과 본문의 문맥적 연관성 분석하여 이를 기반으로 분류 모델을 파인튜닝 하였다. 또한, 사전학습된 T5(Text-to-Test Transfer Transformer)와 BART (Bidirectional and Auto-Regressive Transformers)를 기반으로 기사의 본문에서 핵심 내용을 한 문장 단위로 요약하도록 파인튜닝 하였고, 이를 정제된 기사 제목으로 간주하였다. 이후 모델들의 성능을 비교 분석하여 각 과제별 최적의 모델을 채택하였으며, 이를 통해 낚시성 기사로 인한 독자의 혼란을 줄이고 뉴스 소비 과정의 신뢰성과 효율성을 높이는데 기여하고자 한다.

주요용어: 낚시성 기사 탐지, 제목 생성, 사전학습 모델, 뉴스 신뢰성

[a]이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2022-NR068754).

[b](31499) 충청남도 아산시 배방읍 호서로 79번길 20, 호서대학교 데이터사이언스학과, 석사과정.
suyeon_24@naver.com

[c]교신저자: (31499) 충청남도 아산시 배방읍 호서로 79번길 20, 호서대학교 빅데이터AI학부, 교수.
sthan@hoseo.edu

An Asymptotically Efficient Closed-form Estimator for the Rician Distribution[a]

Keunwoo Park[b*] · Hyoung-Moon Kim[c]

Summary: We study fast and numerically stable estimation of the Rician parameters in the signal–dominant (mid–to–high SNR) regime that arises in many practical links. We propose a two–step, closed–form procedure: (i) a polynomial moment map from low–order sample moments produces a $\sqrt{n}$ –consistent initializer, and (ii) a single information–Newton (one–step Newton–Raphson) update at that point attains the same asymptotic covariance as the maximum–likelihood estimator (MLE) under standard regularity conditions. Because the method requires no iterative optimization (unlike the MLE), the runtime scales linearly with the sample size. In simulations targeting mid–to–high SNR(e.g. -6dB to 12dB), absolute bias and root–mean–squared error are essentially indistinguishable from iterative MLE, yet runtime is substantially smaller. Results on 5.9,GHz vehicular channel data show a goodness–of–fit that is practically identical to MLE. While low–SNR boundary cases are reported for completeness, the method is designed as a fast, MLE–accurate default for the mid–to–high SNR regime relevant to large–scale and real–time deployments.

Keywords: Rician distribution, closed-form estimator, Newton-Raphson, asymptotic efficiency

[a]This work was in part supported by the National Research Foundation of Korea (NRF), grant funded by the Korean government (MSIT)(RS-2024-00357199).

[b]Seoul, Korea. Applied Statistics, Konkuk University

[c]Corresponding author: Seoul, Korea. Applied Statistics, Konkuk University. hmk966a@gmail.com

Audio-Based Classification of Bee and Environmental Noise Using a Wasserstein Support Histogram Machine[a]

Seyong Park[b*] · Ilsuk Kang[c]

Summary: Honeybees are among the most important pollinators, with approximately 75% of the world's top 100 crops relying on pollinators, according to the Food and Agriculture Organization (FAO). Thus, honeybees play a vital role in maintaining agricultural ecosystems. However, recent reports of Colony Collapse Disorder (CCD) have highlighted a drastic decline in honeybee populations, emphasizing the need for continuous monitoring of their condition. In this study, we aimed to distinguish between segments containing pure bee sounds ("Bee") and those contaminated with external environmental noise ("NoBee") using audio data. The audio signals were transformed into various time–frequency representations, such as the Mel-spectrogram, MFCC, and Chromagram. To overcome the limitations of SVM classification, we employed the Support Histogram Machine (SHM) with a kernel function based on the Wasserstein distance. Furthermore, previous studies used data derived from the same audio files simultaneously in the training, validation, and testing sets, which can lead to overestimated performance. We adopted a Group Cross-Validation approach based on individual audio files. As a result, the proposed method achieved an average classification accuracy of 84.8%, demonstrating its reliability for audio-based monitoring and classification of bee sounds.

Keywords: Support Histogram Machine, Time–frequency representations, Wasserstein distance kernel, Group cross-validation

[a]This work was in part supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2024-00440787).

[b](Gaesin-dong, Chungbuk National University) 1, Chungdae-ro, Seowon-gu, Cheongju-si, Chungcheongbuk-do, 28644, Republic of Korea

[c](Gaesin-dong, Chungbuk National University) 1, Chungdae-ro, Seowon-gu, Cheongju-si, Chungcheongbuk-do, 28644, Republic of Korea. isk1218@chungbuk.ac.kr

Debiased Controllable Text Generation with Classifier-guided Rectified Flow Language Model[a]

Jiwon Park[b*] · Sooyeon Kim[c] · Weonyoung Joo[d]

Summary: Controllable text generation aims to generate texts that satisfy specific conditions. However, spurious correlations in the training data can lead to biased generation results. For instance, when generating toxic texts with the text generation models, race- or gender-related keywords often appear in the generated texts. While existing works mainly focus on improving the relevance between generated sentences and target conditions, such undesired biases can be problematic as they unintentionally reinforce harmful stereotypes. In light of this, we propose a novel controllable text generation method based on a rectified flow language model, tailored to mitigate the spurious correlations between target attributes and unwanted features. We first pre-train the latent rectified flow language model, consisting of an encoder-decoder structure. Next, we adversarially train auxiliary classifiers in the latent space to remove the correlation between spurious features (e.g., the appearance of race- or gender-specific words) and true class semantics (e.g., toxicity of the text). Finally, using the trained classifiers, we conditionally generate samples by providing the classifier guidance to the rectified flow language model. Experiments on the CivilComments-WILDS and MultiNLI datasets demonstrate that our proposed method enables debiased generation and alleviates the relevance of sensitive attributes while maintaining text quality.

Keywords: Controllable Text Generation, Non-autoregressive Text Generation, Debiased Learning

[a]This work was supported by (1) the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00217022); (2) the Basic Science Research Program (Priority Research Institute) through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1A6A1A10039823); and (3) the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00334477).

[b]Graduate student, Department of Statistics, EWHA Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. jiwonpark0107@ewha.ac.kr

[c]Graduate student, Department of Statistics, EWHA Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. kim.sooyeon@ewha.ac.kr

[d]Corresponding Author: 5Assistant professor, Department of Statistics, EWHA Womans University, 2, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. weonyoungjoo@ewha.ac.kr

A Novel Discrete Parametric Survival Model based on the Alternative Discrete Hazard Rate[a]

Hyojin Park[b]* · Hyunju Lee[c]

Summary: In this study, we propose a novel discrete parametric additive hazard model based on the concept of the alternative discrete hazard rate. Unlike conventional discrete-time or Cox models that employ a multiplicative hazard structure, the proposed approach provides direct and interpretable measures of covariate effects on the alternative discrete hazard. In addition, the relative survival derived from the model offers a clear cumulative interpretation of covariate effects over time, analogous to that of the additive hazard model in the continuous setting. The proposed model encompasses a wide range of specific parametric models by assuming various baseline discrete distributions, including the discrete Weibull distribution. For illustration, the model flexibly accommodates right-censored survival data, and we demonstrate its superior fit compared with alternative models.

Keywords: Additive discrete hazard model, Alternative discrete hazard rate, Discrete Weibull distribution, Discrete Modified Weibull distribution, Relative Survival

[a]This work was in part supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00240817).
[b]Dept. of Statistics, Hankuk University of Foreign Studies, Yongin, 17035, Rep. of Korea. 202101624@hufs.ac.kr
[c]Dept. of Statistics, Hankuk University of Foreign Studies, Yongin, 17035, Rep. of Korea. hyunjulee@hufs.ac.kr

Predicting ICU Length of Stay Using Functional Linear Regression on Continuous Physiological Signals[a]

Junyoung Bae[b*] · Ilsuk Kang[c]

Summary: Predicting the length of stay (LOS) in intensive care units (ICUs) is vital for effective hospital resource allocation and patient management. This study proposes a functional regression framework that leverages continuous physiological signals to predict ICU LOS using the MIMIC-IV database. Time-series data from the first 72 hours after ICU admission-including heart rate, blood pressure (systolic and diastolic), respiratory rate, and oxygen saturation-were averaged hourly and represented as smooth functions. Functional principal component analysis (FPCA) was applied to capture temporal variations, and the resulting component scores were used as functional predictors. These features, along with scalar covariates such as age, gender, marital status, and race, were incorporated into a functional regression model. The model effectively captured dynamic physiological patterns influencing LOS and provided interpretable insights into patient outcomes. Evaluation based on RMSE and MAE showed that the proposed approach outperformed traditional regression using aggregated data, highlighting the potential of functional data analysis for robust ICU outcome prediction.

Keywords: Functional Data Analysis, Functional Regression, ICU length of stay, MIMIC-IV Database

[a]This work was in part supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2024-00440787).

[b](Gaesin-dong, Chungbuk National University) 1, Chungdae-ro, Seowon-gu, Cheongju-si, Chungcheongbuk-do, 28644, Republic of Korea

[c](Gaesin-dong, Chungbuk National University) 1, Chungdae-ro, Seowon-gu, Cheongju-si, Chungcheongbuk-do, 28644, Republic of Korea. isk1218@chungbuk.ac.kr

Wafer Map Defect Type Classification via a Graph Neural Network with Patch Embedding and k-Nearest Neighbor Graph Construction[a]

Taehyeok Seo[b*] · Ilsuk Kang[c]

Summary: In the semiconductor manufacturing process, wafer map defect pattern classification is essential for identifying process anomalies and improving yield management. We propose a Graph Neural Network (GNN)-based approach for wafer defect type classification using the WM-811K dataset. Each wafer map is converted into a graph structure by partitioning the image into patches, where each patch is regarded as a node. Edges are then constructed using a k-nearest neighbor (K-NN) algorithm to connect spatially related patches. The proposed model consists of two key components: (1) A graph-level feature extraction module composed of stacked Graph Attention Network (GAT) blocks that learn inter-node relationships, and (2) a prediction module implemented as a fully connected neural network. Experimental results based on accuracy, precision, recall and F1-score show that our model outperforms baseline models. By representing wafer maps as graph structure, the proposed approach provides greater flexibility in capturing complex local patterns and their spatial relationships, making it effective for robust wafer defect prediction in real semiconductor fabrication.

