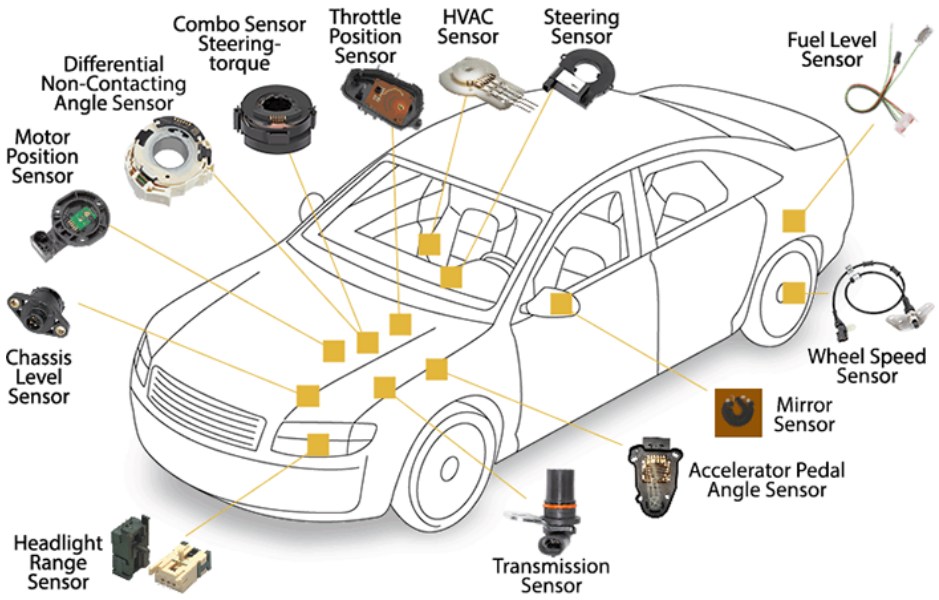
Variational Inference of Joint Models using Multivariate Gaussian Convolution Processes

Xubo Yue, Ph.D. Student
Raed Al Kontar, Assistant Professor

The University of Michigan, Ann Arbor

May 20, 2019

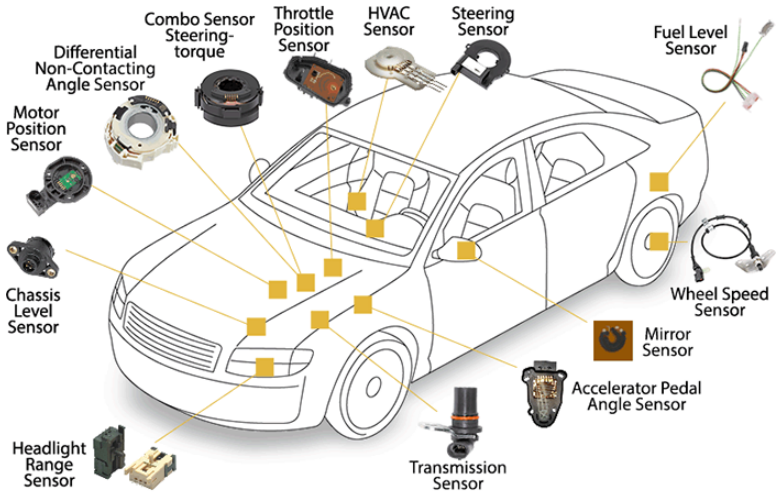
Signal Measurement Time-series Data



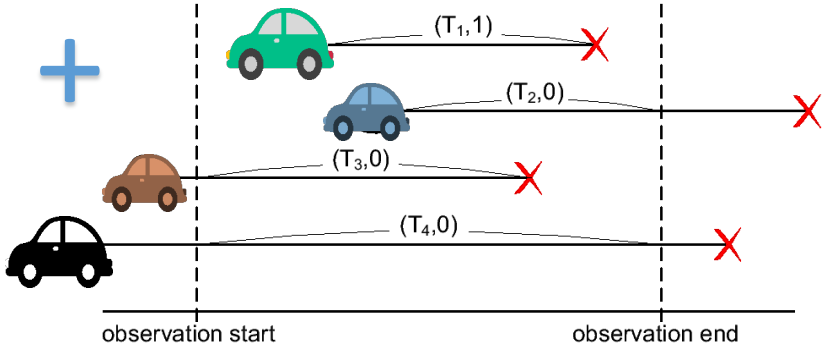
https://www.raypcb.com/the-schematic-diagram-of-automotive-sensor-signal-conditioning/
http://clipart-library.com/cartoon-vehicle-cliparts.html

Background

Time-series Data



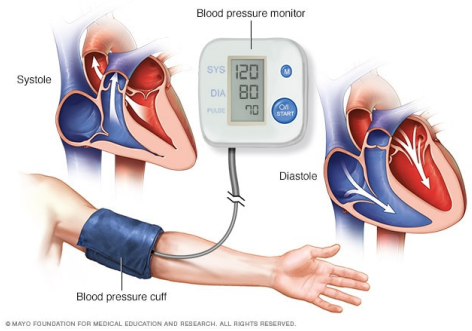
+

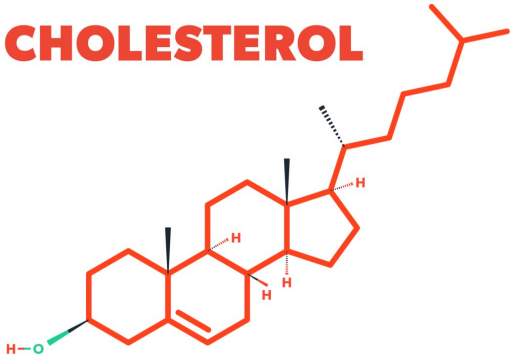


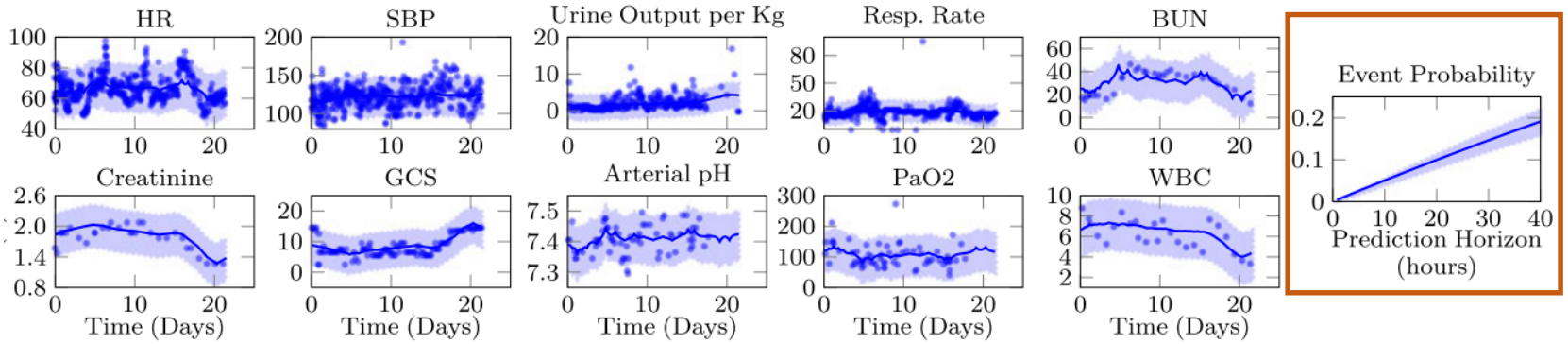
Lifetime Data

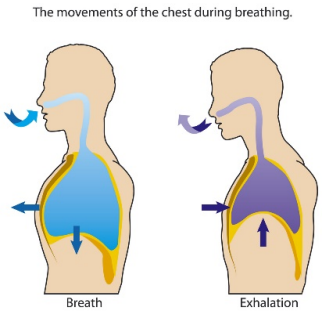
https://www.raypcb.com/the-schematic-diagram-of-automotive-sensor-signal-conditioning/
http://clipart-library.com/cartoon-vehicle-cliparts.html

