Drift vs Shift: Decoupling Trends and Changepoint Analysis

David S. Matteson

Cornell University
matteson@cornell.edu
National Institute of Statistical Sciences
director@niss.org
https://doi.org/10.1111/rssb.12325
https://arxiv.org/abs/2011.09437
https://arxiv.org/abs/2201.06606
with Peter Wu, Toryn Schafer, Sean Ryan, and NSF support

NISS Partnerships

→ National Institute of Statistical Sciences (NISS)

- Research non-profit, founded by the ASA, IBS, IMS, and more
- High-impact research for **science, engineering, health,** and **society**
- Data intensive research for evidence-based decisions & policy
- Leverage expertise from 50+ cross-sector affiliate organizations
- **Virtual events** & **regional hubs**

→ Deep Collaboration

- Total Survey Error & Blended Data @ GWU - NISS Hub (left)
- NISS/FCSM AI in Federal Government Series
- AI Day for Federal Statistics: A CNSTAT Public Event (right)
- NISS New Researcher Network Events & Capstone Projects

→ Premier Virtual Events

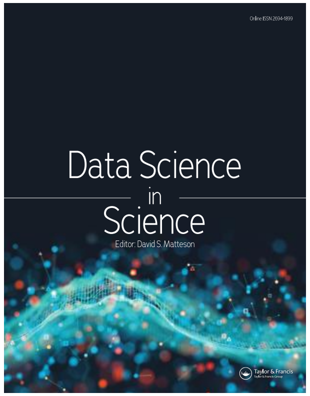
- AI, Statistics & Data Science in Practice:
  - Veridical Data Science: Bridging the Gap in Education and Research
- NISS-COPSS Leadership Webinars:
  - Linking Statistical Theory & Scientific Inquiry
- Virtual Career & Grad Fairs
- Serving Society Forums
  - Gun Violence, Human Trafficking, Opioid Crisis

THE NATIONAL INSTITUTE OF STATISTICAL SCIENCES

NISS | National Institute of Statistical Sciences

CoLab

- Statistics Serving Society
- AI in Statistics
- Collaborative Data Science
- Health & Environment
- Visualization
- New Researchers Network



NISS partners: to strengthen new and ongoing research and training activities - avoid duplication
NISS leads: when there are critical but unmet needs or as important emergent topics arise

NISS Collaboration: Statistical Agencies

- AI in Federal Government
- Study Reports and Expert Opinions
- Task Force Implementation Teams
- Research Staff Contracting
- Methods Evaluation and Best Practice
- Independent Validation
- Advanced Training
- Expert Panel and Attestation

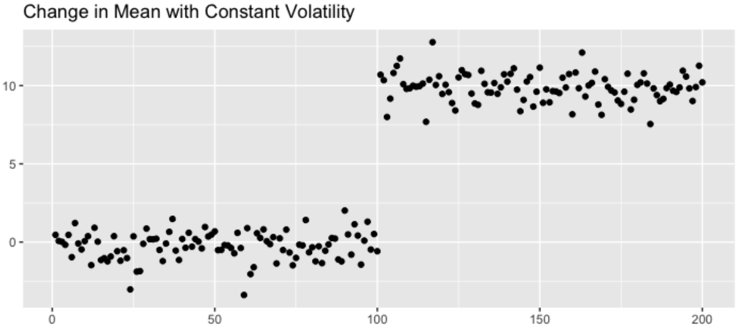
Introduction

- **Goal**: Distinguishing between global or macro patterns and local or 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, the underlying 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**.

