STEM
STEMS
STEAM
STEM H
STEM Topics in Data Science with Applications

Norou Diawara
Mathematics and Statistics Department, ODU
REYES, July 15th, 2020
STEM is worldwide
- Studied at University Cheick Anta Diop
- Then Universite du Havre, in France
- Then Auburn University
- Offered teaching position @ Old Dominion University
What is needed to succeed in a STEM field?

- CRITICAL THINKING
- PROFICIENCY IN SOLVING NON-ROUTINE PROBLEMS
- TEAMWORK
Why is STEM needed?

- Future
- Environmental
- Health
- Economics
- Behavior
- Accountability
- International Community

SHORTAGE OF NURSES, DOCTORS, ENGINEERS, MATHEMATICIANS, STATISTICIANS, DATA SCIENTISTS

STEM JOBS WILL INCREASE BETTER PAY THAN EVER

ROBOTS WILL SOON TAKE OVER MANY AUTONOMOUS JOBS

IMPROPER USE OF TECH CAN LEAD TO DISASTER

SHORTAGE OF NURSES, DOCTORS, ENGINEERS, MATHEMATICIANS, STATISTICIANS, DATA SCIENTISTS
Resources?

Grants for STEM

• Advancing Informal STEM Learning (AISL) National Science Foundation. ...

• Science Education Partnership Award (SEPA) National Institutes of Health. ...

• Bay Watershed Education and Training Program (B-WET) ...

• Innovative Technology Experiences for Students and Teachers (ITEST) ...

• Laboratory Equipment Donation Program (LEDP)

• Universities and Colleges
STEM

- Science
- Technology
- Engineering
- Math
- Statistics & Data Science

STEM → STEMS
STEM → STEAM

- Health
- Arts

STEM-H
STEAM
Numbers and data

Prime Numbers
Data Functions
Algorithms
Numbers and data

Data & design

R. A. Fisher: capture variations, causality
Blocking, Randomization, Replication, and Resampling

\[ Y = \beta + \gamma + \beta \gamma + \epsilon \]

Moran, 1950
Martin & Opepen, 1975
Stoyan & Stoyan, 1998
Møller & Díaz-Avalos, 2009
Vaillant, et al., 2011

A point process is a random point field with each point representing an event of interest.

Mathematician

Connie Palm
Palm calculus

Interested in studying the properties of a chosen point of it
- conditional probabilities must be calculated
- the conditional distribution provides a flexible tool in patterns

Kathrine Johnson

Methed & Hicke, 2014
Neuroscience Division
@ Ely Lilly
Center for Biomedical Imaging
Data @ Emory University
Problem of Interest

We consider two marginally gamma distributions $X$ and $Y$. Put a linear relation thru a latent random variable (r.v.) $Z$ independent of $X$.

More precisely, set

$$Y = aX + Z, \quad a > 0.$$  \hspace{1cm} (1)

Applications

- Medical field
- pair of organs
- Engineering
- failure of second component initiated by the first
The distribution of $Z$ takes different forms and contains information about the relationships between different variables' values.

1. Z is continuous gamma only when $a$ takes a special form.
2. The expressions for $Z$
   - $\alpha_1 = \alpha_2 = \alpha \in \mathbb{N}$,
   - or $\exists \alpha_1 \leq \alpha_2$,
   - or $\not\exists \alpha_1 < \alpha_2$ with $[\alpha_1 + 1] = N \leq \alpha_2$. 
Forecasting under Structural Changes in Non-linear Time Series

**Motivation**

- Partially observed time series models are studied under various conditions, e.g., State Space Models (Durbin and Koopman, 2001), Dynamic Models (West and Harrison, 1997), and Hidden Markov Models (Cappe et al. 2005).
- Challenging: Formulation of conditional distribution and reconstruction of the unobserved components from the observed values.
- More challenging: if the variance is not constant, changing over intervals.
- Common in financial, economic, and environmental time series data.
Forecasting

- Forecasting for multiple regression where different covariates have different structures (Gagnon and Kedem, 2010); Forecasting based on Bayesian approach assuming possibility of breakpoints in the forecast horizon (Pesaran, Pettenuzzo and Timmermann, 2005).
- We developed the method based on joint density of mixture distribution of combined error data.
- Parameters are estimated by using EM algorithm.
- Estimation is improved by using bootstrap combined with EM algorithm.

