Greater Data Science Cooperative (GDSC)

Cornell University & University of Rochester TRIPODS

gdsc.cornell.edu

- PIs:
  - David S. Matteson (Stat, CU)
  - Mujdat Cetin (EE, UR)
- CoPIs:
  - Aaron Wagner (EE, CU)
  - Alex Iosevich (Math, UR)
  - Daniel Gildea (T-CS, UR)
  - Daniel Stefankovic (T-CS, UR)
  - David Bindel (T-CS, CU)
  - Gennady Samorodnitsky (Math, CU)
  - Tongtong Wu (Stat, UR)
  - Qing Zhao (EE, CU)

- Topological Data Analysis
- Data Representation
- Network & Graph Learning
- Decisions, Control & Dynamic Learning
- Diverse & Complex Modalities
- *COVID-19 working group*

- Research Studios and Workshops
- Machine Learning in Medicine (Virtual) Seminars
- Machine Learning in Medicine 2020 Symposium
- Rochester (Area) Data Science Consortium
- Healthcare Data Science Modules
- Research Experiences for Undergraduates (virtual)
- New Journal: Data Science in Science
- Collaboration with other HDR institutes
  - PRISM for Transdisciplinary Systemic Risk (sites.google.com/view/prism-prj)
  - Atomic Level Structural Dynamics in Catalysts (alsdcgroup.wordpress.com)

***MISSION:** Motivated by today's greatest foundational data science challenges arising in medicine, healthcare, and beyond, our vision is to develop a mathematical foundation that integrates trans-disciplinary perspectives and enables application that can ultimately benefit everyone worldwide.*

- *Contact: matteson@cornell.edu*

HDR TRIPODS

GREATER DATA SCIENCE COOPERATIVE (GDSC)

A ROCHESTER & CORNELL COOPERATIVE INSTITUTE

Cornell University | University of Rochester

gdsc.cornell.edu

NSF Awards 1934985 & 1934962

June 2021. Contact: matteson@cornell.edu

MISSION

Motivated by today's greatest foundational data science challenges arising in medicine, healthcare, and beyond, our vision is to develop a mathematical foundation that integrates trans-disciplinary perspectives and enables applications that can ultimately benefit everyone worldwide.

Research Focus

(i) Topological Data Analysis. The challenges that high-dimensional, incomplete, and noisy data present are great, but in many applications, exploiting the topological nature of the problem is possible. GDSC aims to develop new fundamental methods and theory to rigorously explore the promise of this unique approach.

(ii) Data Representation. Data compression, embeddings, and dimension reduction play a fundamental role in data science. Inspired by new core challenges in biomedical imaging, genomics, and neural-spike training data, GDSC aims to develop novel source models and distortion measures, and ultimately seek a unifying theoretical framework across domains and disciplines.

(iii) Network & Graph Learning. Many of the fundamental challenges in applying data science to non-homogeneous populations are best explored through a network or graph structure. GDSC aims to develop new techniques for parameter-dependent eigenvalue problems in spectral community detection, density-estimation methods on networks, and a theoretical framework for time-varying graphical models to study dynamic variable relations in time-evolving networks.

(iv) Decisions, Control & Dynamic Learning. Sequential decisions are high-stakes in medicine. GDSC aims to utilize systems and control-engineering methods to improve health and disease management and develop new foundational theories and methods for label-efficient active learning and dynamic treatment regimes.

(v) Diverse & Complex Modalities. Big data is complex data, and major new innovations are needed. GDSC aims to develop theoretical frameworks for inference under computational and privacy constraints and for high-dimensional data without parametric model assumptions. Text, image, and audio data present further challenges. To address such challenges, GDSC aims to explore transition systems for graph parsing of natural language and new fusion approaches for fully multimodal analysis.

GDSC Leadership

- PIs:
  - David S. Matteson (Stat, CU)
  - Mujdat Cetin (EE, UR)
- CoPIs:
  - Aaron Wagner (EE, CU)
  - Alex Iosevich (Math, UR)
  - Daniel Gildea (T-CS, UR)
  - Daniel Stefankovic (T-CS, UR)
  - David Bindel (T-CS, CU)
  - Gennady Samorodnitsky (Math, CU)
  - Tongtong Wu (Stat, UR)
  - Qing Zhao (EE, CU)

Founding GDSC Key Personnel

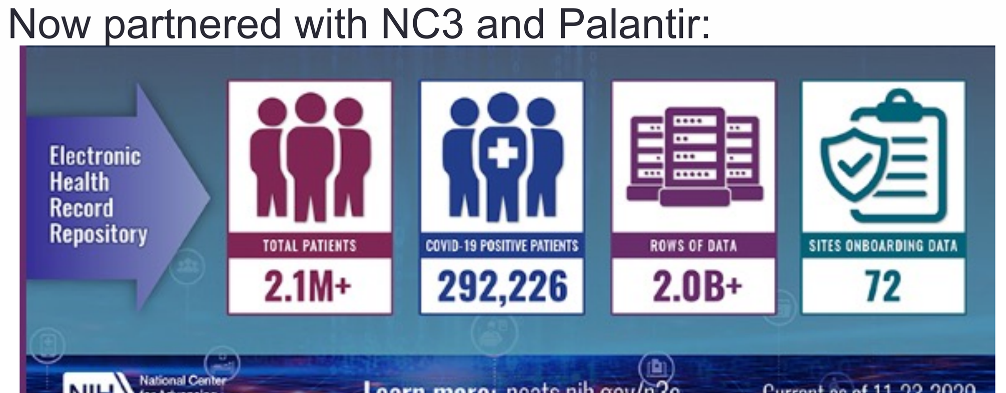
- Rochester:
  - Ajay Anand
  - Andrew McDavid
  - Beilei Xu
  - Chenliang Xu
  - Edgar Bernal
  - Gonzalo Mateos Buckstein
  - Gaurav Sharma
  - Robert Strawderman
  - Wendi Heinzelman
- Cornell:
  - Samprit Banerjee
  - Joe Guinness
  - Mert Sabuncu
  - Jayadev Acharya
  - Chris De Sa
  - James Booth
  - David Ruppert
  - *Mahsa Shoaran*

