Book of Abstracts

Symposium on the Statistical Analysis of Distributional Data

21 August 2026

UNSW Sydney

Contents

Analyses of Contemporary Data Sets

Lynne Billard

University of Georgia, United States of America

Abstract

Contemporary data sets bring challenges, even if only to try to reduce them to manageable sizes. This usually requires some form of aggregation, with that aggregation producing observations as hypercubes instead of the traditional points in p-dimensional space. Early analyses used some form of classical surrogates; but we know these typically are unsatisfactory. More recent results have focused on using the complete information instead of the simpler quick-and-easy classical surrogate approaches. This review attempts to trace developments over the years and thence consider possible future directions.

Spatial Stochastic Frontier Model with Latent Classes

Kai (David) Du

University of Wollongong, Australia

Abstract

This paper proposes a spatial stochastic frontier model with latent classes to address two empirical regularities in efficiency analysis: spatial interdependence and unobserved technological heterogeneity. Existing approaches typically ignore spatial spillovers when modelling heterogeneity or assume homogeneous technology when accounting for spatial dependence. Our framework integrates a spatial autoregressive structure with latent class panel analysis, allowing for class-specific production frontiers, spatial lag parameters, and inefficiency distributions. We estimate the model using the Expectation-Maximization (EM) algorithm for computational stability and derive class-specific marginal effects decomposed into direct and indirect (spillover) components. Monte Carlo simulations confirm the estimator's finite-sample reliability, showing that parameter biases are small and diminish with sample size, while coverage probabilities cluster around the selected level of significance. An empirical application to productivity of 137 countries in 2000–2019 reveals that the inefficient class exhibits positive and well-identified spatial dependence ($\hat{\rho}_2 = 0.05$–0.26 across subsamples and specifications), reflecting greater reliance on technology adoption from neighbours than advanced economies. Economic freedom is strongly associated with latent class membership, and for developing economies, indirect spatial spillovers act as a performance drag.

Random Forests for Histogram-valued Responses with Symbolic Mixed-type Predictors

Jaejik Kim

Sungkyunkwan University, South Korea

Abstract

This study considers the problem of predicting a histogram-valued response from symbolic predictors represented by histogram- and/or modal-valued data. Histogram-valued observations may be obtained by aggregating classical data or may be observed directly in the form of histograms. Such observations provide useful information by characterizing the distributional properties of the underlying objects. To effectively predict histogram-valued responses, we propose a random forests approach, which is widely recognized for its strong predictive performance in machine learning. As the base learner, we first develop a regression tree designed for symbolic data by partitioning the predictor space using features extracted from the moments of individual histogram-valued observations, together with an efficient search algorithm for modal-valued observations. We further introduce a variable importance measure for symbolic predictors to facilitate the interpretation of the proposed random forests. The performance of the proposed method is evaluated through extensive simulation studies, and its practical usefulness is demonstrated through an application to real-world traffic volume data.

Symbolic Data Methods for Evaluating Center-Level Factors in Care Delivery Research

Jennifer Le-Rademacher

Mayo Clinic, United States of America

Abstract

Exponential smoothing (ES) often outperforms other techniques in time series forecasting across a wide range of data-generating processes. While ES has traditionally been applied to time series in $\mathbb{R}$, this paper extends the methodology to distributional time series, where each observation is a probability distribution on $\mathbb{R}$. The primary contribution of this work is twofold. First, we propose a principled and intuitive generalization of ES within the Wasserstein space, which retains the exceptional parsimony of classical ES. Second, we theoretically and empirically demonstrate that the smoothing parameter can be consistently estimated by minimizing a Wasserstein distance. Applications to distributional time series of high-frequency financial returns and household electricity demands confirm the practical effectiveness of our Wasserstein ES model.

Wasserstein Exponential Smoothing

Takuro Matsubara

University of Sydney, Australia

Abstract

Exponential smoothing (ES) often outperforms other techniques in time series forecasting across a wide range of data-generating processes. While ES has traditionally been applied to time series in $\mathbb{R}$, this paper extends the methodology to distributional time series, where each observation is a probability distribution on $\mathbb{R}$. The primary contribution of this work is twofold. First, we propose a principled and intuitive generalization of ES within the Wasserstein space, which retains the exceptional parsimony of classical ES. Second, we theoretically and empirically demonstrate that the smoothing parameter can be consistently estimated by minimizing a Wasserstein distance. Applications to distributional time series of high-frequency financial returns and household electricity demands confirm the practical effectiveness of our Wasserstein ES model.

Bayesian mixture models for ultrafine particle concentrations using histogram-valued data

Conor Murphy

UNSW Sydney, Australia

Abstract

Modern environmental monitoring produces increasingly large datasets for which fitting flexible Bayesian models to every observation can become computationally prohibitive. Motivated by the complex distributional behaviour of ultrafine particle number concentration measurements, we propose a Bayesian hierarchical framework that combines time-varying Gaussian mixture models with symbolic likelihoods based on histogram-valued data, enabling inference on the whole distribution utilising only distributional summaries for model fitting. A simulation study demonstrates that our proposed approach adequately recovers the underlying model while achieving substantial computational savings. The methodology is applied to the ultrafine particle number concentration measurements from the Queensland University of Technology UPTECH study, revealing temporal changes in multimodality that are not captured by models based solely on mean concentration.

