Bayesian Change Point Detection for Mixed Data with Missing Values

Alexander C Murph, Curtis B Storlie
University of NC at Chapel Hill, Mayo Clinic Rochester Department of Quantitative Health Sciences

ABSTRACT

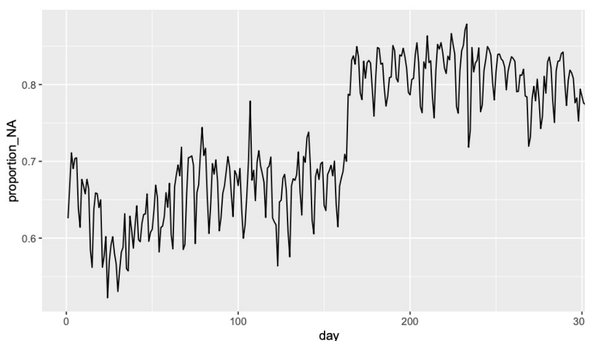
BACKGROUND

When a predictive model is in production, it must be monitored over time to ensure that its performance does not suffer from drift or abrupt changes to data. Typically, this is done by evaluating the algorithm's predictions to outcome data and ensuring that the algorithm maintains an acceptable level of accuracy over time. However, it is far preferable to learn about major changes in the input data that could affect the model's performance in real-time, long before learning that the performance of the model itself has dropped by monitoring outcome data. Thus, there is large need for robust, real-time monitoring of high dimensional input data over time.

OBJECTIVES AND METHODS

Here we consider the problem of change point detection on high-dimensional longitudinal data with mixed variable types and missing values. We do this by fitting an array of Mixture Gaussian Graphical Models to groupings of homogeneous data in time, called regimes, which we model as the observed states of a Markov process with unknown transition probabilities. The primary goal of this model is to identify when there is a regime change, as this indicates a significant change in the input data distribution. To handle the messy nature of real-world data which has mixed continuous/discrete variable types, missing data, etc., we take a Bayesian latent variable approach. This affords us flexibility to handle missing values in a principled manner, while simultaneously providing a way to encode discrete and censored values into a continuous framework. We take this approach a step further by encoding the missingness structure, which allows our model to then detect major changes in the patterns of missingness, in addition to the structure of the data distributions themselves. We assess our approach on simulated data and apply it to an in-production model for the need for a palliative care consult at Mayo Clinic Rochester.

FIGURE 1: Observed Proportion of Missing Values in Lipase



Proportion of missing values observed in the Lipase variables over 300 days in 2020-2021 at the Mayo Clinic in Rochester. In addition to shifts in the data distribution, regime changes can be driven by changes in the missingness structure. A regime change occurred around day 160.

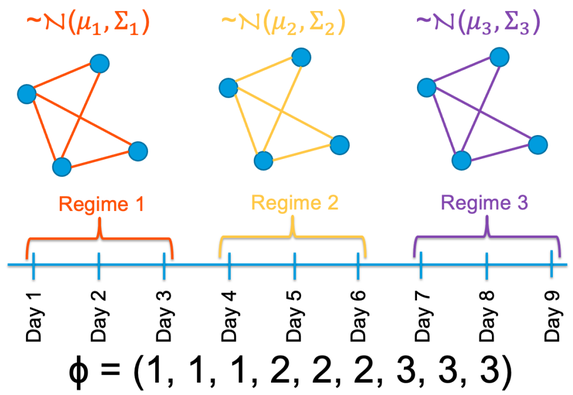
Thank you to the Mayo Kern Center for the Science of Health Care Delivery, and to the RC Bose Travel award at UNC Chapel Hill for conference funding!

MODEL

*WE LEARN AN ARRAY OF **RANDOM GRAPHS** ASSIGNED TO REGIMES*
*DATA ARE THEN DISTRIBUTED BASED ON THEIR **REGIME ASSIGNMENT***

- We take a fully Bayesian approach to learning a vector of *regime assignments* ϕ, to which we assign a prior according to a Markov Process with unknown transition probabilities. Samples from this vector are simulated via a Merge-Split-Swap algorithm similar to [1].
- The data *within a regime* is modeled according to a *Mixture Gaussian Graphical Model* (Mixture GGM), where the unknown graph structure is assumed static across regimes.
- Mixture GGM parameters, Markov process transition probabilities, and regime assignments all learned under a Gibbs framework.
- Graph structures across regimes are learned with a Double Reversible Jump Metropolis-Hastings Algorithm [2].

FIGURE 2: Example of Model Fit



Example of a single fit of our model, with three regimes over the course of nine days. A Mixture GGM is fit to each regime. The regime vector parameter ϕ encodes the regime structure of this fit.

MESSY DATA?

NO PROBLEM.

OUR APPROACH ANTICIPATES MAJOR MODERN DATA CHALLENGES FACED IN REAL-WORLD APPLICATIONS

DATA ISSUES AND MODEL CHALLENGES	SOLUTIONS
• Missing Data • Mixed Data (discrete, binary, continuous) • Censored Values	Bayesian Latent Variables
• HUGE data (n in millions, p ~= 250)	Sparsity Assumption on Mixture GGM
• Lack of information for some priors	Bayesian Hierarchical Modeling
• Learning parameters for a GGM	Double Reversible Jump Metropolis Hastings, Conjugate priors.