Keywords: Semiconductor Manufacturing, Wafer Map Defect Classification, Graph Neural Network, Patch-level Node Embedding, K-Nearest Neighbor Graph

[a]This work was in part supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2024-00440787)

[b](Gaesin-dong, Chungbuk National University) 1, Chungdae-ro, Seowon-gu, Cheongju-si, Chungcheongbuk-do, 28644, Republic of Korea. taehyeok72@chungbuk.ac.kr

[c](Gaesin-dong, Chungbuk National University) 1, Chungdae-ro, Seowon-gu, Cheongju-si, Chungcheongbuk-do, 28644, Republic of Korea. isk1218@chungbuk.ac.kr

Censored Broken Adaptive Ridge Rank Regression via Induced Smoothing Approach[a]

Suyeon Seon[b*] · Dipankar Bandyopadhyay[c] · Taehwa Choi[d]

Summary: Broken adaptive ridge (BAR) penalty approximates L_0 regularization through iterative reweighting of L_2 penalties. This penalty possesses both the oracle property and the grouping effect for highly correlated covariates which are the main desirable theoretical properties in penalized regression. In this study, we propose the BAR-penalized linear rank regression via induced smoothing for the semiparametric accelerated failure time model with complex censored data. To achieve the tractable computation and the reliable inference, we adopt an induced smoothing approach to the Gehan-type rank estimator. For scalable penalization, we develop a cyclic coordinate descent algorithm that minimizes the penalized objective function and estimates the regression coefficients in a coordinate-wise manner. We further extend the proposed method to more complex survival data, such as multivariate partly interval-censored data. Under mild conditions, the proposed estimator satisfies both the oracle property and the grouping effect. Numerical studies compare our approach with several well-known penalties and demonstrate its superior selection accuracy and estimation efficiency across various scenarios. An application to real clinical data further demonstrate the practical usefulness of the proposed method.

Keywords: Broken adaptive ridge regression, induced smoothing, Accelerated failure time model, Clustered data

[a]This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (RS-2024-00340298).

[b]Master's student, School of Mathematics, Statistics and Data Science, Sungshin Women's University, 2, Bomun-ro 34da-gil, Seoul 02844, South Korea220256021@sungshin.ac.kr

[c]Professor, Department of Biostatistics, School of Public Health, Virginia Commonwealth University, Richmond, Virginia, USA. dbandyop@vcu.edu

[d]Assistant Professor, School of Mathematics, Statistics and Data Science, Sungshin Women's University, 2, Bomun-ro 34da-gil, Seoul 02844, South Korea. tchoi@sungshin.ac.kr

Estimating Population-level Influenza Incidence from Sentinel Surveillance Data Using an INGARCH Model

Changdae Son[a*] · Yeonsu Lee[b] · Hyojung Lee[c]

Summary: Influenza is a seasonal infectious disease requiring timely prediction and analysis for effective public health response. In Korea, surveillance relies on two main data sources. Influenza-Like Illness (ILI) data, collected through sentinel surveillance as sample-based data, reflect overall trends rather than the total magnitude of infection. In contrast, confirmed case data cover all influenza patients nationwide, offering a more accurate measure of outbreak scale but lacking real-time availability. This study aims to develop a model that predicts population-level confirmed cases using real-time ILI data. Previous approaches using scaling parameters have mainly focused on retrospective estimation, limiting their use for forecasting. We extend this approach into an INGARCH (INteger-valued Generalized Autoregressive Conditional Heteroskedasticity) model to enable the prediction of future influenza incidence. To address data limitations, we apply a zero-inflated distribution, and account for time-varying variance using negative binomial and generalized Poisson distributions in the likelihood formulation. This framework enhances prediction performance under incomplete or noisy data and can be extended to incorporate alternative data sources for forecasting confirmed influenza cases.

Keywords: Influenza, Bayesian Inference, Forecasting

[a]80, Daehak-ro, Buk-gu, Daegu 41566, Korea. Department of Statistics, Kyungpook National University. cdson@knu.ac.kr
[b]80, Daehak-ro, Buk-gu, Daegu 41566, Korea. Department of Statistics, Kyungpook National University. lys010512@knu.ac.kr
[c]80, Daehak-ro, Buk-gu, Daegu 41566, Korea. Department of Statistics, Kyungpook National University. hjlee@knu.ac.kr

Sharpening Variance Estimation: An Empirical Bayes Approach under Mean-variance Dependence

Chaewon Song[a*] · Hoyoung Park[b†]

Summary: We propose a two-dimensional nonparametric maximum likelihood estimation (2D-NPMLE) framework within the Empirical Bayes (EB) paradigm for the joint estimation of mean and variance parameters. Unlike conventional F-modeling, which assumes equal replicates, and G-modeling, which allows unequal sample sizes yet assumes independence between the two parameters, our approach accommodates heterogeneous sample sizes and explicitly models mean–variance dependence. The method jointly estimates the empirical distribution of the location parameter and scale parameter, denoted as (θ_i, σ_i^2) through likelihood maximization over a bivariate grid using convex optimization. This procedure guarantees convergence to a global optimum, yielding posterior mean estimators for both parameters. The resulting EB posteriors are further used to construct simultaneous confidence intervals for the means via double-shrinkage techniques, improving both precision and coverage. Extensive simulation studies under diverse dependence structures confirm the superior accuracy and adaptability of the proposed 2D-NPMLE compared to existing variance estimation approaches. In a real-data application using the MAIHDA framework, our method effectively captured heterogeneity in body mass index (BMI) across social strata with unequal subgroup sizes, demonstrating its practical utility in complex multilevel data. Overall, the proposed framework offers a flexible and robust tool for modern variance estimation and small-area inference under heteroscedastic and unbalanced settings.

Keywords: Confidence Interval, Empirical Bayes, Heteroscedasticity, Mean-Variance Dependence, Nonparametric Maximum Likelihood Estimation

[a]Department of Statistics, Sookmyung Women's University. chaewon914@sookmyung.ac.kr
[b]Department of Statistics, Sookmyung Women's University. hyparks@sookmyung.ac.kr

전이학습 기반 이항형 타겟 주성분분석[a]

심대희[b*] · 이은령[b]

요약: 본 연구는 보조 패널 데이터의 정보를 활용하여 타겟 패널의 잠재 요인 구조를 추정하는 타겟 주성분분석(Target-PCA)을 이항형(binary) 타겟 변수로 확장한 Binary Target-PCA via Transfer Learning 방법론을 제안한다. 기존 Target-PCA는 연속형 타겟 변수를 전제로 설계되어, 사회과학·추천시스템 등에서 빈번히 등장하는 이항형 또는 범주형 데이터에는 직접 적용하기 어렵다는 한계를 가진다. 이를 해결하기 위해 본 연구는 범주형 변수를 정량화하는 다양한 방법 중 FACTALS(Factor Analysis by Alternating Least Squares) 기반 접근을 채택하여, 범주형 타겟 변수를 잠재 연속형 행렬로 변환하고 이를 Target-PCA 프레임워크에 통합하였다. 또한 Target-PCA의 가중치 구조를 활용하여 보조 패널과 타겟 패널 간의 정보 전이 및 요인 추정을 수행하고, 결측치가 존재하는 불균형 패널 구조에서도 안정적인 2차 모멘트 추정과 요인 복원이 가능함을 확인하였다. 제안된 Binary Target-PCA는 명목형과 순서형 정량화에 적용 가능하며, 기존 Target-PCA가 연속형 변수에 한정되었던 분석 틀을 범주형 변수로 확장함으로써, 보다 일반화된 전이학습 기반 요인 추정 프레임워크를 제시한다.

주요용어: 타겟 주성분분석, 이항형 타겟 변수

[a]This work is supported by National Research Foundation of Korea grant funded by the Korean government (no. NRF2022R1A2C1012798, RS-2025-02216235).

[b](03063) 서울시 종로구 성균관로 25-2, 성균관대학교 통계학과

다양한 차원축소기법을 이용한 SCFI 선행지표 개발 및 성능 비교[a]

안근찬[b*] · 김재직[c]

요약: 상하이 컨테이너 운임지수(Shanghai Containerized Freight Index; SCFI)는 글로벌 해운 시장의 핵심 지표로 활용되나, 단기 스팟 운임(spot rate)만을 반영하여 시장 상황을 후행적으로 나타내는 특징을 갖는다. 이에 반해 실제 해운업계의 운송 계약은 대부분 3~6개월 선행하여 체결되기 때문에 SCFI 지표와 실무적 의사결정 간의 불일치가 발생하는 문제가 존재한다. 이에 본 연구에서는 이러한 SCFI의 후행성을 보완하고 시장 참여자들에게 장기적이고 전략적인 통찰을 제공할 수 있는 새로운 선행지표 개발을 목표로 한다. 선행지표 개발을 위해 최근 시장의 심리와 기대의 신속한 반영을 위해 뉴스 데이터가 주목을 받고 있다는 점에 착안하여 해운업계와 관련된 방대한 뉴스 데이터와 주요 거시 경제 데이터들을 활용한다. 선행지표 개발에 있어 고차원인 뉴스 데이터의 '차원의 저주 (curse of dimensionality)' 문제를 해결하고 핵심 정보를 효과적으로 추출하기 위해 본 연구는 다양한 선형 투영(linear projection) 차원축소방법들을 이용한다. 최종적으로 본 연구에서 개발된 선행지표들의 성능과 유효성은 실제 SCFI와의 비교를 통해 검증된다.

주요용어: 상하이 컨테이너 운임지수, 선행지수, 차원축소, 뉴스 데이터

[a]이 논문은 정부(과학기술정보통신부) 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. RS-2024-00407300, RS-2025-16067563).

[b](03063) 서울시 종로구 성균관로 25-2, 성균관대학교 통계학과, 석사과정

[c](03063) 서울시 종로구 성균관로 25-2, 성균관대학교 통계학과. 교수. jaejik@skku.edu

Machine Learning-based Prediction of Meal Attendance: A Case Study Using Chungbuk National University Cafeteria Data[a]

Yoonseong An[b*] · Ilsuk Kang[c]

Summary: Accurate prediction of meal attendance in university dormitory cafeterias with limited budgets plays a crucial role in enhancing student satisfaction and reducing food waste. To address this issue, this study focused on the dormitory cafeteria at Chungbuk National University (CBNU) as a real-world case for data-driven meal demand prediction. While current predictions rely heavily on the dietitian's experience, inaccurate estimates often result in declined meal quality and profitability. Raw data were collected from the CBNU dormitory cafeteria, and menu information was vectorized using TF-IDF, which quantifies the relative importance of each menu item. Additionally, contextual factors influencing meal attendance, such as holidays and weekdays, were incorporated into predictive models, including LightGBM and LSTM. The proposed model achieved high predictive performance, with an R^2 value of 0.883 and a Mean Absolute Error (MAE) of 0.05. Moreover, the results of feature importance and SHAP analyses were consistent with intuitive expectations for each variable. Therefore, implementing a data-driven learning model for meal demand prediction can provide substantial benefits to cafeteria operations. It is expected that accurate predictions will improve the efficiency of meal services and contribute positively to the environment.

