Uma abordagem baseada em similaridade empírica para o estimador de Kaplan-Meier

Isabel de Castro Beneyto

Orientador: Prof. Dr. Leandro Chaves Rêgo
Coorientador: Prof. Dr. Anselmo R. Pitombeira Neto

isabelcastro@alu.ufc.br

Programa de Pós-Graduação em Modelagem e Métodos Quantitativos
Universidade Federal do Ceará

22 de maio de 2024

UNIVERSIDADE FEDERAL DO CEARÁ

Sumário

UNIVERSIDADE FEDERAL DO CEARÁ

Análise de Sobrevivência

Área dedicada ao estudo de dados temporais e à realização de previsões sobre eventos futuros.

Figura: Nuvem de palavras com termos relacionados à análise de sobrevivência.

Fonte: [1]

Tempo de Falha

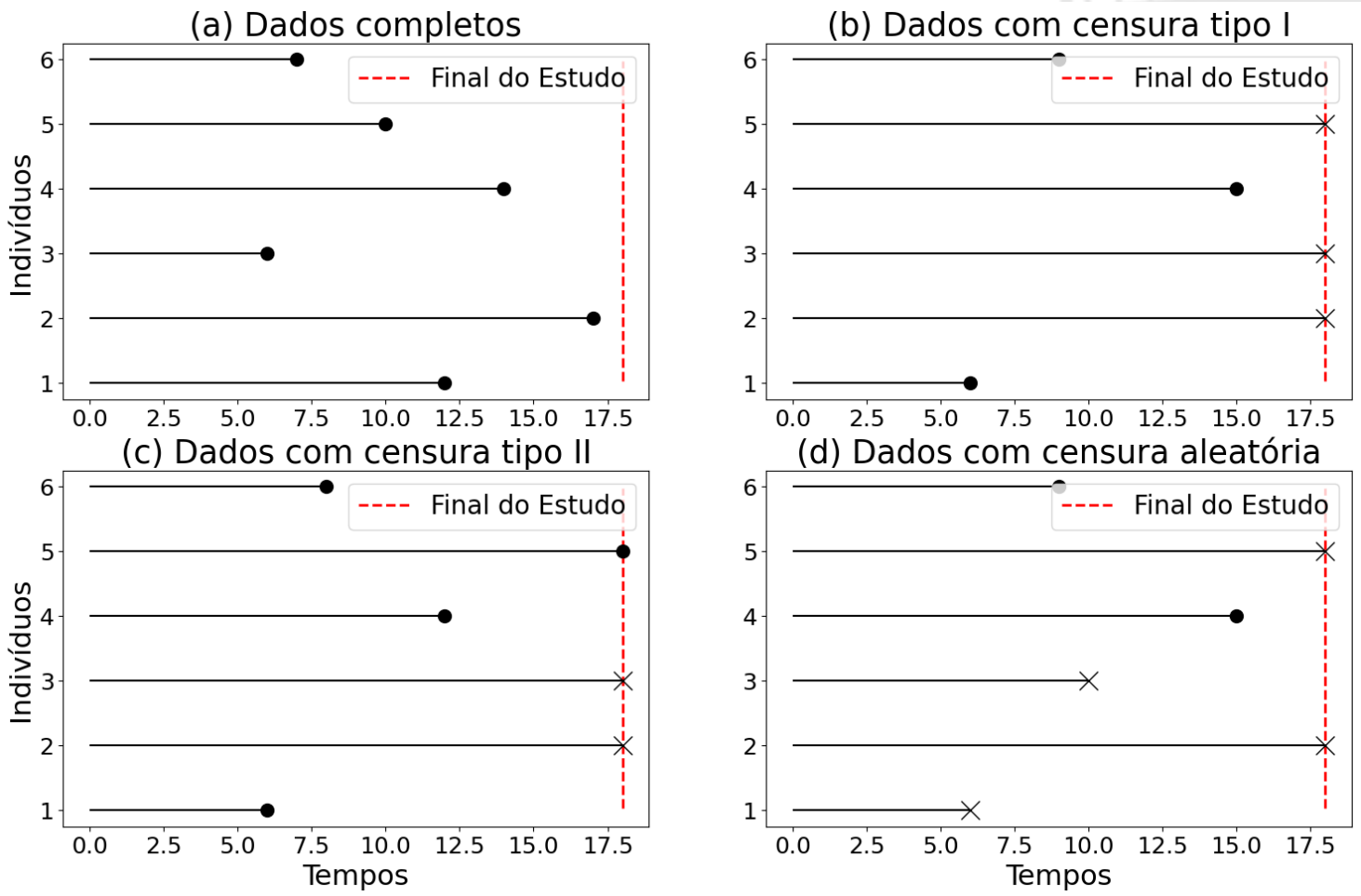
1. **Tempo inicial:**
   Indica o início do estudo.

2. **Escala de medida:**
   Tempo real de observação.

3. **Evento de interesse:**
   Evento que vem a ser a falha.

Censuras

Figura: Diferentes tipos de censura



Representação dos dados

Os dados de sobrevivência para o indivíduo i ($i = 1, ..., n$) sob estudo, são representados pelo par $(t_i, \delta_i, \mathbf{x}_i)$ sendo t_i o tempo de falha ou de censura, $\mathbf{x}_i$ é o vetor de covariáveis e δ_i a variável indicadora de falha ou censura, isto é,

$$\delta = \begin{cases} 1, & \text{se } t_i \text{ é um tempo de falha} \\ 0, & \text{se } t_i \text{ é um tempo censurado.} \end{cases} \quad (1)$$

Função de Sobrevivência

A função de sobrevivência $S(t)$ é definida como a probabilidade de uma observação não falhar até um certo tempo t. Isto é descrito como

$$S(t) = P(T > t), \tag{2}$$

onde T é uma variável aleatória representando o tempo de falha.

