43rd Annual Meeting of Alberta Statisticians

Book of Abstracts

MacEwan University

Edmonton, Alberta, Canada

September 23, 2023

Contents

Saturday, September 23, 2023

University of Alberta

Yan Yuan, University of Alberta

Developing and Validating Risk Prediction Models for Premature Menopause in Female Childhood Cancer Survivors

Female childhood cancer survivors are at much higher risk of premature menopause (a.k.a. primary ovarian insufficiency - POI), which causes infertility and the accompanied wide range of post-menopausal symptoms. To counsel survivors, accurate risk estimates at different ages are needed. In this talk, I will discuss our recently completed study where we used data from 7891 participants in the childhood cancer survivors study to develop and 1349 survivors in the St Jude Life study (SJLIFE) to externally validate prediction algorithms based on age-specific logistic regression and XGBoost, as well as the added contributions of relevant polygenic risk scores (PRSs). Given cancer diagnosis and treatment information, these POI risk prediction algorithms perform excellently (AUROCs ranging from 0.91-0.96 in SJLIFE) and can inform decisions regarding fertility preservation interventions among female childhood cancer survivors.

Sharandeep Singh Pandher, University of Alberta

Shrinkage Methods in the Generalized Autoregressive Moving Average Models

In this study, we propose pretest and shrinkage methods for the estimation of regression coefficients of the generalized autoregressive moving average (GARMA) model, which are pervasive for modeling binary and count time series data. This model accommodates a set of covariates in addition to the ARMA parameters. We want to estimate regression and ARMA parameters when some of the regression parameters may belong to a subspace. We apply the maximum partial likelihood method to obtain the unrestricted maximum partial likelihood estimator(UMPLE) and also the restricted maximum partial likelihood estimator (RMPLE) for the model with parameter restriction and then present the improved pretest and shrinkage estimators. We establish the asymptotic distributional biases and risks of the proposed estimators and evaluate their relative performance with respect to the UMPLE. The methodology is investigated using simulation studies and then demonstrated using a real data example.

Alexander Melnikov, University of Alberta

Stochastic Regression Analysis via Martingale Approach

The talk has a goal to show how martingale methods are helpful in further developments of classical regression analysis. We will talk about a general regression model with martingale-like errors which works simultaneously in both discrete and continuous time. For such a model we will study the Least Squared estimates (LS-estimate). Their strong consistency will be derived from the strong Large Numbers Law (LNL) for martingales. Their asymptotic normality will be stated using a martingale version of the Central Limit Theorem (CLT). We pay attention to sequential LS-estimates with their fixed accuracy property. We explain how these estimates work for testing statistical hypotheses.

University of Calgary

Afrin Rumana, University of Calgary

Trends in Financial Markets: Uncovering the Distribution Of Intensity and Duration

In financial investment, market trends are ubiquitous. Put simply trending markets are characterized by changes in price that are persistent in time. In this proposed research, we are interested in understanding the global properties of trending markets ex-post, as there is a shortage of research in this direction. The primary goal of our study is to provide a reliable approach for categorizing financial market trends by defining their strength and persistence. However, the noisy characteristics of financial data and the hidden character of a true market trend make this endeavor nontrivial. Towards this end, We use resampling techniques and establish empirical labeling algorithms in parallel with Hidden Markov Models and Bayesian filtering to estimate the underlying structure and dynamics of market trends. From our results, we can comment on market trend intensity and duration across various financial markets and asset classes. Here, we focus on labeling trends —as opposed to identifying them in real-time—as this can provide valuable diagnostic information ex-post about how the macroeconomic conditions of the market influence the dynamics and characteristics of trends.

Danika Lipman, University of Calgary

A Bayesian Variable Selection Model for Semi-Continuous Response Using Gaussian Process

To my knowledge, there is not a statistical method that can perform variable selection in a setting where there is a semi-continuous response with a non-linear relationship to predictor variables. I have developed a two-part model to accommodate a semi-continuous response,

that uses Gaussian processes to capture the non-linear relationship between input variables and outcomes. To perform variable selection, Bayesian variable selection is induced in both parts of the model through the construction of the kernel matrices. I have employed the Nystrom approximation for kernel matrices to reduce the computational complexity that occurs when working with kernel matrices and large sample sizes. In this presentation I will present my model, an overview of the algorithm used for estimation, and results from simulation studies. The simulation studies demonstrate the time saving benefits of the Nystrom method, and the model's ability to select active variables from a large number of candidate features. From simulations, I determine my method is competitive in prediction and variable selection with methods such as elastic net, and other methods that capture non-linearity such as random forests, and gradient boosted trees.

Haixu Wang, University of Calgary

Nonlinear Prediction of Functional Time Series

We propose a nonlinear prediction (NOP) method for functional time series. Conventional methods for functional time series are mainly based on functional principal component analysis or functional regression models. These approaches rely on the stationary or linear assumption of the functional time series. However, real data sets are often non-stationary, and the temporal dependence between trajectories cannot be captured by linear models. Conventional methods are also hard to analyze multivariate functional time series. To tackle these challenges, the NOP method employs a nonlinear mapping for functional data that can be directly applied to multivariate functions without any preprocessing step. The NOP method constructs feature space with forecast information, hence it provides a better ground for predicting future trajectories. The NOP method avoids calculating covariance functions and enables online estimation and prediction. We examine the finite sample performance of the NOP method with simulation studies that consider linear, nonlinear and

non-stationary functional time series. The NOP method shows superior prediction performances in comparison with the conventional methods. Three real applications demonstrate the advantages of the NOP method model in predicting air quality, electricity price and mortality rate.

MacEwan University

Ryan Trang, MacEwan University

Learning Partial Differential Equations with fMRI Data

We propose a novel approach to investigate the existence of a lower dimensional, invertible representation of BOLD fMRI data that can be modeled as a partial differential equation (PDE). Non-linear dimensionality reduction is applied to obtain an embedding of the fMRI data before a transformer neural network is used to generate the solution of the PDE over the range of observed values in our dataset. Finally, we leverage symbolic regression (PDE-Find) to learn the governing PDE. We believe that a global model of the brain that can be derived end-to-end via machine learning, as a departure from conventional region of interest (ROI) based analysis, may provide additional insights into topics such as the study of functional connectivity, or group differences facing neuroimaging researchers.

David Thiessen, MacEwan University

Using Working Models to Improve Estimation Efficiency for Cox Regression with Missing Data

The unified approach (Chen & Chen, 2000) is a method to improve estimation efficiency in regression models with missing data by extracting information from individuals with partially observed variables. Working models are defined as nested submodels of the regression model and individuals with partially observed data can be used in estimating

these working models. Survival analysis is a natural application for the unified approach, as missingness in the response variable is handled by ordinary methods for censored data and nested models can easily be defined. In this talk, we review the application of the unified approach to the Cox proportional hazards model (Thiessen, Zhao, & Tu, 2022) and demonstrate the application with an experimental R package.

Cristina Anton, MacEwan University

Cluster-Weighted Models for Functional Data

We present a family of parsimonious cluster-weighted models. The parameters of these models are estimated using an expectation maximization (EM) algorithm. We apply the proposed method for simulated and real-world data.