Covariance Specification and Missing Data Imputation During Periods of Prolonged Missingness with Bayesian Dynamic Linear Models

NESS, 34th meeting

October 1, 2021

Jacob Strock, MS

Gavino Puggioni, PhD

Problem overview and applied context

DLM overview

Role of covariance specification on inference

Simulation and analysis overview

Results

Discussion

Problem overview and applied context

DLM overview

Role of covariance specification on inference

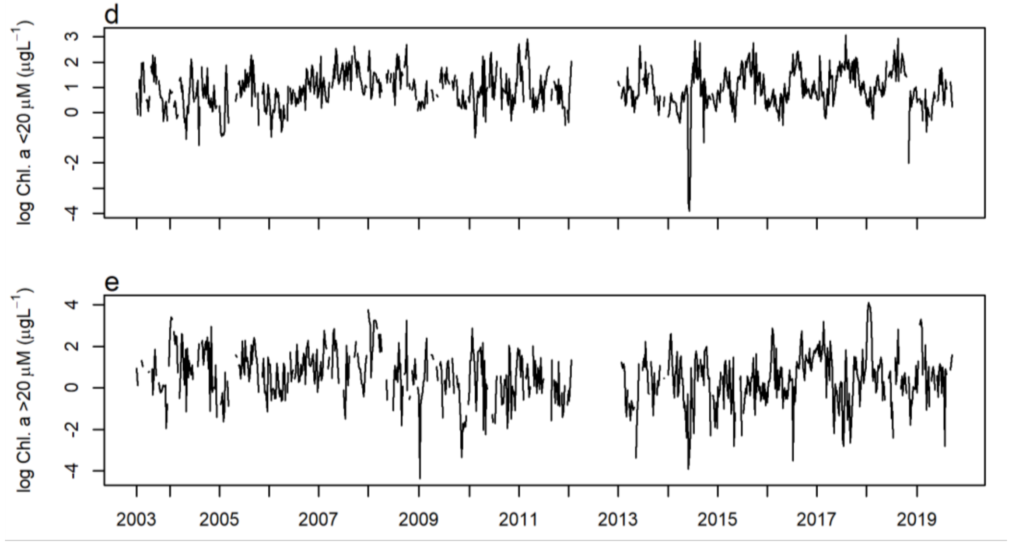
Simulation and analysis overview

Results

Discussion

Background: troubles with missing data in time series

- Component of Masters thesis work in Statistics at University of Rhode Island:
  - Investigating multistage DLM modeling structures in application to long-term environmental monitoring
- Long-term environmental monitoring data
  - 17 years of weekly resolution data on biological, physical, and chemical data
    - Temperature
    - Nitrogen
    - Chlorophyll (algal pigment)
- Applied research questions:
  - How are have each of these features changed with policy decisions on nutrient pollution?
  - How have they affected each other?
    - What are long-term and seasonal patterns?
    - How do biological and chemical components relate?

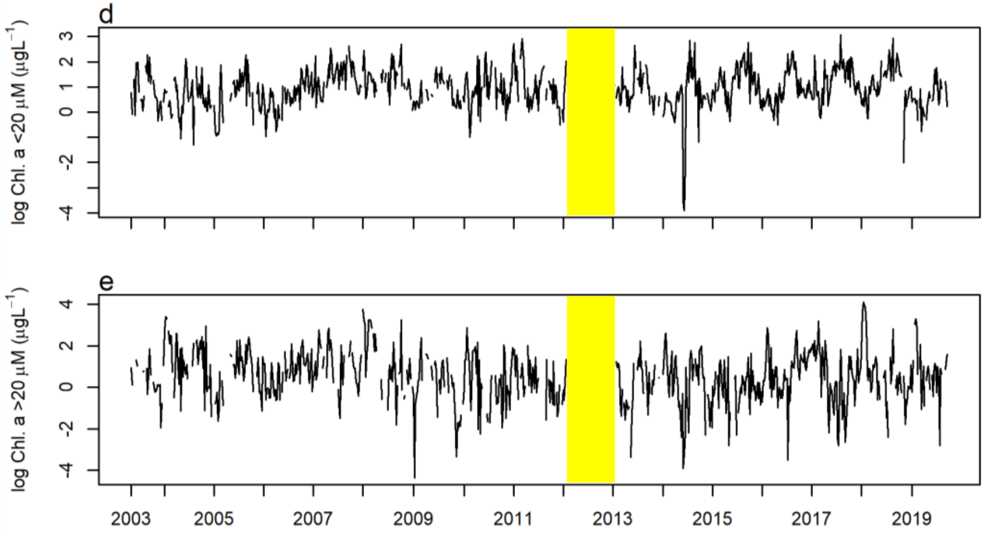


Background: troubles with missing data in time series

1. Major gaps in data that were a concern for making accurate inference on these patterns:
    - Causes engine issues on monitoring boat 
    - Gaps in funding $$$
    - Anomalous weather events 
    - Periods of extended missingness have capacity to impact inferences on the full data series
2. Heteroskedasticity in the series

- Intertwined:
    - Time-varying covariance structure impacts imputation
    - Extended periods of missing data impact inference on covariance structure

Missingness Length	Frequency
1	61
2	6
4	1
48	1

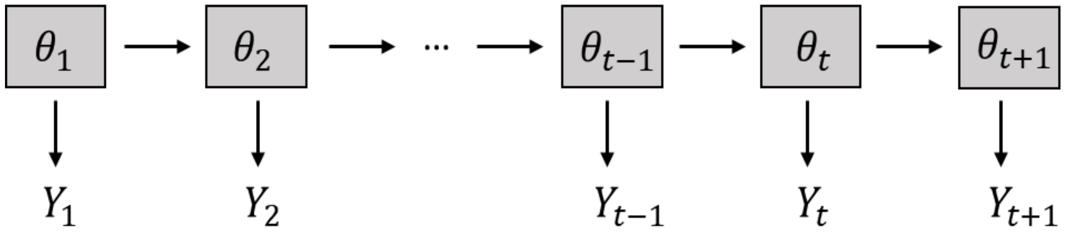


Objectives:

- Compare model selection criteria in data with heteroskedasticity and extensive missingness
- Compare performance of covariance specification during extended periods of missingness

- Problem overview and applied context
- DLM overview
- Role of covariance specification on inference
- Simulation and analysis overview
- Results
- Discussion

