Comparing Bias and MSE of Kaplan-Meier estimators

Ying Dai
Lisa Wilson
Alejandra Castillo

ST 625 Project

Framework

Give lifetime data:

$$X_1, X_2, ..., X_n \overset{iid}{\sim} F$$

$$C_1, C_2, ..., C_n \overset{iid}{\sim} G$$

where $X_i \perp C_i$

Goal: Estimate F, though G complicates this task.
Define:

$$T_i = min(X_i, C_i)$$

$$\delta_i = \begin{cases} 1 & \text{if } X_i < C_i \\ 0 & \text{if } X_i \geq C_i \end{cases}$$

Data observed: (T_i, δ_i) with times of death: $t_1, ..., t_k$

Motivation

The goal of this project was to compare the performance of three estimators, used to estimate the survival function of lifetime data. The metrics for comparison used were bias and mean squared error.

- Kaplan–Meier estimator (Efron)
- Kaplan–Meier estimator (Gill)
- S^*

Estimators

Kaplan–Meier estimator (1958)

$$\hat{S}_{KM}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \leq t} \left(1 - \frac{d_i}{Y_i}\right) & \text{if } t_1 \leq t \end{cases}$$

Kaplan–Meier estimator (Efron) (1967)

$$\hat{S}_E(t) = \begin{cases} S_{KM}(t) & \text{if } t < t_k \\ 0 & \text{if } t_k \leq t \end{cases}$$

Kaplan–Meier estimator (Gill)

$$\hat{S}_G(t) = \begin{cases} S_{KM}(t) & \text{if } t < t_k \\ S_{KM}(t_k) & \text{if } t_k \leq t \end{cases}$$

Estimators

$$P(T > t) = P(min(X, C) > t) = P(X > t)P(C > t)$$
$$\Leftrightarrow \frac{P(T > t)}{P(C > t)} = P(X > t)$$

Use Empirical Estimators to obtain:

$$S^*(t) = \frac{\frac{1}{n}\Sigma_{i=1}^{n}\mathbb{1}(T_i > t)}{P(C > t)}$$

Simulations

We use simulation work to compare finite-sample properties of three estimators.

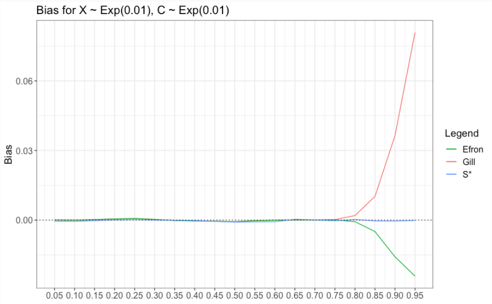
- All simulations conducted in R
- Sample size used: 100
- Replications in each simulation: 10000

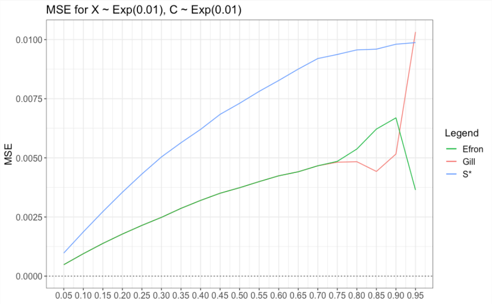
Part I : Simulations with different distributions for X and C

We chose the following distributions for X and C:

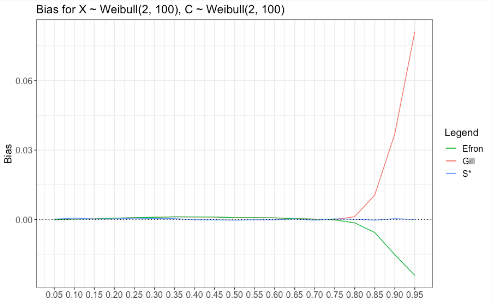
- Exponential($\lambda = 0.01$)
- Weibull($\alpha = 2$, $\beta = 100$)
- Normal($\mu = 150$, $\sigma = 50$)

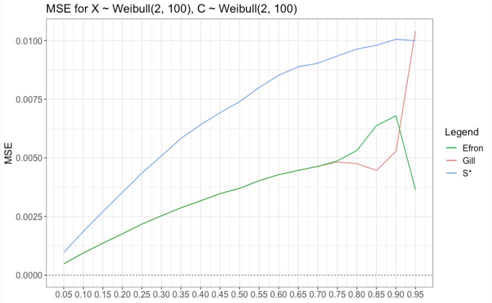
X & C from same distribution: $X \sim \text{Exp}(0.01)$, $C \sim \text{Exp}(0.01)$