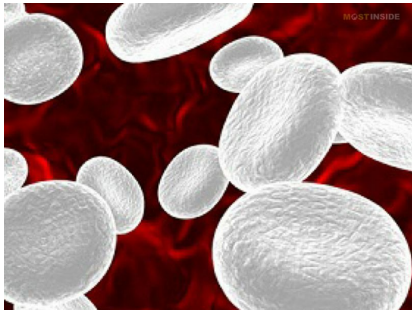
Background







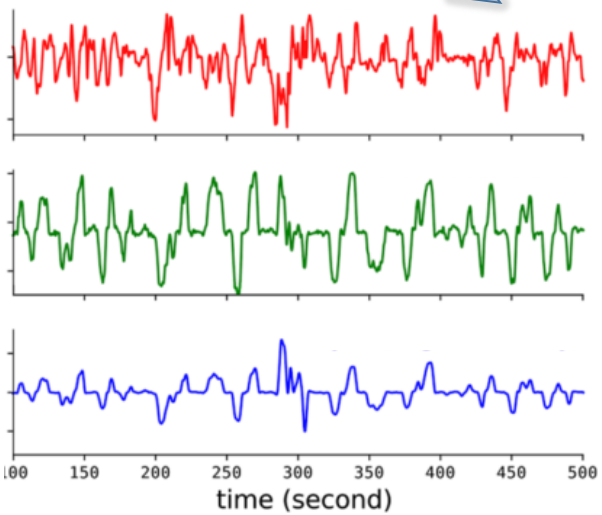


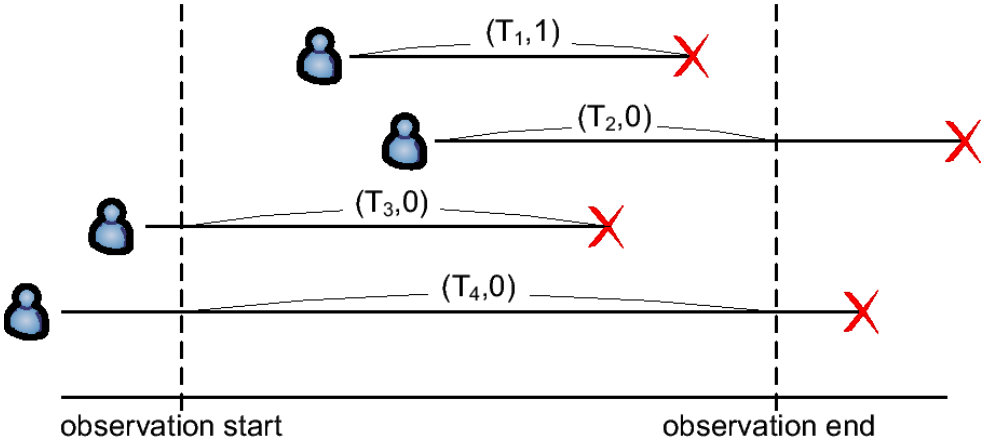


https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8013802

Background

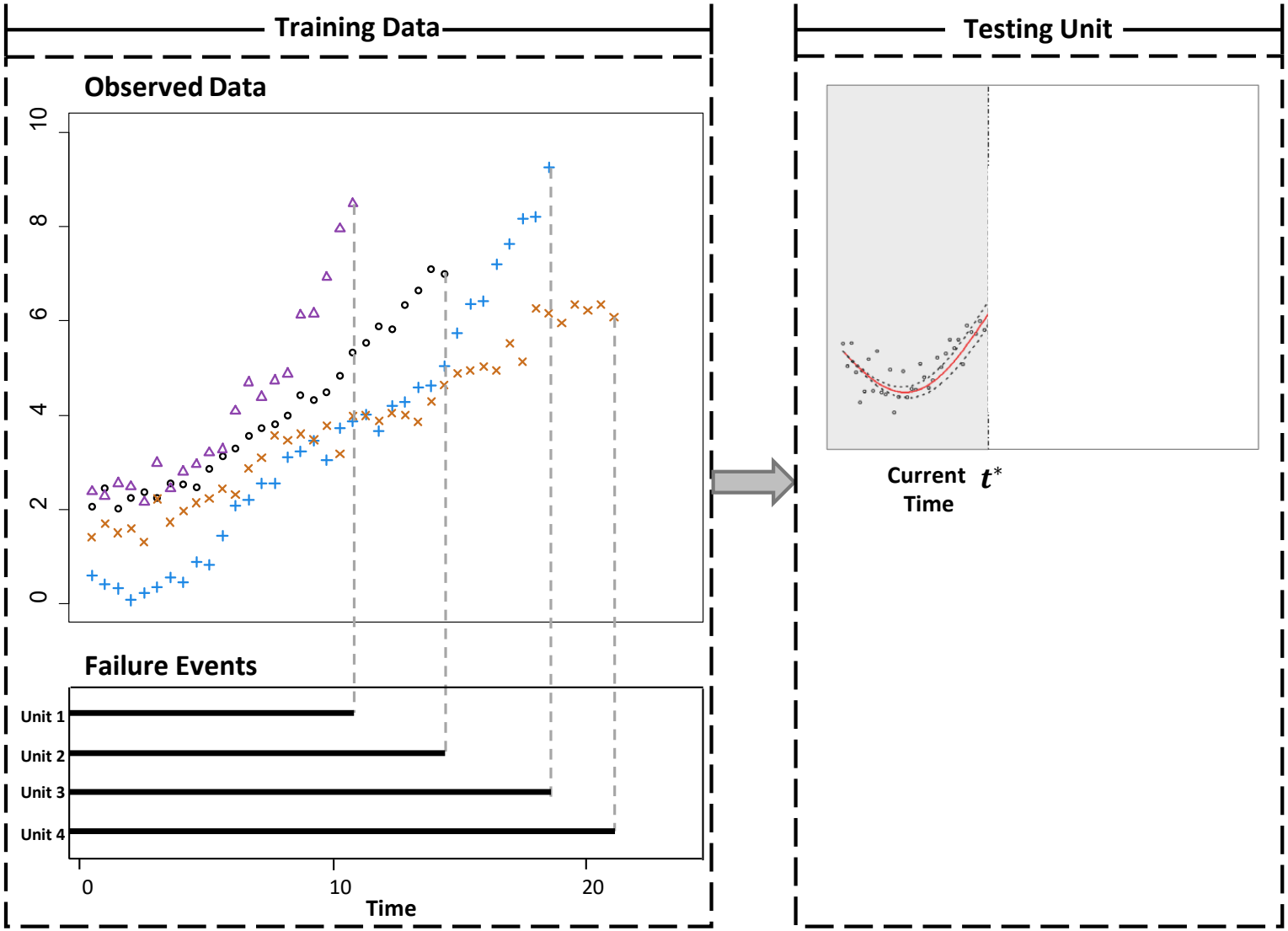
Time-series Data + Survival Data



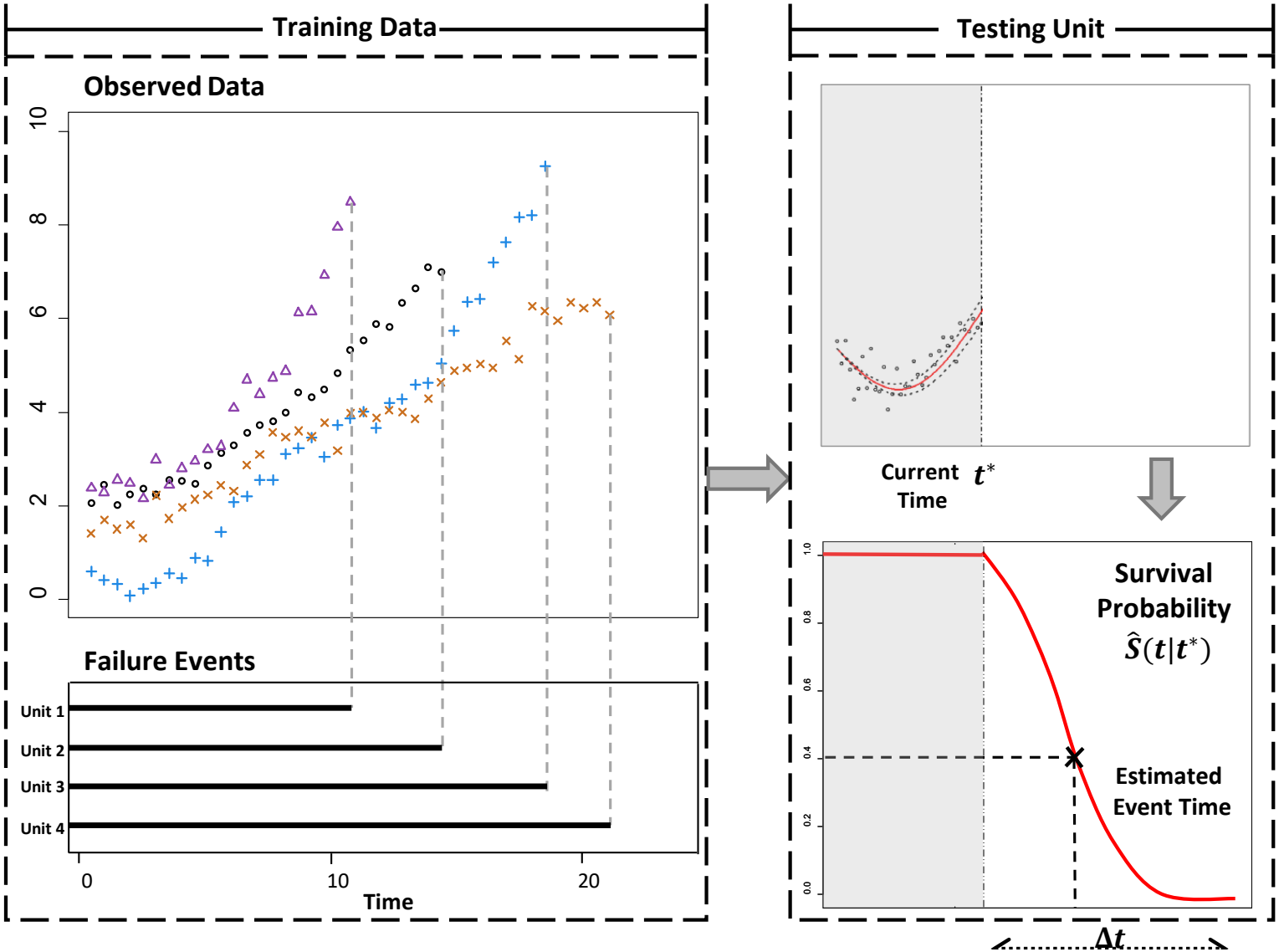


https://www.sentiance.com/2017/09/25/passenger-and-driver/
https://labblog.uofmhealth.org/industry-dx/for-optimizing-patient-data-track-and-share-every-moment

Framework


Training Data
Observed Data
0
2
4
6
8
10
Failure Events
Unit 1
Unit 2
Unit 3
Unit 4
0
10
20
Time
Testing Unit
Current Time
t^*

Framework


Training Data
Observed Data
Failure Events
Unit 1
Unit 2
Unit 3
Unit 4
Time
Testing Unit
Current Time t^*
Survival Probability $\hat{S}(t|t^*)$
Estimated Event Time
Δt

Data Structure

Data: $\{\boldsymbol{\mathcal{D}}_i\}_{i=1}^N = \{(V_i, \delta_i, \boldsymbol{w}_i, \boldsymbol{Y}_i)\}_{i=1}^N$.

- Outcome: occurrence of an event or losing information (censored).
- Event time $V_i = \min\{C_i, T_i\} \in \mathbb{R}_+$.
- Event indicator $\delta_i \in \{0, 1\}$.
- Fixed features $\boldsymbol{w}_i$ (e.g., gender).
- Observed longitudinal data $\boldsymbol{Y}_i$.

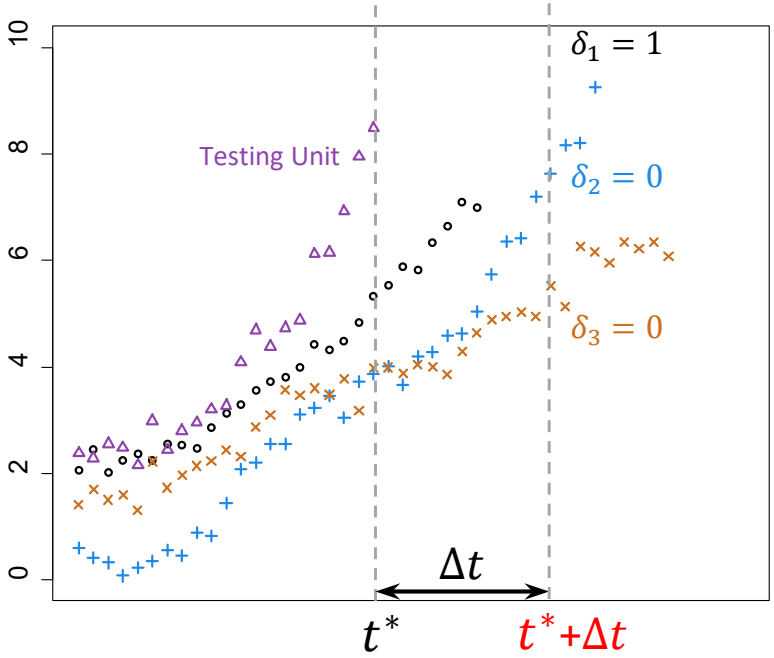
How?

- What is the most intuitive approach to analyze these data?

Survival Data - Logistic Regression/SVM

Time-series Classification

- Event data is transformed into binary labels $\delta_i = 1/0$ denoting whether units failed or not within the time interval $[t^*, \Delta t + t^*]$.
- The time-fixed covariate w_i and the last observed signal measurement at t^* are used as the model predictors.



Unit	Event Indicator δ at $t^* + \Delta t$	Longitudinal Signal Y at t^*
1	1	5.00
2	0	3.80
3	0	3.85

Problems?