State Space modeling: the Dynamic Linear Model



$$\left.\begin{aligned} Y_t &= \boldsymbol{F}_t\boldsymbol{\theta}_t + v_t, & v_t &\sim N(0, \boldsymbol{V}) \\ \boldsymbol{\theta}_t &= \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + w_t, & w_t &\sim N(0, \boldsymbol{W}) \end{aligned}\right\} \begin{aligned} &\text{Observation equation} \\ &\text{State equation} \end{aligned}$$

$$\left.\begin{aligned} &\boldsymbol{\theta}_0 \sim N(\boldsymbol{m}_0, \boldsymbol{C}_0) \\ &\boldsymbol{V} \sim IG(a_v, b_v) \\ &\boldsymbol{W} \sim IW(a_w, b_w) \end{aligned}\right\} \text{Priors}$$

Variable	Definition
Y_t	Datum/Data
F_t	Observational matrix
G_t	Evolutional Matrix
θ_t	Unobserved state
V_t	Observational variance/covariance
W_t	Evolutional variance/covariance

Extensive literature and history:

- Kalman, R. E. 1960. “A New Approach to Linear Filtering and Prediction Problems.” *Journal of Fluids Engineering, Transactions of the ASME* 82 (1): 35–45. https://doi.org/10.1115/1.3662552.
- West, Mike, and Jeff Harrison. 1997. *Bayesian Forecasting and Dynamic Models*. 2nd ed. Verlag New York: Springer. https://doi.org/10.1007/b98971.

- **Why?**
- Flexible structure
- Components are additive
  - ➢ long-term trend, season, regression components separately
- Any parameter and component can time vary
- Interpolation via inference
- Quantify uncertainty in missingness and states

Kalman Filtering and Smoothing

1. One-step -ahead predictive distribution of the latent state, $f(\theta_t|y_{1:t-1}) = N(a_t, R_t)$, where:
$$a_t = E(\theta_t|y_{1:t-1}) = G_t m_{t-1},$$
$$R_t = Var((\theta_t|y_{1:t-1}) = G_t C_{t-1} G_t' + W_t$$

2. One-step -ahead predictive distribution of the observation, $f(Y_t|y_{1:t-1}) = N(f_t, Q_t)$, where:
$$ft = E(Y_t|y_{1:t-1}) = F_t a_t$$
$$Qt = Var(Y_t|y_{1:t-1}) = F_t R_t F_t' + V_t$$

3. The filtered distribution of the latent state, $f(\theta_t|y_{1:t}) = N(m_t, C_t)$, where:
$$m_t = E(\theta_t|y_{1:t}) = a_t + R_t F_t' Q_t^{-1} e_t,$$
$$C_t = Var(\theta_t|y_{1:t}) = R_t - R_t F_t' Q_t' F_t R_t,$$
$$e_t = Y_t - ft$$

4. The smoothed distribution of the latent state, $f(\theta_t|y_{1:T}) = N(s_t, S_t)$, where:
$$s_t = E(\theta_t|y_{1:T}) = m_t + C_t G_{t+1}' R_{t+1}' (s_{t+1} - a_{t+1}),$$
$$S_t = C_t - C_t G_{t+1}' R_{t+1}^{-1} (R_{t+1}) R_{t+1}^{-1} G_{t+1} C_t$$

Kalman Filtering and Smoothing with Missing Data

1. One-step -ahead predictive distribution of the latent state, $f(\theta_t|y_{1:t-1}) = N(a_t, R_t)$, where:
$$a_t = E(\theta_t|y_{1:t-1}) = G_t m_{t-1},$$
$$R_t = Var((\theta_t|y_{1:t-1}) = G_t C_{t-1} G_t' + W_t$$

2. One-step -ahead predictive distribution of the observation, $f(Y_t|y_{1:t-1}) = N(f_t, Q_t)$, where:
$$ft = E(Y_t|y_{1:t-1}) = F_t a_t$$
$$Qt = Var(Y_t|y_{1:t-1}) = F_t R_t F_t' + V_t$$

3. The filtered distribution of the latent state, $f(\theta_t|y_{1:t}) = N(m_t, C_t)$, where:
$$m_t = E(\theta_t|y_{1:t}) = a_t + R_t F_t' Q_t^{-1} e_t,$$
$$C_t = Var(\theta_t|y_{1:t}) = R_t - R_t F_t' Q_t' F_t R_t,$$
$$e_t = Y_t - ft$$

4. The smoothed distribution of the latent state, $f(\theta_t|y_{1:T}) = N(s_t, S_t)$, where:
$$s_t = E(\theta_t|y_{1:T}) = m_t + C_t G_{t+1}' R_{t+1}'(s_{t+1} - a_{t+1}),$$
$$S_t = C_t - C_t G_{t+1}' R_{t+1}^{-1}(R_{t+1}) R_{t+1}^{-1} G_{t+1} C_t$$

Kalman Filtering and Smoothing with Missing Data

1. One-step -ahead predictive distribution of the latent state, $f(\theta_t|y_{1:t-1}) = N(a_t, R_t)$, where:
$$a_t = E(\theta_t|y_{1:t-1}) = G_t m_{t-1},$$
$$R_t = Var((\theta_t|y_{1:t-1}) = G_t C_{t-1} G_t' + W_t$$

2. One-step -ahead predictive distribution of the observation, $f(Y_t|y_{1:t-1}) = N(f_t, Q_t)$, where:
$$ft = E(Y_t|y_{1:t-1}) = F_t a_t$$
$$Qt = Var(Y_t|y_{1:t-1}) = F_t R_t F_t' + V_t$$

3. The filtered distribution of the latent state, $f(\theta_t|y_{1:t}) = N(m_t, C_t)$, where:
$$m_t = E(\theta_t|y_{1:t}) = a_t + R_t F_t' Q_t^{-1} e_t,$$
$$C_t = Var(\theta_t|y_{1:t}) = R_t - R_t F_t' Q_t' F_t R_t,$$
$$e_t = Y_t - ft$$

4. The smoothed distribution of the latent state, $f(\theta_t|y_{1:T}) = N(s_t, S_t)$, where:
$$s_t = E(\theta_t|y_{1:T}) = m_t + C_t G_{t+1}' R_{t+1}' (s_{t+1} - a_{t+1}),$$
$$S_t = C_t - C_t G_{t+1}' R_{t+1}^{-1} (R_{t+1}) R_{t+1}^{-1} G_{t+1} C_t$$

Problem overview and applied context

DLM overview

Role of covariance specification on inference

Simulation and analysis overview

Results

Discussion

Specification of Evolutional Covariance and Inference

- Static covariance
  - $W_0 \quad \sim \ Inverse - Wishart\ (\nu_0, S_0)$
  - $W|Y_{1:T}, V, \theta_{1:T} \quad \sim \ Inverse - Wishart\ (\nu_0 + T, S_0 + RSS_\theta)$

- Discounted covariance

$$R_t = Var((\theta_t | y_{1:t-1}) = G_t C_{t-1} G_t' + W_t$$
$$R_t = P_t + W_t$$
$$W_t = \frac{1-\delta}{\delta} P_t$$