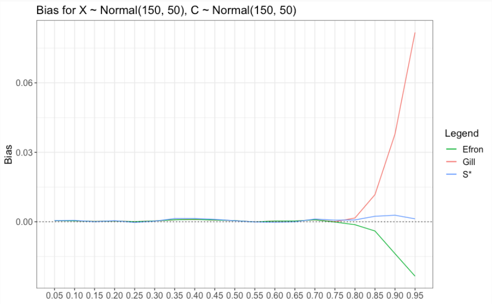


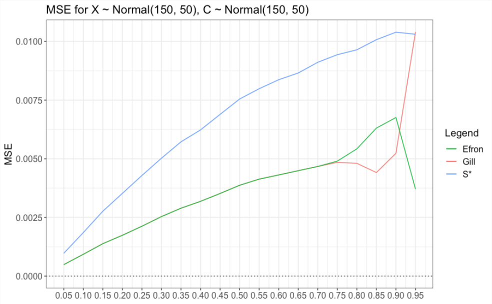
X & C from same distribution: $X \sim$ Weib(2, 100), $C \sim$ Weib(2, 100)



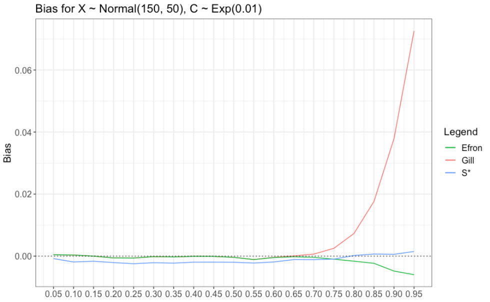


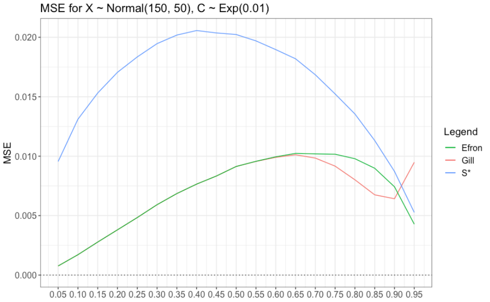
X & C from same distribution: X $\sim$ N(150, 50), C $\sim$ N(150, 50)



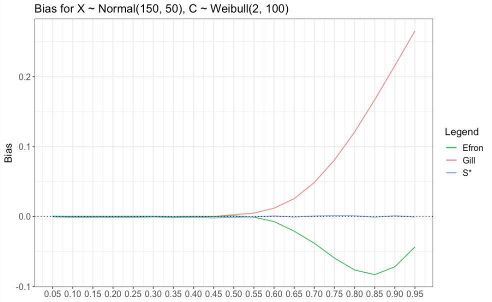


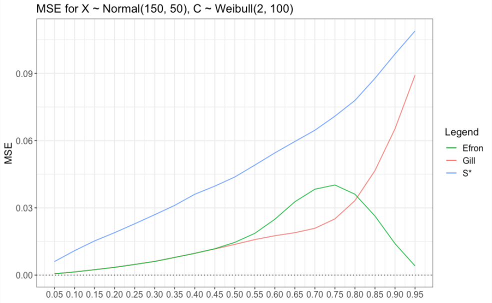
$P(C > t) > P(X > t)$ for larger t: X $\sim$ N(150, 50), C $\sim$ Exp(0.01)



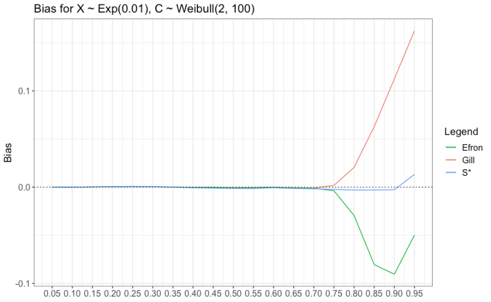


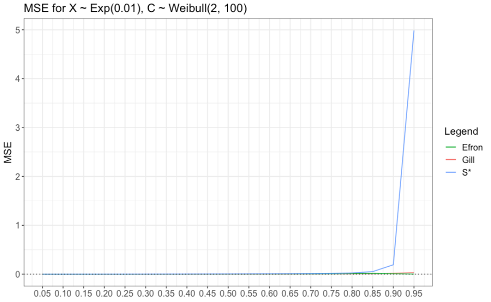
$P(C > t) \leq P(X > t)$ for larger t: X $\sim$ N(150, 50), C $\sim$ Weib(2, 100)