GDSC Research Associates & Collaborators

- Toryn Schafer
- Michael Jauch
- Sean Ryan
- Marie Duker
- Andrew Thomas
- Elaine Hill
- Victor Hernandez
- Dongmei Li
- Zhengwu Zhang

COVID-19 Working Group

- Wagner, A. B., Hill, E. L., Ryan, S. E., Sun, Z., Deng, G., Bhadane, S., Martinez, V. H., Wu, P., Li, D., Anand, A., Acharya, J., & Matteson, D. S. (2020). **Social distancing merely stabilized COVID-19 in the US**. *Stat (International Statistical Institute)*, e302. Advance online publication. https://doi.org/10.1002/sta4.302
  - Now partnered with NC3 and Palantir:

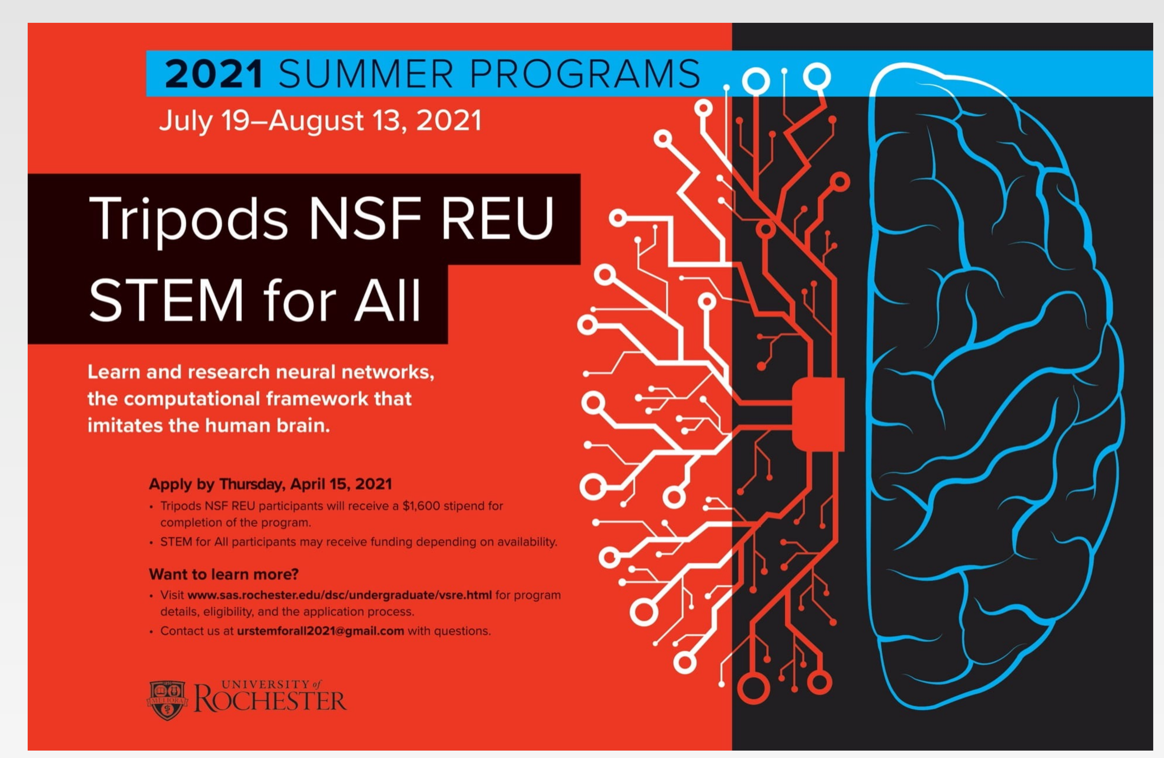


GDSC: Machine Learning in Medicine

- Machine Learning in Medicine (Virtual) Seminars
  - Monthly Series
  - Recruiting Speakers

- Machine Learning in Medicine 2021 Symposiums:
  - Virtual symposium in January 2021
  - In-person symposium in Oct/Nov 2021
  - @ Weill Cornell Medicine

GDSC: Grad for All 2020 & 2021



- Empower & inspire all interested students from Western New York area to pursue advanced degrees.
- Targeting traditionally under-represented STEM groups.
- Advising, information, skills and training needed to succeed in graduate school, academic careers, and industry.
- Coursework, Research and Mentoring.
- Year 1 program simplified demographic breakdown: *8 Women, 2 Hispanics, 1 African-American, 7 Cornell, 6 UR, 1 Geneseo CC, 1 Rochester Institute of Technology.*
- https://web.math.rochester.edu/people/faculty/iosevich/stemforall2020.html

Additional GDSC Activities

- Postdoctoral Researchers and Graduate Students
- Research Workshops
  - Annual research conferences (SciML in 2021)
  - Annual GDSC research "studio"
  - MLIM++
- REU: Grad for All (Summers)
- Rochester (Area) Data Science Consortium
- Healthcare Data Science Modules
- Teach-the-Teacher: High School Data Science Outreach
- Rotating Research Short Course
- CAMSAP'19 tutorial: Connecting The Dots: Identifying Network Structure Of Complex Data Via Graph Signal Processing
- University Rochester Goergen Institute for Data Science
- Cornell Center for Data Science for Enterprise & Society.