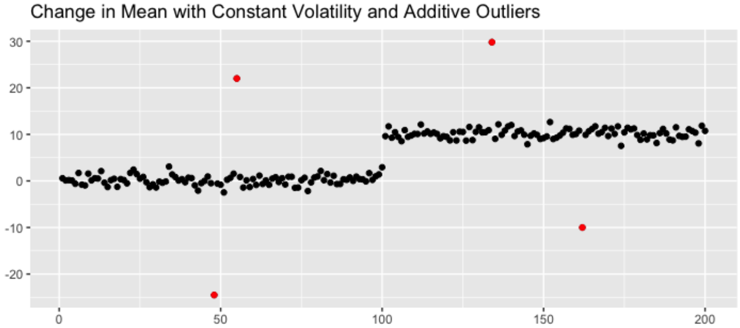
Tools

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

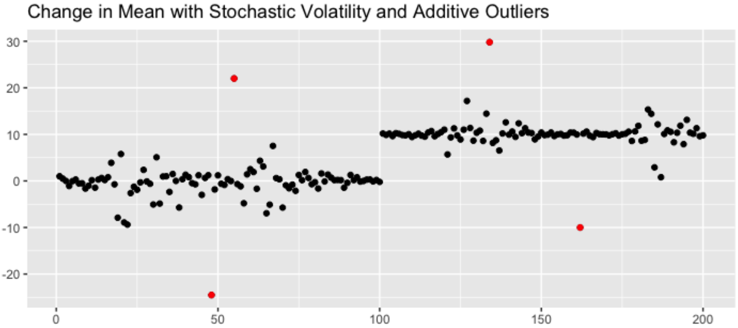
Change in Mean



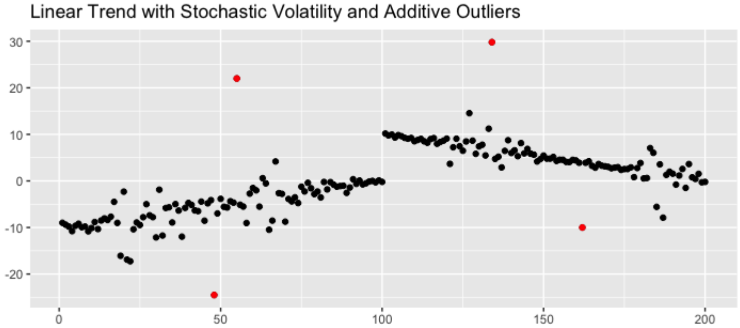
with Additive Outliers



with Stochastic Volatility



with Time Trends

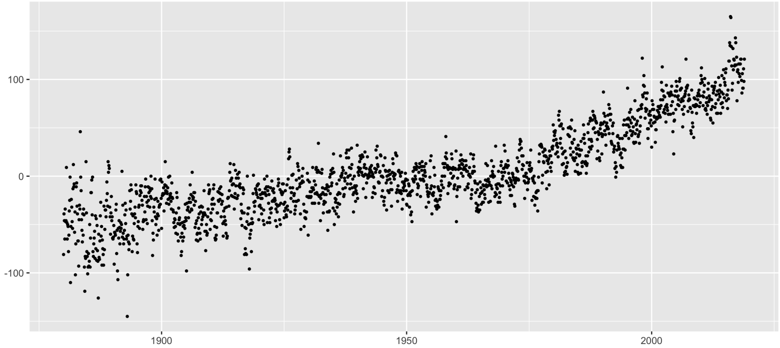


Challenges I

- Change point algorithms struggle with outliers and heterogeneity
- Outliers significantly skew the data distribution, and they violate the commonly required Gaussian noise assumption
- Heterogeneity typically leads to high/low volatility periods; many algorithms over-predict the number of change points in high volatility periods

The Real World I: Global Land Surface Air Temperature

Monthly global land surface air temperature from 1880 to 2018.

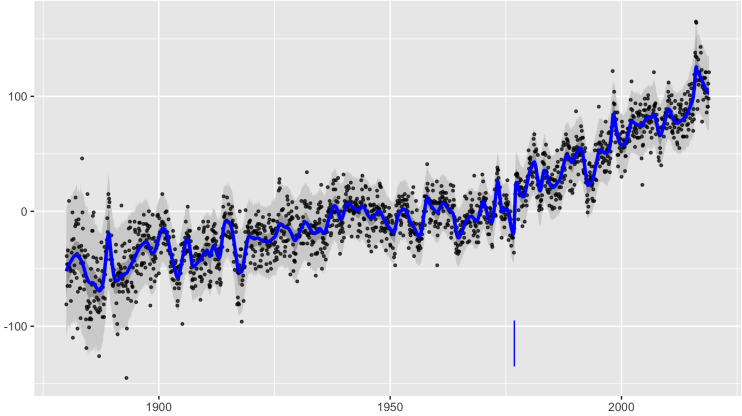


Challenges II

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

Drift vs Shift: Global Land Surface Air Temperature

A Bayesian framework to estimate the time trend



- Drift: locally smooth; outlier and heterogeneity robust; uncertainty quantification
- Shift: structural change detection; outlier, heterogeneity and drift robust; uncertainty quantification ⟶