Evolutional Covariance specification:

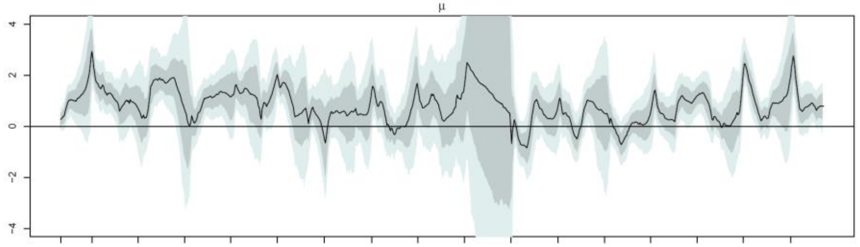
- One-step -ahead predictive distribution of the latent state, $f(\theta_t | y_{1:t-1}) = N(a_t, R_t)$, where:

$$a_t = E(\theta_t | y_{1:t-1}) = G_t m_{t-1},$$
$$R_t = Var((\theta_t | y_{1:t-1}) = G_t C_{t-1} G_t' + W_t$$

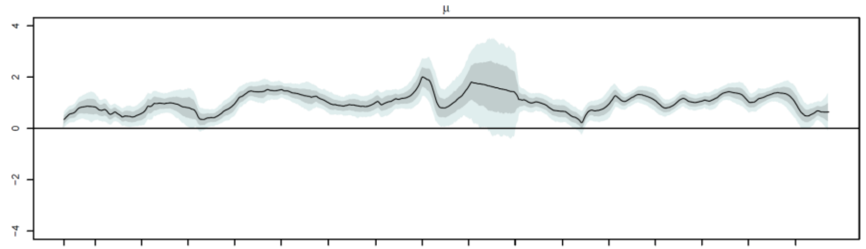
$$W_t = \frac{1-\delta_i}{\delta_i} P_{i,t},$$
$$P_{i,t} = G_t C_{t-1} G_t'$$

- Discount factor δ_i, describes the loss of information between time-steps
  - $\delta_i = 1$ is equivalent to a static model (no evolutional change in the states)
  - Commonly specified between 0.85 and 0.999
- Can be sampled or fixed and compared across models
- Major parameter reduction
- Accommodates evolution in covariance

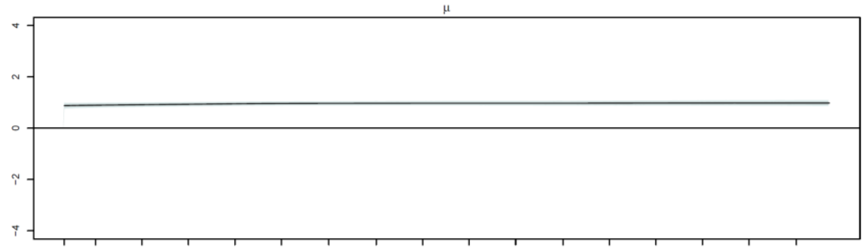
$\delta = 0.8$



$\delta = 0.9$



$\delta = 0.999$

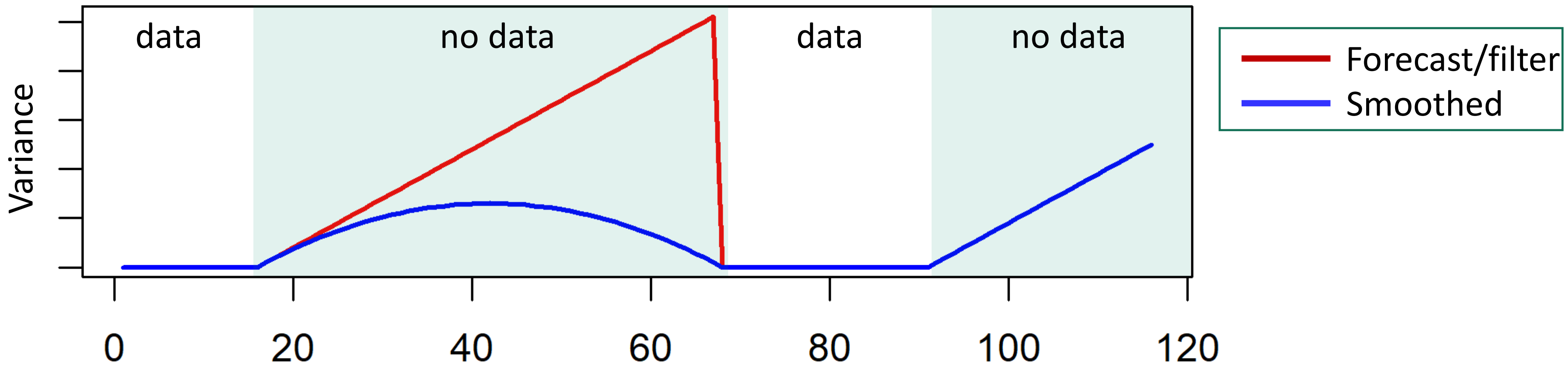


Uncertainty and Covariance

- In a forecast, information is lost at a linear loss rate relative to W_t following recursive forecasting:

$$m_t = E(\theta_t | y_{1:t-1}) = G_t m_{t-1},$$
$$C_t = Var((\theta_t | y_{1:t-1}) = G_t C_{t-1} G_t' + W_t, \qquad t = j, \ldots, k$$

- Smoothing (backward recursion) is not affected so in missingness periods:

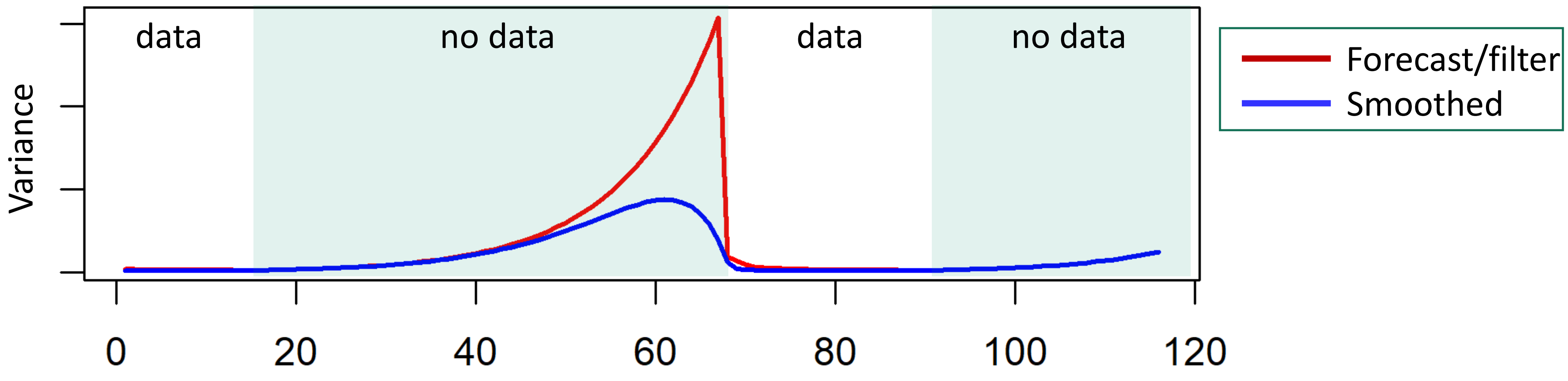


- In a forecast, information is lost at an exponential loss rate relative to W_t following recursive forecasting:

$$
\begin{aligned}
m_t &= E(\theta_t|y_{1:t-1}) = G_t m_{t-1}, \\
C_t &= Var((\theta_t|y_{1:t-1}) = G_t C_{t-1} G_t' + W_t, \qquad\qquad t = j, \ldots, k \\
W_t &= \frac{1-\delta_i}{\delta_i} P_{i,t}, \\
P_{i,t} &= G_t C_{t-1} G_t'
\end{aligned}
$$

$$\Rightarrow \quad C_t(k) = \frac{G^k C_t G'^k}{\delta^k}$$

- Smoothing (backward recursion) is not affected so in missingness periods:

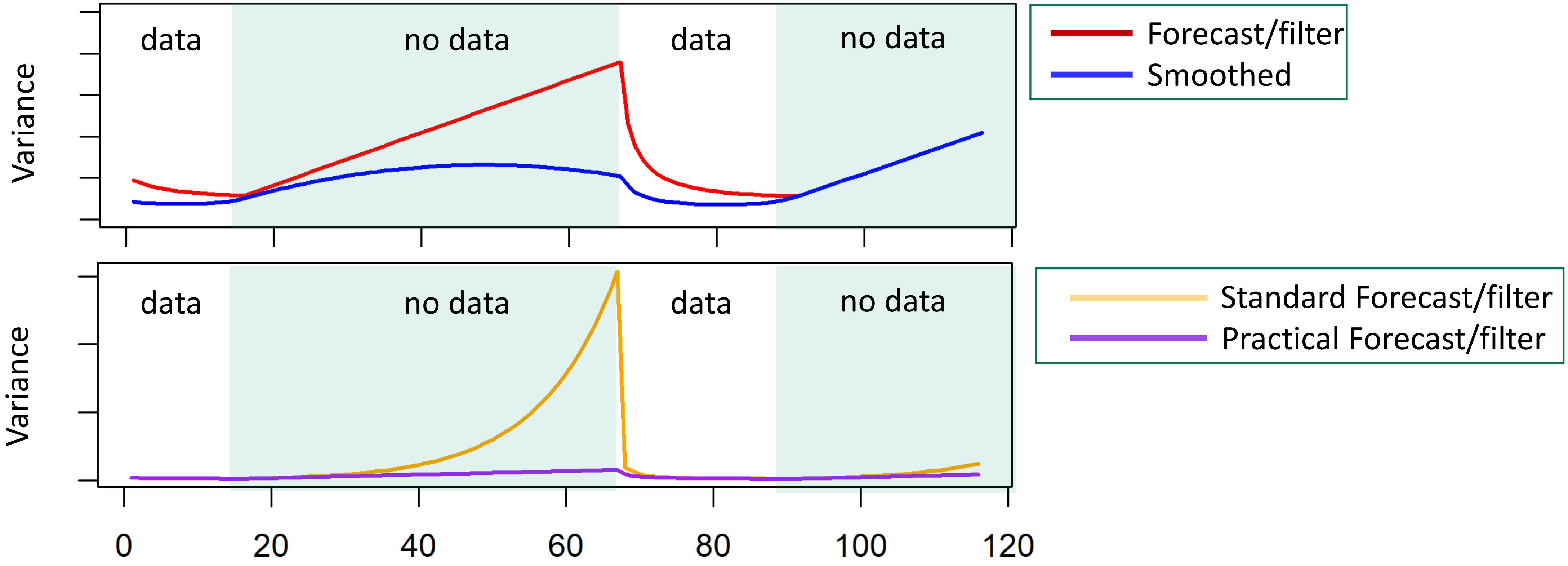


Practical Evolutional Covariance specification:

- Practical discounting constrains information loss to a linear rather than exponential rate during longer period missingness

$$R_t(k) = \frac{G^k C_t G'^k}{\delta^k}$$
$$if\ k > 1, \quad R_t = G^{k-1} C_{t+1} G'^{\,k-1}$$

- Harrison and West 1997
- In essence, fix W_t after forecast beyond 1 step
- Greatly constrains uncertainty where discount factors are low



Problem Overview

<u>In long-term monitoring, challenged by data with:</u>

1. Long periods of missingness
2. Non-stationarity on covariance (static covariance not appropriate)

- Kalman filtering and smoothing & discount factors are common tools to handle time series with missingness
- Discount common to handle covariance
- But evaluation of discounting in extended periods of missing data is lacking

<u>Objective</u>

- What is optimal criteria for selecting a discount factor?
  - If simulate missing periods with given discount factor, which criteria best recovers discount factor?
- What is the optimal covariance specification in forecasting during extended data missingness?
  - Compare criteria when missingness is introduced. What gives best inference to missing data/
  - How does practical vs standard discounting impact model selection when you have extended periods of missingness