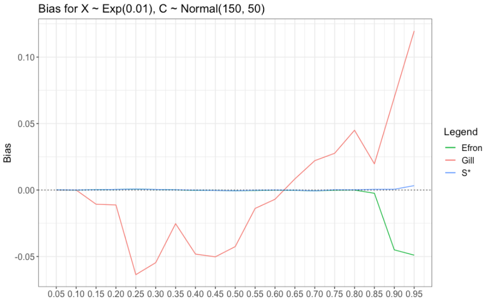


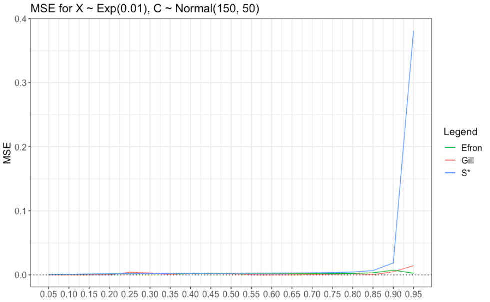
$P(C > t)$ small for large t: X $\sim$ Exp(0.01), C $\sim$ Weibull(2, 100)



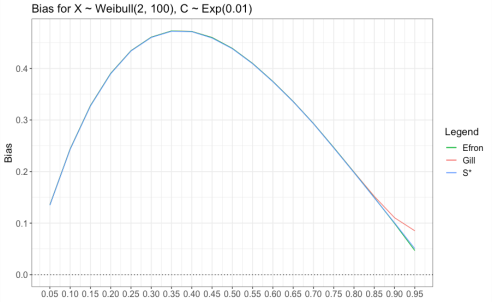


$P(C > t)$ small for large t: X $\sim$ Exp(0.01), C $\sim$ Normal(150, 50)

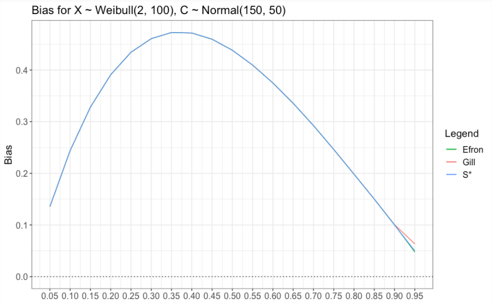


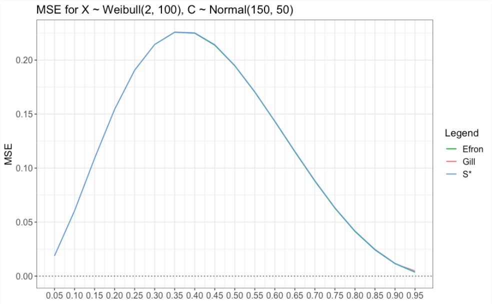


$P(C > t) \geq P(X > t)$ for larger t: X $\sim$ Weibull(2, 100), C $\sim$ Exp(0.01)



$P(C > t) \geq P(X > t)$ for larger t: X $\sim$ Weib(2, 100), C $\sim$ N(150, 50)





Comments

- In most of these settings, $S^*(t)$ provides improvement in terms of bias, particularly for higher times t, but is more variable than the KM estimators.
- $S^*(t)$ performs essentially the same as the KM estimators when $X \sim Weibull(2, 100)$ and $C \sim Exp(0.01)$ or $C \sim Normal(150, 50)$.