Estimador de Kaplan-Meier

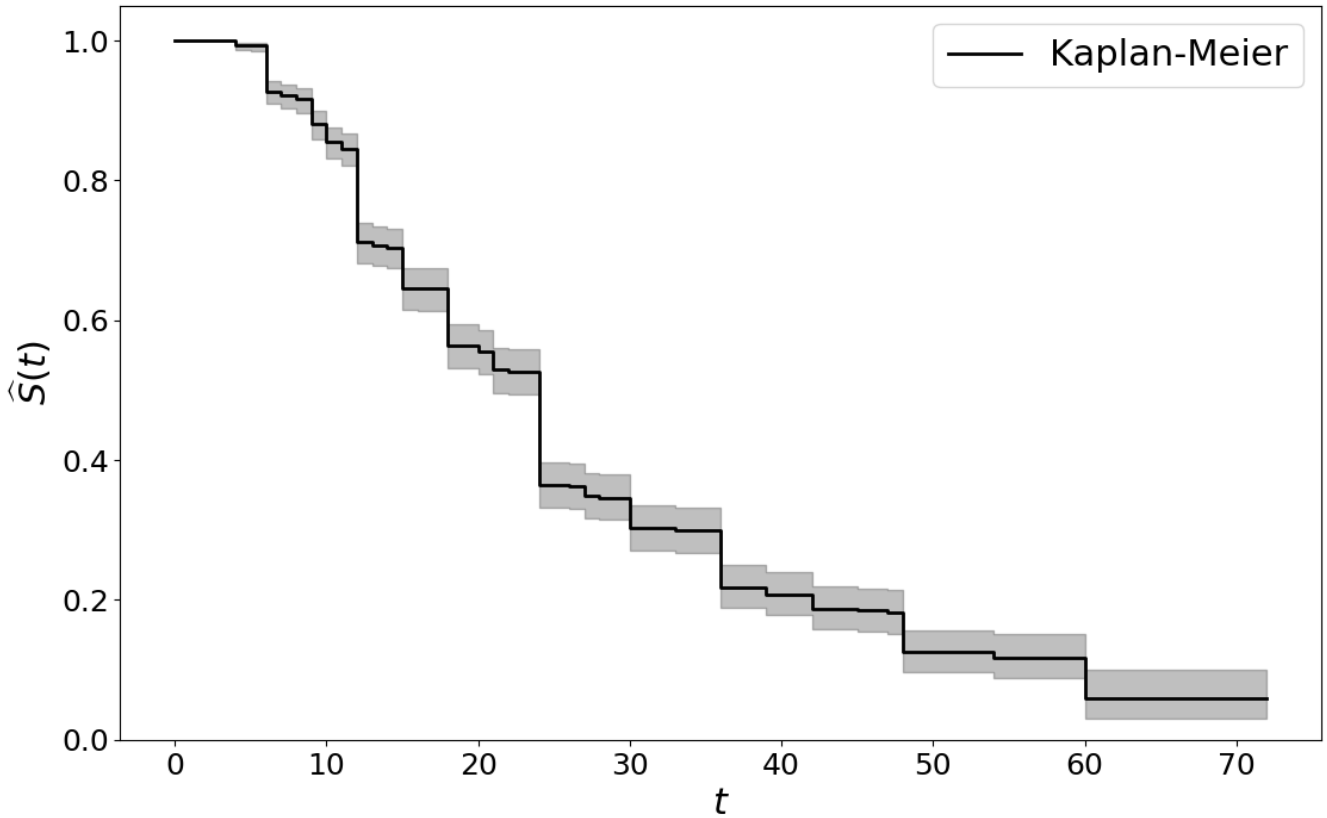
Considere que n indivíduos sob estudo no período de acompanhamento têm falhas ou censuras em tempos distintos t'_i, para $i \in \{1, 2, , \ldots, n'\}$, com $n' \leq n$, como sendo o i-ésimo menor valor no conjunto $\{t_1, t_2, \ldots, t_n\}$. A probabilidade de estar vivo em t é calculada por

$$\widehat{S}(t) = \prod_{i=1}^{n'} \left(1 - \frac{d_i}{g_i}\right)^{\delta_i \mathbb{1}\{t'_i \leq t\}}, \tag{3}$$

em que:

- $d_i = \sum_{j=1}^{n} \delta_j \mathbb{1}\{t_j = t'_i\}$ é o número de falhas em t'_i, $i = 1, ..., n'$
- $g_i = \sum_{j=1}^{n} \mathbb{1}\{t_j \geq t'_i\}$ é o número de indivíduos sob risco em t'_i, $i = 1, ..., n'$.

Figura: Estimativa de Kaplan-Meier para os dados de risco de CREDIT elaborada utilizando a biblioteca *lifelines* disponível em Python [2].



Similaridade Empírica

O método de similaridade empírica combina observações passadas de $\mathbf{x}$ e y com os valores atuais de $\mathbf{x}$ para gerar uma avaliação de y através de uma média ponderada por similaridade [3].

O preditor baseado em similaridade de y_j, dada uma função de similaridade s, é definido como:

$$\widehat{y}_j = \frac{\sum_{i \neq j} s(\mathbf{x}_i, \mathbf{x}_j) y_i}{\sum_{i \neq j} s(\mathbf{x}_i, \mathbf{x}_j)}. \tag{4}$$

Inicialmente, buscamos funções de similaridade baseadas em distâncias ponderadas [4]. Possíveis candidatos naturais para essa função são:

$$s(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} e^{-d_w(\mathbf{x}_i, \mathbf{x}_j)} & (\mathrm{EX}), \\ \frac{1}{1+d_w(\mathbf{x}_i, \mathbf{x}_j)} & (\mathrm{FR}), \end{cases} \tag{5}$$

em que serão usadas funções de similaridade do tipo exponencial (EX) e fracionário (FR).

Dado uma base de dados com m covariáveis, onde m_1, m_2, e m_3 são as quantidades de covariáveis quantitativas, categóricas nominais, e categóricas ordinais, respectivamente, e $m_1 + m_2 + m_3 = m$. Dado o vetor de pesos $\mathbf{w} \in \mathbb{R}_+^m$, calculamos a distância total como:

$$\begin{aligned} d_w &= d_w^Q + d_w^{CN} + d_w^{CO} \\ &= \sum_{l=1}^{m_1} w_l (x_i^l - x_j^l)^2 + \sum_{l=m_1+1}^{m_1+m_2} w_l \mathbb{1}(x_i^l, x_j^l) + \sum_{l=m_1+m_2+1}^{m} \frac{w_l (x_i^l - x_j^l)^2}{(x_{max}^l - x_{min}^l)^2}. \end{aligned} \tag{6}$$

Objetivos

- Propor o estimador de Kaplan-Meier baseado em Similaridade (SBKM), uma adaptação do estimador de Kaplan-Meier (KM) que utiliza funções de similaridade.
- Demonstrar a aplicação do estimador proposto em dados reais.

Estimador de Kaplan-Meier baseado em similaridade

Considere uma base de dados com observações $(t_i, \delta_i, \mathbf{x}_i)$, para $i = 1, 2, ..., n$. Para um vetor de características $\mathbf{x} = (x^1, ..., x^m)$, a função de sobrevivência condicional $S(t|\mathbf{x})$ será estimada por

$$\widehat{S}(t|\mathbf{x}) = \prod_{i=1}^{n'} \left[1 - \frac{\sum_{j=1}^{n} s_w(\mathbf{x}, \mathbf{x}_j)\delta_j \mathbb{1}\{t_j = t_i'\}}{\sum_{j=1}^{n} s_w(\mathbf{x}, \mathbf{x}_j)\mathbb{1}\{t_j \geq t_i'\}} \right]^{\delta_i \mathbb{1}\{t_i' \leq t\}}. \quad (7)$$

Para o caso especial em que $s_w(\mathbf{x}, \mathbf{x}_j) = 1$ para todo j, retomamos a Equação 3.

Para estimar os valores dos parâmetros w, podemos utilizar o método de máxima verossimilhança empírica [5],

$$L(w) = \prod_{k=1}^{n} [\widehat{P}(t_k|\mathbf{x}_k)]^{\delta_k} [\widehat{S}(t_k|\mathbf{x}_k)]^{1-\delta_k}, \tag{8}$$

em que $\widehat{S}(t_k|\mathbf{x}_k)$ é a própria curva de sobrevivência estimada pelo SBKM e $\widehat{P}(t_k|\mathbf{x}_k)$ é a probabilidade de falha em t_k dado o vetor de covariáveis $\mathbf{x}_k$.