Problem overview and applied context

DLM overview

Role of covariance specification on inference

Simulation and analysis overview

Results

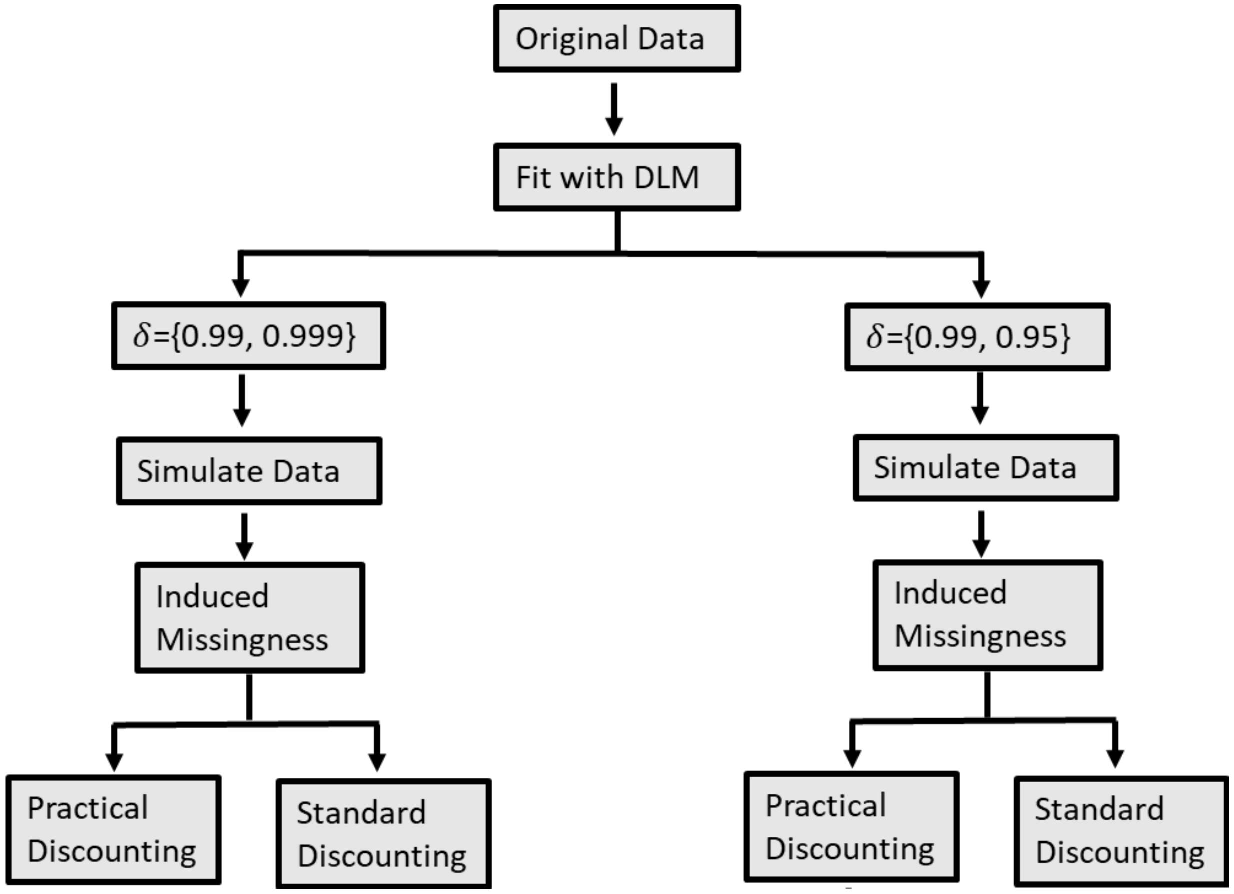
Discussion

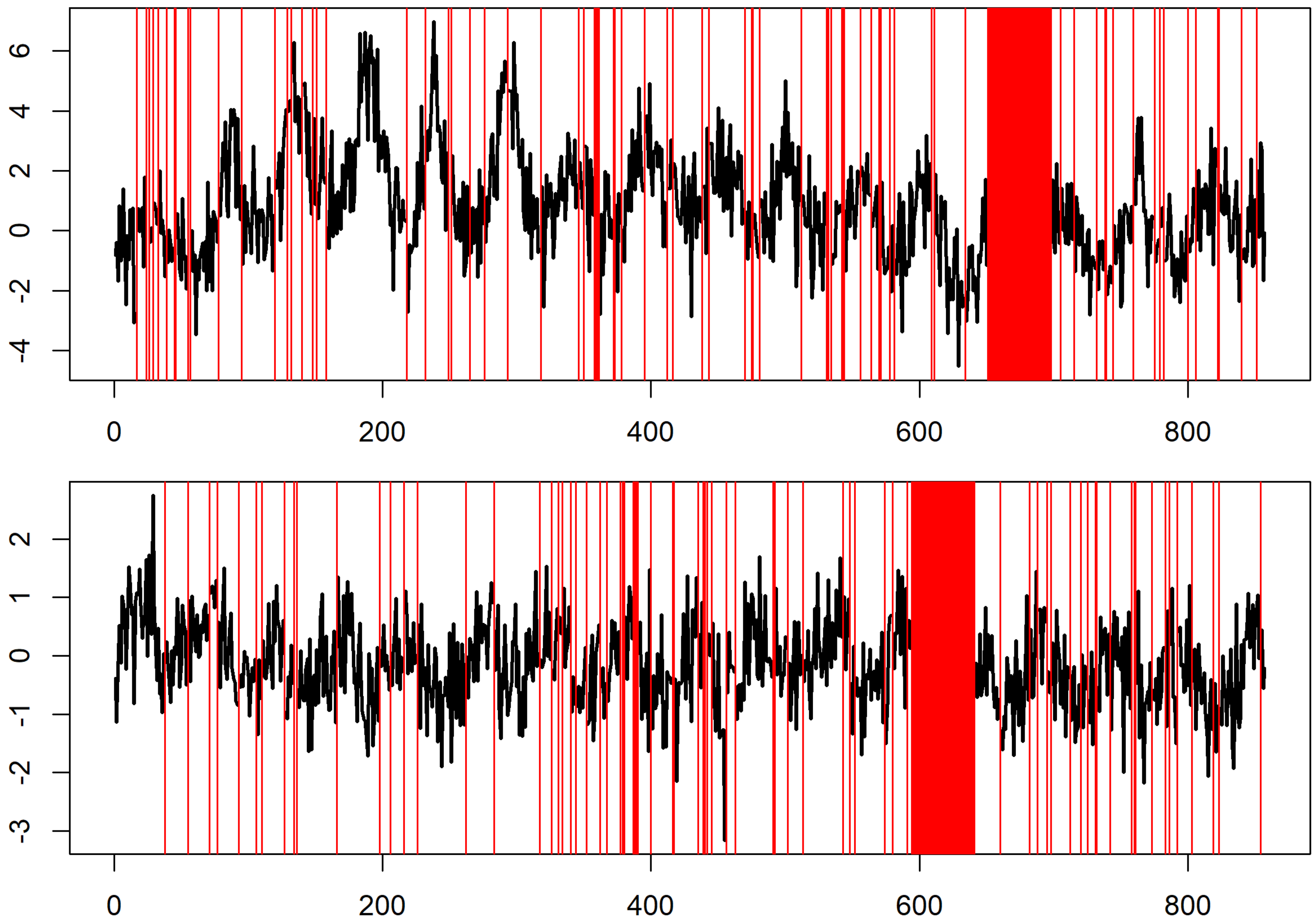
Simulation:

- Goal: simulate data with similar properties to real data but with known parameter values
- Fit a DLM with dynamic intercept and regression to original data with two sets of discount factors
  - {0.999, 0.99}
  - {0.99, 0.95}
- Calculate posterior mean of V and W_t
- Recursively draw latent states and observed values until the original data length is reached

$$\begin{aligned} Y_t &= \boldsymbol{F}_t \boldsymbol{\theta}_t + v_t, & v_t &\sim N(0, \boldsymbol{V}) \\ \boldsymbol{\theta}_t &= \boldsymbol{G}_t \boldsymbol{\theta}_{t-1} + w_t, & w_t &\sim N(0, \boldsymbol{W}) \end{aligned}$$

- In a copy of each simulated series, randomly introduce missingness with frequency patterns identical to the original data set




6
4
2
0
-2
-4
0
200
400
600
800
2
1
0
-1
-2
-3
0
200
400
600
800

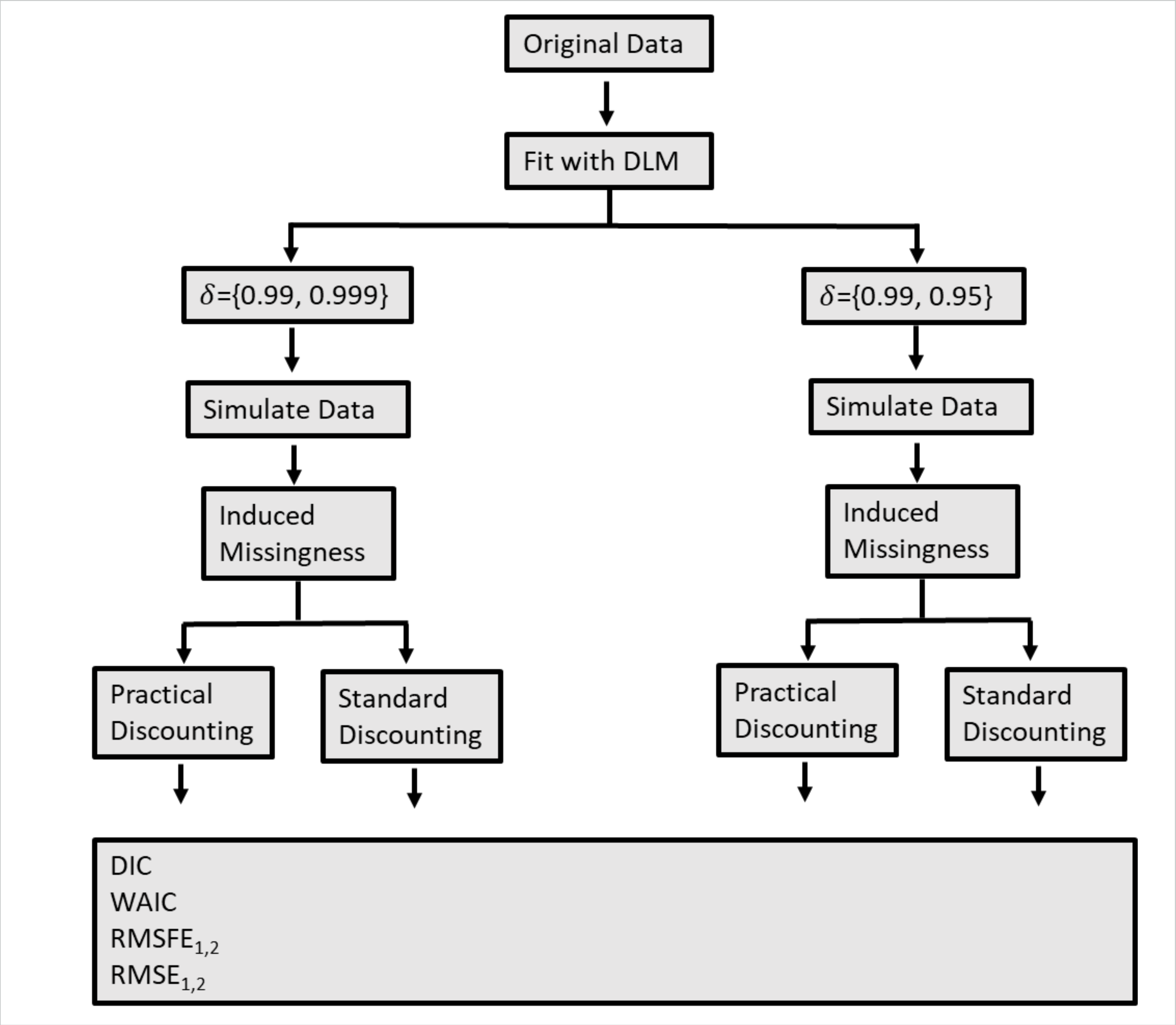
To inform model selection, several indices were considered:

- 6 criteria emphasizing difference between:
  - Within and out of sample accuracy:
    - DIC_4, WAIC vs. RMSE, RMSFE
  - Posterior mean point estimate vs integration over predictive density
    - $RMSE_1$ vs $RMSE_2$
    - $RMSFE_1$ vs. $RMSFE_2$
- RMSE calculated in missing data periods
- RMSFE calculated for each time step
- DIC_4 and WAIC calculated in periods of observed data

Metric	Version	Calculation
DIC_4		(Celeux et al. 2006)
$WAIC$		(Watanabe 2010)
$RMSE$	1	$RMSE_{r,1} = \sqrt[2]{\frac{\sum_{t=1}^{n}(y_t - \boldsymbol{F}_{t-1}\boldsymbol{\theta}_{t-1,i})^2}{n}}$
	2	$RMSE_{t,2} = \sqrt[2]{\frac{\sum_{i=1}^{\mathrm{r}}(y_t - \boldsymbol{F}_{t-1}\boldsymbol{\theta}_{t-1,i})^2}{r}}$
$RMSFE$	1	$RMSFE_{r,1} = \sqrt[2]{\frac{\sum_{t=1}^{n}(y_t - \boldsymbol{F}_{t-1}\boldsymbol{\theta}_{t-1,i})^2}{n}}$
	2	$RMSFE_{t,2} = \sqrt[2]{\frac{\sum_{i=1}^{r}(y_t - \boldsymbol{F}_{t-1}\boldsymbol{\theta}_{t-1,i})^2}{r}}$

Evaluation:

- Fit each series with practical/standard discounting
  - MCMC 10000
  - Burn-in 1000
- Calculate:
  - DIC_4
  - WAIC
  - $RMSFE_1$
  - $RMSFE_2$
  - $RMSE_1$
  - $RMSE_2$
- For RMSFE and RMSE metrics, calculate all pairwise posterior probabilities: $P(Metric_{\delta_A} < Metric_{\delta_{-A}})$
- Compare performance metrics and optimal discount sets for each data set and discounting method