A Decoupling Approach

- A two-step Bayesian + ML 'decoupling' method
- Identify changepoints within any dynamic linear model (DLM)
- First, DLM de-noises
- Second, regularized loss on posterior distribution identifies changepoints (with uncertainty quantification!)
- No hard selection within the time varying model

A Simple DLM

$$
\begin{aligned}
y_t &= x_t'\beta_t + \epsilon_t, & \epsilon_t &\sim N(0, \sigma^2_{\epsilon,t}) \\
\triangle^D \beta_t &= \omega_t & \omega_t &\sim N(0, \sigma^2_\omega)
\end{aligned}
$$

Changepoints from Posterior Samples

Let $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, ..., \boldsymbol{\beta}^{(K)}$ denote K MCMC draws of $\{\beta_t\}$

Changes in the kth sample path, $\{\beta_t^{(k)}\}$?

Find them using ML!

For example:

$$L_\lambda^{(k)}(\widetilde{\boldsymbol{\beta}}) = ||\boldsymbol{W}^{1/2}(\boldsymbol{X} \circ \boldsymbol{\beta}^{(k)} - \boldsymbol{X} \circ \widetilde{\boldsymbol{\beta}})||_2^2 + q_\lambda(\widetilde{\boldsymbol{\beta}}) \qquad (1)$$

where $q_\lambda()$ is a penalty function to somehow shrink $\tilde{\boldsymbol{\beta}}$

$\boldsymbol{W} = \text{diag}(w_1, ..., w_n)$ might be diag w/ weights for each obs

Loss function induces a second level shrinkage on the coefficients

Changepoints from Posterior Samples

Let $\beta^{(1)}, \beta^{(2)}, ..., \beta^{(K)}$ denote K MCMC draws of $\{\beta_t\}$

$$L_\lambda^{(k)}(\widetilde{\beta}) = ||\boldsymbol{W}^{1/2}(\boldsymbol{X} \circ \beta^{(k)} - \boldsymbol{X} \circ \widetilde{\beta})||_2^2 + q_\lambda(\widetilde{\beta})$$

$$\hat{\beta}_\lambda^{(k)} = argmin\ L_\lambda^{(k)}(\widetilde{\beta})$$

Now take posterior mean...any problem?

Decoupled Loss

Let $\beta^{(1)}, \beta^{(2)}, ..., \beta^{(K)}$ denote K MCMC draws of $\{\beta_t\}$

Let $\bar{\beta}$ denote the posterior mean of the K draws.

We define the decoupled loss as:

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

$$\hat{\beta}_\lambda = argmin\ L_\lambda(\widetilde{\beta})$$

When is this a good approximation to Bayes optimal procedure?

Weight Matrix

- For $\boldsymbol{W} = \text{diag}(w_1, ..., w_n)$, the classic choice for WSL is inverse of the variance
- Use posterior variance estimates at each time-step

$$w_i = \overline{\sigma_{\epsilon,i}^{-2}} \qquad \text{for } i = 1, ..., n \tag{3}$$

- Induces smaller l_2-norm for obs with larger variance, etc.

Penalty Function

One choice, given λ, for selection of changepoints:

$$q_\lambda(\tilde{\beta}) = \lambda \sum_t \frac{1}{|\psi_t|} | \triangle^D \beta_t |$$

where $\psi_t = \frac{1}{K} \sum_{i=1}^{K} \triangle^D \beta_t^{(k)}$. ψ_t is used as a local linear approximation to push the l_1 penalty toward l_0

$\triangle^D$ is the differencing operator

Changepoint Selection

- With the decoupled approach, we have a sequence of loss functions indexed by the parameter λ.
- As $\lambda \to 0$, then there would be no enforcement of sparsity and every point will be treated as a changepoint. As $\lambda \to \infty$, all $\{\triangle^D \tilde{\beta}_t\}$ will be 0 and no changepoint will be detected.
- For a particular number of changepoints, there exist a corresponding range of λ values.
- We will use uncertainty quantification to understand the trade-off between fit and number of changepoints.