- Fail to consider the functional form of longitudinal data: only use summary statistics.
- Unable to impute missing data.
- Fortunately, we have joint model (2008)!

Statistical Model

Joint Model

$$h_i(t) = h_0(t)\exp[\boldsymbol{\gamma}^T \boldsymbol{w}_i + \beta f_i(t)].$$

- $h_i(t)$: hazard; instantaneous failure rate at t.
- Data: $\boldsymbol{\mathcal{D}}_i = (V_i, \delta_i, \boldsymbol{w}_i, \boldsymbol{Y}_i)$.
- $\boldsymbol{\gamma}, \beta$: regression coefficient.
- $f_i(t)$: underlying latent values. Modeled as mixed-effect model.

$$y_i(t) = f_i(t) + \epsilon_i(t).$$

- $h_0(t)$: baseline hazard.

Parametric Model

Parametric Joint Model

- Linear mixed-effect model:

$$y_i(t) = f_i(t) + \epsilon_i(t) = \boldsymbol{z}^T(t)\boldsymbol{b}_i + \epsilon_i(t).$$

- e.g., $y_i(t) = b_{i0} + b_{1t} + b_{i2}t^2$.

Parametric Model

Parametric Joint Model

- Linear mixed-effect model:

$$y_i(t) = f_i(t) + \epsilon_i(t) = \boldsymbol{z}^T(t)\boldsymbol{b}_i + \epsilon_i(t).$$

- e.g., $y_i(t) = b_{i0} + b_{1t} + b_{i2}t^2$.
- Likelihood function is intractable.
- Plug estimated parameters into the survival model (two-stage inference).

Parametric Model

Parametric Joint Model

- Linear mixed-effect model:

$$y_i(t) = f_i(t) + \epsilon_i(t) = \boldsymbol{z}^T(t)\boldsymbol{b}_i + \epsilon_i(t).$$

- e.g., $y_i(t) = b_{i0} + b_{1t} + b_{i2}t^2$.
- Likelihood function is intractable.
- Plug estimated parameters into the survival model (two-stage inference).

Remark

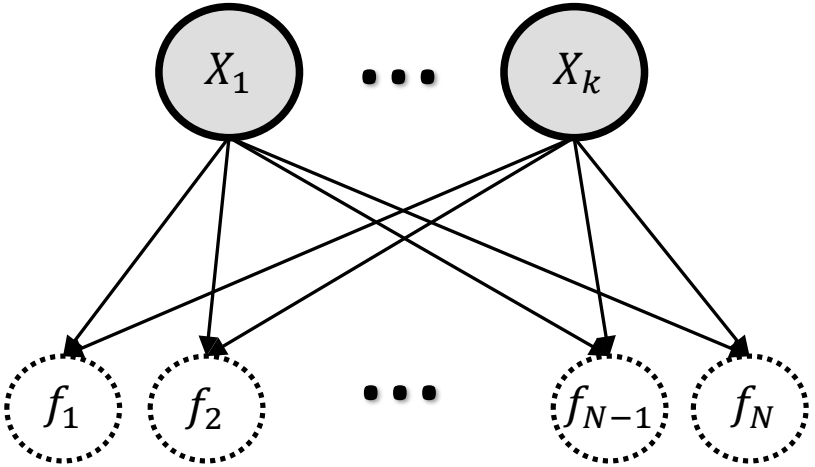
- Strong parametric assumption.
- Fail to propagate the uncertainty from time-series data to the survival component.

First Contribution

Contribution 1

- Introduce a flexible non-parametric mixed-effect model.
- Flexible: ability to handle heterogeneous data.
- Scalable: efficient for big data.