Keywords: Meal Demand Forcasting, TF-IDF, LightGBM, LSTM, Predictive Analytics

[a]This work was in part supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2024-00440787).

[b](Gaesin-dong, Chungbuk National University) 1, Chungdae-ro, Seowon-gu, Cheongju-si, Chungcheongbuk-do, 28644, Republic of Korea

[c](Gaesin-dong, Chungbuk National University) 1, Chungdae-ro, Seowon-gu, Cheongju-si, Chungcheongbuk-do, 28644, Republic of Korea. isk1218@chungbuk.ac.kr

Detecting Multi-Writer Mixtures in Handwritten Documents via Multiple Instance Learning[a]

Deok-Kwan Yang[b*] · Soyoung Park[c]

Summary: This study targets intra-document, locally mixed handwriting (partial forgery/ghostwriting) and formulates the task in a Multiple Instance Learning (MIL) framework, where a document is a bag and words or short sentence windows are instances. In Stage-1, we extract ArcFace-based 256-dim word embeddings. In Stage-2, we define instances with Win=1 (word) or Win=5 (contiguous 5-word window via L2-mean pooling) and construct anchor–partner mixed bags. In Stage-3, we fairly compare Attention-MIL, Gated-MIL, DSMIL, TransMIL, and a Mean-Pooling baseline under a common protocol, and conduct sensitivity analyses over bag size (N) (10–50), mixture ratio (p) (10–50%), and the Win=1↔Win=5 shift. Results show that MIL models are consistently superior to the baseline; best performance appears at larger (N) (e.g., 50), while lower (p) (sparser mixing) leads to notable degradation. Win=5 leverages sentence/span context and is comparable or superior to Win=1 in some settings. The study provides a model comparison and design guidelines for Win/N/p, offering a practical basis for screening and flagging local forgeries within a single document.

Keywords: Multiple Instance Learning, Multi-Writer Detection, Handwritten Document Analysis, Partial Forgery Detection

[a]This work was in part supported by the Korea Research Foundations, Korea, under grant KRF-0000
This research was supported by research funds from the 2025 Pusan National University University Innovation Support Project.
[b](Gasan-dong, IT Premier Tower) 88, Gasan digital 1-ro, Geumcheon-gu, Seoul 08590, Korea. offic@kss.or.kr
[c]Room 220, Research & Lab Building, College of Natural Sciences, 2, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan 46241, Republic of Korea, Statistics, Pusan National University. ydk9266722@gmail.com

CycleGAN-Turbo 기반 i2i 전처리를 이용한 악천후 조건 도메인 적응 효과 분석

오승환[a]*

요약: 본 연구에서는 악천후(야간/우천/야간우천/안개) 환경에서의 객체 탐지 성능 저하를 완화하기 위해 입력 공간(image-level) 기반 비지도 도메인 적응(UDA)을 적용 및 평가하였다. day-clear 도메인으로만 파인튜닝한 YOLOv11x를 source-only 설정으로 고정하고, CycleGAN-Turbo 기반의 image-to-image(i2i) 전처리를 통해 네 개 도메인(night-sunny, dusk-rainy, night-rainy, day-foggy)에서의 성능을 COCO mAP@[.50:.95]와 mAP@50으로 정량 평가하였다. 또한 크기별S/M/L) 및 클래스 단위 분석을 통해 성능 개선의 분포를 추적하였다. (1) 도메인별 최적 변환 강도(step)가 상이하여 단일 고정치로는 최적화가 어려웠다. (day_foggy-5k, night_sunny-8k, dusk_rainy-9k, night_rainy-4k). (2) 도메인 간 갭이 큰 경우 유의미한 개선이 나타났으며, dusk_rainy(+0.066/+0.103)와 night_rainy(+0.048/+0.108)은 뚜렷한 성능 향상을 보였으나, day_foggy(+0.004/+0.008)은 미미했고 night_sunny(-0.029/-0.043)는 하락하였다. (3) 개선 효과는 Large(L) 및 일부 Medium(M) 객체에 집중되었으며 Small(S)은 거의 개선되지 않았다. dusk_rainy와 night_rainy는 각각 L에서 +0.100, +0.092 (mAP@[.50:.95])의 향상을 보였다. 클래스 관점에서도 car, bus, truck, person의 L/M 구간에서 개선이 두드러졌으며 bike, motor, rider의 S 구간은 여전히 난점으로 남았다. (4) FID는 초기 성능 상승과 동조하지만, 최소 FID 시점이 mAP@[.50:.95] 및 mAP@50의 최대 지점과 일치하지 않았다. 이는 FID가 변환 품질의 필요조건 지표로는 유용하지만, 최적 체크포인트 선택에는 검증 mAP가 더 적합함을 의미한다. (5) 해상도 축소(예: resize 256)는 일관된 성능 하락을 유발했다. (dusk_rainy mAP@[.50:.95] 0.265→0.247, mAP@50 0.426→0.396) 종합하면, detector의 파라미터를 변경하지 않는 plug-and-play 형태의 UDA 전처리로서 가시성 저하 도메인(우천/야간우천/안개)과 대형 객체에 특히 효과적인 반면, night-sunny와 같이 분포 드리프트가 크게 작용하는 도메인에서는 오히려 성능 저하가 발생하였다.

주요용어: CycleGAN-Turbo, 비지도 도메인 적응(UDA), Object Detection

[a]고려대학교

On Robust M-estimations on Riemannian Manifolds

Jihyun Ryu[a*] · SungKyu Jung[b]

Summary: Robust statistical estimation is essential when data are contaminated with outliers or heavy-tailed noise. Many contemporary datasets arise on nonlinear manifolds, where geometric structure must be explicitly respected. In this work, we investigate Riemannian robust M-estimators, combining the influence-function–based robustness of classical M-estimators with the intrinsic geometry of Riemannian manifolds. This synthesis addresses both the statistical need for robustness and the geometric necessity of manifold-aware analysis. From a computational perspective, we contribute efficient implementations for computation of manifold-valued M-estimators within the open-source geomstats library, enabling automatic Riemannian gradients and customized robust loss functions. Our framework is evaluated across multiple manifolds—including spheres, hyperbolic spaces, and the cone of symmetric positive definite (SPD) matrices, through extensive simulations. In particular, we study the calibration of cutoff parameters for downweighting outliers and propose ranges that achieve approximately 95% of relative efficiency under Gaussian-like distributions, analogous to Euclidean benchmarks. For heavy-tailed distributions, M-estimators are shown to have greater efficiency than the usual Fréchet mean. We demonstrate the practical utility of Riemannian robust M-estimators through real-world applications such as earthquake epicenter on the Earth and wind direction fields. In particular, robust M-estimators yield more stable and reliable estimates than the Fréchet mean. Our results establish robust M-estimation as a principled, computationally tractable, and empirically effective tool for manifold-valued data analysis.

Keywords: robust M-estimators, manifold-valued data, statistical software, etc

[a]Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea. ktf10a@snu.ac.kr

[b]Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea. sungkyu@snu.ac.kr

Robust Support Vector Machines with M-estimator-driven Loss Functions[a]

Mingyeong Kim[b] · Jia Yoo[c] · Jin Hee Yoon[d*]

Summary: Support Vector Machines (SVMs) are powerful classifiers with strong generalization ability, yet their performance deteriorates under outliers and noisy labels due to the quadratic penalty of the conventional L_2 loss. To enhance robustness, this study investigates SVM formulations that incorporate M-Estimator-driven loss functions, which bound the influence of extreme samples through non-quadratic residual penalties. Specifically, several representative M-estimators—including Huber, Cauchy, Welsch, and Fair functions—are embedded into the SVM framework and optimized via standard quadratic programming. Each formulation provides a distinct balance between sensitivity and resistance to noise by reshaping the loss gradient near large residuals. Experiments on multiple UCI benchmark datasets with varying outlier ratios (0–5%) demonstrate that M-Estimator-based SVMs consistently outperform the standard L_2-SVM in both classification accuracy and robustness. Among the tested models, the Cauchy- and Welsch-based SVMs achieve the smallest accuracy degradation as contamination increases, highlighting their stable margin properties. These results confirm that M-Estimator-driven SVMs offer a simple yet effective means to improve resilience against outliers without altering the fundamental SVM structure, providing a practical and interpretable framework for robust classification in noisy environments or existence of outliers.

Keywords: Support Vector Machine, Outliers, Robustness, M-Estimator

[a]This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2024-00351610).
[b]Dept. Mathematics and Statistics, Sejong Univ., 209, Neundongro, Gwanjin-gu, Seoul, Korea. alsrud8238@gmail.com
[c]Dept. Mathematics and Statistics, Sejong Univ., 209, Neundongro, Gwanjin-gu, Seoul, Korea. ujia0220@gmail.com
[d]Dept. Mathematics and Statistics, Sejong Univ., 209, Neundongro, Gwanjin-gu, Seoul, Korea. jin9135@sejong.ac.kr

이질성 탐색을 위한 적응형 벌점화 기반 군집 가중 모형

이다인[a*] · 서병태[a]

요약: 다양한 데이터에서 잠재적 이질성을 포착하기 위한 통계적 접근으로 혼합 회귀 모형(Finite Mixture Regression, FMR)이 제안되어 왔다. 그러나 기존의 모형은 공변량 분포가 군집 형성에 기여하는 상황을 충분히 설명하지 못하며, 고차원 환경에서의 변수 선택과 군집별 구조 차이 탐색에도 한계가 있다. 본 연구는 이러한 한계를 보완하기 위해 반응변수와 공변량의 결합 분포를 통합적으로 모형화하는 적응형 벌점화 군집가중모형(Cluster-Weighted Model with Heterogeneity Pursuit and Adaptive Lasso; CWM-HP-AL)을 제안한다. 제안 모형은 공통 효과와 군집 특이 효과를 분리한 회귀계수 표현과 적응형 ℓ_1 벌점을 결합하여, 관련 변수와 군집 간 이질성을 동시에 식별할 수 있도록 설계되었다. 또한 공변량의 분포 차이를 반영하기 위해 Gaussian Parsimonious Covariance Model(GPCM)의 14가지 공분산 구조를 고려하고, EM 알고리즘과 Bregman 좌표하강법을 결합한 효율적인 추정 절차를 제시하였다. 모의실험 결과, 제안 모형은 단순한 조건에서는 기존 방법과 유사한 수준의 성능을 보이면서도, 공변량 분포 차이·상관 잡음·분산 불균형이 존재하는 복잡한 환경에서는 모수 추정, 변수 선택, 군집화 성능 모두에서 일관된 우수성을 나타냈다. 이는 CWM-HP-AL이 기존 혼합 회귀 기반 모형의 장점을 유지하면서도 공변량 분포 정보를 효과적으로 통합하여 보다 정교한 이질성 탐색을 가능하게 함을 보여준다.