Diagnostic Tools

We first compute the "projected posterior" [**Woody˙2020**] to quantify uncertainty of the changepoint estimates.
For a given value λ, let η_λ denote the time indices which $\{\triangle^D \tilde{\beta}_t \neq 0\}$.
The kth projected posterior is then given by:

$$R_\lambda^{2(k)} \equiv 1 - \frac{||\boldsymbol{W}^{1/2}(\boldsymbol{X} \circ \boldsymbol{\beta}^{(k)} - \boldsymbol{X} \circ \boldsymbol{\beta}_{\eta_\lambda}^{(k)})||^2}{||\boldsymbol{W}^{1/2}(\boldsymbol{X} \circ \boldsymbol{\beta}^{(k)} - \boldsymbol{X} \circ \mu_{\boldsymbol{\beta}^{(k)}})||^2}$$

k is the kth MCMC posterior draw, $\mu_{\boldsymbol{\beta}^{(k)}}$ is the mean of $\boldsymbol{\beta}^{(k)}$.
This metric is a similar R-squared in that it measures the amount of variation explained by the projected posterior $\boldsymbol{\beta}_\eta$ for each of the MCMC draws.

Changepoint Selection

The projected posterior maps $\{\beta^{(k)}, i = 1, ..., k\}$ from each of the k MCMC iteration onto the best fitted model given the changepoints. The diagnostic tool is then given by:

$$R^2_\lambda = \frac{1}{K} \sum_{k=1}^{K} \frac{||\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k)}_\eta||^2}{||\boldsymbol{\beta}^{(k)} - \bar{\boldsymbol{\beta}}^{(k)}||^2}$$

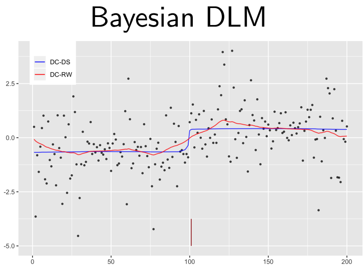
where $\bar{\boldsymbol{\beta}}^{(k)} = \frac{1}{n} \sum_{t=1}^{n} \beta^{(k)}_t$. This metric is a similar measurement to R^2 where it measures the amount of variation explained by the estimated signal $\boldsymbol{\beta}_\eta$ for each of the MCMC draws.

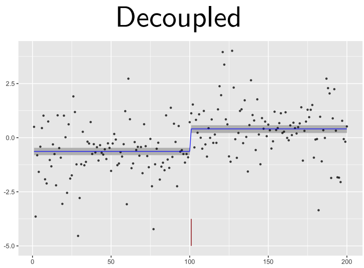
For selecting the optimal value of λ, we will select the lowest number of change point which $E[R^2_\lambda]$ exceed a certain threshold.

Advantages

- By separating the process of modeling and inference, the framework allow us to fit a highly complex Bayesian model to deal with intricacies of the data such as outliers and heterogeneity.
- The approach can provide uncertainty quantification in amount of variation explained by different number of changepoints.
- By combining the flexibility and robustness of a Bayesian framework with the hard thresholding selection of a regularized loss estimator, the approach can adapt changepoint analysis to higher order trends, regression and multivariate settings.

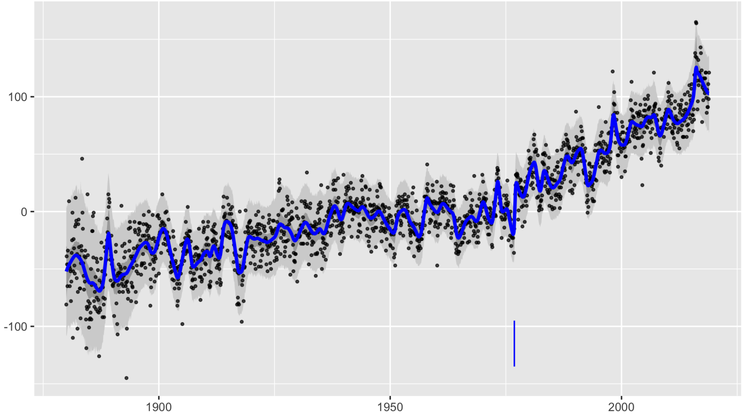
Simulation Example: Stochastic Volatility





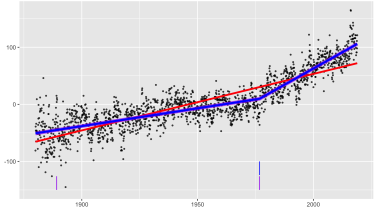
Global Land Surface Air Temperature Example