SIMULATION RESULTS

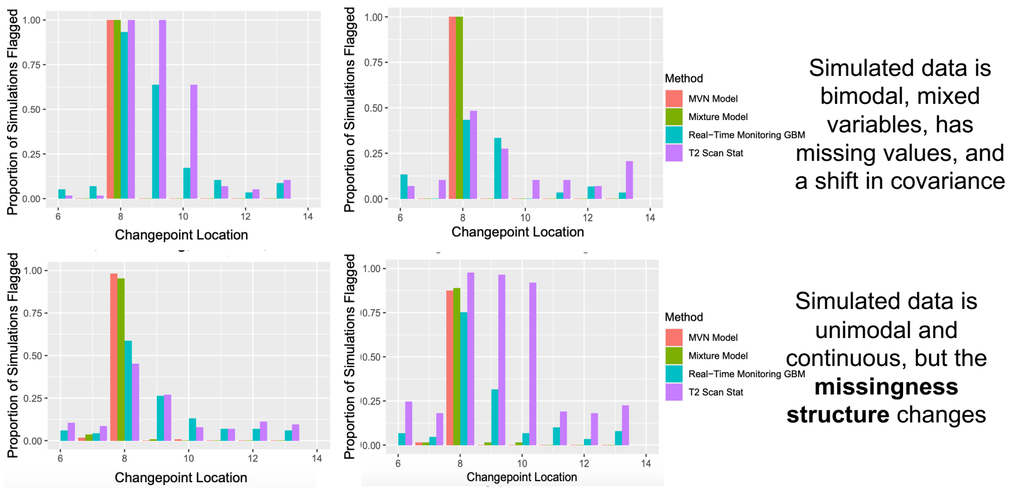
FIGURE 3: Simulation Study of Changepoint Model vs. Hotelling T^2 Scan Statistic

Simulated data is unimodal, with cont. variable types, missing values, with a mean shift

Simulated data is bimodal, mixed variables, has missing values, and a shift in covariance

Simulated data is unimodal, mixed variables, has missing values, and a shift in covariance

Simulated data is unimodal and continuous, but the **missingness structure** changes

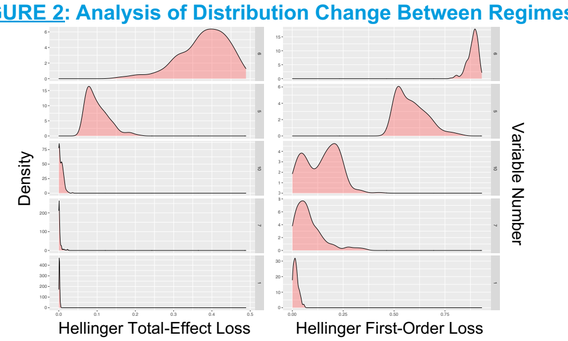


Proportion of times over 300 simulated data sets where a changepoint was located; the true changepoint is at timepoint 8. We compare our Changepoint model with regular GGMs, our method with Mixture GGMs, and a Hotelling T^2 Scan statistic with a 3-day memory.

POSTERIOR ANALYSIS

*ONCE A **CHANGEPOINT IS LOCATED:***
*HOW DO WE DETERMINE **WHAT CHANGED?***

FIGURE 2: Analysis of Distribution Change Between Regimes



Total-Effect and First-Order Loss of Hellinger Distance between distributions before and after known changepoint. The only change in the true data distributions is a mean shift in only variables 5 & 6, which this method correctly identifies.

Dist'n Before: X	Marginal Dist'n **w/o** Variable i: X \ i	Total-Effect Loss of i: 1 – H(X \ 1, Y \ 1)/H(X,Y)
Dist'n After: Y		
Hellinger Distance: H(X,Y)	Marginal Dist'n **of** Variable i: X : i	First-Order Loss of i: H(X : i, Y : i)/H(X,Y)

CONCLUSIONS

- Our method **outperforms the Hotelling T^2 Scan Statistic** both in terms of True Positive and False Negative Rates.
- After determining a changepoint, our Posterior Analysis identifies ***in what way*** the data changed.
- Now that the performance of this method has been strongly verified, it will be applied to in-practice models at the Mayo Clinic in Rochester. A preliminary implementation will be to the model in [3].
- Our method identifies changes in a dataset's ***missingness structure*** and can therefore **find changes not directly tied to the distribution of the data**.

REFERENCES

1. Martínez, A. F. and Mena, R. H. (2014) On a Nonparametric Change Point Detection Model in Markovian Regimes. *Bayesian Analysis*, 9, 823 – 858
2. Lenkoski, A. (2013) A direct sampler for g-wishart variates. *Stat*, 2, 119–128.
3. Murphree, D. H., Wilson, P. M., Asai, S. W., Quest, D. J., Lin, Y., Mukherjee, P., Chhugani, N., Strand, J. J., Demuth, G., Mead, D., Wright, B., Harrison, A., Soleimani, J., Herasevich, V., Pickering, B. W. and Storlie, C. B. (2021) Improving the delivery of palliative care through predictive modeling and healthcare informatics. *Journal of the American Medical Informatics Association*, 28, 1065–1073.

CONTACT

Alexander C Murph
University of NC at Chapel Hill
acmurph@live.unc.edu