Exemplo fictício

Suponha que temos uma base de dados $(t_i, \delta_i, \mathbf{x}_i)$, para $i = 1, 2, ..., 4$, onde $\mathbf{x}_i = (x_i^1, x_i^2)$ é um vetor de 2 covariáveis associadas à i-ésima observação.

id	Tempo	Falha	Estágio	Idade
1	2.5	0	1	76
2	3.8	1	4	40
3	7.0	1	2	54
4	10.0	0	3	68
5	-	-	1	72

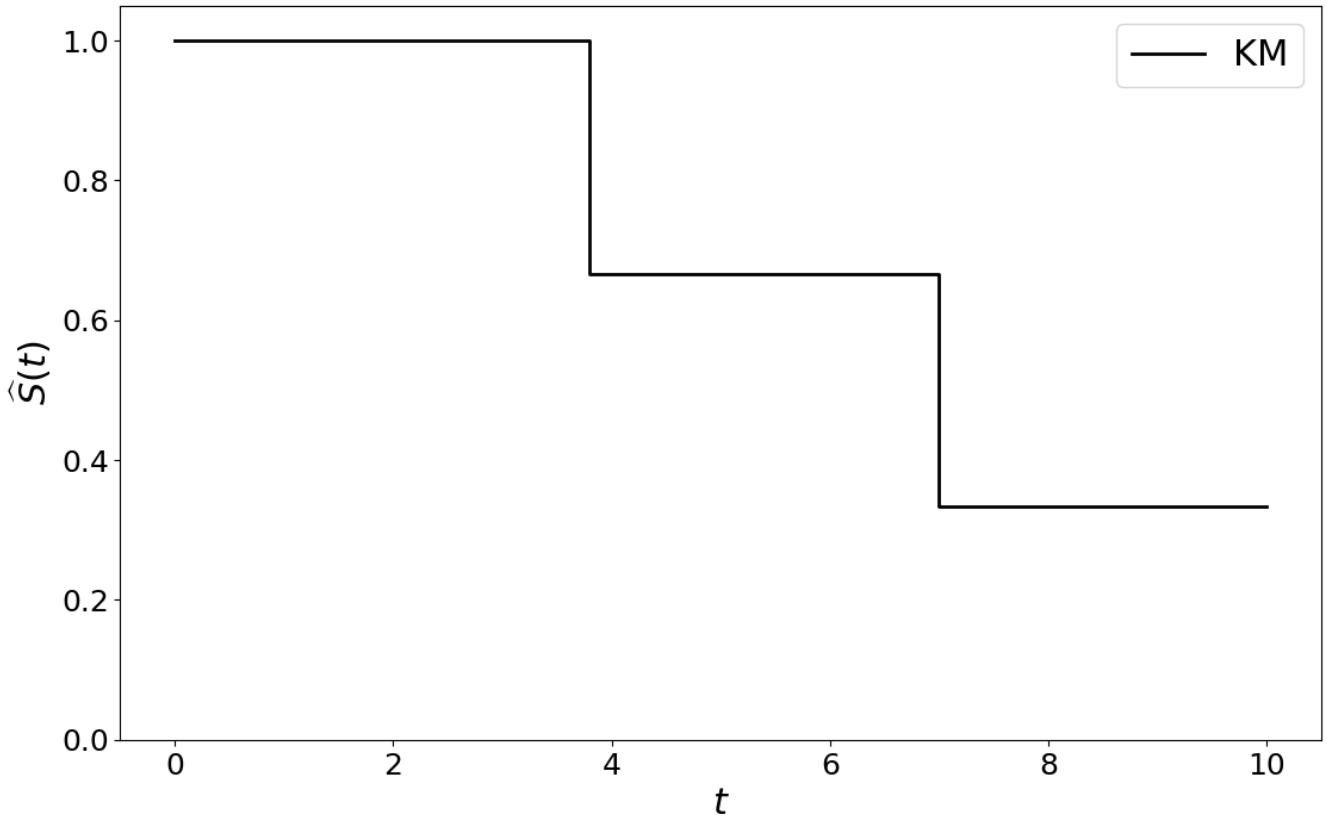
Tabela: Dados fictícios

Para assegurar que todas as covariáveis contribuam igualmente para a medida de distância, optamos por padronizar a covariável numérica Idade.

id	Tempo	Falha	Estágio	Idade
1	2.5	0	1	1
2	3.8	1	4	0
3	7.0	1	2	0.388
4	10.0	0	3	0.777
5	-	-	1	0.888

Tabela: Dados fictícios normalizados

Figura: Estimativa de Kaplan-Meier para os dados fictícios elaborada utilizando a biblioteca lifelines disponível em Python.



Para o vetor de covariáveis $\mathbf{x}_5 = (1, 0.888)^\intercal$, estimamos a função de sobrevivência para um paciente com essas características usando a Equação 7, onde $n' = 4$.

$$\widehat{S}(t|\mathbf{x}_5) = \prod_{i=1}^{4} \left(1 - \frac{\sum_{j=1}^{4} s_w(\mathbf{x}_5, \mathbf{x}_j)\delta_j \mathbb{1}\{t_j = t_i'\}}{\sum_{j=1}^{4} s_w(\mathbf{x}_5, \mathbf{x}_j)\mathbb{1}\{t_j \geq t_i'\}}\right)^{\delta_i \mathbb{1}\{t_i' \leq t\}} \tag{9}$$

$$= \left(1 - \frac{s_w(\mathbf{x}_5, \mathbf{x}_2)}{s_w(\mathbf{x}_5, \mathbf{x}_2) + s_w(\mathbf{x}_5, \mathbf{x}_3) + s_w(\mathbf{x}_5, \mathbf{x}_4)}\right)^{\mathbb{1}\{t_2' \leq t\}}$$

$$\left(1 - \frac{s_w(\mathbf{x}_5, \mathbf{x}_3)}{s_w(\mathbf{x}_5, \mathbf{x}_3) + s_w(\mathbf{x}_5, \mathbf{x}_4)}\right)^{\mathbb{1}\{t_3' \leq t\}}.$$

Para calcular esse exemplo, a função de similaridade será fixada da forma:

$$s_w(\mathbf{x}_i, \mathbf{x}_j) = e^{-d_w(\mathbf{x}_i, \mathbf{x}_j)}, \tag{10}$$

onde d_w representa a distância total obtida a partir da Equação 6, e é calculada da seguinte forma:

$$\begin{aligned} d_w &= d_w^{CO} + d_w^{Q} \\ &= \frac{w_1(x_i^1 - x_j^1)^2}{(x_{max}^1 - x_{min}^1)^2} + w_2(x_i^2 - x_j^2)^2, \end{aligned} \tag{11}$$

Para estimar os valores dos parâmetros w, podemos utilizar o método de maximização da verossimilhança definido na Equação 8