Multivariate Gaussian Convolution Process (MGCP)

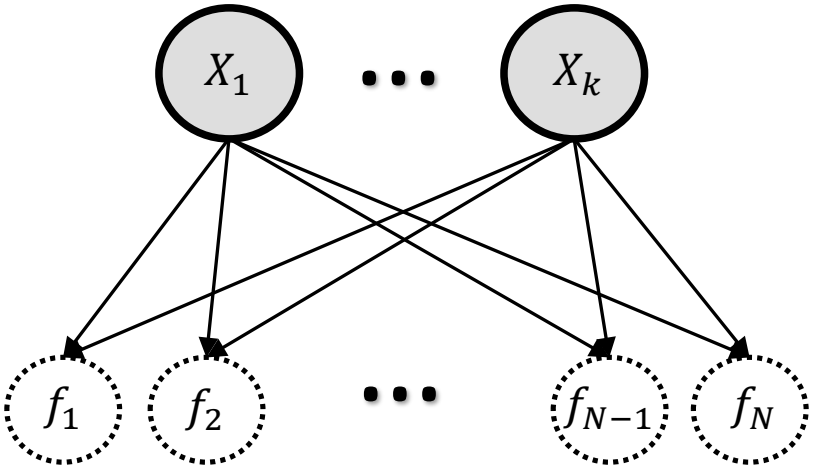


Model

- MGP:

$$\begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_N(t) \end{bmatrix} = \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_N(t) \end{bmatrix} + \begin{bmatrix} \epsilon_1(t) \\ \epsilon_2(t) \\ \vdots \\ \epsilon_N(t) \end{bmatrix}$$

Multivariate Gaussian Convolution Process (MGCP)



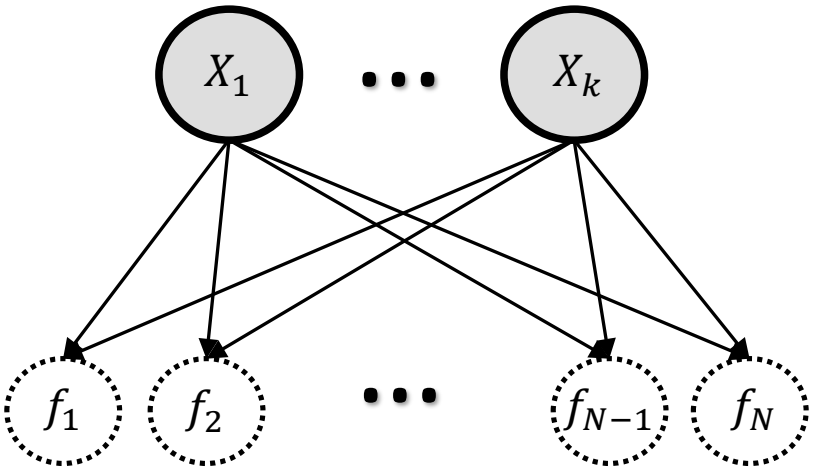
Convolution and Smoothing Kernels

- Each signal is expressed as a convolution of latent functions drawn from a GP. The Convolution of GP is GP.

$$f_i(t) = \sum_{k=1}^{K} \int_{\mathbb{R}} G_{i,k}(t-u) X_k(u) du.$$

- Bayesian interpretation. Alternative to neural network, etc.

Multivariate Gaussian Convolution Process (MGCP)



Contribution 1 (Recap)

- Introduce a flexible and scalable non-parametric mixed-effect model.
- Flexible: ability to handle heterogeneous data by shared latent functions.
- Scalable: efficient for big data.

MGCP - Sparse

Complexity

- Time complexity $\mathcal{O}(n^3)$, n is number of data points for all individuals.
- Take several hours even for small data.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{1,1} & \cdots & \Sigma_{1,n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n,1} & \cdots & \Sigma_{n,n} \end{bmatrix} \Longrightarrow \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \Sigma_{1,1} & \cdots & \Sigma_{1,n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n,1} & \cdots & \Sigma_{n,n} \end{bmatrix}^{-1}$$

100 units with 300 observations:

N= 30,000

MGCP - Sparse

Contribution 1 (Recap)

- Introduce a flexible and scalable non-parametric mixed-effect model.
- Flexible: ability to handle heterogeneous data.
- Scalable: efficient for big data.

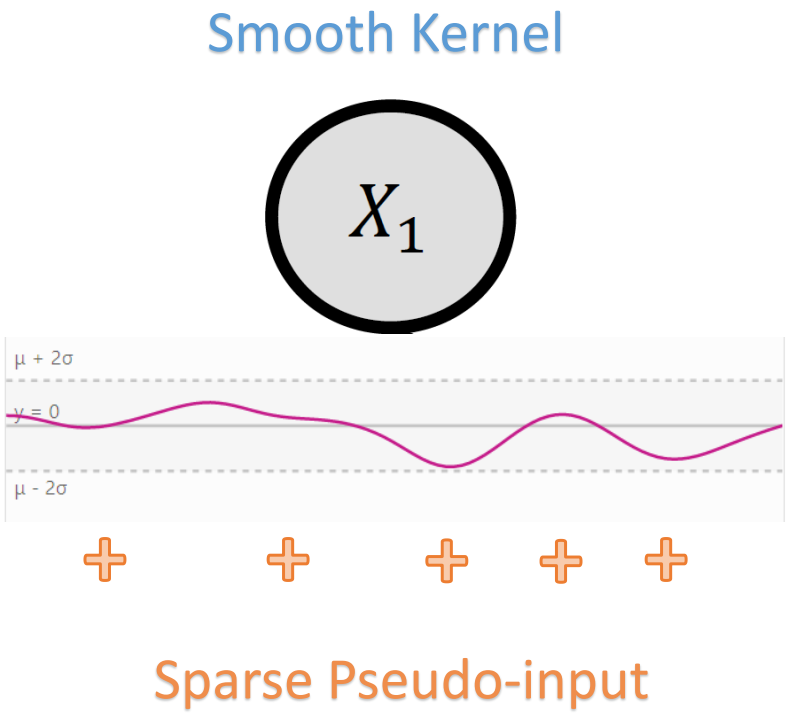
Pseudo-inputs

- Introduce $M(\ll n)$ pseudo-inputs from the latent functions denoted as $\boldsymbol{X}_k(\mathcal{Z}) = [X_k(z_1), \ldots, X_k(z_M)]^T$ where $\mathcal{Z} = \{z_i\}_{i=1}^M$.
- Time complexity $\mathcal{O}(n^3) \rightarrow \mathcal{O}(nM^2)$ (by the Woodbury formula).

MGCP - Sparse

Pseudo-inputs

- M pseudo-inputs to the latent functions.
- Use finite samples from latent functions.



Short Summary

MGCP+Survival Model→ Joint Model!

- Apply the sparse MGCP to the mixed-effect model.
- Joint Model:

$$h_i(t) = h_0(t) \exp\left[\boldsymbol{\gamma}^T \boldsymbol{w}_i + \beta \sum_{k=1}^{K} \int_{\mathbb{R}} G_{i,k}(t-u) X_k(u) du\right].$$

- Likelihood function is intractable.