Part II: Mixture Gamma

We next assume a Mixture Gamma for the Censoring Distribution

- Let $Y_i \sim$ Gamma (α_i, β_i)
- for p_i, $\sum_{i=1}^{n} p_i = 1$
- $C = \sum_{i=1}^{n} p_i Y_i$

Part II: Mixture Gamma

We choose three types of Mixture Gamma for the Censoring Times:

- weights: $p_1 = 0.1, p_2 = 0.3, p_3 = 0.6$
- alphas: $\alpha_1 = 10, \alpha_2 = 20, \alpha_3 = 40$
- betas: $\beta_1 = \beta_2 = \beta_3 = 1.5$

- weights: $p_1 = p_2 = p_3 = \frac{1}{3}$
- alphas: $\alpha_1 = 10, \alpha_2 = 20, \alpha_3 = 40$
- betas: $\beta_1 = 7, \beta_2 = 2, \beta_3 = 1$

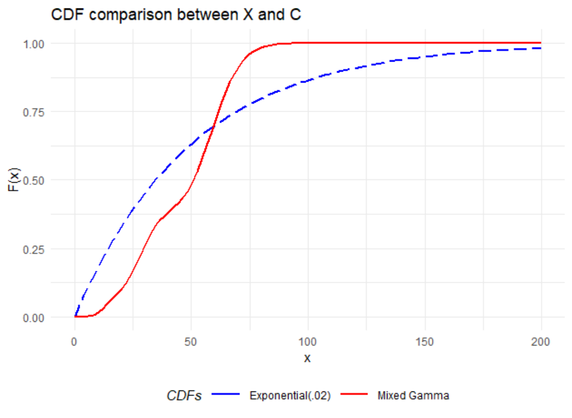
- weights: $p_1 = 0.1, p_2 = 0.3, p_3 = 0.6$
- alphas: $\alpha_1 = 10, \alpha_2 = 20, \alpha_3 = 40$
- betas: $\beta_1 = \beta_2 = \beta_3 = 1.5$

We choose three types of distributions for the Survival Times:

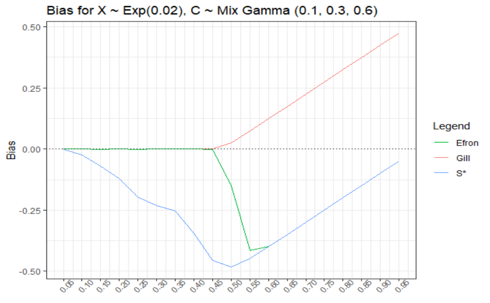
- Exponential($\lambda = 0.02$)
- Weibull($\alpha = 0.5$, $\beta = 25$)
- Normal($\mu = 50$, $\sigma = 50$)

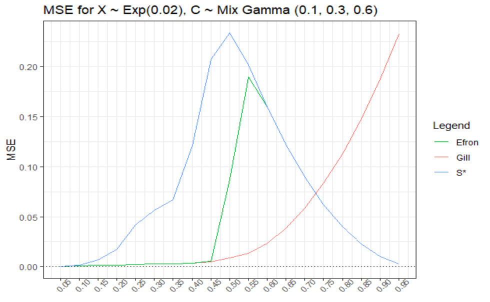
$X \sim \text{Exp}(0.02)$, $C \sim$ Mixture Gamma

- weights: $p_1 = 0.1, p_2 = 0.3, p_3 = 0.6$
- alphas: $\alpha_1 = 10, \alpha_2 = 20, \alpha_3 = 40$
- betas: $\beta_1 = \beta_2 = \beta_3 = 1.5$



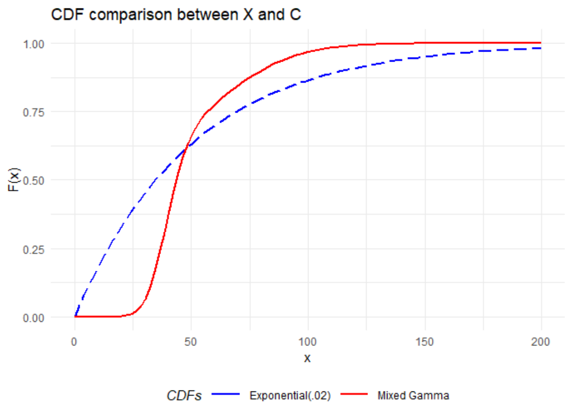
$X \sim \text{Exp}(0.02)$, $C \sim$ Mixture Gamma