주요용어: 군집가중모형, 이질성 탐색, 적응형 벌점화 기법

[a](03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과

함수형 회귀 분석을 통한 리튬 이온 배터리 용량 예측

이도윤[a] · 안경민[b]

요약: 리튬 이온 전지(Lithium-ion battery)는 높은 에너지 밀도와 낮은 자가방전율로 인해 휴대용 전자기기, 전기자동차 등 다양한 에너지 저장 장치에 널리 활용되고 있는 대표적인 2차 전지이다. 그러나 반복적인 충·방전 주기는 전해질 증발과 내부 팽창 등의 문제를 초래하여 용량 저하를 유발하고, 이는 심각한 안전 문제로 이어질 수 있다. 따라서 리튬 이온 배터리의 용량 예측은 유지보수 및 안정성 확보를 위해 필수적이다. 본 연구에서는 함수형 데이터 분석(Functional DataAnalysis, FDA)에 기반한 새로운 용량 예측 방법을 제안한다. 이 접근법은 이산적으로 축적된 모니터링 데이터를 연속적인 함수로 표현하여, 기존의 데이터 기반 모델로는 포착하기 어려운 진폭(amplitude) 및 위상(phase) 변동을 효과적으로 반영한다. 특히, 상온뿐 아니라 저온 및 고온 등 다양한 온도 조건에서의 용량 저하를 분석 대상으로 하였으며, 함수형 선형 회귀 모형(functionallinear regression model)과 함수형 주성분 회귀 모형(functional principalcomponent regression model)을 적용하였다. 제안된 FDA 기반 방법은 동일 조건에서 학습된 딥러닝 모델 대비 매우 우수한 예측력을 보였다.

주요용어: 딥러닝, 리튬이온 배터리, 함수형 주성분 분석, 함수형 회귀분석

[a]계명대학교 통계학과
[b]서울여자대학교 데이터사이언스학과

Adaptive Weighted Total Variation Penalty for Precise Change Point Detection[a]

Dong-Young Lee[b*] · Kwan-Young Bak[c] · Jae-Hwan Jhong[d]

Summary: Total variation-based methods, such as the fused lasso, are standard for change point detection but are impaired by issues like local monotonicity. To address these limitations, this study comparatively analyzes the fused lasso with two alternative methodologies. These methodologies are the Penalized Regression Spline (PRS) and a novel Adaptive Penalized Regression Spline (APRS) that incorporates data-adaptive weights. Using an efficient coordinate descent algorithm, we evaluated the three methods on both simulated and real-world data. The comparative analysis, conducted under various Signal-to-Noise Ratio (SNR) conditions, revealed that APRS demonstrated superior performance, particularly in minimizing false positives.

Keywords: B-splines, Coordinate Descent Algorithm, Fused Lasso, Weighted Total Variation

[a]The work of Kwan-Young Bak was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00342014 and RS-2022-00165581).
The work of Jae-Hwan Jhong was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00342014 and RS-2024-00440787).

[b]Master's student, Department of Statistics, Chung Buk National University, Cheongju 28644, Korea. dtsy5891@gmail.com
[c]Assistant professor, School of Mathematics, Statistics and Data Science, Sungshin Women's University, Seoul 02844, Korea. kybak@sungshin.ac.kr
[d]Associate professor, Department of Statistics, Chung Buk National University, Cheongju 28644, Korea. jjh25@cbnu.ac.kr

Robust Gene Set Testing with Integrated Methylome Signals: A Plug-in Framework Leveraging Network Topology[a]

Minhyuck Lee[b*] · Hokeun Sun[c]

Summary: DNA methylation is an epigenetic modification that regulates gene expression and it plays a crucial role in disease progression and cancer development. Although epigenome-wide association studies (EWAS) have identified numerous differentially methylated CpG sites, conventional gene set testing approaches remain limited by the following three main challenges: (i) insufficient robustness in defining significant CpGs, (ii) reliance on mean methylation levels, which overlooks variability in methylation, and (iii) lack of integration of biological network information. We propose a robust CpG score–based gene set enrichment framework that incorporates both mean and variance signals into gene-level scores and evaluates pathway enrichment via a threshold-free fast GSEA procedure. To further capture the biological context, we introduce a plug-in weighting scheme based on gene network topology, including degree-weighted and local effect weighted betweenness centrality measures. Importantly, the plug-in design allows users to flexibly incorporate diverse sources of biological insight as gene-level weights, making the framework broadly adaptable to different research contexts. Simulation studies and application to TCGA breast cancer (BRCA) methylation data demonstrate that the proposed method improves detection power while maintaining false discovery rates at a reasonable level, and reveals biologically meaningful pathways overlooked by conventional testing. This plug-in framework enhances the robustness, flexibility, and interpretability of DNA methylation–based gene set testing, offering a versatile strategy to decode gene regulation from epigenome-wide methylation profiles.

Keywords: DNA methylation, Gene set enrichment analysis, Genetic network, Differential variability

[a]This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-16065086).
[b]Master's Student, Department of Statistics, Pusan National University, Busan 46241, Korea. dlalsgur4273@pusan.ac.kr
[c]Professor, Department of Statistics, Pusan National University, Busan 46241, Korea. hsun@pusan.ac.kr

3D Measurement of Vehicle Dent Using Deep Learning-based Segmentation and Mobile LiDAR

Seonjae Lee[a*] · Yonghak Lee[b]

Summary: Dents are the most frequently occurring form of vehicle damage during collisions, characterized by three-dimensional structural deformations in which the surface is inwardly displaced. Traditional vehicle dent measurement methods rely on manual measurements and 2D image analysis that lack depth information, while high-precision 3D scanner-based approaches are impractical due to their high cost. To overcome these limitations, this study proposes an automated dent measurement framework that integrates a deep learning-based segmentation model with a mobile LiDAR sensor. The proposed framework first detects dent regions in RGB images using the YOLOv11-seg model, and then projects the identified 2D boundaries onto a 3D point cloud acquired from the LiDAR sensor to obtain real-world coordinates. The width and height of the dents are computed by measuring Euclidean distances between corresponding points, while the depth is determined by estimating a reference plane from surrounding undamaged regions and calculating point-to-plane distances. The proposed method enables quantitative 3D measurement of vehicle dents and demonstrates significant potential for applications in damage assessment, repair cost estimation, and insurance claim processing.

Keywords: 3D point cloud, mobile LiDAR sensor, Image segmentation

[a]Department of Data Science, University of KONKUK, 120, Neungdong-ro, Gwangjin-gu, Seoul, Republic of Korea
[b]Department of Statistics, University of KONKUK, 120, Neungdong-ro, Gwangjin-gu, Seoul, Republic of Korea

FastKRR: An R Package for Efficient Kernel Ridge Regression with RcppArmadillo[a]

Gyeongmin Kim[b] · Seyoung Lee[c*] · Miyoung Jang[d] · Dongha Kim[e] · Kwan-Young Bak[f]

Summary: A basic kernel ridge regression is computationally expensive, as the cost grows cubically with the dataset size. This is due to the inversion of the n×n kernel matrix, which requires O(n^3) operations. To address this issue, we propose an approach that approximates the kernel matrix and leverages the Woodbury formula to achieve faster computation. In particular, we implement three approximation strategies—Nyström, pivoted Cholesky, and random Fourier features—to efficiently solve large-scale regression problems. The algorithms are written in C++ with RcppArmadillo and integrated into an R package, providing fast and scalable computation. For hyperparameter tuning, we utilize the CVST package to reduce the time required for parameter search. Furthermore, our method is integrated into the tidymodels ecosystem, providing a modern and seamless workflow for machine learning in R.

Keywords: Kernel ridge regression, kernel approximation, Armadillo, parallel computing

[a]This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00342014).

[b](02844) 2, Bomun-ro 34 Da-gil, Seongbuk-gu, Seoul, Korea, Master's student, Department of Statistics, Sungshin Women's University. 220254001@sungshin.ac.kr

[c](02844) 2, Bomun-ro 34 Da-gil, Seongbuk-gu, Seoul, Korea, Master's student, Department of Statistics, Sungshin Women's University. 220254005@sungshin.ac.kr

[d](02844) 2, Bomun-ro 34 Da-gil, Seongbuk-gu, Seoul, Korea, Master's student, Department of Statistics, Sungshin Women's University. 220254006@sungshin.ac.kr

[e](02844) 2, Bomun-ro 34 Da-gil, Seongbuk-gu, Seoul, Korea, Assistant professor, School of Mathematics, Statistics and Data Science, Sungshin Women's University. dongha0718@sungshin.ac.kr

[f](02844) 2, Bomun-ro 34 Da-gil, Seongbuk-gu, Seoul, Korea, Assistant professor, School of Mathematics, Statistics and Data Science, Sungshin Women's University. kybak@sungshin.ac.kr

LSTM-Autoencoder 기반 하이브리드 모델을 활용한 봄철 서울 지역 지면온도 예측 개선 연구

이양우[a*] · 김찬수[b]

요약: 지면온도는 기후 시스템 내 물리적, 생물학적, 화학적 과정을 모니터링하는데 핵심적인 변수이고, 일사 흡수 및 장파 복사 등의 기상 조건과 지표면 특성에 민감하게 반응하여 시공간적으로 큰 변동성을 보여준다. 본 연구에서는 2018년부터 2022년 서울 지역의 봄철동안 수집된 1시간 간격의 종관기상관측자료를 활용하여 기온, 풍속, 습도, 이슬점 온도, 일조시간, 일사량의 변수를 통해 지면온도를 예측하기 위해 의사결정나무, 선형회귀, 랜덤포레스트, 그래디언부스팅의 머신러닝 기법을 적용하였다. 초기 예측 결과에서 고온의 지면 온도를 과소 추정하는 경향이 확인되었으며, 이는 고온 등 극한 상황에서 기존 모델의 예측 성능이 제한적임 보여준다. 이러한 한계를 보완하기 위해 LSTM-Autoencoder 모델를 결합한 하이브리드 예측 기법을 적용하였다. 적용 결과 봄철 고온 지면온도에서 대한 예측의 정확도가 향상되었으며, 이상값 기반 예측 기법이 기상 및 환경 데이터를 활용한 예측 모델링에 효과적으로 적용될 수 있음을 볼 수 있다.