Short Summary

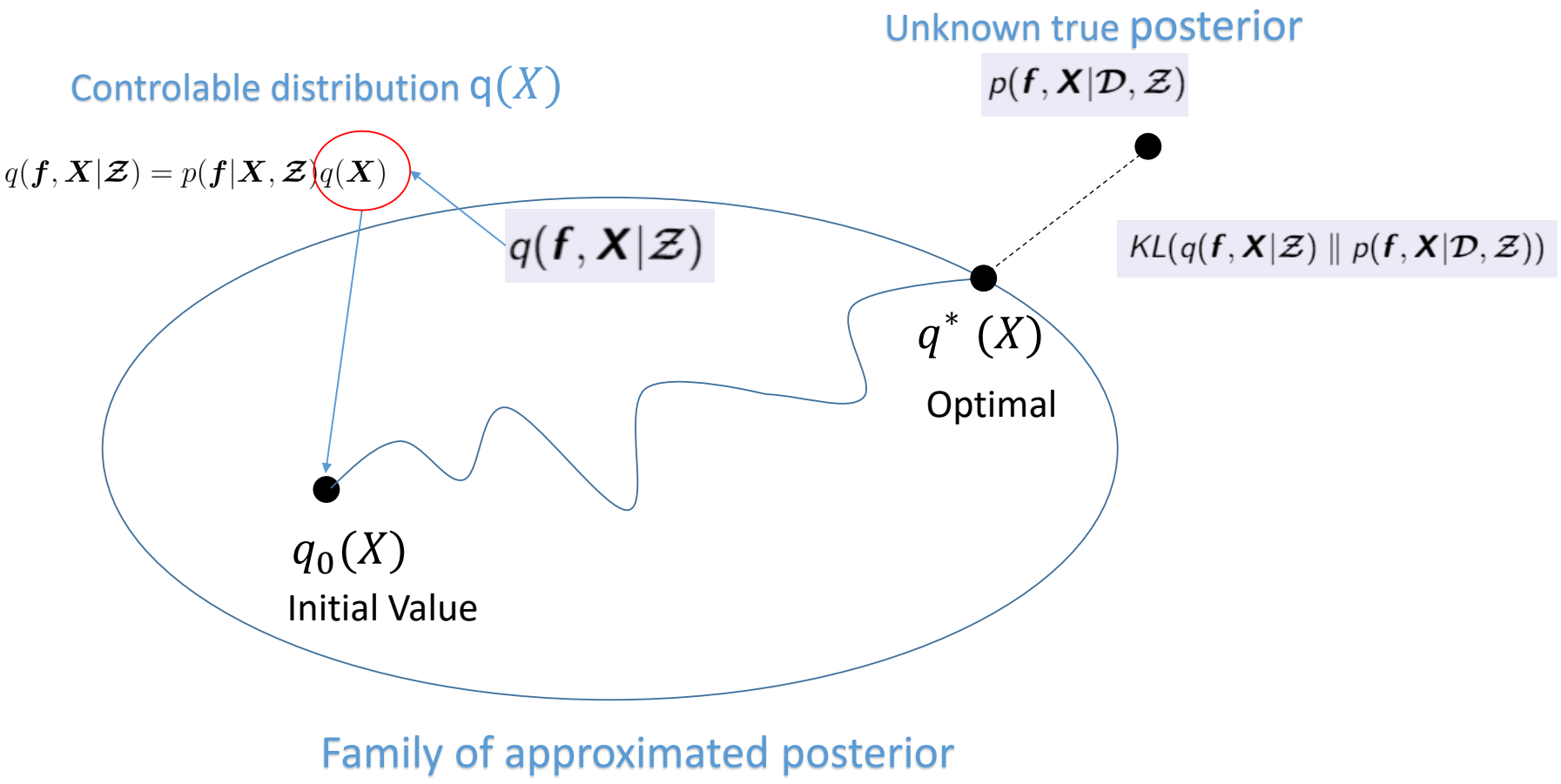
Contribution 1 (Recap)

- Introduce a flexible and scalable non-parametric mixed-effect model.
- Flexible: ability to handle heterogeneous data.
- Scalable: efficient for big data.

Contribution 2

- Introduce a variational inference framework to jointly estimate all parameters simultaneously.
- Big pic: find the lower bound of the likelihood function.
- VI provides a penalized likelihood function and a sparse approximation to the covariance matrix.

Joint Model and Variational Inference - 1



https://nips.cc/Conferences/2016/Schedule?showEvent=6199

Joint Model and Variational Inference - 2

Variational Inference

- Variational distribution $q(\boldsymbol{f}, \boldsymbol{X}|\mathcal{Z}) = p(\boldsymbol{f}|\boldsymbol{X}, \mathcal{Z})q(\boldsymbol{X})$.

$$
\begin{aligned}
& KL(q(\boldsymbol{f}, \boldsymbol{X}|\mathcal{Z}) \parallel p(\boldsymbol{f}, \boldsymbol{X}|\mathcal{D}, \mathcal{Z})) \\
&= \log p(\mathcal{D}) - \int q(\boldsymbol{f}, \boldsymbol{X}|\mathcal{Z}) \log \frac{p(\boldsymbol{f}, \boldsymbol{X}, \mathcal{D}|\mathcal{Z})}{q(\boldsymbol{f}, \boldsymbol{X}|\mathcal{Z})} d\boldsymbol{X} d\boldsymbol{f} \\
&= \log p(\mathcal{D}) - \mathcal{L} \geq 0.
\end{aligned}
$$

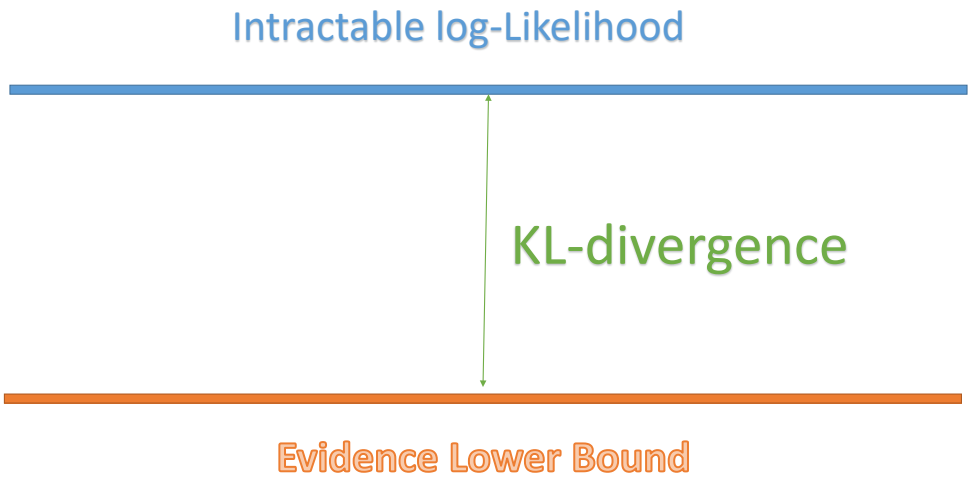
- Select the optimal distribution $q(\boldsymbol{X})$ to maximize $\mathcal{L}$.

Joint Model and Variational Inference - 2

Variational Inference

- Variational distribution $q(\boldsymbol{f}, \boldsymbol{X}|\mathcal{Z}) = p(\boldsymbol{f}|\boldsymbol{X}, \mathcal{Z})q(\boldsymbol{X})$.

$$
\begin{aligned}
& KL(q(\boldsymbol{f}, \boldsymbol{X}|\mathcal{Z}) \parallel p(\boldsymbol{f}, \boldsymbol{X}|\mathcal{D}, \mathcal{Z})) \\
&= \log p(\mathcal{D}) - \int q(\boldsymbol{f}, \boldsymbol{X}|\mathcal{Z}) \log \frac{p(\boldsymbol{f}, \boldsymbol{X}, \mathcal{D}|\mathcal{Z})}{q(\boldsymbol{f}, \boldsymbol{X}|\mathcal{Z})} d\boldsymbol{X} d\boldsymbol{f} \\
&= \log p(\mathcal{D}) - \mathcal{L} \geq 0.
\end{aligned}
$$



Evidence Lower Bound

Lemma

$$\mathcal{L}^* = \log[\mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{I} + \boldsymbol{Q})] + \zeta + \sum_{i=1}^{N} \Big\{\delta_i \big\{ \log h_0(V_i) + \boldsymbol{\gamma}^T \boldsymbol{w}_i + \beta \mu_i(V_i) \big\}$$
$$- \int_0^{V_i} h_0(u) \exp[\boldsymbol{\gamma}^T \boldsymbol{w}_i] \exp\big[\beta[\mu_i(u) + \frac{1}{2}\sigma_i^2(u)]\big] du \Big\}.$$

- $\boldsymbol{Q} = \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{X}} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{f}}$ and $\zeta = -\frac{1}{2\sigma_\epsilon^2} \mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})$.
- $\mu_i(u) = \boldsymbol{K}_{f_i(u),\boldsymbol{X}} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1} \sigma_\epsilon^{-2} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}} (\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}} + \sigma_\epsilon^{-2} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{f}} \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{X}})^{-1} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{f}} \boldsymbol{Y}$.
- $\sigma_i^2(u) = \boldsymbol{K}_{f_i(u),f_i(u)} - \boldsymbol{K}_{f_i(u),\boldsymbol{X}} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1} (\boldsymbol{I} - \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}} (\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}} + \sigma_\epsilon^{-2} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{f}} \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{X}})^{-1} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}) \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1} \boldsymbol{K}_{\boldsymbol{X},f_i(u)}$.

Sparse

Pseudo-inputs

- M (e.g., 5-10) pseudo-inputs to the latent functions.