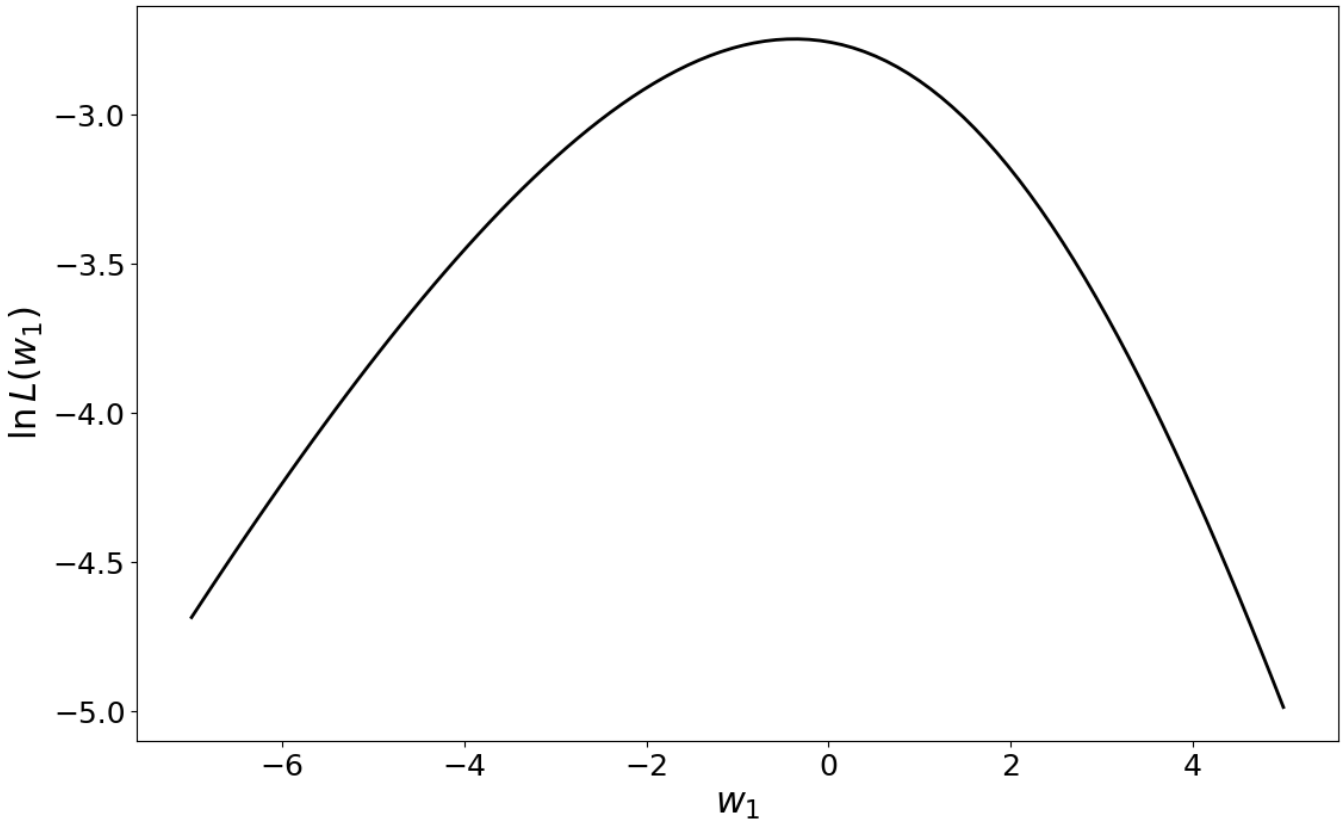
$$L(w) = \widehat{S}(t_1|\mathbf{x}_1)\widehat{P}(t_2|\mathbf{x}_2)\widehat{P}(t_3|\mathbf{x}_3)\widehat{S}(t_4|\mathbf{x}_4). \tag{12}$$

Calculando o logaritmo natural desta função, obtemos

$$\begin{aligned} \ln[L(w)] &= -\ln[1 + e^{-(0.444w_1+0.150w_2)} + e^{-(0.111w_1+0.604w_2)}] \\ &- \ln[e^{-(0.444w_1+0.150w_2)} + 1 + e^{-(0.111w_1+0.151w_2)}] \\ &- \ln[e^{-(0.111w_1+0.604w_2)} + e^{-(0.111w_1+0.151w_2)} + 1]. \end{aligned} \tag{13}$$

Assumimos, então, que $w_1 + w_2 = 1$.

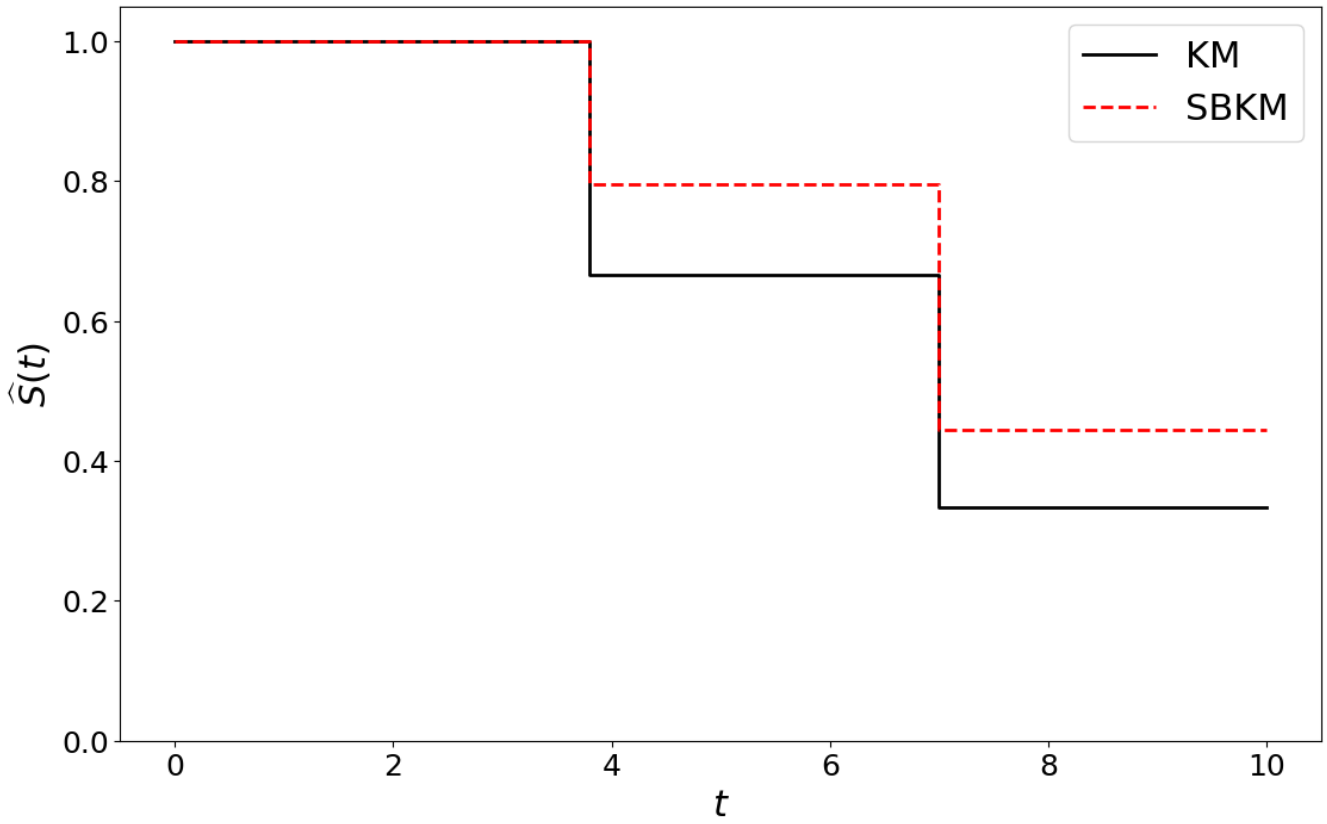
Figura: Representação gráfica do logaritmo da função de verossimilhança empírica.



Dito isso, obtemos $w_1 = 0$ e $w_2 = 1$. Assim, obtemos a função de sobrevivência $\widehat{S}(t|\mathbf{x}_5)$

$$\widehat{S}(t|\mathbf{x}_5) = (0.796)^{\mathbb{1}\{t_2' \leq t\}}(0.556)^{\mathbb{1}\{t_3' \leq t\}} \tag{14}$$

Figura: Curva de sobrevivência obtida para o 5º paciente pelo estimador SBKM em comparação com a estimativa de KM.