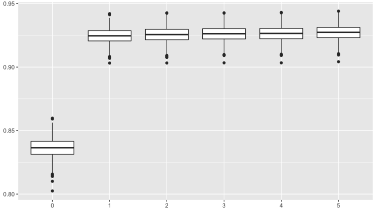
We consider monthly global land surface air temperature from 1880 to 2018. The Bayesian TVP fit of the data can be seen as follows:



Global Land Surface Air Temperature Example

The resulting decoupled approach estimate for different number of changepoints can be seen as follows:



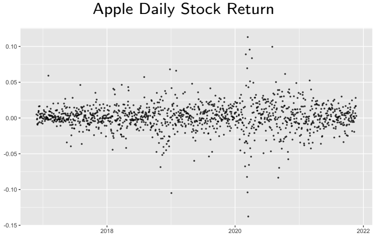


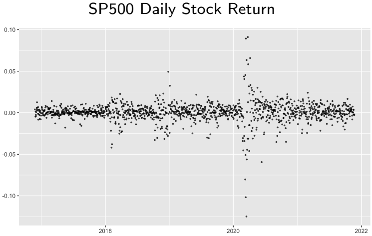
Final Conclusions

- A framework for selecting changepoint from posteriors produced by Bayesian time-varying parameter models.
- By separating the process of trend modeling and changepoint analysis, the framework allows for fitting of an arbitrary complex model to deal with intricacies inherent in data.
- Can be extended to deal with outliers, heterogeneity, higher order changes in trend, changes in regression coefficients and changes in multivariate data.
- Full theoretical justifications are an exciting challenge.

Financial Time Series

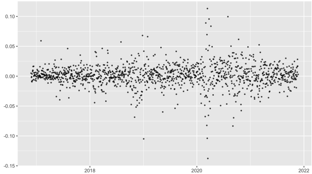
We analyze changes in regression relationship between daily Apple returns and the market returns across the last 5 years.



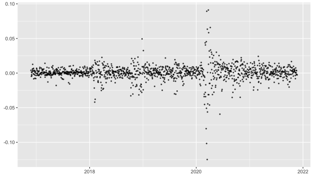


Financial Time Series

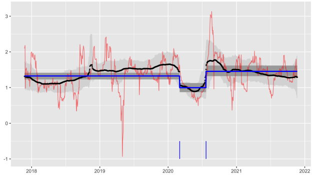
Apple Daily Stock Return



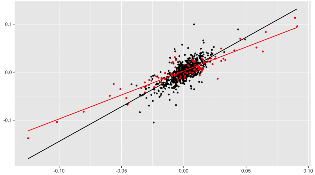
SP500 Daily Stock Return



Decoupled Fit vs Rolling OLS for $\{\beta_t\}$



Scatter of Paired Returns



We two changepoints at location 03/02/2020 and 07/22/2020. These dates mark the beginning of the pandemic where significant volatility occurred within the stock market. During this time period, the market movement became the dominated trend which led to coefficient of beta to be close to 1.

Part 1) Model Based Solution
ABCO: Adaptive Bayesian Changepoints w/ Outliers

- Model the underlying time-varying signal
- Distinguishing local trends (drift) from major changes (shift)
- Detect change points in the presence of stochastic volatility and outliers
- Flag outliers in data based on an outlier scoring metric

ABCO

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

$$y_t = \beta_t + \zeta_t + \epsilon_t$$

- Time trend or signal (β_t)
- Additive outlier (ζ_t)
- Heteroskedastic noise (ϵ_t)

Trend or Drift Modeling

ABCO utilizes a class of priors called global-local shrinkage priors on the D*th* order difference of the state variable $\{\beta_t\}$

$$y_t = \beta_t + \zeta_t + \epsilon_t$$

$$\begin{aligned} \triangle^D \beta_t &= \omega_t & \omega_t &\sim N(0, \tau_\omega^2 \lambda_{\omega,t}^2) \\ h_t &:= log(\tau_\omega^2 \lambda_{\omega,t}^2) & & \\ h_{t+1} &= \mu + (\phi_1 + \phi_2 s_t)(h_t - \mu) + \eta_{t+1} & \eta_{t+1} &\sim Z(\alpha, \beta, 0, 1) \end{aligned}$$

$\triangle^D$ = D*th* difference ($D = 1$ or 2).