Selected Research Highlights

- Blanca, A., Chen, Z., Stefankovic, D., and Vigoda, E. (2020). Hardness of Identity Testing for Restricted Boltzmann Machines and Potts models. Proceedings of Machine Learning Research, vol 125, pages 514-529. PMLR.
- Davidow, M. and Matteson, D. S. (2020). Factor analysis of mixed data for anomaly detection. preprint arXiv:2005.12129.
- Ekmekci, C., & Cetin, M. (2021). Model-Based Bayesian Deep Learning Architecture for Linear Inverse Problems in Computational Imaging.
- Frank, A.-S. J., Matteson, D. S., Solvang, H. K., Lupattelli, A., and Nordeng, H. (2020b). Extending balance assessment for the generalized propensity score under multiple imputation. Epidemiologic Methods, 9(1).
- Gelsinger, M. L., Tupper, L. L., and Matteson, D. S. (2019). Cell line classification using electric cell-substrate impedance sensing (ecis). The International Journal of Biostatistics, 16(1).
- McDavid, A., Corbett, A. M., Dutra, J. L., Straw, A. G., Topham, D. J., Pryhuber, G. S., ... & Holden-Wiltse, J. (2021). Eight practices for data management to enable team data science. Journal of Clinical and Translational Science, 5(1).
- Saboksayr, S. S., Mateos, G., & Cetin, M. (2021). Online Discriminative Graph Learning from Multi-Class Smooth Signals. preprint arXiv:2101.00184.
- Tang, B. and Matteson, D. S. (2021). Graph-based continual learning. ICLR 2021 preprint arXiv:2007.04813.
- Wu, H., & Matteson, D. S. (2020). Adaptive Bayesian Changepoint Analysis and Local Outlier Scoring. preprint arXiv:2011.09437.
- Zhang, W., Grin, M., and Matteson, D. S. (2020). Modeling nonlinear growth followed by long-memory equilibrium with unknown change point. preprint arXiv:2007.09417.

Drift vs Shift: Decoupling Trends & Changepoint Analysis

David S. Matteson, with Haoxuan Peter Wu & Sean Ryan
Cornell University (TRIPODS w/URochester) & the National Institute of Statistical Sciences (NISS)

Introduction

- **Goal**: Distinguish global/macro patterns from local/micro fluctuations
- 'Drift' describes the micro-level evolution of a process. This may appear as variation about gradual trends.
- 'Shifts' refer to discontinuities, rapid changes, or major breaks in trend. These represent macro-level changes in a process.
- Both might be mechanistically or stochastically generated and/or modeled. However, causes of shifts are typically different from those of drift.
- While understanding such differences is a prime objective, this first requires distinguishing: **Drift vs Shift**.

Tools include:

- Trend Filtering
- Stochastic Volatility
- Outlier Detection
- Dynamic/Adaptive Shrinkage
- Dynamic Linear Models (DLM)
- Change Point Analysis
- Bayesian (Time Series) Analysis
- Machine Learning (Regularization)

Challenges

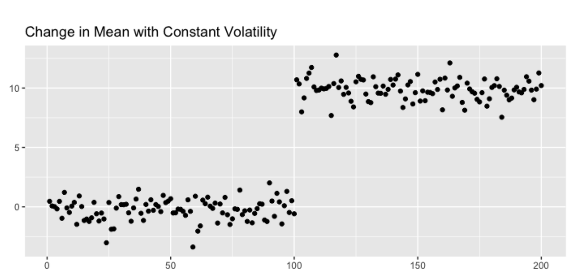


Fig 1 Change in Mean

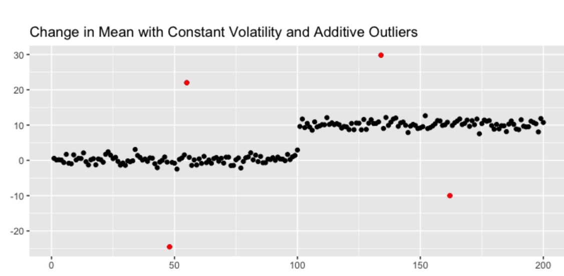


Fig 2 Adaptive Outliers

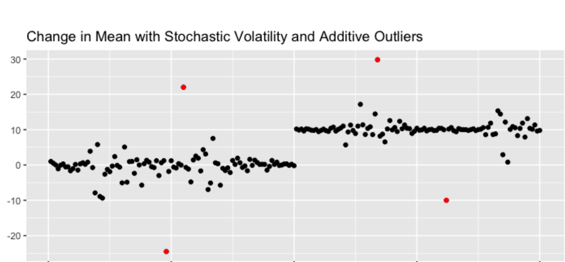


Fig 3 Stochastic Volatility

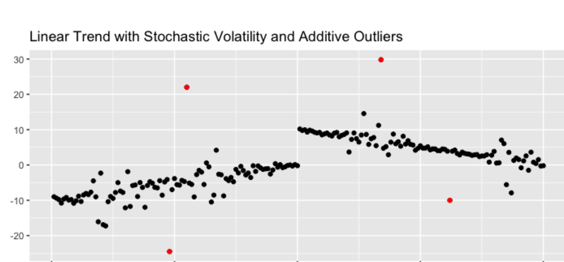


Fig 4 Linear Trend

- **Outliers** violate common Gaussian noise assumptions.
- **Heterogeneity** leads to over-prediction of changepoints.