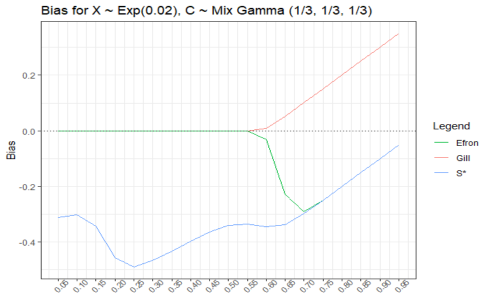


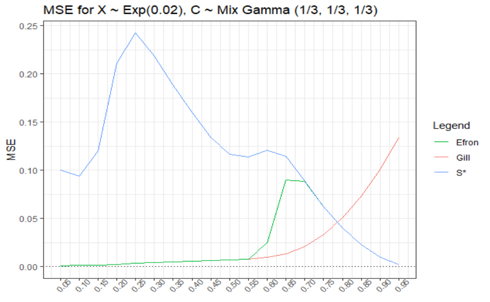
$X \sim \text{Exp}(0.02)$, $C \sim$ Mixture Gamma

- weights: $p_1 = p_2 = p_3 = \frac{1}{3}$
- alphas: $\alpha_1 = 10, \alpha_2 = 20, \alpha_3 = 40$
- betas: $\beta_1 = 7, \beta_2 = 2, \beta_3 = 1$



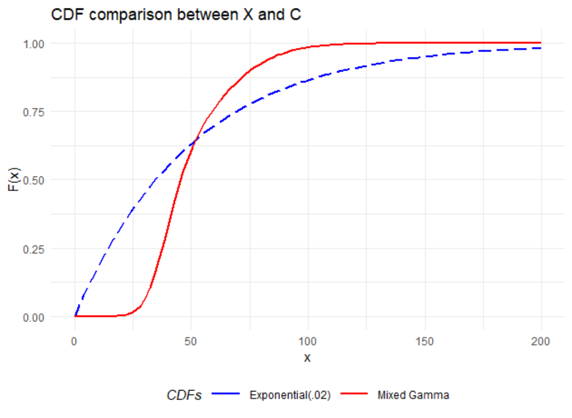
X ∼ Exp(0.02), C ∼ Mixture Gamma



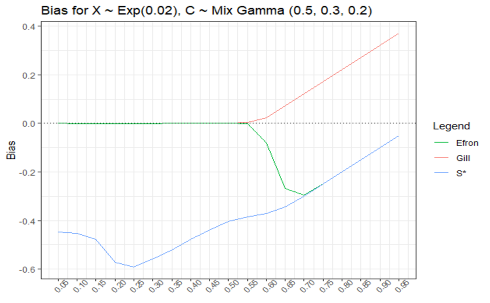


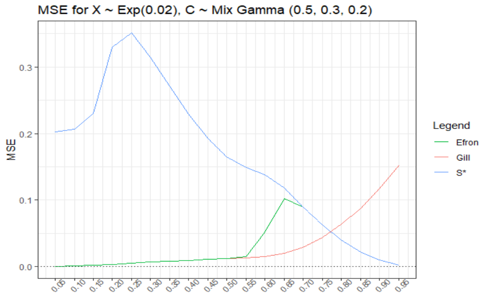
$X \sim \text{Exp}(0.02)$, $C \sim$ Mixture Gamma

- weights: $p_1 = 0.5, p_2 = 0.3, p_3 = 0.2$
- alphas: $\alpha_1 = 10, \alpha_2 = 20, \alpha_3 = 40$
- betas: $\beta_1 = 6, \beta_2 = 2, \beta_3 = 1$



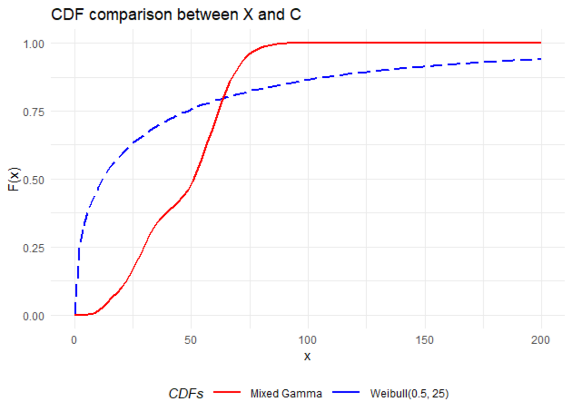
X $\sim$ Exp(0.02), C $\sim$ Mixture Gamma