Suppose a "Dynamic Shrinkage Process" for $\{\beta_t\}$ (Kowal *et al.* 2018): $Z(\alpha = \frac{1}{2}, \beta = \frac{1}{2}, 0, 1)$

For changepoints: additional threshold s_t needed

Drift Modeling in the Presence of Changepoints

Indicator $\{s_t\}$ 'activates' at changepoints.

Linked to log variance $\{h_t\}$ of 'changes' in mean trend $\{\beta_t\}$

$$y_t = \beta_t + \zeta_t + \epsilon_t$$

$$\triangle^D \beta_t = \omega_t$$

$$h_{t+1} = \mu + (\phi_1 + \phi_2 s_t)(h_t - \mu) + \eta_{t+1}$$

$$s_t = \begin{cases} 1 & \text{if } \log(\omega_t^2) > \gamma \\ 0 & \text{if } \log(\omega_t^2) \leq \gamma \end{cases}$$

Threshold level, γ, determines changepoint cutoff.

Thresholding 'resets' h_{t+1} (via ϕ_2) after isolated changepoint at t.

The Additive Outlier Component

Outliers $\{\zeta_t\} \sim$ "horseshoe+" shrinkage prior (Bhadra *et al.* 2017)

$$y_t = \beta_t + \zeta_t + \epsilon_t$$

$$\begin{aligned}
(\zeta_t | \sigma_{\zeta,t}) &\sim N(0, \sigma^2_{\zeta,t}) \\
(\sigma_{\zeta,t} | \tau_\zeta, \eta_{\zeta,t}) &\sim C^+(0, \tau_\zeta \eta_{\zeta,t}) \\
\tau_\zeta &\sim C^+(0, \sigma_{\tau,\zeta}) \\
\eta_{\zeta,t} &\sim C^+(0, \sigma_{\eta,\zeta})
\end{aligned}$$

$C^+ =$ half-Cauchy distribution & $\sigma_{\tau,\zeta}, \sigma_{\eta,\zeta}$ are hyper-parameters.

Outlier Scoring

$$y_t = \beta_t + \zeta_t + \epsilon_t$$

Propose locally adaptive outlier score based: '%' posterior variance from outlier component:

$$o_t = \widetilde{E}\left(\frac{\sigma^2_{\zeta,t}}{\sigma^2_{\zeta,t} + \sigma^2_{\epsilon,t}}\right)$$

where $\widetilde{E}$ denotes the posterior expectation.

Outlier score o_t ranges 0 to 1.

Scores rank outliers, user may threshold to label them.

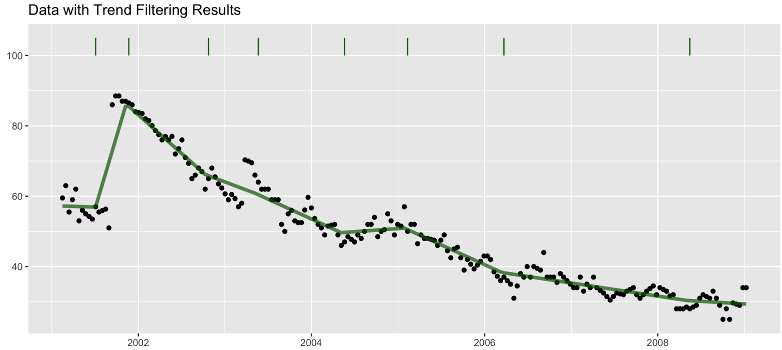
The Measurement Error or Noise Component

The measurement error variance $\{\sigma^2_{\epsilon,t}\}$ follows a stochastic volatility model of order 1 (Kim *et al.* 1998).