주요용어: 지면온도, 머신러닝기법, LSTM-Autoencoder, 하이브리드 예측

[a](32588) 충청남도 공주시 공주대학로 56, 공주대학교 응용수학과, 석사과정. h993318@naver.com
[b]공주대학교, 응용수학과, 교수

시간적 정보를 반영한 대규모 언어모델 기반 오디오 캡셔닝 품질 향상

이윤정[a*] · 임창원[b]

요약: 오디오 캡셔닝(Audio Captioning)은 주어진 음향 신호 내 사건과 배경을 자연어로 기술하는 태스크로, 멀티모달 학습의 주요 연구 주제로 주목받고 있다. 그러나 기존 모델은 주로 정적인 장면 요약에 초점을 맞추어, 사건의 발생 순서나 지속 시간 등 시간적 정보(Temporal Information)를 충분히 반영하지 못하는 한계가 있었다. 본 연구의 목적은 시간적 서술 구조를 강화하여 오디오 캡션의 자연스러움과 사실성을 높이는 것이다. 이를 위해 사전학습된 오디오 인코더로부터 사건 단위의 특징을 추출한다. 기본 구조로는 CNN14 기반 오디오 인코더와 BART 언어모델을 결합한 베이스라인 모델을 사용하였으며, 여기에 대규모 언어모델(LLM)을 활용해 이러한 특징을 시간 순서에 맞게 결합 및 서술하도록 학습시켰다. 또한 LLM의 생성 문장을 기반으로 시간적 연결성과 논리적 일관성을 강화하기 위한 Temporal Prompting 기법을 적용하였다. Clotho 데이터셋을 이용해 평가한 결과, 제안한 방법은 기존 베이스라인 대비 Temporal F1 점수에서 유의한 향상을 보였다. 본 연구는 LLM의 언어 생성 능력을 활용해 시계열적 정보를 통합함으로써, 오디오 캡셔닝의 시간적 이해와 기술 정확도를 동시에 개선할 수 있음을 보여준다.

주요용어: 오디오 캡셔닝, LLM, 시간적 정보, 멀티모달 학습, CNN14, BART

[a](06974) 서울특별시 동작구 흑석로 84, 중앙대학교 통계데이터사이언스학과. dlwogus9713@cau.ac.kr
[b](06974) 서울특별시 동작구 흑석로 84, 중앙대학교 통계데이터사이언스학과. clim@cau.ac.kr

Empirical Bayesian Estimation of Prior Distributions Using NPMLE for Predicting Batting Averages in KBO

Euichae Lee[a*] · Hoyoung Park[b†]

Summary: In this study, we apply the Empirical Bayesian framework to estimate prior distributions using Nonparametric Maximum Likelihood Estimation (NPMLE). All parameters of the observed data distribution are treated as unknown quantities, and their distributions are estimated through NPMLE within the Empirical Bayesian framework. This approach allows for adaptive estimation of distributions based on the data, providing robust and accurate estimations. Due to these characteristics, the method has become widely used in various applications. We demonstrate the effectiveness of our method through simulation studies and validate its performance by applying it to predict the batting averages of players in the Korea Baseball Organization (KBO) using data from the 2023 and 2024 seasons. Our results show that the proposed methodology provides accurate predictions and demonstrates strong performance in real-world scenarios.

Keywords: Batting Average Prediction, Empirical Bayesian Framework, Nonparametric Maximum Likelihood Estimation, Prior Distribution Estimation

[a]Department of Statistics, Sookmyung Women's University. dldmlco99@sookmyung.ac.kr
[b]Department of Statistics, Sookmyung Women's University. hyparks@sookmyung.ac.kr

Bayesian Networks for Analyzing Intersectional Fairness: A Structure-based Perspective

이재은[a]*, 황범석[b]

Summary: Algorithmic fairness measures can detect disparities in model predictions, but they do not tell us how those disparities actually come about. They also struggle when multiple protected characteristics—such as race and gender together—interact to produce intersectional bias. We address both problems with a Bayesian network framework for structural fairness auditing. The framework has two main parts. First, we develop an Order-Graph MCMC procedure that uses 2-Tier Priors to build causal structures by blending patterns seen in the data with causal relationships known from domain expertise. Second, we introduce the Indirect Fairness Score (IFS), which measures how much each pathway in the causal structure contributes to unfairness by blocking a pathway and observing the change in conditional probabilities. We applied our methods to the UCI Adult Income dataset. The MCMC procedure had a slightly worse BIC than greedy search methods, but the causal structures it produced were much more stable—arc support values were near \(1.0\)—and contained fewer unnecessary edges. When we examined demographic parity and equalized odds, we found disparities reaching \(0.412\) in groups where both sex and age were factors. Breaking down these numbers with IFS revealed that different pathways in the causal network had opposite effects on women: some routes created disadvantages, while other routes actually benefited them. What makes our approach useful is that it does not just indicate that bias exists—it shows which specific causal pathways are responsible for it. This enables targeted interventions on the mechanisms causing disparities, rather than broad adjustments aimed only at improving aggregate fairness metrics.

Keywords: Bayesian Network (BN), Causal Inference, Indirect Fairness Score (IFS), Intersectional fairness

[a]중앙대학교 통계데이터사이언스 학과. yizenxnii@naver.com
[b]중앙대학교 통계데이터사이언스 학과

Time-to-event BOIN Design Incorporating Low-grade Toxicity Information

Yerim Im[a*] · Soyeon Park[a] · Yong Zang[b] · Ying Yuan[c] · Ick Hoon Jin[a,d]

Summary: In Phase I clinical trials, the Time-to-Event Bayesian Optimal Interval (TITE-BOIN) design mitigates delays from late-onset toxicities but overlooks predictive early signals from low-grade toxicities (LGTs). We propose a novel Bayesian framework that extends TITE-BOIN by prospectively incorporating binary LGT data. A key contribution is a comprehensive joint likelihood function that distinctly models three data scenarios: complete observations, partially pending data, and fully pending data. Within this framework, the overall dose-limiting toxicity (DLT) rate is parameterized as a composite of the LGT rate and the conditional probability of a subsequent DLT. For robust posterior inference with pending data, we develop a new imputation logic based on the conditional expectation of unobserved DLT outcomes, tailored for the two distinct missingness patterns. The model parameters are estimated via Markov Chain Monte Carlo (MCMC), and their posterior means are subsequently used to compute the imputed DLT values. By leveraging early LGT signals through this rigorous likelihood-based approach, the design provides a more nuanced and timely assessment of a dose's toxicity profile, enhancing trial safety and leading to a more robust methodology for determining the maximum tolerated dose (MTD).

Keywords: Bayesian adaptive clinical trial, Missing data imputation, Survival analysis

[a]Department of Statistics and Data Science, Yonsei University, South Korea
[b]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, USA
[c]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, USA
[d]Department of Applied Statistics, Yonsei University, South Korea

ResNet-BiGRU with Conditioned Query-based Cross-attention and Weighted Loss for Automated Chagas Disease Detection from 12-Lead ECG[a]

Hyuno Im[b*] · Nahyun Lee[c] · Taeyoung Kang[d] · Taehwan Kim[d] · Donggun Kim[e] · Donggyu Lee[e] · Seungsang Oh[e] · Wuming Gong[f] · Il-Youp Kwak[g]

Summary: Our team CAUETUMN, a participating team in the 2025 PhysioNet/CinC Challenge, investigates whether integrating physiologically interpretable and deep sequence representations, enhanced by Query-based cross-attention, can improve Chagas disease detection from 12-lead ECGs. We combine three streams: (i) a 4-layer 1D ResNet for local morphology extraction; (ii) a bidirectional GRU with gated attention for long-range temporal context; and (iii) handcrafted R-peak morphology and demographic features (age, sex). The auxiliary features are projected into a high-dimensional query space to attend over sequence embeddings, enabling integration of temporal patterns. Raw ECGs undergo baseline wander removal with an OC/CO morphological filter. The model is trained on SaMi-Trop, PTB-XL, and CODE-15% dataset using a loss function that incorporates both class-specific weights to address label imbalance and group-specific weights to account for dataset-level distribution differences. Final Challenge score on hidden test set was 0.218, ranking 17th among the 41 eligible participating teams.

Keywords: Electrocardiogram, Chagas disease, Deep learning

[a]This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00341075, RS-2023-00208284).
[b]Department of Applied Statistics, Chung-Ang University, 84, Heukseok-ro, Dongjak-gu, Seoul 06974. imhyuno@cau.ac.kr
[c]Department of Smart Cities, Chung-Ang University, 84, Heukseok-ro, Dongjak-gu, Seoul 06974
[d]Department of Applied Statistics, Chung-Ang University, 84, Heukseok-ro, Dongjak-gu, Seoul 06974
[e]Department of Mathematics, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul 02841
[f]Cardiovascular Division, Department of Medicine, University of Minnesota, 2231 6th St SE, MN 55455, USA
[g]Department of Applied Statistics, Chung-Ang University, 84, Heukseok-ro, Dongjak-gu, Seoul 06974. ikwak2@cau.ac.kr

Music Tjerapy and Gardening Enhances Post-discharge Adherence in Psychiatric[a]

Wonsik Jung[b,c] · Insu Jan[d,e] · Hyeonjeong Lim[f*] · Subin Jeong[g] · Hyun Ju Lee[c] · Gahui Han[h] · Shin-Young Park[c] · Hoyeong Jeong[c] · Yaehyeon Kim[c] · Seoeun Lee[c] · Eunjee Lee[i] · Kwang-Yeon Choi[b,c]

Summary: Maintenance pharmacotherapy after acute hospitalization for schizophrenia, bipolar disorder, and major depression is essential for relapse prevention and functional recovery, yet medication adherence —measured by the proportion of days covered (PDC) and medication possession ratio (MPR)—averages only about 50%, ranging from 20% to 80%. While interventions such as medication education and motivational interviewing can improve adherence, their effects are often short-lived without continuous support. Horticultural and music therapy have been reported to reduce depression and anxiety, but their long-term effects on adherence remain unclear. This study evaluated the impact of these therapies on medication adherence and outpatient treatment retention among psychiatric inpatients at Chungnam National University Hospital. From July 2022 to July 2023, 139 patients who received weekly horticultural and music therapy were compared with 278 matched controls hospitalized between July 2020 and June 2022, matched 1:2 by gender, age, and diagnosis. After excluding those without post-discharge visits, 387 patients were analyzed. Medication adherence, assessed by MPR and PDC over 12 months, was analyzed using repeated-measures ANOVA, and outpatient retention was evaluated with a Cox proportional hazards model. The treatment group demonstrated significantly higher MPR at 4, 6, and 8 months after discharge, although the difference diminished at 10 and 12 months. Age and music therapy participation were significant predictors of adherence and retention, with young and middle-aged patients and music therapy participants showing greater adherence and longer outpatient follow-up. These findings suggest that music therapy during hospitalization enhances post-discharge medication adherence and continuity of care, particularly among younger and middle-aged psychiatric patients.