Large Matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \Sigma_{1,1} & \cdots & \Sigma_{1,n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n,1} & \cdots & \Sigma_{n,n} \end{bmatrix}$$

Approximation

$$(\sigma_\epsilon^2 I + K_{f,X} K_{X,X}^{-1} K_{X,f})^{-1}$$
$$= \sigma_\epsilon^{-2} I - \sigma_\epsilon^{-4} K_{f,X} (K_{X,X} + \sigma_\epsilon^{-2} K_{X,f} K_{f,X})^{-1} K_{X,f}$$

$\mathcal{O}(n^3) \rightarrow \mathcal{O}(nM^2)$

$$\Sigma = K_{f,X} K_{X,X}^{-1} \; K_{X,f}$$

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

Sparse Low-rank Matrix

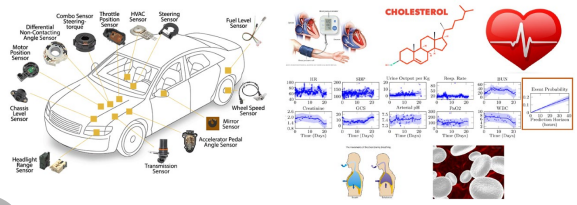
Woodbury Formula

Short Summary

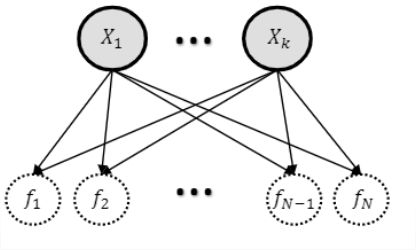
Joint Model

$$h_i(t) = h_0(t)\exp[\gamma^T w_i + \beta f_i(t)].$$

Application



Map

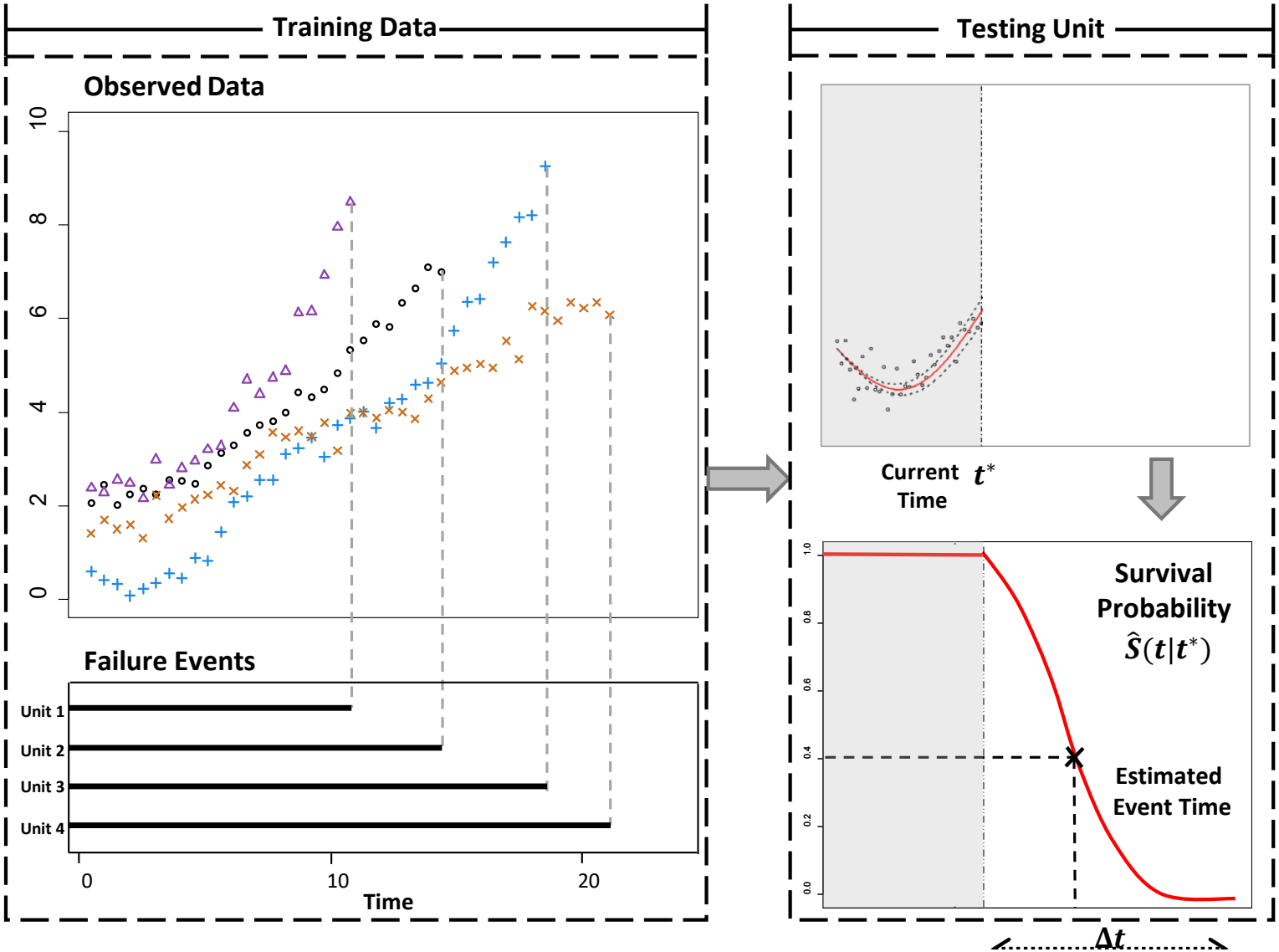


MGCP

$$\mathcal{L}^* = \log[\mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{I} + \boldsymbol{Q})] + \zeta + \sum_{i=1}^{N} \{\delta_i \{ \log h_0(V_i) + \gamma^T \boldsymbol{w}_i + \beta \mu_i(V_i)\}$$
$$- \int_0^{V_i} h_0(u) \exp[\gamma^T \boldsymbol{w}_i] \exp\left[\beta[\mu_i(u) + \frac{1}{2}\sigma_i^2(u)]\right] du\}.$$

Inference

Short Summary



Event Prediction

Survival Probability Prediction

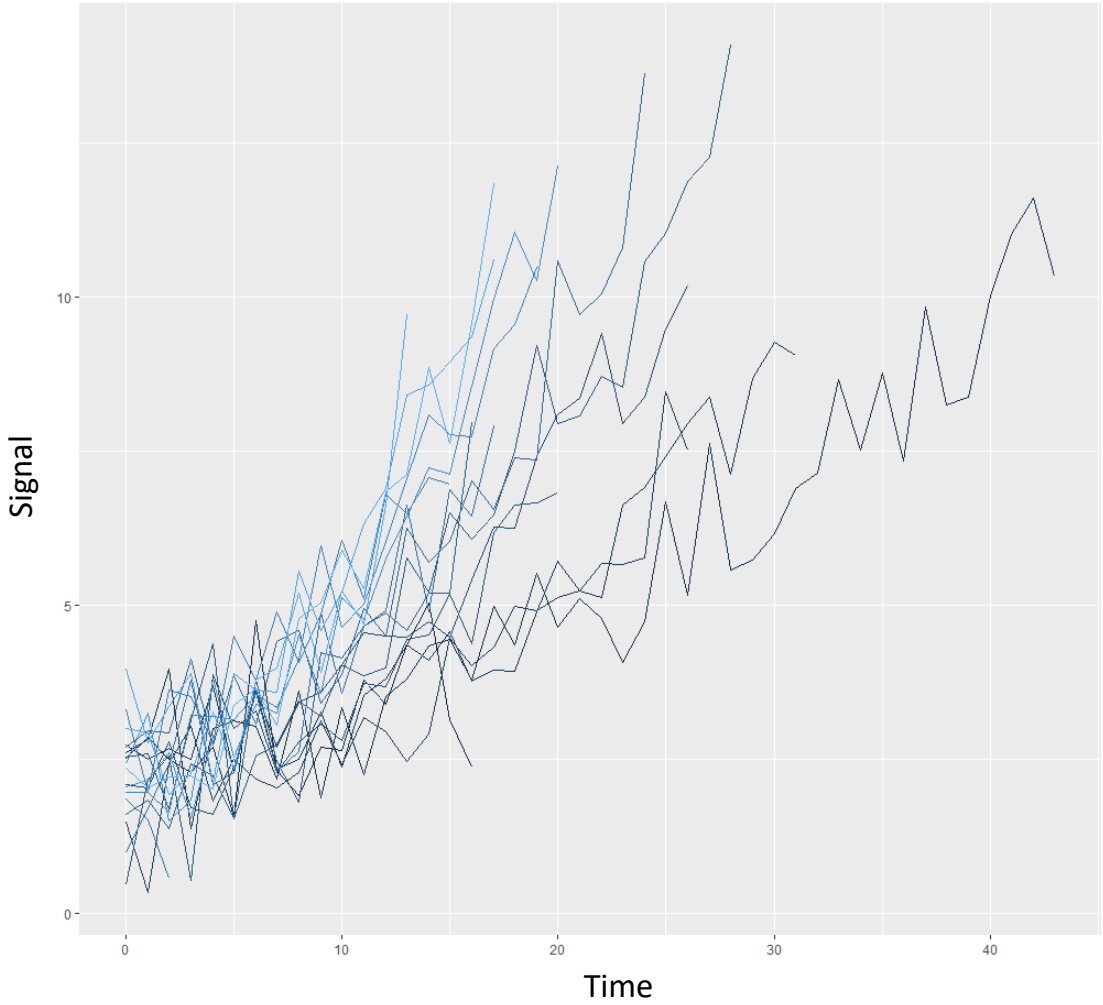
$$\hat{P}_{\Delta t} = \hat{S}(t^* + \Delta t | t^*, \boldsymbol{w}_N, \boldsymbol{f}_N) = \frac{\hat{S}(t^* + \Delta t | \boldsymbol{w}_N, \boldsymbol{f}_N)}{\hat{S}(t^* | \boldsymbol{w}_N, \boldsymbol{f}_N)}$$
$$= \exp\{-\int_{t^*}^{t^*+\Delta t} \hat{h}_0(u) \exp\left[\hat{\gamma}^T \boldsymbol{w}_N + \hat{\beta} \boldsymbol{A} \boldsymbol{D}^{-1} \boldsymbol{Y}\right] du\}.$$