Dados de Sobrevivência

Para ilustrar os métodos abordados neste trabalho, utilizamos dois conjuntos de dados reais: CREDIT e SUPPORT

Dados	Número de amostras	Número de covariáveis	Taxa de censura
CREDIT	1000	17	30.0%
SUPPORT	8873	14	31.97%

Tabela: Descrição das características dos conjuntos de dados após o pré-processamento.

Métricas de Avaliação

- **Índice de concordância (CI):** Avalia a ordenação relativa dos tempos de sobrevivência dos indivíduos.
  CI varia entre 0 e 1, onde CI $= 1$ corresponde à melhor previsão do modelo.

- **Brier Score (BS):** Extensão do erro quadrático médio para dados censurados à direita.
  Um BS mais próximo de 0 indica melhor desempenho preditivo.

Base de dados CREDIT

- A base foi dividida em conjuntos de treino, validação e teste, contendo 560, 140 e 300 amostras, respectivamente.
- Foram selecionadas duas covariáveis mais relevantes para as previsões,“amount” e “installment_rate”, pelo método *Stepwise Forward*.

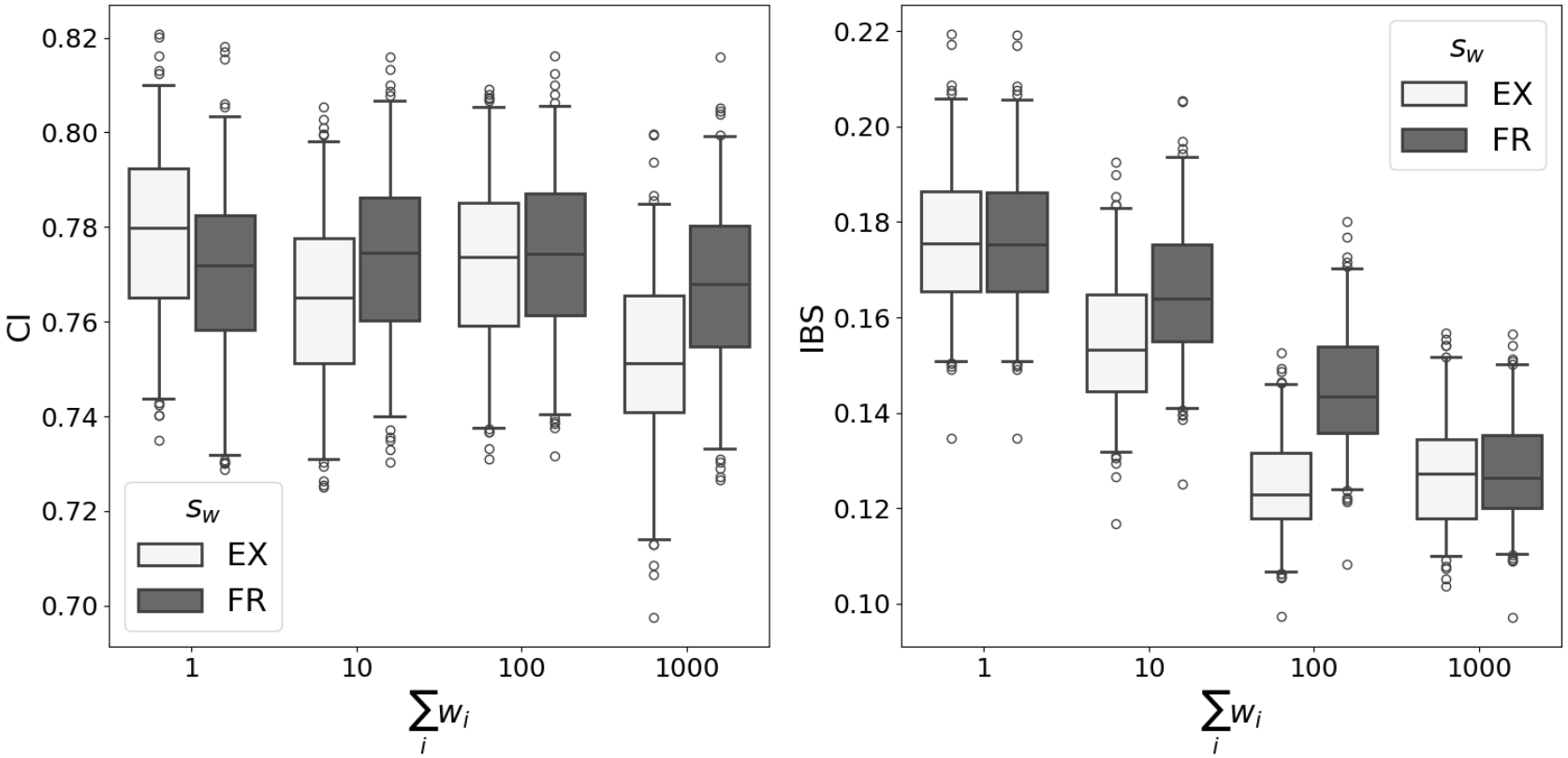
Normalização dos Pesos

$\sum_i w_i$	s_w	w_1	w_2	$\sigma(t_m)$
1	EX	0.28	0.72	0.16 (0.14, 0.18)
	FR	0.38	0.62	0.17 (0.15, 0.19)
10	EX	6.01	3.99	2.72 (2.15, 3.32)
	FR	5.64	4.36	1.20 (1.04, 1.34)
100	**EX**	**75.65**	**24.35**	**7.60 (6.69, 8.47)**
	FR	**65.09**	**34.91**	**3.43 (3.06, 3.73)**
1000	EX	953.83	46.17	9.43 (8.30, 10.33)
	FR	763.58	236.42	6.05 (5.53, 6.53)

Tabela: Pesos estimados w_1, w_2 e desvio padrão dos tempos de falha estimados $\sigma(t_m)$ para diferentes condições de normalização $\sum_i w_i$ e funções de similaridade s_w.

- Observa-se que a normalização dos pesos parece influenciar a variabilidade das previsões.
- A função de similaridade EX parece gerar estimativas mais variáveis em quase todos os cenários.

Figura: Métricas de performance do estimador (CI e IBS) para diferentes condições de normalizações $\sum_i w_i$ e funções de similaridade s_w.



Tempo Estimado de Falha

$\hat{t}$	s_w	CI	$\sigma(\hat{t})$
t_m	**EX**	**0.772 (0.738, 0.805)**	**7.60 (6.69, 8.47)**
	FR	**0.773 (0.741, 0.806)**	**3.43 (3.06, 3.73)**
$t_{0.5}$	EX	0.763 (0.727, 0.797)	8.13 (7.01, 9.33)
	FR	0.761 (0.725, 0.795)	3.49 (3.15, 3.82)

Tabela: Métrica de avaliação CI e desvio padrão dos tempos de falha estimados $\sigma(\hat{t})$ utilizando o tempo médio t_m e o tempo mediano $t_{0.5}$ para diferentes funções de similaridade s_w

- Os resultados indicam que não há diferenças significativas ao optar por qualquer uma das formas para estimar t.