Problem overview and applied context

DLM overview

Role of covariance specification on inference

Simulation and analysis overview

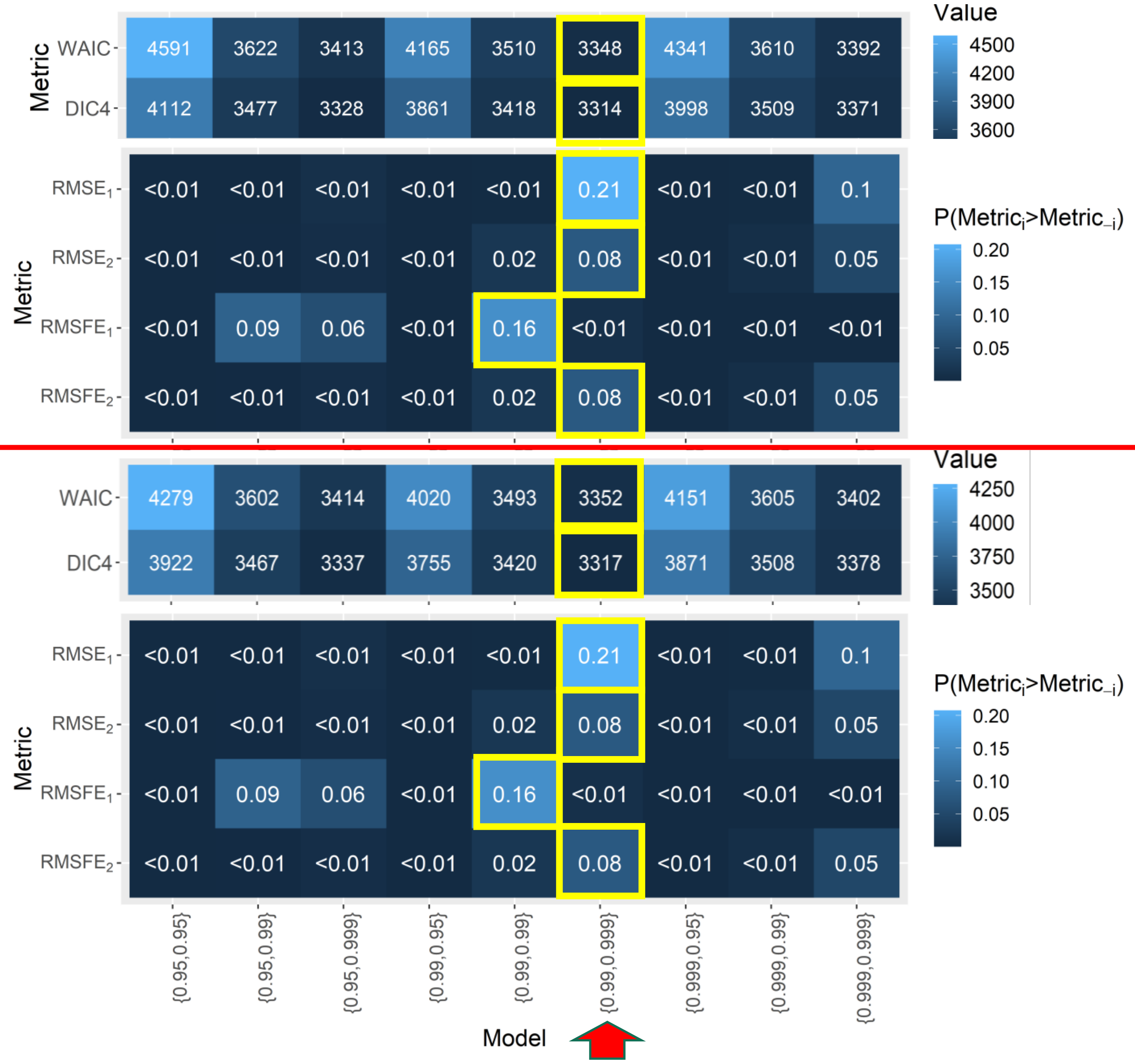
Results

Discussion

High Discount Data Generation Model (0.99, 0.999)