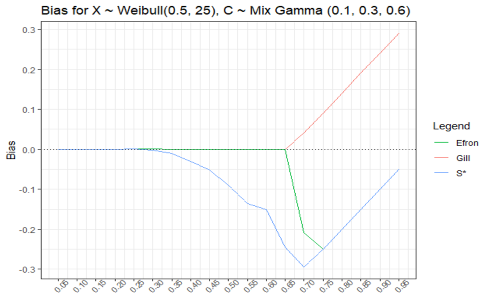


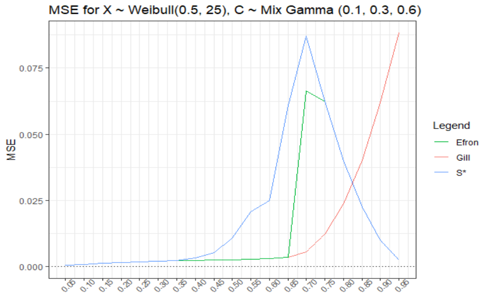
$X \sim$ Weibull(0.5, 25), $C \sim$ Mixture Gamma

- weights: $p_1 = 0.1, p_2 = 0.3, p_3 = 0.6$
- alphas: $\alpha_1 = 10, \alpha_2 = 20, \alpha_3 = 40$
- betas: $\beta_1 = \beta_2 = \beta_3 = 1.5$



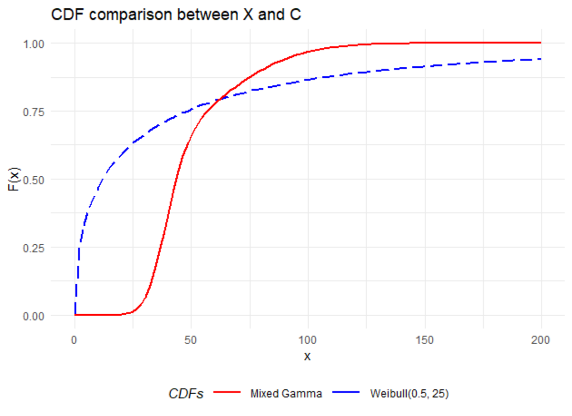
X $\sim$ Weibull(0.5, 25), C $\sim$ Mixture Gamma



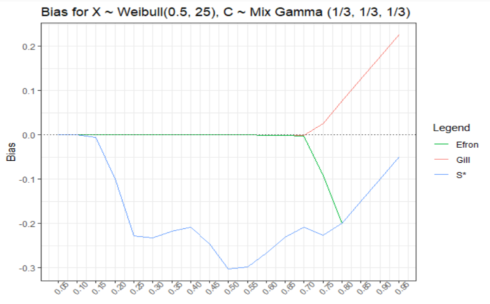


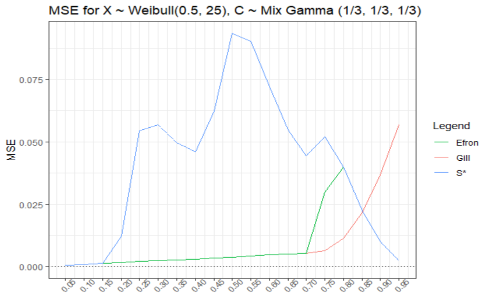
$X \sim$ Weibull(0.5, 25), $C \sim$ Mixture Gamma

- weights: $p_1 = p_2 = p_3 = \frac{1}{3}$
- alphas: $\alpha_1 = 10, \alpha_2 = 20, \alpha_3 = 40$
- betas: $\beta_1 = 7, \beta_2 = 2, \beta_3 = 1$



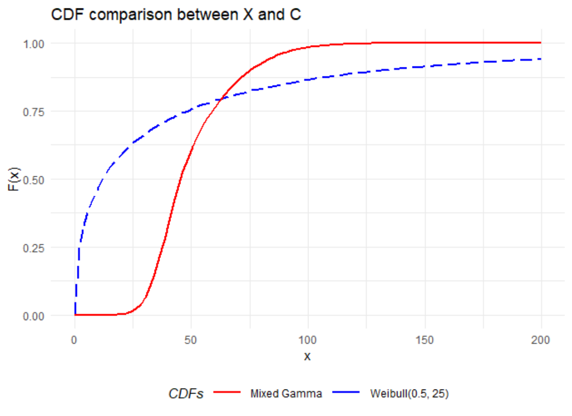
X $\sim$ Weibull(0.5, 25), C $\sim$ Mixture Gamma