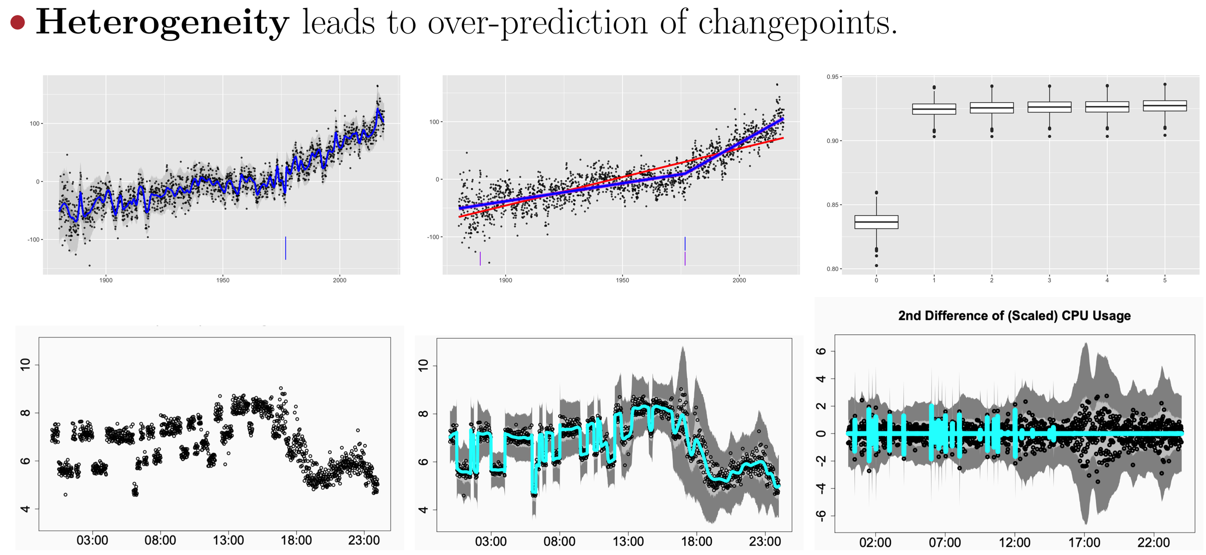


Fig 5 Global land surface air temperature (top); CPU cloud usage (bottom).

- Real world data has complex patterns and trends.
- Outliers and heterogeneity are the norm.
- Nature of changepoints ambiguous.

Solutions

- Model based ABCO: Adaptive Bayesian Changepoints w/ Outliers[1].
- A two-step Bayesian 'decoupling' method developed via DLM[2].

ABCO Model

Given a time series $\{y_t\}$, ABCO supposes the decomposition:

$$y_t = \underset{\substack{\text{mean}\\ \text{signal}}}{\beta_t} + \underset{\substack{\text{additive}\\ \text{outlier}}}{\zeta_t} + \underset{\substack{\text{heteroskedastic}\\ \text{noise}}}{\epsilon_t}$$

- **Trend Signal $\{\beta_t\}$**

Ref "Dynamic Shrinkage Process" [3], ABCO uses global-local shrinkage priors on the Dth order difference ($\triangle^D$, $D = 1, 2$) on the state variable $\{\beta_t\}$:

$$\triangle^D \beta_t = \omega_t, \qquad \omega_t \sim N(0, \tau_\omega^2 \lambda_{\omega,t}^2 = e^{h_t}),$$
$$h_{t+1} = \mu + (\phi_1 + \phi_2 s_t)(h_t - \mu) + \eta_{t+1}, \qquad \eta_{t+1} \sim Z(\alpha, \beta, 0, 1).$$

Z-distribution: log inverted Beta; heavy left-tail

- **Changepoint:** threshold γ & indicator $s_t = \begin{cases} 1 & \text{if } \log(\omega_t^2) > \gamma \\ 0 & \text{if } \log(\omega_t^2) \leq \gamma \end{cases}$.

- **Additive Outlier $\{\zeta_t\}$**

The outlier term $\{\zeta_t\}$ follows a "horseshoe+" shrinkage prior:

$$(\zeta_t|\sigma_{\zeta,t}) \sim N(0, \sigma_{\zeta,t}^2)$$
$$(\sigma_{\zeta,t}|\tau_\zeta, \eta_{\zeta,t}) \sim C^+(0, \tau_\zeta \eta_{\zeta,t})$$
$$\tau_\zeta \sim C^+(0, \sigma_{\tau,\zeta})$$
$$\eta_{\zeta,t} \sim C^+(0, \sigma_{\eta,\zeta})$$

with half-Cauchy $C^+(\cdot)$ and prior shrinkage hyper-parameters $\sigma_{\tau,\zeta}, \sigma_{\eta,\zeta}$.

- **Outlier:** custom cutoff & locally adaptive score $o_t := \widetilde{E}\left(\frac{\sigma_{\zeta,t}^2}{\sigma_{\zeta,t}^2 + \sigma_{\epsilon,t}^2}\right)$.