Prediction of f

$$p(f_N(T^*)|\boldsymbol{Y}) = \int p(f_N(T^*)|\boldsymbol{X}) p(\boldsymbol{X}|\boldsymbol{Y}) d\boldsymbol{X}$$
$$= \mathcal{N}(\boldsymbol{A}\boldsymbol{D}^{-1}\boldsymbol{Y}, \boldsymbol{K}_{f_N(T^*), f_N(T^*)} - \boldsymbol{A}\boldsymbol{D}^{-1}\boldsymbol{A}^T),$$

- $\boldsymbol{A} = \boldsymbol{K}_{f_N(T^*), \boldsymbol{X}} \boldsymbol{K}_{\boldsymbol{X}, \boldsymbol{X}}^{-1} \boldsymbol{K}_{\boldsymbol{X}, \boldsymbol{f}}$ and $\boldsymbol{D} = \boldsymbol{Q} + \sigma_\epsilon^2 \boldsymbol{I}$.

Simulation - Training Units (Parametric)

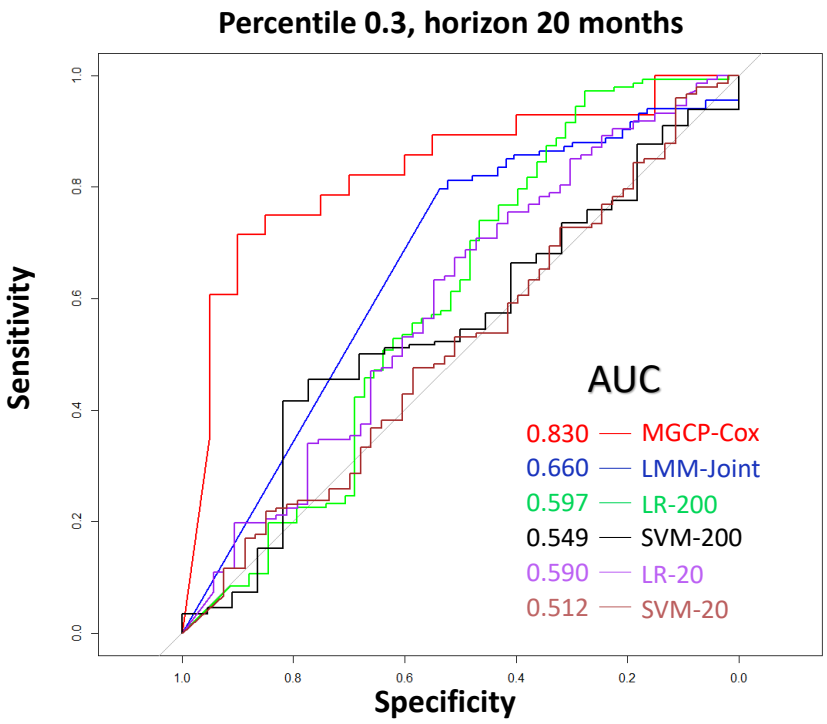


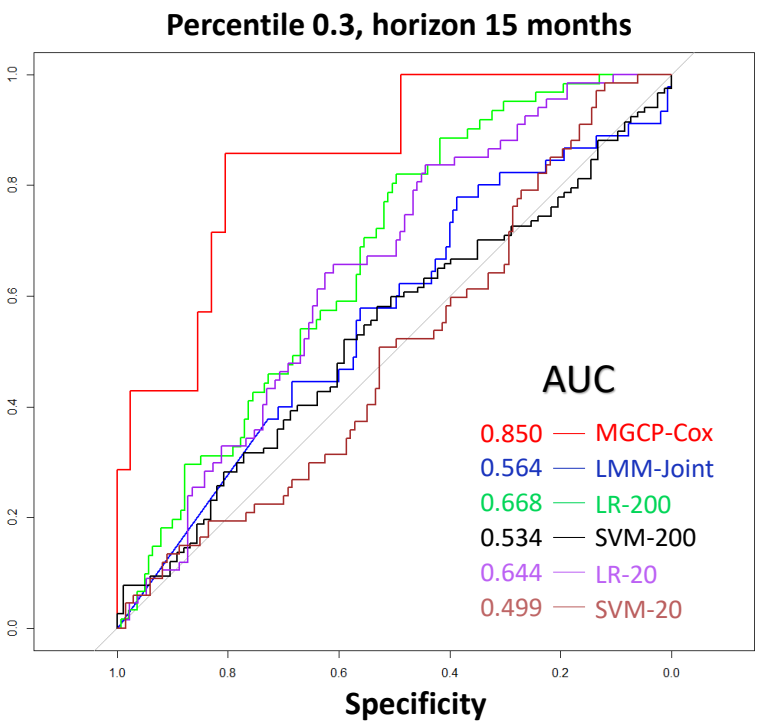
$$y = b_{i0} + b_{i1}t^{1.2} + b_{i2}t^{1.7} + \epsilon_i(t).$$

Simulation - Classification Problem

- Sample size $N = 20$.
- Predict survival probability within a future time interval Δt.
- Transform probabilities into binary variables by a certain classification threshold.
- Performance is evaluated the the Receiver Operating Characteristic (ROC) curve.
- Baseline methods: logistic regression ($N = 200$), SVM ($N = 200$), parametric joint model (LMM-Joint) ($N = 30$).

Early-stage (30% observation)