[a]This material was based on work partially supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant(RS-2022-NR068754). This research was supported by the National Research Foundation of Korea(NRF) grant funded by the Korean government (MSIT) (No. 2019M3E5D1A02068548; 2020R1C1C1011378).

[b]Department of psychiatry, College of medicine, Chungnam National University, Daejeon, Republic of Korea
[c]Department of psychiatry, Chungnam National University Hospital, Daejeon, Republic of Korea
[d]Department of statistics and data science, Chungnam National University, Daejeon, Republic of Korea
[e]Korea Bioinformatics Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon, Republic of Korea
[f]Institute of Natural Sciences, Chungnam National University, Daejeon, Republic of Korea
[g]Biomedical Research Institute, Seoul National University Hospital, Seoul, Republic of Korea
[h]Department of mathematics, Daejeon, Chungnam National University, Republic of Korea
[i]Department of information and statistics, Chungnam National University, Daejeon, Republic of Korea

서포트 벡터 머신에 대한 효율적인 변수 선택법[a]

전재우[b*] · 김재직[c]

요약: 서포트 벡터 머신 (Support Vector Machine; SVM)은 분류 문제에서 강력한 성능을 보이지만, 커널 공간 (kernel space)을 이용하는 특성으로 인해 변수의 중요도를 추정하기 어렵고 변수 선택 또한 쉽지 않다. SVM에 대한 기존 변수 선택 방법들은 적은 변수의 수를 가정하여 교차검증 (cross validation)을 통해 예측률을 구하고 이를 기준으로 후진제거 (backward elimination)를 진행하는 것이 일반적이었다. 하지만 이러한 방식은 고차원 데이터 적용에 있어 매우 오랜 계산 시간을 요구한다는 문제를 갖는다. 이 문제를 해결하기 위해 본 연구에서는 스크리닝(screening)과 재귀적 변수 제거 (Recursive Feature Elimination; RFE)를 결합한 효율적인 변수 선택 방법을 제안한다. 제안하는 방법은 스크리닝 단계에서 SVS (Sufficient Variable Screening)를 도입하여 비선형 관계와 조건부 효과를 모두 고려하여 변수의 개수를 줄이고, 최종 변수 선택 단계에서는 기존 SVM-RFE가 변수들의 중요도 순위만 제공했던 한계를 넘어 최종 변수 선택 기준을 제시한다. 이를 통해 본 방법은 계산 시간을 획기적으로 줄일 수 있으며, 모형 스스로 최적의 변수 집합을 결정하는 것이 가능하다. 제안한 방법의 성능은 다양한 모의실험과 실제 유전자 데이터를 통해 검증된다.

주요용어: 서포트 벡터 머신, 변수 선택, 고차원 데이터, 비선형 효과

[a]이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. RS-2024-00407300, RS-2025-16067563).
[b](03063) 서울시 종로구 성균관로 25-2, 성균관대학교 통계학과, 석사과정
[c](03063) 서울시 종로구 성균관로 25-2, 성균관대학교 통계학과, 교수. jaejik@skku.edu.

Wind Power Site Assessment Via Copula-based Modeling[a]

Jimin Jeon[b*] · Jaehong Jeong[c]

Summary: We propose a two-stage probabilistic framework to identify optimal locations for wind power in South Korea, a region with high wind variability. In the first stage, a spatial copula model generates seasonal wind speed predictions across specific quantiles, capturing spatial dependencies among neighboring sites. The second stage uses a vine copula to refine these predictions by modeling conditional dependencies across quantiles. Candidate sites are evaluated based on received power, confidence interval length, and exceedance probability. Our analysis reveals that high-priority sites are concentrated near Yeosu, Namhae, and Geoje along the southern coast of the Korean Peninsula. Secondary locations are distributed along the broader southwestern and southeastern coasts, including areas near Ulsan, Pohang, and eastern Jeju.

Keywords: Wind power, Site assessment, Copula-based modeling, Spatial copula

[a]This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00345167).
[b]Department of Applied Statistics, Hanyang University, Seoul, 04763, South Korea
[c]Department of Mathematics, Hanyang University, Seoul, 04763, South Korea

BLMM과 baseline eGFR 정보를 활용한 eGFR 예측 모델 개발 및 검증[a]

정민재*[b] · 오만숙[c]

요약: 배경: 만성 신장 질환(CKD)은 질병의 진행을 예측하는 것에 있어 조기 진단이 필수적이다. 추정 사구체 여과율(eGFR)의 종단적 변화(eGFR slope)는 신기능 저하의 핵심 지표이지만, 변화 양상은 개인마다의 편차가 크다. 본 연구는 베이지안 선형 혼합 모델(BLMM)을 활용하여 한국인 코호트에서 기저 시점의 임상 정보와 바이오마커를 기반으로 미래 eGFR 궤적을 예측하고, 개인의 첫 eGFR 측정값을 이용해 예측 정확도를 향상시키는 전략의 유용성을 검증하고자 한다. 방법: 한국인유전체역학조사사업(KoGES)의 안산-안성 코호트 데이터에서 기저 eGFR이 60 이상이고 단백뇨가 없는 2404명을 대상으로 분석을 수행했다. eGFR의 종단적 변화를 예측하기 위해 베이지안 선형 혼합 모델을 구축하였으며, 개인별(ID)로 다른 절편과 시간(RelativeYear)에 따른 기울기를 갖는 임의 효과를 설정했다. 모델의 예측 성능은 두 단계로 평가했다: (1) 기저 시점의 임상 및 바이오마커 정보만으로 미래 eGFR 예측, (2) 개인의 기저 eGFR 측정값을 이용, 보정한 후 예측. 모델 예측 성능은 평균 제곱근 오차(RMSE), 결정계수(R^2) 등 여러 지표를 활용하여 평가하였다. 결과: 고정 효과 분석 결과, 연령, 성별, 당뇨, 요산, C-펩타이드, 호모시스테인이 eGFR 수준과 유의미한 연관성을 보였다. 예측 성능 평가에서, 기저 정보만으로 예측했을 때 검증 데이터셋에 대한 R^2는 0.249였다. 그러나 개인의 기저 eGFR 값으로 예측을 보정한 후, R^2는 0.647로 대폭 향상되었고 RMSE는 9.48에서 6.50로 감소했다. 결론: 본 연구는 임상 및 환경 요인을 통합한 베이지안 선형 혼합 모델과 기저 eGFR 값을 이용한 사후 보정 전략이 개인의 미래 신기능 변화 궤적을 예측하는 데 효과적임을 입증했다. 이는 현실진료 환경이나 공공 건강 데이터 기반의 조기 고위험군 선별 도구로서의 활용 가능성을 시사한다.

주요용어: BLMM, eGFR, 개인화 예측, KOGES 코호트

[a]본 연구는 대한민국 질병관리청의 CODA (Clinical & Omics Data Archive)를 통해 제공된 유전 및 건강 정보 자료를 바탕으로 수행되었습니다.
본 연구는 한국정부(과학기술정보통신부, MSIT)에서 지원하는 한국연구재단(NRF)의 연구비 지원으로 이루어졌습니다.
[b](03760) 서울시 서대문구 이화여대길 52, 이화여자대학교 통계학과. eileenjung@ewha.ac.kr
[c](03760) 서울시 서대문구 이화여대길 52, 이화여자대학교 통계학과. msoh@ewha.ac.kr

장벽을 반영한 SPDE 기반 시공간 모형을 이용한 한국 수질 오염 분석

정아영[a*]

요약: 수질 오염은 시기와 지역에 따라 특성이 달라지므로, 이를 정확히 이해하고 효과적으로 관리하기 위해서는 시공간적 분석이 필요하다. 본 연구에서는 2020년부터 2023년까지 4년간의 전국 수질측정망 데이터를 활용하여 시공간 분석을 수행하고자 한다. 공변량으로는 경지 면적, 산림 면적, 축산 두수, 하수처리장 방류량, 월평균 기온, 월합 강수량, 도시–농촌 구분, 고도, 경사, 상·중·하류 구분 등을 고려하였다. 하천 시스템은 복잡한 구조를 가지며, 육상 지형은 물리적 장벽으로 작용하기 때문에, 이러한 장벽을 반영한 확률 편미분방정식(SPDE) 기반의 시공간 모형을 적용하고자 한다. 분석은 R의 INLA(Integrated Nested Laplace Approximation) 패키지를 활용한 베이지안 추론 기반으로 수행되며, 사후 평균 추정치와 특정 기준 초과 확률의 사후 추정값을 지도화하여 수질 오염의 공간적 패턴을 시각화하고자 한다.

주요용어: 수질, 장벽 모델, INLA, 시공간 통계

[a]이화여대 통계대학원

Log-Linear BART를 이용한 병원 물품 소비량 분포 예측[a]

류현지[b] · 정용화[b*] · 김재직[c]

요약: 일반적으로 물류 및 재고 관리 분야에서 물품들의 소비량은 카운트 데이터 (count data)의 형태로 관측되고, 이를 예측하여 과잉 재고 및 재고 부족으로 인한 비용 손실을 최소로 하는 것은 이 분야에서 핵심적인 일 중 하나이다. 병원의 경우 환자들을 위해 사용되는 다양한 물품들에 대한 재고 관리는 병원 경영에 있어 매우 중요한 문제이고, 이 물품들의 소비량을 예측하는 것은 이 문제의 핵심이다. 다만, 병원의 경우 환자들의 긴급한 치료 및 수술 등의 이유로 재고 부족이 훨씬 더 치명적인 상황이므로 평균적인 소비량을 예측하기보다는 그 상한 (upper bound) 예측에 더 관심이 있다. 이 상한을 예측하기 위해 카운트 데이터의 특징인 이산성과 과산포를 고려한 로그-선형 베이지안 가법 회귀 나무 (Log-Linear BART) 모형이 최근 개발되었고, 본 연구에서는 이 모형을 병원 물품 소비량 데이터에 적용하여 그 상한을 예측하고자 한다. 또한 개별 물품들의 예측 결과들을 기반으로 전체 물품에 대해 재고 부족이 발생하는 물품의 비율을 조절할 수 있는 방안을 제안한다.