- **Heteroskedastic Noise $\{\epsilon_t, \sigma_{\epsilon,t}^2\}$**

The noise $\{\sigma_{\epsilon,t}^2\}$ follows a stochastic volatility model of order 1.

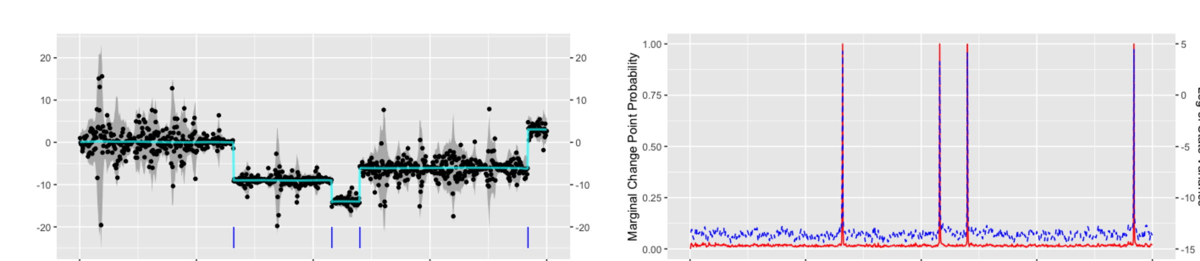
$$y_t = \beta_t + \zeta_t + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma_{\epsilon,t}^2),$$
$$\log(\sigma_{\epsilon,t}^2) = \mu_\epsilon + \phi_\epsilon(\log(\sigma_{\epsilon,t}^2) - \mu_\epsilon) + \xi_{\epsilon,t}, \qquad \xi_{\epsilon,t} \sim N(0, \sigma_\xi^2).$$

- **Dynamic Regression Generalizations**

Set $\boldsymbol{x}_t = (x_{1,t}, ..., x_{p,t})$ as p predictors at time t, and $\boldsymbol{\omega}, \boldsymbol{h}, \boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{\eta}$ analogously.

$$y_t = \boldsymbol{x}_t' \boldsymbol{\beta}_t + \zeta_t + \epsilon_t \qquad \triangle^D \boldsymbol{\beta}_{t+1} = \boldsymbol{\omega}_t$$
$$\omega_{j,t} \sim N(0, \tau_{\omega,0}^2 \tau_{\omega,j}^2 \lambda_{\omega,j,t}^2 = e^{h_{j,t}}) \qquad \boldsymbol{h}_{t+1} = \boldsymbol{\mu} + (\boldsymbol{\phi}_1 + \boldsymbol{\phi}_2 s_t)(\boldsymbol{h}_t - \boldsymbol{\mu}) + \boldsymbol{\eta}_{t+1}$$

ABCO Simulations



Algorithms	Rand Avg. ↑	Adj. Rand Avg. ↑	Avg. Diff. CP ↓	Avg. Dist. to True ↓
ABCO	**0.938**	**0.873**$_{(0.022)}$	**0.51**	**0.03**
E.Divisive	0.848	0.655$_{(0.022)}$	3.68	58.45
Pelt	0.730	0.361$_{(0.016)}$	10.88	128.36
WBS	0.686	0.352$_{(0.021)}$	106.65	144.45
BCP	0.841	0.595$_{(0.024)}$	22.00	60.16

Fig 6 Length-1000 time series with shift and SV(1) variance model

$$\log(\sigma_{\epsilon,t}^2) = \phi_\epsilon \log(\sigma_{\epsilon,t-1}^2) + \alpha_t, \qquad \alpha_t \sim N(0, \sigma_\alpha^2), \qquad \phi_\epsilon = 0.9, \qquad \sigma_\alpha = 0.4.$$

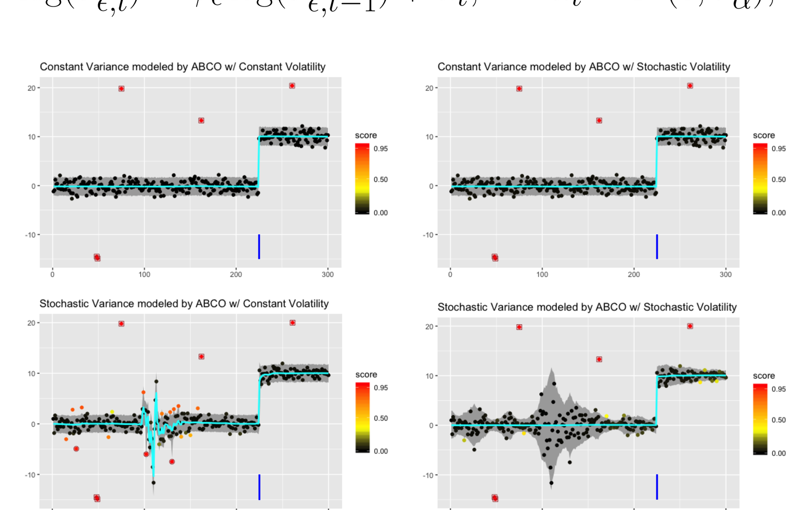


Fig 7 Robustness plus Outlier Scoring.