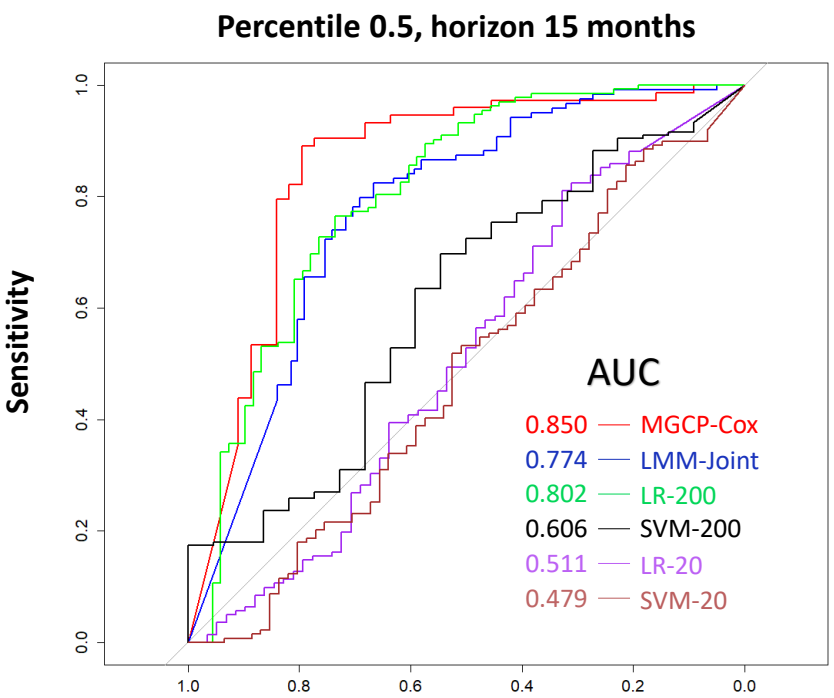


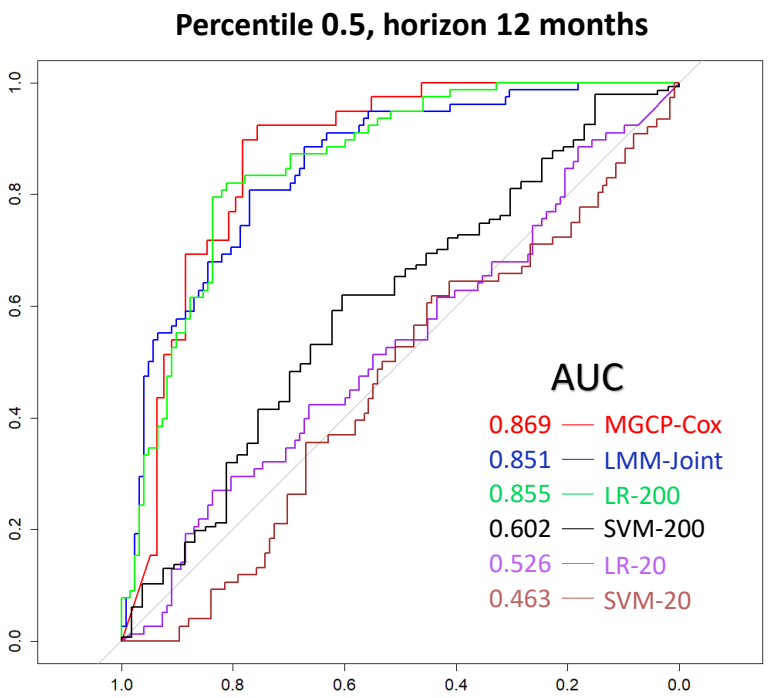


Discussion

- Accurately detect failure event at early stage.
- Important. Avoid economic loss!

Middle-stage (50% observation)

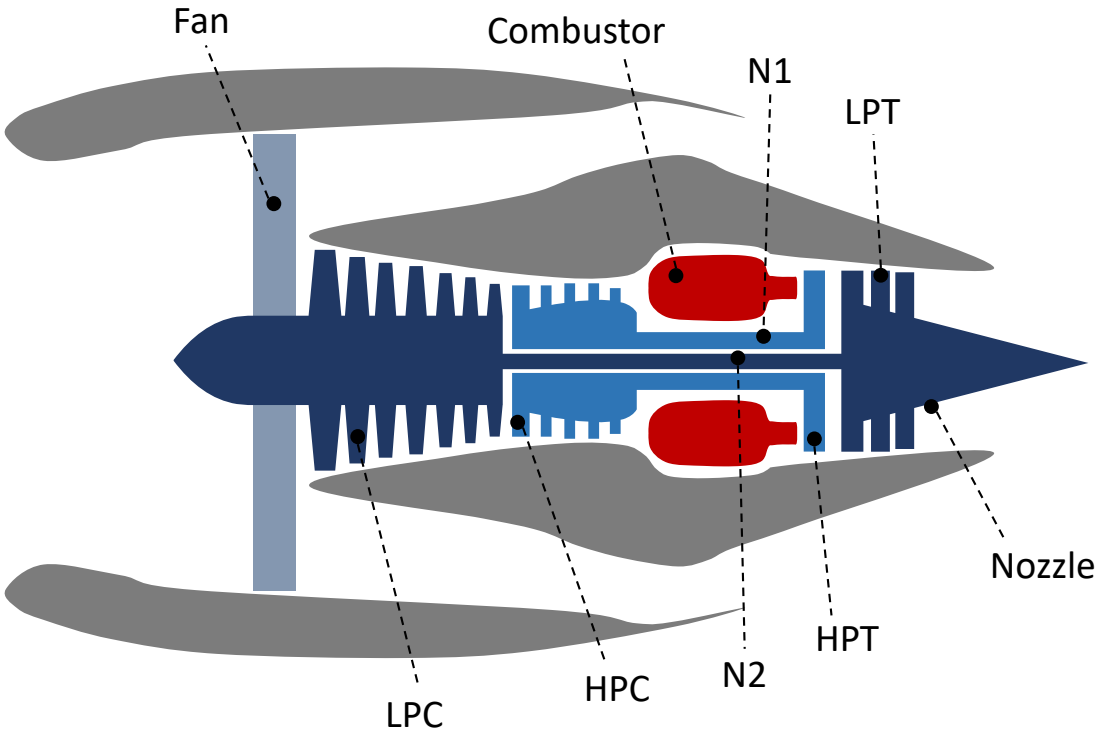




Discussion

- Require small sample size. Cost-effective.
- Outperform parametric model.

NASA Data - Description



Symbol	Description	Units
T2	Total temperature at fan inlet	°R
T24	Total temperature at LPC outlet	°R
T30	Total temperature at HPC outlet	°R
T50	Total temperature at LPT outlet	°R
P2	Pressure at fan inlet	psia

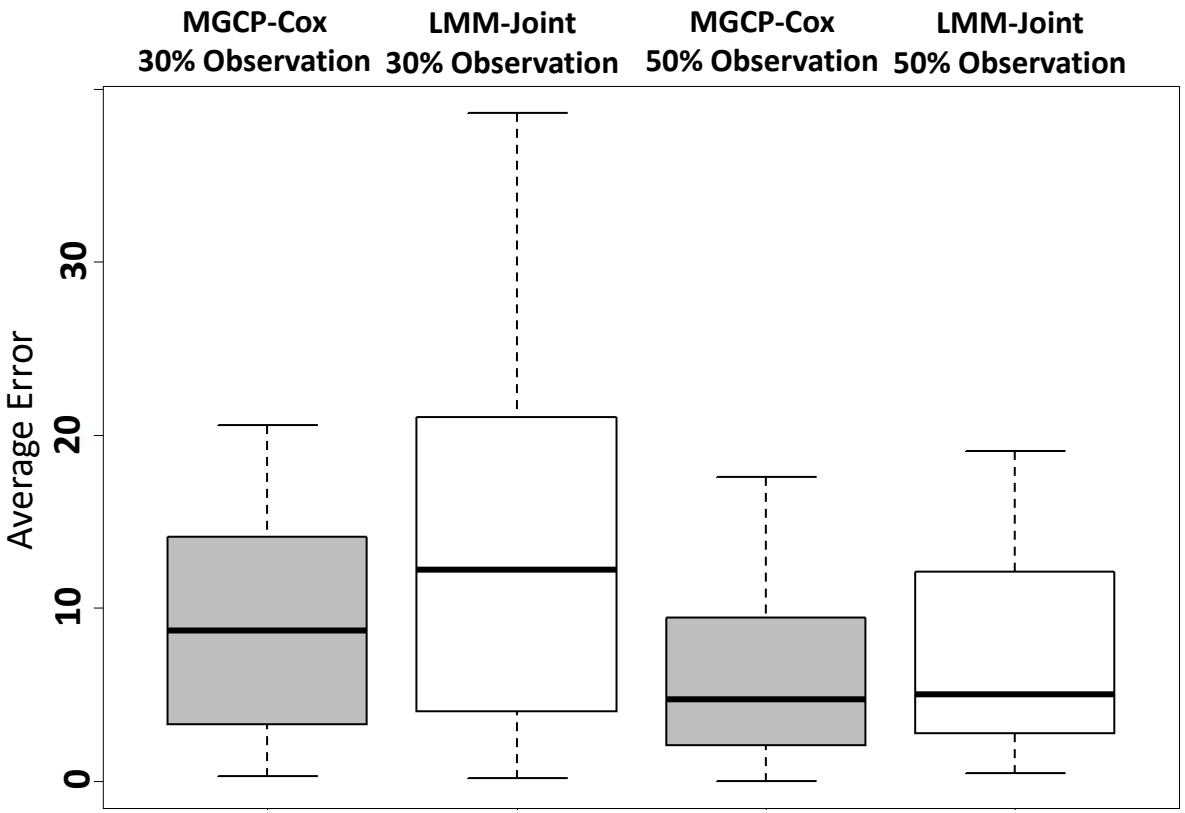
NASA Data - Description

- Failure time data of aircraft turbofan engines and degradation signals from sensors mounted on these engines.
- Time domain: 0-300.
- The mean remaining lifetime

$$\widehat{mrl}(t^*) = \int_{t^*}^{\infty} \hat{S}(u|t^*, \boldsymbol{w}_{test}, \boldsymbol{f}_{test})du.$$

- Evaluation: absolute error $AE = |rl_j - \widehat{mrl}_j|$ where rl_j is the true remaining lifetime of the testing unit.

NASA Data - Result



Conclusion

Contribution

- Relieve parametric assumption: introduce a flexible and scalable non-parametric mixed-effect model.
- Introduce a variational inference framework to jointly estimate all parameters simultaneously.

Impact

- Accurate, requires small sample-size.
- Applicable to healthcare system, automobile system, or even large and complex connected system.