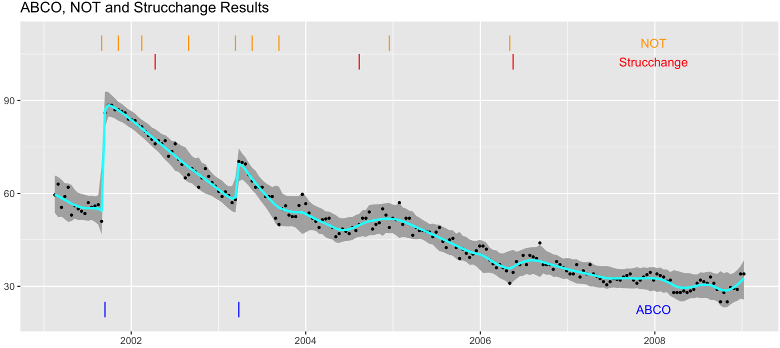
$$\begin{aligned} y_t &= \beta_t + \zeta_t + \epsilon_t & \epsilon_t &\sim N(0, \sigma^2_{\epsilon,t}) \\ \log(\sigma^2_{\epsilon,t+1}) &= \mu_\epsilon + \phi_\epsilon(\log(\sigma^2_{\epsilon,t}) - \mu_\epsilon) + \xi_{\epsilon,t+1}; & \xi_{\epsilon,t} &\sim N(0, \sigma^2_\xi) \end{aligned}$$

ABCO adds robustness to heterogeneity via SV(1) model for measurement error.

Real World 3: George W. Bush Approval Rating



Real World 3: George W. Bush Approval Rating



ABCO Summary

- ABCO: great flexibility and many possible extensions
- Utilizes state-space approach and global-local shrinkage priors
- ABCO detects local trends, in presence of additive outliers and heterogeneity

Time Series Analysis

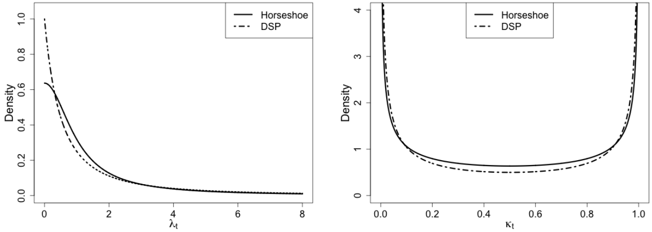
Define $z_{t,h} := \phi^h \eta_{t-h}$. Infinite order Moving Average representation of v_t results in:

$$\begin{aligned} v_t &= \mu + \phi(v_{t-1} - \mu) + \eta_t \qquad\qquad \eta_t \sim Z(1/2, 1/2, 0, 1) \\ &= \mu + \sum_{h=0}^{\infty} \phi^h \eta_{t-h} \\ &= \mu + \sum_{h=0}^{\infty} z_{t,h}. \end{aligned}$$

As shown in B-N82, $z_{t,h}$ is a scaled Hyperbolic Secant Distribution, with $E(z_{t,h}|\phi) = 0$ and $Var(z_{t,h}|\phi) = (|\phi|^h \pi)^2$. Consider $z_t := \sum_{h=0}^{\infty} z_{t,h}$.

If $|\phi| = 0.5$, $\eta_{t-h} \overset{i.i.d}{\sim} Z(1/2, 1/2, 0, 1)$, and $z_{t,h} := \phi^h \eta_{t-h}$, then $z_t := \sum_{h=0}^{\infty} z_{t,h} \overset{a.s}{\to} Logistic(0, 2)$.

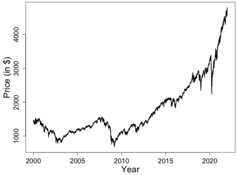
Time Series Analysis



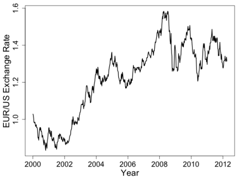
(a) Prior densities on λ_t (b) Prior densities on κ_t

Figure: Comparisons of prior densities on λ_t and κ_t between Horseshoe Prior and the stationary distribution of Dynamic Shrinkage Prior with $\eta_t \overset{i.i.d}{\sim} Z(1/2, 1/2, 0, 1)$, $\phi = 1/2$ and $\mu = 0$.