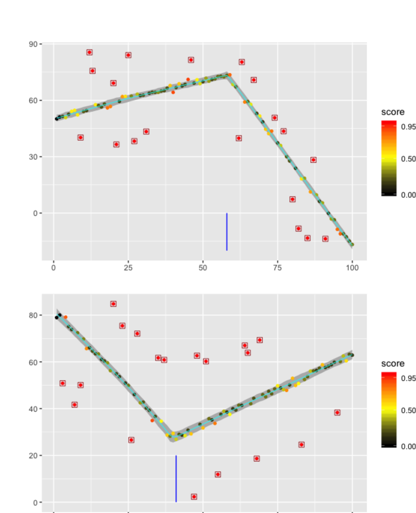


Fig 8 Linear Meetup Model.

ABCO Applications

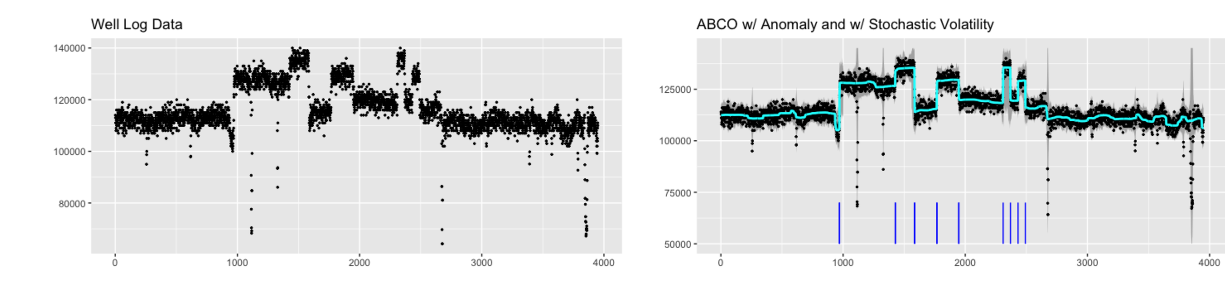


Fig 9 On Well-Log data, nuclear magnetic response within rock formations, originally published in [4] as a good framework for changepoint detection.

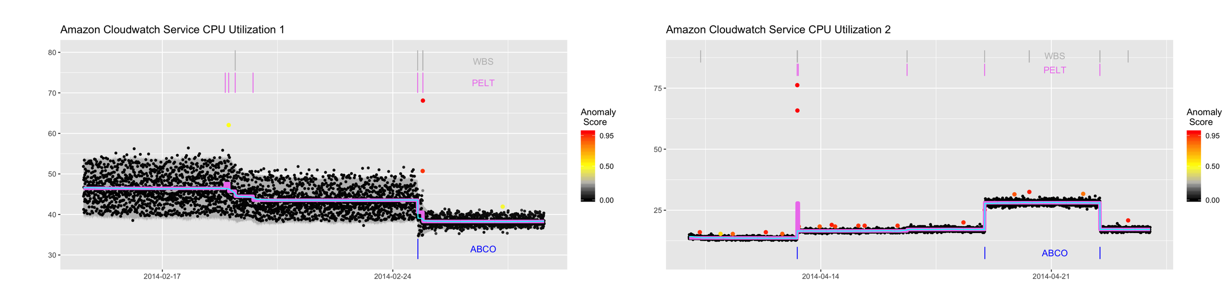


Fig 10 On Amazon Cloudwatch Service CPU Utilization data.

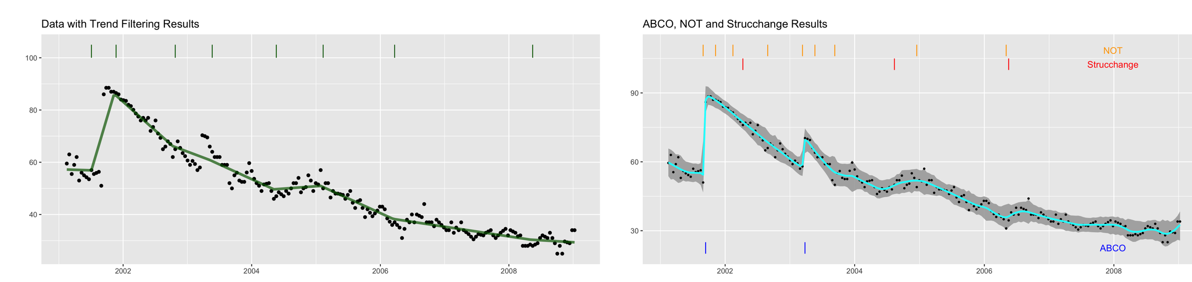


Fig 11 On George W. Bush Approval Rating data.

Decoupling Approach

- **Dynamic Linear Models (DLM)**

Given a time series $\boldsymbol{Y} = (y_1, ..., y_n)'$, a predictor series $\boldsymbol{X} = (x_1, ..., x_n)'$,

$$y_t = x_t \beta_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_{\epsilon,t}^2),$$
$$\triangle^D \beta_t = \omega_t, \qquad \omega_t \sim N(0, \sigma_\omega^2).$$

- **Decoupled Regularized Loss**

Denote $\bar{\boldsymbol{\beta}}$ as the posterior mean of k MCMC draws $\{\boldsymbol{\beta}^{(i)}, i = 1, ..., k\}$ of $\{\beta_t\}$.

$$\textbf{Decoupled loss: } L_\lambda(\widetilde{\boldsymbol{\beta}}) = ||\boldsymbol{W}^{1/2}(\boldsymbol{X} \circ \bar{\boldsymbol{\beta}} - \boldsymbol{X} \circ \widetilde{\boldsymbol{\beta}})||_2^2 + q_\lambda(\widetilde{\boldsymbol{\beta}}).$$

- $\boldsymbol{W} = \text{diag}(w_1, ..., w_n)$ is diagonal with weights for each measurement being $$w_i = 1/\bar{\sigma}_{\epsilon,i}^2, \text{ for } i = 1, ..., n.$$
- Penalty function $q_\lambda()$ induces sparsity into $\tilde{\boldsymbol{\beta}}$ with form $$q_\lambda(\tilde{\boldsymbol{\beta}}) = \lambda \sum_t \frac{1}{|\psi_t|} | \triangle^D \beta_t|,$$

where $\psi_t = \frac{1}{k} \sum_{i=1}^k \triangle^D \beta_t^{(i)}$ and $D = 1, 2$ controls the type of change.