Distância Euclidiana Ponderada (DEP)

Redefinimos a Distância Euclidiana Ponderada como:

$$d_w(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^{m} w_l (x_i^l - x_j^l)^2 \right)^q . \tag{15}$$

q	s_w	w_1	w_2	$\sigma(t_m)$
0.25	EX	52.27	47.73	1.64 (1.49, 1.77)
	FR	49.91	50.09	0.64 (0.60, 0.68)
0.5	EX	64.56	35.44	4.91 (4.27, 5.54)
	FR	56.25	43.75	1.44 (1.31, 1.55)
1.0	EX	75.65	24.35	7.60 (6.69, 8.47)
	FR	65.09	34.91	3.43 (3.06, 3.73)
2.0	EX	83.99	16.01	8.24 (7.32, 9.11)
	FR	74.22	25.78	6.60 (5.77, 7.35)
4.0	FR	82.22	17.78	7.97 (7.02, 8.83)

Tabela: Pesos estimados w_1, w_2 e desvio padrão dos tempos de falha estimados $\sigma(t_m)$ para diferentes valores de q e funções de similaridade s_w.

- Nota-se uma correlação positiva entre o parâmetro q e a variabilidade dos tempos de falha estimados, indicada pelo desvio padrão σ.

Distância de Minkowski Ponderada (DMP)

A Distância de Minkowski Ponderada pode ser expressa da seguinte maneira:

$$d_w(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^{m} w_l |x_i^l - x_j^l|^p \right)^{1/p} \tag{16}$$

p	s_w	w_1	w_2	$\sigma(t_m)$
0.25	FR	88.23	11.77	6.83 (6.15, 7.62)
0.5	FR	85.01	14.99	5.79 (5.24, 6.46)
1.0	EX	89.49	10.51	9.79 (8.57, 10.78)
	FR	66.74	33.26	3.28 (3.06, 3.51)
2.0	EX	64.56	35.44	4.91 (4.27, 5.55)
	FR	56.25	43.75	1.44 (1.31, 1.55)
4.0	EX	54.9	45.1	1.48 (1.29, 1.64)
	FR	53.58	46.42	0.72 (0.65, 0.78)

Tabela: Pesos estimados w_1, w_2 e desvio padrão dos tempos de falha estimados $\sigma(t_m)$ para diferentes valores de p e funções de similaridade s_w.

- Ao analisar o desvio padrão dos tempos estimados de falha, notamos uma correlação negativa com o valor de p.

d_w		s_w	CI	IBS
DMP	$p = 0.25$	FR	0.702 (0.658, 0.745)	0.142 (0.122, 0.168)
	$p = 0.5$	FR	0.718 (0.671, 0.759)	0.138 (0.119, 0.165)
	$p = 1.0$	EX	0.748 (0.708, 0.781)	0.134 (0.112, 0.160)
		FR	0.744 (0.704, 0.778)	0.146 (0.125, 0.172)
	$p = 2.0$	EX	0.770 (0.737, 0.802)	0.136 (0.116, 0.159)
		FR	0.750 (0.711, 0.783)	0.162 (0.138, 0.190)
	$p = 4.0$	EX	0.767 (0.735, 0.798)	0.162 (0.139, 0.190)
		FR	0.746 (0.709, 0.778)	0.170 (0.145, 0.198)
DEP	$q = 0.25$	EX	0.754 (0.717, 0.787)	0.160 (0.136, 0.188)
		FR	0.712 (0.672, 0.748)	0.171 (0.146, 0.199)
	$q = 0.5$	EX	0.770 (0.737, 0.802)	0.136 (0.116, 0.159)
		FR	0.750 (0.711, 0.783)	0.162 (0.138, 0.190)
	$q = 1.0$	**EX**	**0.772 (0.738, 0.805)**	**0.125 (0.107, 0.146)**
		FR	**0.773 (0.741, 0.806)**	**0.145 (0.124, 0.170)**
	$q = 2.0$	EX	0.773 (0.738, 0.807)	0.124 (0.105, 0.146)
		FR	0.771 (0.737, 0.805)	0.128 (0.110, 0.149)
	$q = 4.0$	FR	0.772 (0.737, 0.806)	0.125 (0.106, 0.146)

Tabela: Métricas de performance do estimador (CI e IBS) para diferentes distâncias definidas pelos valores de p e q, e funções de similaridade s_w.

- Observa-se que não há diferença estatisticamente significativa em nenhuma das métricas relatadas.

Robustez dos Pesos Estimados

- Ao variar o chute inicial para a otimização dos parâmetros, observamos que a variação dos pesos estimados w_1 e w_2 é mínima;
- Optamos por manter o chute inicial como $\mathbf{w} = \mathbf{0}$.

s_w	w_1	w_2	**CI**	**IBS**
EX	**75.65 (75.64, 75.65)**	**24.35 (24.35, 24.36)**	**0.774 (0.774, 0.774)**	**0.125 (0.125, 0.125)**
FR	65.09 (65.08, 65.09)	34.91 (34.91, 34.92)	0.775 (0.775, 0.775)	0.136 (0.136, 0.136)

Tabela: Pesos estimados w_1, w_2 e métricas de performance do estimador (CI e IBS) para diferentes chutes iniciais.

Realizamos uma repetição da técnica de amostragem *bootstrap* com reposição na base de treino, para avaliar a consistência e robustez dos pesos estimados.

Figura: Pesos estimados w_1, w_2 para diferentes formas da função de similaridade s_w.