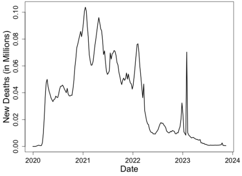
Financial Time Series II



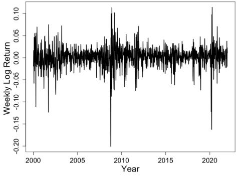
(a) S&P500 Index



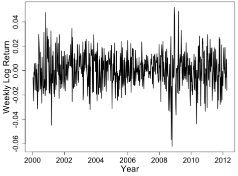
(b) EURO/USD Exchange Rate



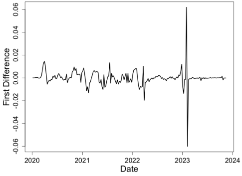
(c) Global COVID-19 Deaths



(d) Weekly Log Return



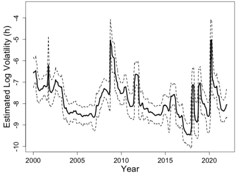
(e) Weekly Log Return



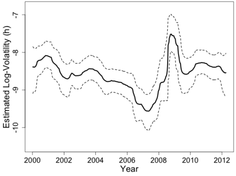
(f) First Difference

Figure: Price of S&P500 between 2012-01-01 and 2021-12-31 (a),

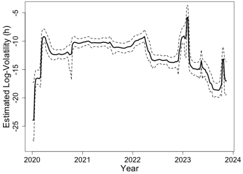
Financial Time Series II



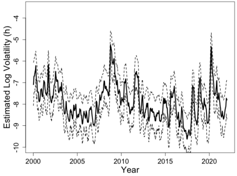
(a) S&P 500 Index (ASV)



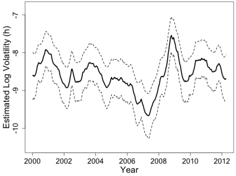
(b) Exchange Rate (ASV)



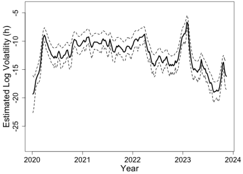
(c) COVID-19 (ASV)



(d) S&P 500 Index (SV)

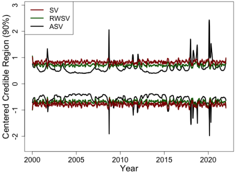


(e) Exchange Rate (SV)

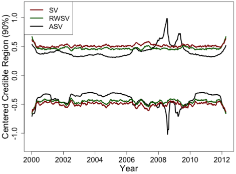


(f) COVID-19 (SV)

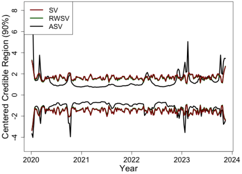
Financial Time Series II



(a) S&P 500 Index



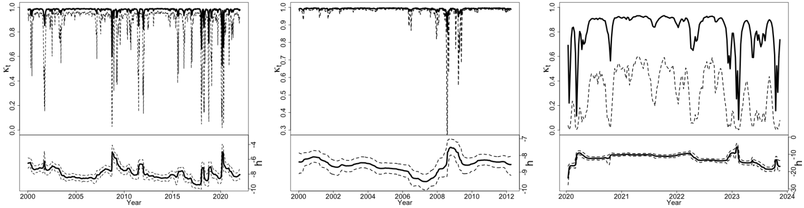
(b) Exchange Rate



(c) COVID-19 Deaths

Figure: The 90% quantile-based credible regions on the log-variance h_t are subtracted by its posterior sample mean. The three datasets include weekly returns on S&P 500 between 2000-01-01 and 2021-12-31, weekly returns on EURO/USD exchange rate between 2000-01-03 and 2012-04-04, and weekly changes in Global COVID-19 death tolls between 2020-01-03 and 2023-11-13. The centered credible regions for Adaptive Stochastic Volatility with Dynamic Shrinkage Processes (ASV-DSP) are drawn in black, the ones based on Stochastic Volatility (SV) model are in dark red and the ones based on Random Walk Stochastic Volatility (RWSV) model are in dark green.

Financial Time Series II



(a) S&P 500 Index (b) Exchange Rate (c) COVID-19 Deaths

Figure: Comparison between the expected shrinkage parameter $\kappa_t := \frac{1}{1+var(\omega_t|\tau,\lambda_t)} = \frac{1}{1+exp(v_t)} = \frac{1}{1+\tau^2\lambda_t^2}$ and expected h_t based on Adaptive Stochastic Volatility with Dynamic Shrinkage Processes (ASV-DSP) estimated on weekly returns on S&P 500 between 2000-01-01 and 2021-12-31, weekly returns on EURO/USD exchange rate between 2000-01-03 and 2012-04-04, and weekly changes in Global COVID-19 death tolls between 2020-01-03 and 2023-11-13, respectively. The dotted lines represent the one-sided 95th and centered 90th percentile credible regions for κ_t and h_t respectively.