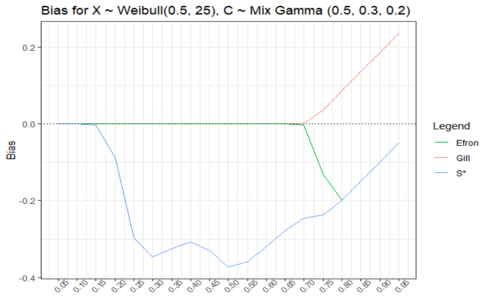


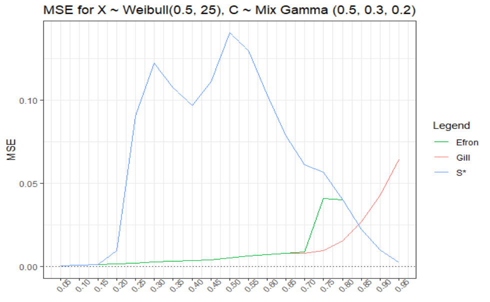
$X \sim$ Weibull(0.5, 25), $C \sim$ Mixture Gamma

- weights: $p_1 = 0.5, p_2 = 0.3, p_3 = 0.2$
- alphas: $\alpha_1 = 10, \alpha_2 = 20, \alpha_3 = 40$
- betas: $\beta_1 = 6, \beta_2 = 2, \beta_3 = 1$



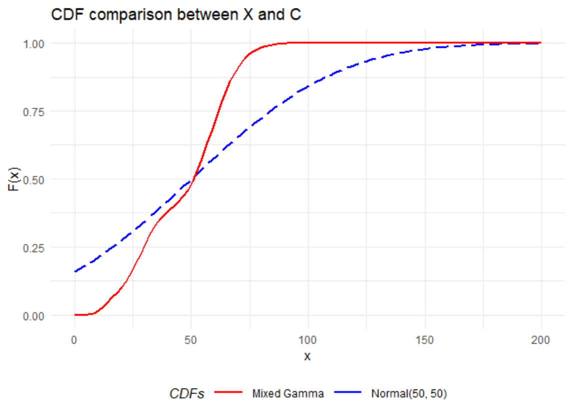
X $\sim$ Weibull(0.5, 25), C $\sim$ Mixture Gamma



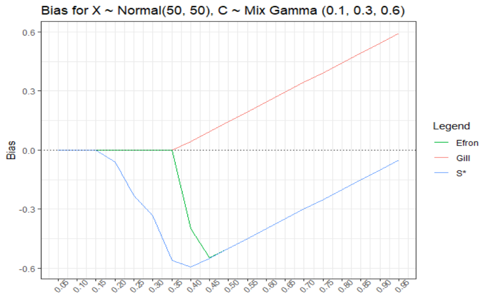


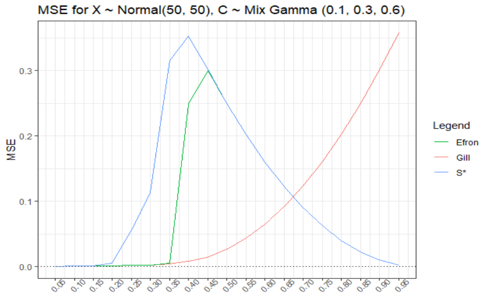
X $\sim$ Normal(50, 50), C $\sim$ Mixture Gamma

- weights: $p_1 = 0.1, p_2 = 0.3, p_3 = 0.6$
- alphas: $\alpha_1 = 10, \alpha_2 = 20, \alpha_3 = 40$
- betas: $\beta_1 = \beta_2 = \beta_3 = 1.5$