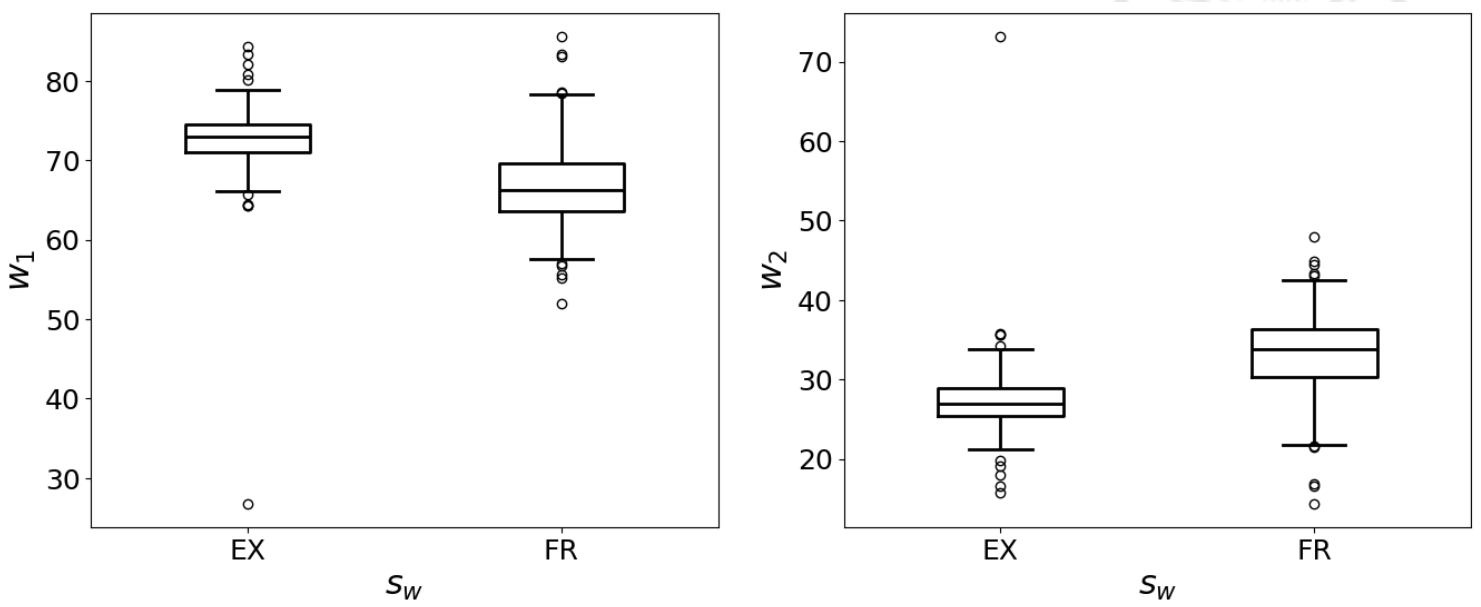
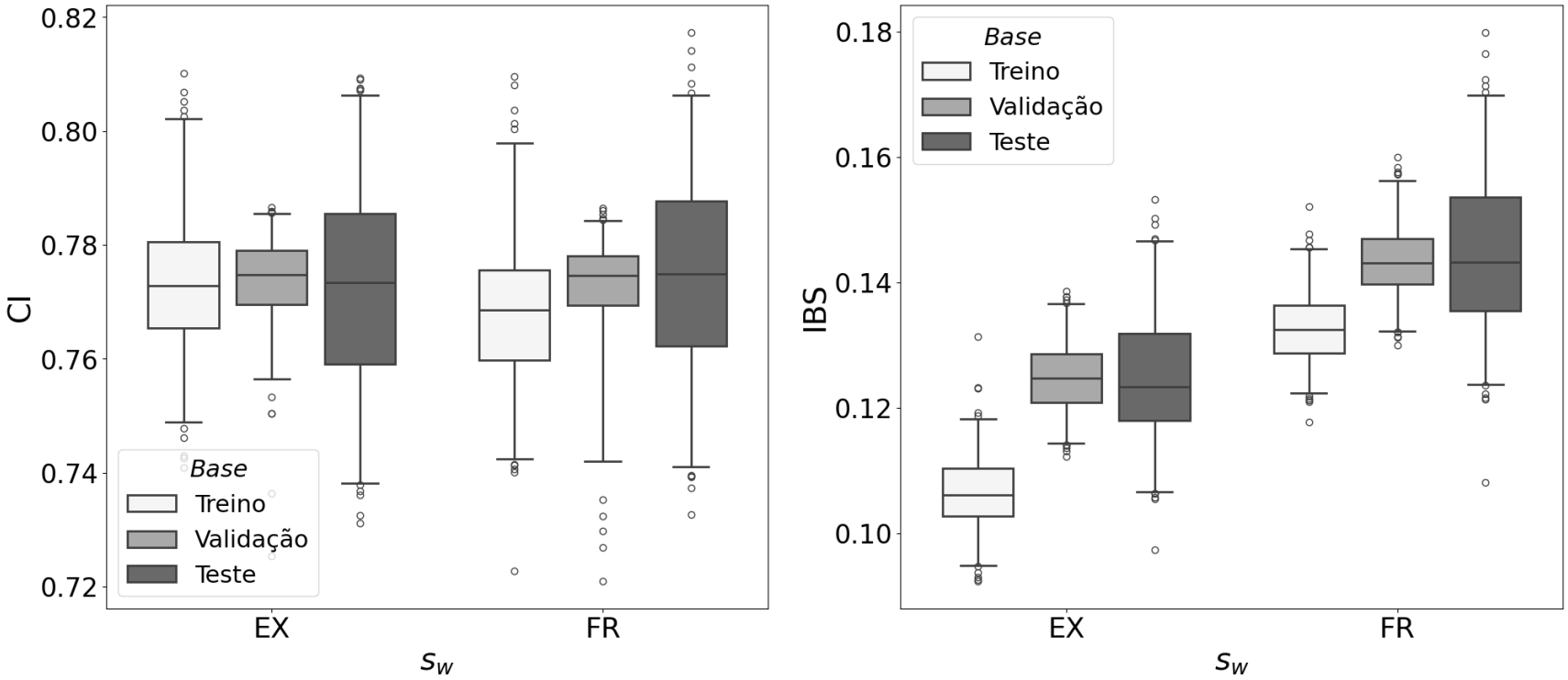


Figura: Métricas de performance do estimador (CI e IBS) em treino, validação e teste para diferentes funções de similaridade s_w.



Modelos de Referência

- O SBKM se mostra competitivo em relação aos demais modelos nas métricas de CI e IBS na base de teste.

	Treino		Validação		Teste	
Modelo	**CI**	**IBS**	**CI**	**IBS**	**CI**	**IBS**
KM	-	-	-	-	0.500 (0.500, 0.500)	0.170 (0.146, 0.198)
SBKM	**0.770**	**0.111**	**0.777**	**0.121**	**0.772 (0.738, 0.805)**	**0.125 (0.107, 0.146)**
COX	0.770	0.123	0.788	0.116	0.775 (0.738, 0.808)	0.118 (0.102, 0.137)
EN-COX	0.769	0.123	0.789	0.117	0.774 (0.739, 0.808)	0.118 (0.102, 0.138)
WEIBULL	0.770	0.121	0.789	0.113	0.776 (0.735, 0.806)	0.115 (0.098, 0.139)
ST	0.764	0.124	0.789	0.128	0.722 (0.688, 0.758)	0.130 (0.110, 0.155)
RST	0.789	0.118	0.793	0.122	0.762 (0.724, 0.798)	0.128 (0.110, 0.151)
GB-COX	0.873	0.080	0.807	0.121	0.726 (0.681, 0.767)	0.137 (0.113, 0.166)

Tabela: Métricas de performance dos modelos (CI e IBS) em treino, validação e teste.

Amostragem

- Para o menor valor de n, os desempenhos em treino são os melhores. Por outro lado, os desempenhos em teste são os piores.
- A partir de $n = 100$, as métricas alcançadas são tão boas quanto quando se utiliza a base de treinamento completa.

			Treino		**Validação**		**Teste**	
n	w_1	w_2	**CI**	**IBS**	**CI**	**IBS**	**CI**	**IBS**
25	61.55	38.45	0.888	0.034	0.646	0.121	0.605 (0.551, 0.655)	0.247 (0.212,0.278)
50	63.39	36.61	0.778	0.115	0.644	0.260	0.643 (0.589, 0.690)	0.173 (0.155, 0.192)
100	66.09	33.91	0.811	0.108	0.752	0.174	0.751 (0.712, 0.784)	0.127 (0.115, 0.143)
250	**70.74**	**29.26**	**0.755**	**0.108**	**0.776**	**0.137**	**0.762 (0.727, 0.794)**	**0.121 (0.103, 0.143)**
560	75.65	24.35	0.770	0.111	0.777	0.121	0.772 (0.738, 0.805)	0.125 (0.107, 0.146)

Tabela: Pesos estimados w_1, w_2 e métricas de performance do estimador (CI e IBS) em treino, validação e teste para amostras de diferentes tamanhos n.

- Observamos que o desempenho de todos os modelos para uma amostra de tamanho $n = 250$ foi semelhante à performance obtida ao utilizar o conjunto de dados original.