- **Changepoint Selection**

Given λ, denote η_λ as the time indices which $\{\triangle^D \tilde{\beta}_t \neq 0\}$.

$$\textbf{Diagnostic tool: } R_\lambda^2 = \frac{1}{k} \sum_{i=1}^k \frac{||\boldsymbol{\beta}^{(i)} - \boldsymbol{\beta}_\eta^{(i)}||^2}{||\boldsymbol{\beta}^{(i)} - \bar{\boldsymbol{\beta}}^{(i)}||^2}$$

where $\bar{\boldsymbol{\beta}}^{(i)} = \frac{1}{n} \sum_{t=1}^n \beta_t^{(i)}$, and the optimal λ determined by least changepoints given $E[R_\lambda^2]$ exceeds a certain threshold.

* **Multiple Predictors & Covariates**

Set predictors $\boldsymbol{X} = \text{blockdiag}(\boldsymbol{x}_1', ..., \boldsymbol{x}_n')$ with $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,p})'$, and covariates $\boldsymbol{Z} = (\boldsymbol{z}_1, ..., \boldsymbol{z}_n)$ with $\boldsymbol{z}_i = (z_{i,1}, ..., z_{i,l})'$.

$$y_t = \boldsymbol{x}_t' \boldsymbol{\beta}_t + \boldsymbol{\alpha}' \boldsymbol{z}_t + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma_{\epsilon,t}^2), \qquad \triangle^D \boldsymbol{\beta}_t = \boldsymbol{\omega}_t, \qquad \boldsymbol{\omega}_t \sim N(0, \Sigma_{\omega,t}).$$

Extend the model with the decoupled loss:

$$L_\lambda(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) = ||\boldsymbol{W}^{1/2}(\boldsymbol{X}\bar{\boldsymbol{\beta}} + \boldsymbol{Z}\bar{\boldsymbol{\alpha}} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}} - \boldsymbol{Z}\widetilde{\boldsymbol{\alpha}})||_2^2 + q_\lambda(\widetilde{\boldsymbol{\beta}}),$$
$$q_\lambda(\widetilde{\boldsymbol{\beta}}) = \lambda \sum_{t=1}^n \sum_{g=1}^G \frac{1}{|\boldsymbol{\psi}_{g,t}|} | \triangle^D \boldsymbol{\beta}_{g,t}|.$$

Decoupling Simulations

Bayesian DLM

Decoupled

Comp. F1-Score

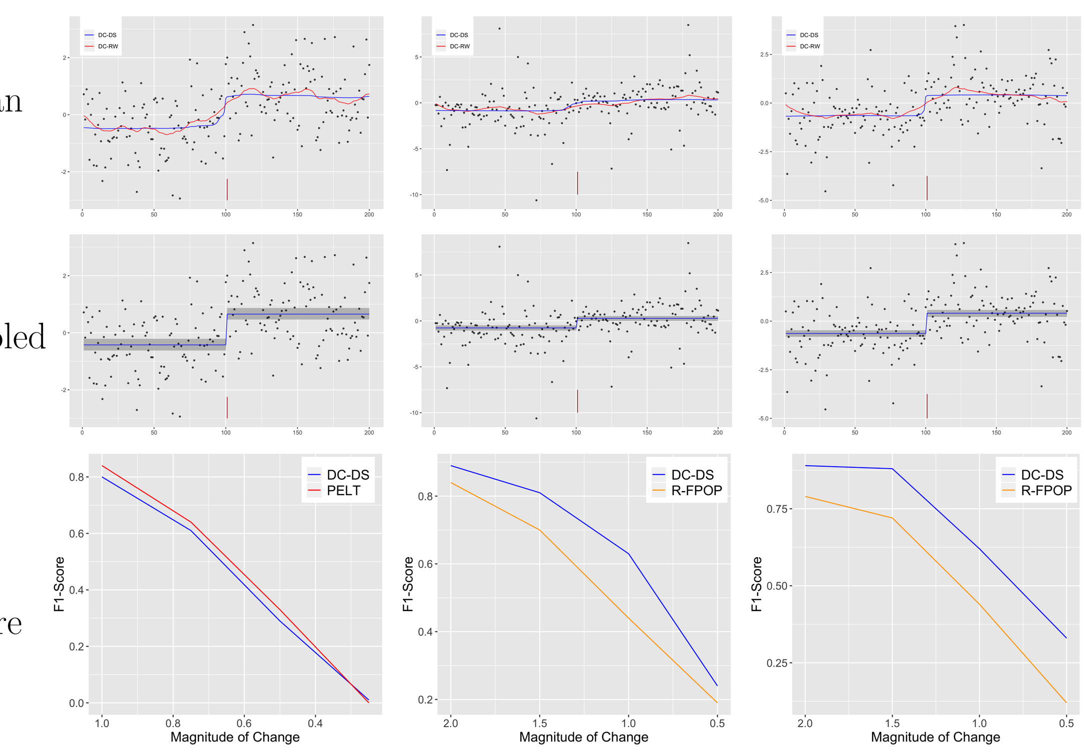


Fig 12 Gaussian Noise (left); Outliers (middle); Stochastic Volatility (right). DC-DS: decoupled results with shrinkage. DC-RW: decoupled with random walk.