Parameter Estimation Using EM Algorithm

General case: Mixture of \( r \) density functions

Let \( \xi \sim f(\xi; \mathbf{p}) = \sum_{j=1}^{r} \rho_j f_j(\xi) \)

\[ \log L(\mathbf{p}; \xi) = \prod_{i=1}^{n} f(\xi_i|\mathbf{p}) = \sum_{i=1}^{n} \log \left( \sum_{j=1}^{r} \rho_j f_j(\xi_i) \right) \quad \text{-> Not solvable in closed form} \]

Define

\[ \mathbf{w} = (w_j, i = 1, \ldots, n; j = 1, \ldots, r) = \begin{cases} 1, & \text{observation } \xi_i \text{ comes from the distribution } f_j \\ 0, & \text{else} \end{cases} \]

Log-Likelihood for complete data:

\[ \log L_{\mathbf{c}}(\mathbf{p}; \xi, \mathbf{w}) = \log \prod_{i=1}^{n} \prod_{j=1}^{r} (\rho_j f_j(\xi_i))^w_j = \sum_{i=1}^{n} \sum_{j=1}^{r} w_j \log \rho_j + \sum_{i=1}^{n} \sum_{j=1}^{r} w_j \log f_j(\xi_i) \]

Conditional Expectation \( E_{\xi_{(k)}}(w_j|\xi) = P_{\xi_{(k)}}(w_j = 1|\xi) = w_j^{(k)} = \frac{p_{(k)}^{(j)}f_j(\xi)}{\sum p_{(k)}^{(j)}f_j(\xi)} \).
Spatio-temporal models

- **Goal:** model for random patterns of points in a high dimensional space (locations) where events of interest (may them be cyber-attacks, contamination sources, etc...) are observed, are accounted for under such processes, with addition of time frequencies.

- **Which measure?** Moran’s statistics

Figure 1: Regular space with 16 subareas

Figure 2: Poisson point process
Spread under regular spacing

Figure 3: Apparent spatial clustering of regular space (16 subareas) through $t = 6$
Evolution of system

Figure 4a: Poisson point process plots at (a) $t = 1$ and (b) $t = 2$

Figure 4b: Poisson point process disc 1 at $t = 3$

Figure 4c: Apparent spatial clustering through $t = 4$
COVID-19 new cases

Figure 5: Line plots of new cases per block
COVID-19 new cases

Figure 6: Shapes / Colors of new cases per block
The use of Markov Decision Processes (MDPs) allows us to project the future discounted utilities for the different attribute-level best-worst pairs. Data is simulated from a real-world example and the discounted utilities for the different attribute-level pairs are projected across five time periods.

Patterns in the discounted utilities suggest that some choice pairs will continuously yield high returns while others yield negative returns. We find that the model of individual choice preferences yields to a balance from several factors/controls at hand.
Other Applications

• Driving safety: the seat belt data
• USAID data (finance)
• Brain functions
• Calculating patterns of certain pandemics
• And more
Seat Belt Use in Virginia

Predicting Seatbelt Use under a Higher Order Structure Design in 2017-2019 Survey Design

Weighted Item Response Theory Model
Figure 7a. Comparisons of seat belt use by year for males and females.

Figure 7b. Comparisons of seat belt use by year for males.

Figure 7c. Comparisons of seat belt use by year for females.
REVIEW OF USAID/DCA DEFAULT HISTORY AND DEVELOPMENT OF DEFAULT ASSUMPTIONS

The data for this project has been provided by USAID’s Development Credit Authority (DCA). It is a panel data set from more than 73 countries and consists of 56 variables in 548,309 data rows, from 1992 to 2012.

\[
\text{Default rate} = \frac{\text{Adjustment Factor} \times \text{Claim amount}}{\text{Transaction Amount}}.
\]
WARF Score, Event Period, Loan Duration and Origination Fee Percent, and then

\[ r = \beta_0 + \beta_e e + \beta_w w + \beta_l l + \beta_f f + \varepsilon \]
Profile plot of default rate with respect to Event Period (e) and WARF Score (w)

Three-dimensional plot of the default rate based on WARF Score (w) and Event Period (e)
Maps of the region showing the key pairwise differences between countries.

The maps were produced in QGIS, an open source GIS software equivalent to ArcGIS. Colors were added to the maps using Microsoft Word.

Figure 4: USAID map for all countries as per https://stories.usaid.gov/usaidmap/?location=africa

https://stories.usaid.gov/usaidmap/?location=africa
Brain image and functions

Figure 9a: Functional near-infrared configuration

Optodes 1, 2, and 15:
one step (up, down, left, right)

Geographical weights:

\[ w_{ij} = e^{-d_{ij}/d} \]

\[
\begin{bmatrix}
4 & 1 & 7 \\
1 & 4 & 8 \\
7 & 8 & 4
\end{bmatrix}
\]
Figure 9b. Moran statistic for Grouping 1 for HbO, HbR, and HbT by stage for days 1 to 12.
Science really rocks

We need each of you
Thank you

Thanks to Dr. González
Thanks to REYES organizers

Support STEM

Questions?
Comments?
Thoughts?
Useful links

• Maple and Mathematica:
  – www.maplesoft.com

• R and SAS Software (free options)
  – https://www.sas.com/
  – http://www.r-project.org/

• Data science challenge: Kaggle
  – www.kaggle.com