Meta-analysis using the Ex-Gaussian Distribution

Jake Olivier

UNSW Sydney, Australia

Abstract

A common approach to meta-analysis is to combine study-level mean and variance estimates as a weighted average assuming the normal distribution with equal means but unequal variances. There are, however, examples where study-level data is skewed and an assumption of symmetry is likely violated. The ex-Gaussian, sometimes called the exponentially modified Gaussian, is a simple mixture of exponential and normal random variables which has absolute skewness in the interval $[0, 2]$. This distribution can be defined by its mean, variance and skewness parameters, or by the parameters of its constituent parts, i.e., exponential and normal parameters. For studies reporting mean, variance, skewness and sample size, we derive an estimator for the ex-Gaussian parameter vector and its asymptotic distribution. A meta-analysis example using summary statistics from 12 studies reporting optical refractive errors will be presented. Refractive errors are known to exhibit negative skewness, and the use of a skewed data-generating process is shown to be an improvement over the Gaussian assumption.

Heteroscedastic Interval-Valued Autoregressive Models

S. Yaser Samadi

Southern Illinois University, United States of America

Abstract

This study proposes a heteroscedastic interval-valued autoregressive model for dependent data observed as intervals, such as high–low financial prices, temperature ranges, or other temporally aggregated symbolic data. In many applications, the uncertainty represented by interval observations is not constant over time: financial ranges widen with market volatility, temperature ranges may vary seasonally, and aggregated measurements may display changing dispersion across time. The proposed heteroscedastic framework preserves the internal structure of each interval by representing interval observations through latent internal mean and variation components. To capture time-varying volatility, the innovation variance is modeled as a smooth function of time. Estimation combines interval-valued composite likelihood ideas with nonparametric kernel smoothing of residual interval variation, leading to an adaptive procedure for estimating the autoregressive dynamics under heteroscedasticity. The resulting framework provides a flexible approach for modeling interval-valued time series when both temporal dependence and changing volatility are present.

Visualizing and forecasting subnational life-table death counts: Gap forecasting methods

Hanlin Shang

Macquarie University, Australia

Abstract

Subnational life-table death counts are highly correlated across time and space and differ by gender. While these associations are helpful in improving forecasts through joint modelling, less attention has been paid to identifying and understanding mortality disparities in gender and regional gaps. We propose a forecasting framework to model and forecast female life-table death counts at the national or subnational level, and to model and forecast the associated gender gap. For either females or males, we could forecast national life-table death counts and the regional gap relative to national data. By combining gender and regional gaps, we also explore the double gap by prioritizing national data and female data. Life-table death counts are unique due to their non-negativity and summability constraints. To address the constraints, we apply a one-to-one transformation, termed cumulative distribution function transformation, to obtain one- to 15-step-ahead forecasts. Using Japanese age-specific life-table death counts between ages 0 and 110+ from 1947 to 2023, we evaluate and compare point and interval forecast accuracy across gender, region, and double gaps. By focusing on these gaps, we can deepen our understanding of the possible factors driving gender and regional mortality variations. (co-authored with Prof. Andrea Nigri, University of Foggia)

Spherically Embedded Time Series with Unknown Trend and Periodic Components

Jiaxen Xu

Macquarie University, Australia

Abstract

Spherically embedded time series with values naturally residing on or can be equivalently mapped to the sphere often exhibit complex non-stationarity driven by latent trend and periodic components. Traditional Euclidean methods fail to account for the sphere's intrinsic non-Euclidean geometry, leaving a critical gap in rigorous modeling and forecasting techniques. To address this methodological gap, we propose a unified geometric framework to analyze nonstationary spherically embedded time series. Central to our approach is a novel nonparametric spherical trend-periodicity decomposition model that uses an optimal-transport-based removal operation to sequentially extract the smooth trend and periodic components while preserving spherical topology. The resulting de-trended and de-seasonalized stationary residuals can be further modeled using a spherical autoregressive model, formalizing a novel trend-periodic spherical autoregressive model. Theoretical foundations for the modeling procedure are established on the consistency under temporal dependence. Extensive simulations corroborate these theoretical guarantees and demonstrate the superior finite-sample predictive performance of the trend-periodic spherical autoregressive model compared to existing spherical autoregressive models and neural network models. Finally, we validate the practical utility of our methodology through applications to electricity generation compositions and bike trip volume profiles, yielding significantly enhanced forecasting accuracy while providing interpretable insights into the underlying structural dynamics.

A Convex Hull Approach to H-Valued Interval-Valued Time Series

Atefeh Zamani

UNSW Sydney, Australia

Abstract

This presentation introduces a framework for analyzing H-valued interval-valued time series, where H is a general Hilbert space. The proposed approach is based on the concept of the convex hull and incorporates the distribution of internal points within each interval-valued observation. This allows the internal structure of interval-valued data to contribute directly to the analysis, rather than treating only the endpoints as informative. The framework is formulated in a general Hilbert space setting and therefore includes, as special cases, real-valued, two-dimensional, and functional random variables and time series.

Within this framework, we investigate how the distribution of internal points affects key moments and dependence characteristics of the process, including the mean, variance, and covariance structure. By introducing random convex combinations of the lower and upper bounds, we obtain a flexible representation that links the geometry of interval-valued observations to their probabilistic properties. The proposed methodology provides a unified way to study interval-valued data in finite- and infinite-dimensional settings and opens the way for more informative modelling and inference for interval-valued stochastic processes.