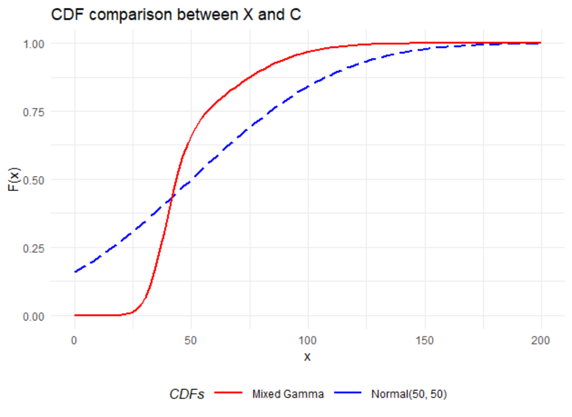
X ∼ Normal(50, 50), C ∼ Mixture Gamma



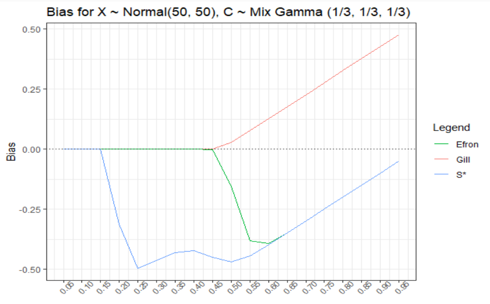


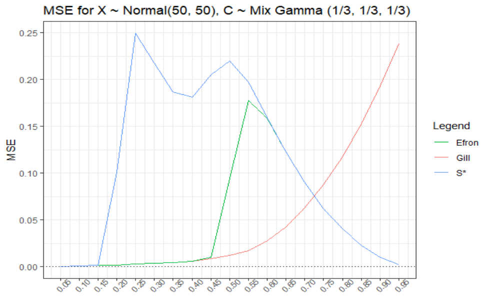
X $\sim$ Normal(50, 50), C $\sim$ Mixture Gamma

- weights: $p_1 = p_2 = p_3 = \frac{1}{3}$
- alphas: $\alpha_1 = 10, \alpha_2 = 20, \alpha_3 = 40$
- betas: $\beta_1 = 7, \beta_2 = 2, \beta_3 = 1$



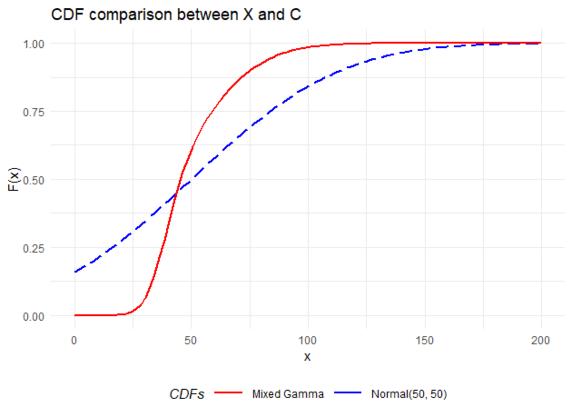
X $\sim$ Normal(50, 50), C $\sim$ Mixture Gamma



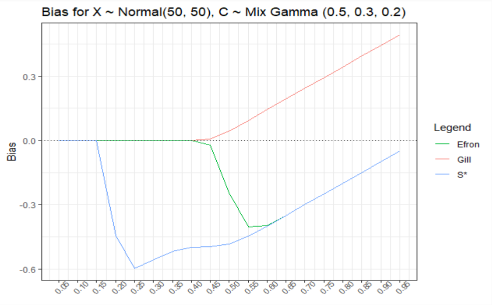


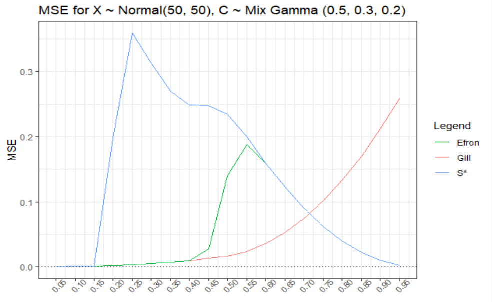
X $\sim$ Normal(50, 50), C $\sim$ Mixture Gamma

- weights: $p_1 = 0.5, p_2 = 0.3, p_3 = 0.2$
- alphas: $\alpha_1 = 10, \alpha_2 = 20, \alpha_3 = 40$
- betas: $\beta_1 = 6, \beta_2 = 2, \beta_3 = 1$



X $\sim$ Normal(50, 50), C $\sim$ Mixture Gamma





Comments

- In all of these simulations Gill has positive bias, Efron has negative bias and S^* has negative bias.
- S^* has more variable bias and MSE. For smaller quantiles, the MSE for S^* increased and decrease for larger quantiles.
- In each simulation the Efron and S^* behaved similarly in terms of bias and MSE. There was never a case where Gill showed a decrease in bias or MSE.
- Efron and S^* had bias and MSE that tended to zero for larger quantiles.
- Efron performed the least terrible with bias smaller magnitude and smaller MSE.

Overall

- From class, we expected the Gill modification to have positive bias and the Efron modification to have negative bias.
- It was interesting to see how variable the S^* estimator was in comparison to the KM estimators.
- One might expect that knowing the distribution of C, the censoring distribution, may contribute to estimators with lower bias and MSE.

Thank you!