주요용어: 카운트 데이터, 베이지안 가법 회귀 나무, 로그-선형 모형, 소비량 예측

[a]이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. RS-2024-00407300, RS-2025-16067563).
[b](03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과, 석사과정
[c](03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과, 교수. jaejik@skku.edu.

Wildlife Abundance Estimation Via Bayesian Hierarchical Model

정우진[a*], 한종건[b], 장원기[c], 장원철[d]

요약: 생태학에서는 특정 지역 내 야생동물 개체수를 정확히 추정하는 것이 중요한 과제이다. Spatial capture–recapture(SCR) 모형은 카메라 트랩(Camera-trap) 데이터를 이용하여 개체의 위치 정보를 통합함으로써 이러한 추정에 널리 활용된다. 본 연구에서는 설악산 인근 지역의 산양 데이터를 대상으로, 개선된 Reversible Jump MCMC(RJMCMC) 알고리즘을 적용하여 개체수를 추정하였다. 또한, 모형 적합 과정에서 발생하는 식별성(identifiability) 문제를 규명하고, 개선된 MCMC 샘플링 기법을 활용하여 다봉후분포(multimodal posterior) 문제를 완화하고자 하였다. 본 연구는 공간적 탐지 모형의 식별성 문제를 탐색적으로 고찰하고, 보다 안정적인 모수 추정에 기여하는 데 목적이 있다.

주요용어: Reversible jump Markov chain Monte Carlo, Spatial Capture-Recapture, Identifiability, Metropolis-Hastings

[a]서울대학교 통계학과. cowzin@snu.ac.kr
[b]서울대학교 통계학과. hih1210@snu.ac.kr
[c]서울대학교 통계학과. kevinjk02@snu.ac.kr
[d]서울대학교 통계학과. wcjang@snu.ac.kr

Extending Logistic PCA to MultiCategory Data with Data-Driven Factor Number Selection[a]

Yujin Jeong[b*] · Eun Ryung Lee[c]

Summary: Factor analysis reduces high-dimensional data into a few latent factors, but applying it to categorical responses is challenging due to their discrete and nonlinear structure. While prior latent-factor research has focused primarily on binary responses, we extend this modeling framework to multinomial and ordinal data, providing a unified approach for multi-category outcomes and incorporating an information-criterion method for automatic factor selection. By balancing model fit and complexity, the IC-based procedure helps prevent both under- and over-extraction of latent structure, improving the stability of factor determination and enhancing recovery of underlying loading patterns. Through simulation studies and real-data analyses, we confirm that the proposed approach offers consistent and reliable performance in both latent structure estimation and factor-number determination.

Keywords: Categorical data, Extending binary to multicategory modeling, Information-criterion-based factor selection, Latent structure estimation

[a]This work is supported by National Research Foundation of Korea grant funded by the Korean government (no. NRF2022R1A2C1012798, RS-2025-02216235).
[b]Graduate Student, Department of Statistics, Sungkyunkwan University
[c]Professor, Department of Statistics, Sungkyunkwan University

KNN Fused LASSO 기반의 시공간 분위수 회귀 (Spatiotemporal Quantile Regression)[a]

정지윤[b*] · 이은령[c]

요약: 도시 수요 데이터는 공간적 근접성과 시간적 변동성을 동시에 지니며, 이러한 특성을 반영하기 위해서는 공간적 상관과 시간적 연속성을 함께 고려한 분석이 필요하다. 본 연구는 분위수 회귀(Quantile Regression)의 강건성과 융합 라쏘(Fused LASSO)의 구조적 제약을 결합한 시공간 분위수 회귀 모형(Spatiotemporal Quantile Regression with KNN Fused LASSO)을 제안한다. 제안 모형은 공간적 인접 관계를 K-최근접이웃(KNN) 그래프로 구성하고, 시간적 순서를 체인 형태의 그래프로 표현하여 각각의 층(layer)에 융합 패널티를 부여함으로써 시공간적 매끄러움(spatial and temporal smoothness)을 동시에 추정한다. 또한 분위수 손실(pinball loss)을 사용하여 이상치나 비대칭 잡음 분포에서도 안정적인 추정을 가능하게 한다. 모수 추정은 ADMM(Alternating Direction Method of Multipliers) 알고리즘을 활용하여 공간 및 시간 제약항을 효율적으로 분리·갱신하며, 각 단계가 폐형식(closed-form)으로 계산되어 대규모 데이터에도 적용 가능하다. 모의실험 결과, 제안 모형은 기존의 평균 회귀나 공간 전용 모형보다 극단 분위수(예: τ=0.1, 0.9) 영역에서 우수한 강건성과 경계 복원 성능을 보였다. 특히 공간적 이질성이 존재하는 구간에서 블록 단위의 군집을 자동으로 형성하여 해석 가능한 시공간 구조를 복원하였다. 이를 통해 제안 모형이 기존 접근법 대비 극단 분위수 영역에서 더욱 향상된 추정 안정성과 구조 복원 능력을 지님을 확인하였다.

주요용어: 시공간 분위수 회귀, K-최근접이웃 융합 라쏘, 교대방향승수법(ADMM), 시공간 클러스터링

[a]This work is supported by National Research Foundation of Korea grant funded by the Korean government (no.NRF2022R1A2C1012798, RS-2025-02216235).
[b](03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과, 대학원생
[c](03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과, 교수

Sequence Kernel Association Test (SKAT) for Genetic Variant Identification Based on MMSE Scores

Jingyeong Jeong[a]* · Eunjee Lee[a]

Summary: Alzheimer's disease (AD) is a representative neurodegenerative disorder characterized by cognitive decline and memory impairment in older adults and is known as the leading cause of dementia. Although both environmental and genetic factors contribute to the onset of AD, conventional genome-wide association studies (GWAS) analyze single nucleotide polymorphisms (SNPs) individually, which limits their ability to capture the collective effects of rare variants or gene–gene interactions. These limitations contribute to the problem of "missing heritability," highlighting the need for advanced statistical approaches that can improve genetic explanatory power in complex trait analyses. Therefore, this study aimed to investigate the genetic association with cognitive function using Mini-Mental State Examination (MMSE) scores and to address the limitations of single-variant analyses by applying the Sequence Kernel Association Test (SKAT), which enables SNP-set based analysis. In this study, we utilized genomic and clinical data from the Alzheimer's Disease Neuroimaging Initiative 2 (ADNI-2) cohort. After performing quality control (QC) and genotype imputation, we secured a set of high-quality genetic variants. Linkage disequilibrium (LD) blocks were constructed to reflect LD structure, and SKAT was applied based on these blocks. As a result, significant association signals were observed on chromosomes 18 and 19, with a particularly notable LD block on chromosome 19 showing statistical significance at $p = 1.93 \times 10^{-6}$. This block was located near the GNA11 gene, which encodes a G-protein subunit involved in intracellular signal transduction. Previous studies have reported that GNA11 participates in vascular regulation and neuronal physiology, suggesting its potential role in the pathophysiology of AD. This study demonstrates the applicability of SKAT as a gene-set based statistical approach in genomic analysis and identifies genetic associations that may not be detectable through single-variant methods. Future work integrating multi-phenotype analysis and functional genomics is expected to further enhance the biological interpretation of these findings.

Keywords: SKAT, Alzheimer's disease, MMSE

[a]Department of information and statistics, Chungnam National University, Daejeon, Republic of Korea

The Effect of Optimization Methods on OOD Detection Performance Based on Uncertainty Measures[a]

Gun-Hak Jin[b*] · Hye-Young Jung[c]

Summary: This study analyzes the effect of optimization methods on uncertainty measures for out-of-distribution (OOD) detection in deep learning models. Sharpness-Aware Minimization (SAM) and Stochastic Gradient Descent (SGD) are compared to evaluate changes in calibration and feature variance across logit-based measures (MSP, Energy-based) and distance-based measures (Mahalanobis, k-NN, DDU). Experimental results show that SAM mitigates over-confidence in predictive probabilities, improving the Expected Calibration Error (ECE). However, feature variance increases, resulting in reduced stability in distance-based OOD detection. These findings suggest that the choice of optimization method significantly affects the behavior of uncertainty measures, revealing a trade-off between calibration improvement and feature-space stability.

Keywords: Sharpness-Aware Minimization, Out-of-Distribution Detection, uncertainty, Calibration

[a]This work was in part supported by the Korea Research Foundations, Korea, under grant KRF-2022 R1F1A1074939.

[b](Hanyang University ERICA Campus) 55, Hanyangdaehak-ro, Sangnok-gu, Ansan-si 15588, Korea. wlsrnsgkr@hanyang.ac.kr

[c](Hanyang University ERICA Campus) 55, Hanyangdaehak-ro, Sangnok-gu, Ansan-si 15588, Korea. hyjunglove@hanyang.ac.kr

Modeling Spatial and Temporal Patterns of Bovine Tuberculosis in Korea: A Negative Binomial and INLA-BYM Approach

Minkyoung Cha[a*]

Summary: Since no effective vaccine for bovine tuberculosis (bTB) has been developed, identifying the factors influencing its incidence is crucial for establishing effective control and prevention strategies. In Korea, variations in bTB infection rates have already been observed across administrative districts (Si/Gun/Gu). Therefore, a statistical disease mapping model that incorporates spatio-temporal correlations and relevant covariates has the potential to contribute to explaining these regional differences. This study aims to perform a spatio-temporal analysis using data from a total of 162 administrative districts, including Metropolitan Cities, from 2015 to 2023. To compare model performance, we conducted Negative Binomial (NB) regression and the Bayesian Integrated Nested Laplace Approximation (INLA) method. Both models include year and region as offsets or random effects to account for spatio-temporal variability. In particular, the INLA approach applies the Besag–York–Mollié (BYM) model, which captures both spatially structured and unstructured random effects.

Keywords: Spatio-temporal, Bovine tuberculosis, Risk factors

[a]Department of Statistics, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. chamink@ewha.ac.kr

Fisher Information and One-step Estimator for Multivariate Logistic Distribution[a]

JungJae-Choi[b*] · Hyoung-Moon Kim[c] · Hyeonwoo-Kim[c]

Summary: We investigate the Gumbel's type I multivariate logistic distribution with location-scale parameters. Exact closed-form expressions for the Fisher information matrix and its inverse are derived using Dirichlet-type II integrals and polygamma identities, with dominant convergence theorem. Based on these results, we propose a consistent method-of-moments estimator and develop a one-step Fisher–scoring estimator that is efficient. The approach provides a tractable alternative to full maximum likelihood while retaining asymptotic efficiency.