Standard

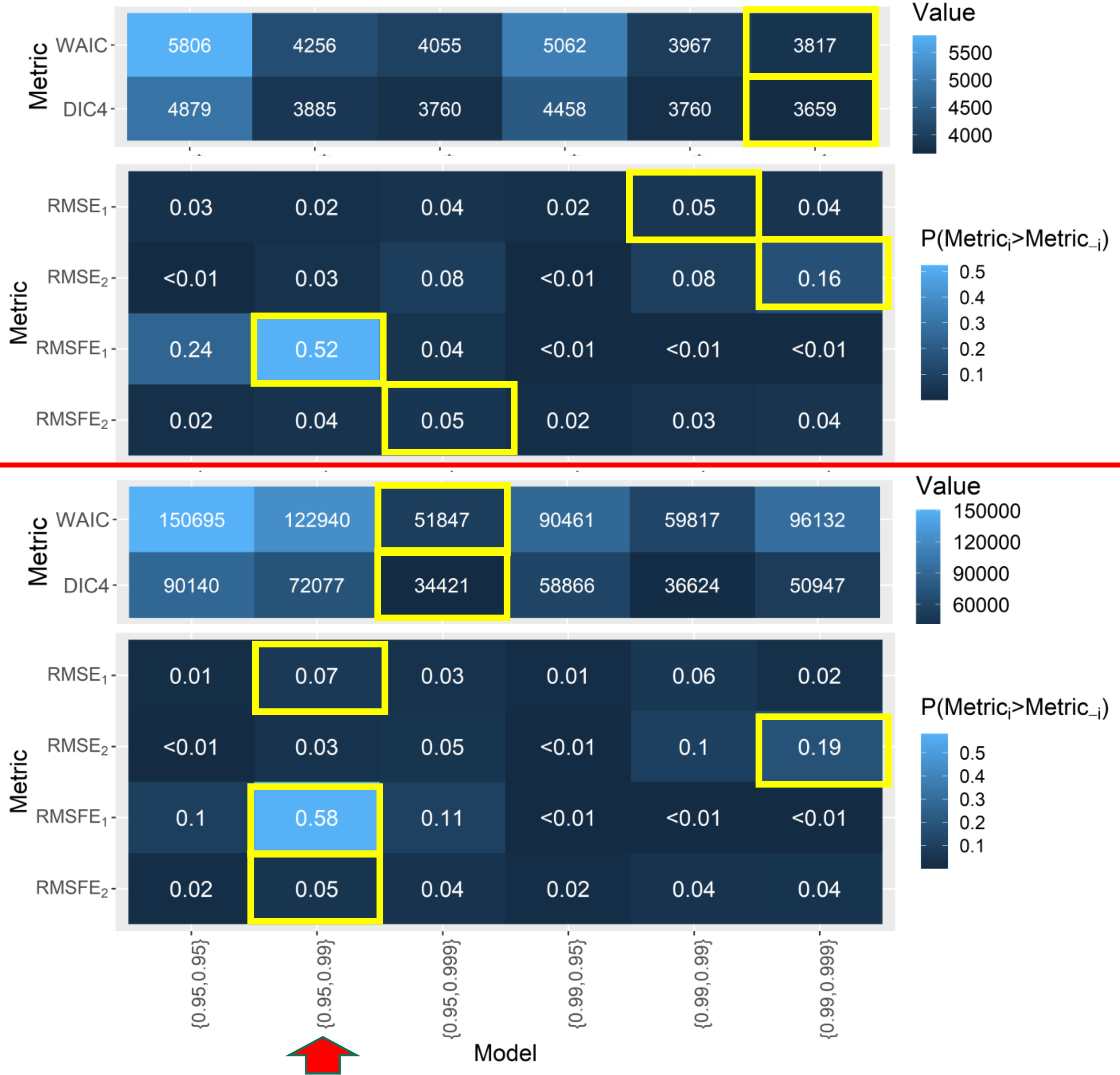
Practical



Low Discount Data Generation Model (0.95, 0.99)

Standard

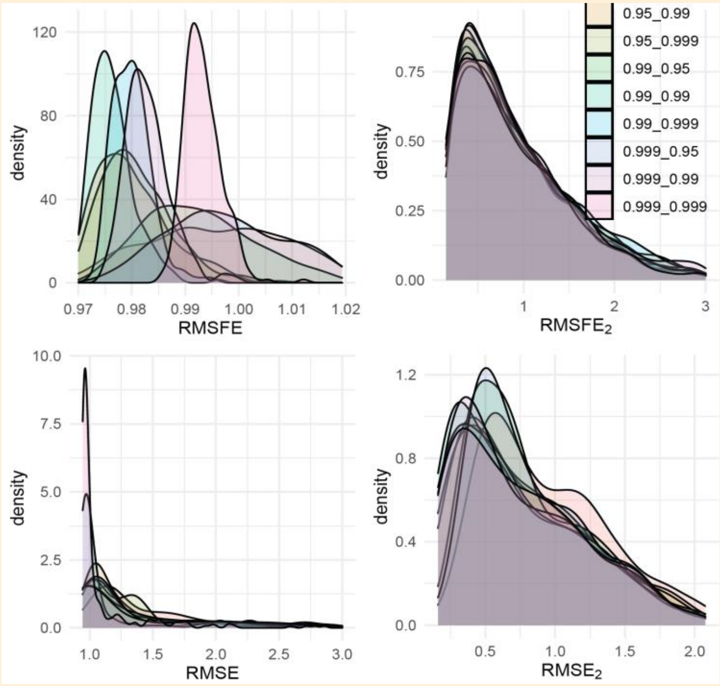
Practical



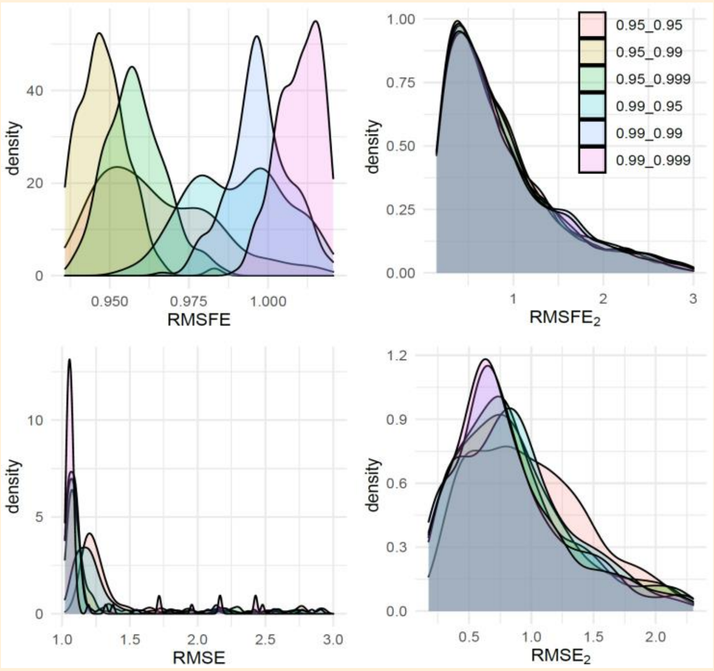
Criteria Evaluation

Data Generation Model	Method	RMSE
0.99, 0.999	Standard	0.95
	Practical	0.95
0.95, 0.99	Standard	0.98
	Practical	0.90

StandardLow



Standard High



- Under data simulated with a pair of high discount factors (0.999, 0.99), all metrics selected within 0.009 of the parameters for data generation
- Under all simulations, RMSFE showed the highest power to discriminate the proper discounting level
- The RMSE suggest that practical discounting will optimize performance in long-periods of missingness.
- DIC supports that practical discounting improves the model fit within sample.
- Although RMSE was a biased metric for model selection, particularly during prolonged periods of missingness, it still had utility in evaluating the performance of practical discounting in data with long period missingness.
- While RMSFE may be the optimal method for discount factor selection, it does not account for performance during long-period missingness as our metric of RMSE does. Therefore, results of RMSE in comparable models with practical and standard discounting provide an evaluation for this imputation method.

Problem overview and applied context

DLM overview

Role of covariance specification on inference

Simulation and analysis overview

Results

Discussion

Take-aways

Relatively consistent and accurate discount factor recovery under conditions of low variability

In conditions of low variability, practical vs. standard discounting:

→ Have minor impact on model fits

→ Does not dramatically impact performance metrics

Under high variability systems selection criteria matters

$RMSFE_1$ shows highest power among error criteria tested

WAIC, DIC4 are highly biased under standard discounting

$RMSFE_{1,2}$ and RMSE show similar performance and power under both practical and standard discounting

Acknowledgements

- Drs. Jing Wu, Susanne Menden-Deuer, Tatiana Rynearson
- NASA
  - NASA STEM -NNX15AI06H

- Kalman, R. E. 1960. “A New Approach to Linear Filtering and Prediction Problems.” *Journal of Fluids Engineering, Transactions of the ASME* 82 (1): 35–45. https://doi.org/10.1115/1.3662552.
- Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of machine learning research, 11(12).
- West, Mike, and Jeff Harrison. 1997. *Bayesian Forecasting and Dynamic Models*. 2nd ed. Verlag New York: Springer. https://doi.org/10.1007/b98971.