Decoupling Applications

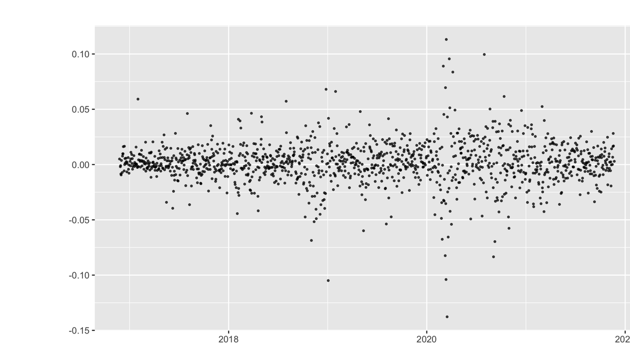


Fig 13 Apple Daily Stock Return

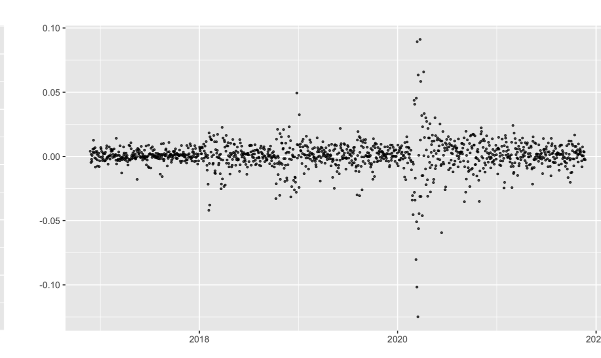


Fig 14 SP500 Daily Stock Return

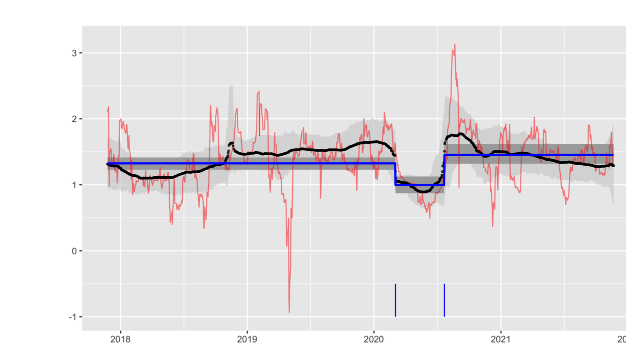


Fig 15 Decoupled vs Rolling OLS $\{\beta_t\}$

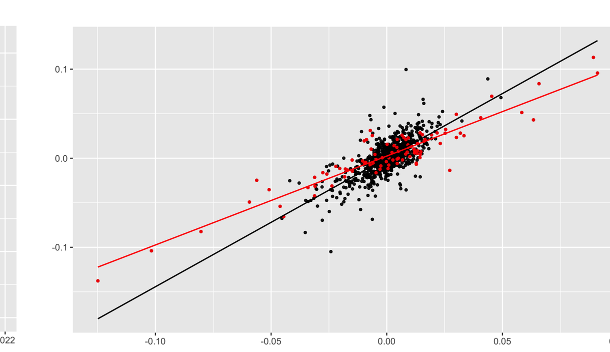


Fig 16 Scatter of Paired Returns

Conclusions

- A framework for inferring changepoints from posteriors produced by Bayesian time-varying parameter models.
- By decoupling trend modeling and changepoint analysis, we allow fitting an arbitrarily complex model to deal with intricacies inherent in data.
- Extensions: higher order trend changes, regression coefficients, multivariate.

References

[1] Haoxuan Wu and David S Matteson. Adaptive bayesian changepoint analysis and local outlier scoring. arXiv preprint arXiv:2011.09437, 2020.

[2] Haoxuan Wu, Sean Ryan, and David S Matteson. Decoupling trends and changepoint analysis. arXiv preprint arXiv:2201.06606, 2022.

[3] Daniel Kowal, David S. Matteson, and David Ruppert. Dynamic shrinkage process. Journal of the Royal Statistical Society: Series B, 2018.

[4] Joseph J. K. Ó Ruanaidh and William J. Fitzgerald. Numerical bayesian methods applied to signal processing. Statistics and Computing, 1996.

[5] Sangjoon Kim, Neil Shephard, and Siddhartha Chib. Stochastic volatility: likelihood inference and comparison with arch models. Review of Economic Studies, 65:361–393, 1998.

[6] David S. Matteson and Nicholas A. James. A nonparametric approach for multiple change point analysis of multivariate data. Journal of the American Statistical Association, 109:334–345, 2014.

[7] Anindya Bhadra, Jyotishka Datta, Nicholas G. Polson, and Brandon Willard. The horseshoe+ estimator for ultra sparse signals. Bayesian Analysis, 12:1105–1131, 2017.

[8] Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. Annals of Statistics, 42:2243–2281, 2014.

[9] Chandra Erdman and John W. Emerson. A fast bayesian change point analysis for the segmentation of microarray data. Bioinformatics, 24:2143–2148, 2008.