Keywords: One-step Estimator, Multivariate logistic distribution, Dirichlet integral, Fisher information matrix

[a]Hyoung-Moon Kim's research was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (RS-2024-00357199).
[b]Department of applied statistics, Konkuk University. boksilboksilchoi@naver.com
[c]Corresponding author: Department of applied statistics, Konkuk University. hmk966a@gmail.com, hmkim@konkuk.ac.kr

A Foundational Study for Explainable Graph Neural Network-based Prediction of Type 2 Diabetes and Gene Pathway Analysis[a]

Hojae Choi[bc*] · Sungkyoung Choi[c]

Summary: Type 2 diabetes (T2D) is a complex metabolic disorder driven by an interplay of genetic, environmental, and lifestyle factors, leading to impaired glucose regulation and increased cardiovascular risk. While genome-wide association studies (GWAS) have identified numerous genetic loci associated with T2D, these studies often overlook the complex gene–gene interactions and biological pathways that contribute to disease risk. To address this gap, we propose a foundational framework for a graph neural network (GNN)–based predictive model that integrates gene–gene interaction networks. We began by performing rigorous quality control (QC) on genotype data. We then conducted a GWAS using multiple statistical approaches—including logistic regression, score tests, and linear mixed models—to identify T2D-associated variants robustly. Moving beyond single-variant analysis, we performed gene-based tests to construct gene-level embeddings. These embeddings serve as the foundational input for the GNN to model complex gene–gene interactions. Finally, we performed functional enrichment analysis using Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) to identify key biological pathways implicated in T2D pathophysiology. Our multi-stage analysis provides a systematic approach for designing node embeddings that are both statistically and biologically informed for future GNN models. This integrative framework bridges the gap between traditional GWAS and network-based machine learning, laying the groundwork for improved genetic risk prediction and a deeper mechanistic understanding of T2D.

Keywords: Type 2 diabetes (T2D), genome-wide association study (GWAS), gene-based analysis

[a]This work was in part supported by the Korea Research Foundations, Korea, under grant RS-2025-25429679 This study was approved by the institutional review board of Hanyang University (IRB no. HYUIRB-202402-010).
[b]Department of Mathematical Data Science, Hanyang University, College of Computing, 55 Hanyang-daehak-ro, Sangnok-gu, Ansan 15588, South Korea. ghwo1015@hanyang.ac.kr
[c]Department of Mathematical Data Science, Hanyang University, College of Computing, 55 Hanyang-daehak-ro, Sangnok-gu, Ansan 15588, South Korea. day0413@hanyang.ac.kr

혼합효과 코시노르 모델을 이용한 CGMacros 데이터 혈당 리듬 분석

허성은[a*], 유재근[b]

요약: 본 연구는 PhysioNet에서 공개된 CGMacros 데이터셋을 활용하여, 자유생활 환경에서 수집된 연속혈당측정(Continuous Glucose Monitoring, CGM) 자료를 혼합효과 코시노르 모형으로 분석하였다. 분석은 리듬만을 포함한 기초 모형(M1), 시간가변 공변량을 추가한 모형(M2), 시간가변과 시간불변 공변량을 모두 포함한 모형(M3)을 단계적으로 확장하는 방식으로 진행되었으며, 리듬과 시간불변 공변량만을 고려한 비교 모형(M4)도 함께 포함하였다. 변수 선택은 전진 선택과 후진 제거 절차를 차례로 수행하였고, 모형 비교는 베이지안 정보 기준을 중심으로 하였다. 그 결과, 심박수와 활동량의 단기 변동, 식사 여부, 그리고 당화혈색소 수치가 혈당 리듬 설명에 핵심적인 역할을 하는 것으로 확인되었다. 최종적으로 M3 모형이 가장 낮은 BIC 값을 보여 최적의 모형으로 선택되었으며, 집단 수준에서는 심박수와 활동량의 변동이 평균수준과 위상에 즉각적인 영향을 주었고, 식사 여부는 평균수준보다는 진폭을 크게 확대시키는 효과가 나타났다. 또한 당화혈색소는 평균수준 상승과 위상 지연에 기여하였다. 개인 수준에서는 리듬 구조가 뚜렷한 참가자일수록 모형 설명력이 높았고, 변동성이 크거나 불규칙한 개인에서는 설명력이 낮게 나타났다. 본 연구는 혼합효과 코시노르 모형에 시간가변 요인과 시간불변 요인을 함께 고려한 혈당 리듬 분석의 구체적 사례를 제시한다.

주요용어: 연속혈당측정, 코시노르 모형, 혼합효과, 시간가변 공변량, 시간불변 공변량

[a](03760) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과, 석사과정. gj0830@ewha.ac.kr
[b](03760) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과, 교수. peter.yoo@ewha.ac.kr

하이브리드 베이지안 네트워크를 활용한 COPD 예측모형: 로지스틱 회귀 기반 CPT 결합

황시아[a*] · 오만숙[b]

요약: 만성폐쇄성폐질환(Chronic Obstructive Pulmonary Disease, COPD)은 국내외 사망과 의료비 부담의 주요 원인이지만 환자의 15%만이 진단을 받는 등 조기 선별이 미흡하다. 흡연·직업·성별 등 다양한 COPD 위험요인 간 인과관계를 파악하기 위해 생성 모델인 베이지안 네트워크(Bayesian Network, BN)가 널리 사용되지만, 판별 정확도는 로지스틱 회귀(Logistic Regression, LR) 같은 판별 모델보다 낮다는 한계가 있다. 위험요인 관계 분석이 질병 관리에 필수적이라 하더라도 선별 능력이 부족하면 임상 활용 가치는 제한적이다. 본 연구는 이러한 문제를 해결하고자 BN 구조는 유지하되 최종 COPD 노드의 조건부확률만 LR로 업데이트하는 하이브리드 베이지안 네트워크(BN-H)를 제안한다. 한국인 중년층을 포함한 KoGES 6차 조사자료(n=3,000)를 분석하였으며, 클래스 불균형(양성:음성≈1:8)은 ADASYN으로 보정하고 PC-stable 알고리즘으로 BN 구조를 학습한 뒤 COPD 노드를 LR로 대체하여 BN-H를 구축하였다. 모형 비교 결과 BN-H는 기존BN 대비 F1-score를 0.254에서 0.338로, Recall을 0.468에서 0.646으로 향상시켜 LR과 동등한 수준의 예측력을 확보하면서도 BN 고유의 인과 해석력을 유지하였다. 이러한 결과는 BN-H가 생성 모델의 설명 가능성과 판별 모델의 예측 성능을 간단히 통합함으로써 COPD 조기 선별의 정확도와 해석 가능성을 동시에 확보할 수 있음을 시사한다.

주요용어: COPD, 베이지안 네트워크, 생성 모델, 판별 모델

[a](03760) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 일반대학원 통계학과, 석사. siiaa@ewha.ac.kr
[b]교신저자, (03760) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과, 교수. msoh@ewha.ac.kr

Time Series Anomaly Detection Using Multi-scale Smoothing Residual Heatmap[a]

Won-Joon Hwang[b*] · Hye-Young Jung[c]

Summary: Various image-based approaches have been proposed for time series anomaly detection. Methods such as the Gramian Angular Field (GAF), Markov Transition Field (MTF), Recurrence Plot (RP), and Spectrogram transform one-dimensional time series into two-dimensional representations for deep learning models to recognize temporal patterns. However, these transformations often cause information distortion and fine-pattern loss, making it difficult to identify the timing and causes of anomalies in the original series. To address these limitations, this study proposes the Multi-Scale Smoothing Residual Heatmap (MSSRH). MSSRH applies multiple smoothing scales to the original time series and visualizes the residuals across the time–scale domain, minimizing information loss and enabling intuitive localization of anomalies. Experiments on ECG anomaly datasets show that the proposed method achieves performance comparable to existing image-based approaches while offering superior interpretability and visual clarity.

Keywords: Time Series Anomarly Detection, Time Series Imaging, Multi-Scale Smoothing Residual Heatmap

[a]This work was in part supported by the Korea Research Foundations, Korea, under grant KRF-2022R1F1074939.
[b](15588) Department of Mathematical Data Science, College of Computing, Hanyang University, 55 Hanyangdaehak-ro, Ansan, Republic of Korea. hwj1616@hanyang.ac.kr
[c](15588) Department of Mathematical Data Science, College of Computing, Hanyang University, 55 Hanyangdaehak-ro, Ansan, Republic of Korea. hyjunglove@hanyang.ac.kr

Detecting Nonlinear Relationships Using Distance Correlation and Optimal Transformation[a]

Sang-Min Park[b*] · Hyoung Moon Kim[c]

Summary: The Pearson correlation coefficient, a standard tool for measuring linear relationships, has clear limitations in detecting the complex nonlinear relationships often found in real-world data. To overcome this limitation, this study proposes and validates a novel analytical methodology that combines the Alternating Conditional Expectation (ACE) algorithm, a non-parametric transformation technique that flexibly learns the hidden structure of data, with distance correlation (dCor), which measures dependence regardless of the relationship's form.Through simulations assuming various nonlinear functions and error distributions that mimic the possibility of outliers, we confirmed that the ACE transformation significantly enhances the detection power of dCor. The improvement was particularly maximized when the error term followed a heavy-tailed or skewed distribution. Furthermore, the key contribution of this study is demonstrating that the dCor improvement metric (ΔdCor) functions not merely as a performance indicator, but as a diagnostic tool for identifying the presence of strong, symmetric nonlinear structures that are completely missed by the Pearson correlation. Analysis of real-world solar power generation data also proved that the proposed methodology is a powerful tool for successfully detecting and linearizing hidden nonlinear relationships.In conclusion, the ACE-dCor combination provides a reliable and effective analytical framework for discovering the hidden dependency structures between variables in complex data and understanding their characteristics.

Keywords: distance correlation, nonlinear dependency, ACE algorithm, detection power

[a]This work was in part supported by the National Research Foundation of Korea (NRF), grant funded by the Korean government (MSIT)(RS-2024-00357199).
[b]Data Science, Konkuk University, Seoul, Korea
[c]Corresponding author: Applied Statistics, Konkuk University, Seoul, Korea. hmk966a@gmail.com

2025년도 한국통계학회 동계 **학술논문발표회 프로시딩**

Proceedings of the KSS Winter Conference 2025

사단 법인 한국통계학회
The Korean Statistical Society