	Treino		Validação		Teste	
Modelo	**CI**	**IBS**	**CI**	**IBS**	**CI**	**IBS**
KM	-	-	-	-	0.500 (0.500, 0.500)	0.179 (0.157, 0.206)
SBKM	**0.755**	**0.108**	**0.776**	**0.128**	**0.762 (0.737, 0.794)**	**0.121 (0.103, 0.143)**
COX	0.758	0.115	0.788	0.120	0.773 (0.738, 0.807)	0.118 (0.104, 0.137)
EN-COX	0.757	0.114	0.788	0.121	0.773 (0.737, 0.807)	0.119 (0.104, 0.136)
WEIBULL	0.758	0.112	0.789	0.118	0.773 (0.738, 0.805)	0.116 (0.099, 0.136)
ST	0.681	0.676	0.116	0.131	0.658 (0.612, 0.694)	0.136 (0.115, 0.160)
RST	0.804	0.110	0.763	0.126	0.741 (0.704, 0.776)	0.127 (0.113, 0.144)
GB-COX	0.871	0.079	0.744	0.141	0.709 (0.668, 0.748)	0.140 (0.118, 0.166)

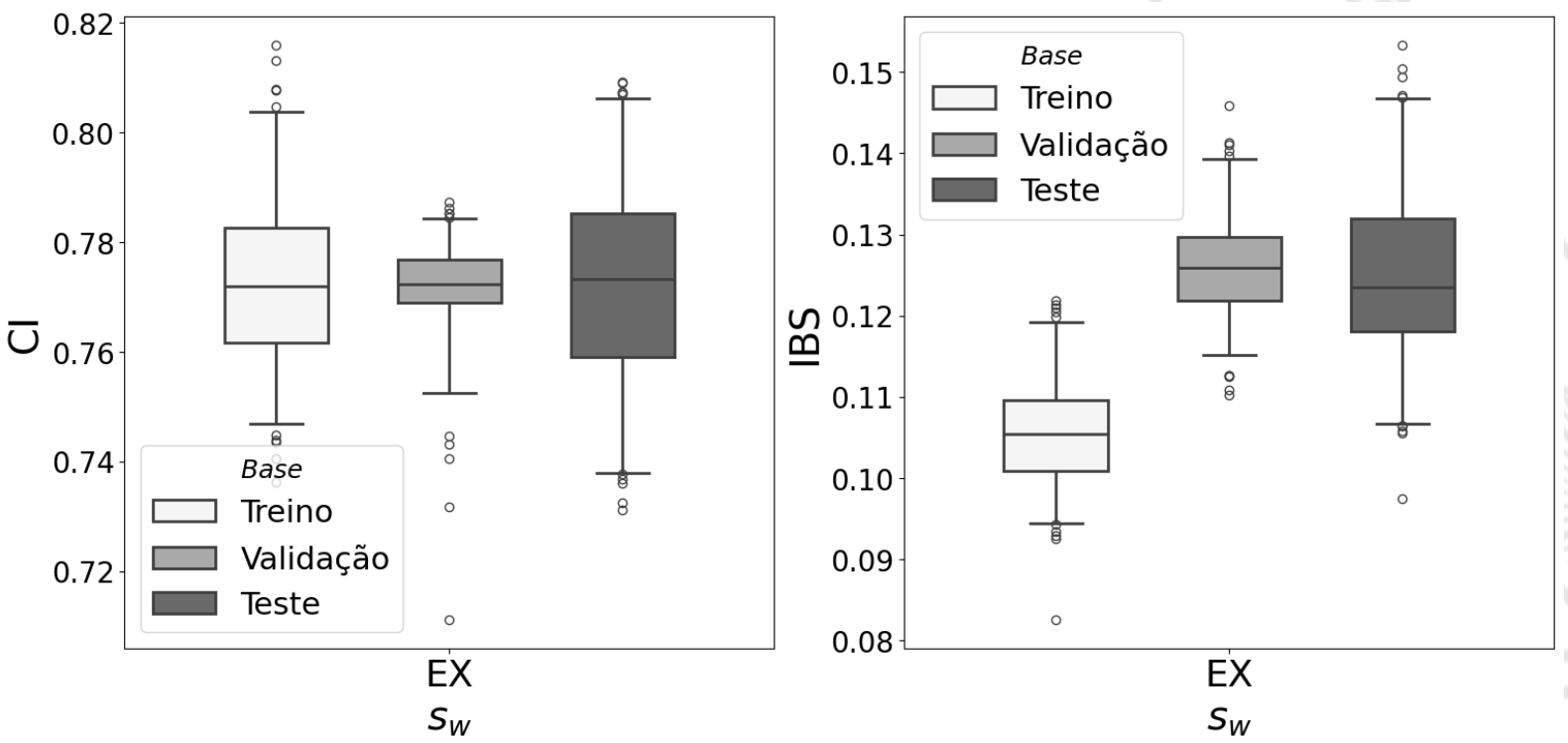
Tabela: Métricas de performance dos modelos (CI e IBS) em treino, validação e teste para uma amostra de $n = 250$.

- Podemos inferir que há uma consistência semelhante tanto nos pesos quanto nas métricas de avaliação quando os pesos são estimados com uma amostra reduzida de $n = 250$, em comparação com a amostra de $n = 560$.

n	w_1	w_2
250	71.93 (67.24, 76.40)	28.07 (23.60, 32.76)
560	72.68 (66.16, 78.88)	27.32 (21.12, 33.84)

Tabela: Pesos estimados w_1, w_2 com seus respectivos intervalos de confiança para amostras de $n = 250$ e $n = 560$.

Figura: Métricas de performance do estimador (CI e IBS) em treino, validação e teste para amostras de tamanho $n = 250$.



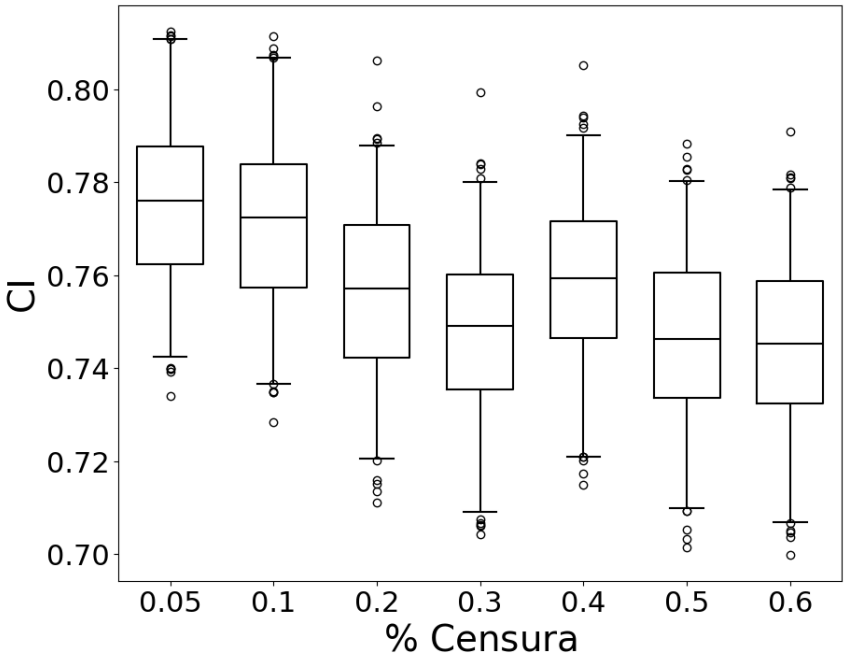
Taxa de Censura

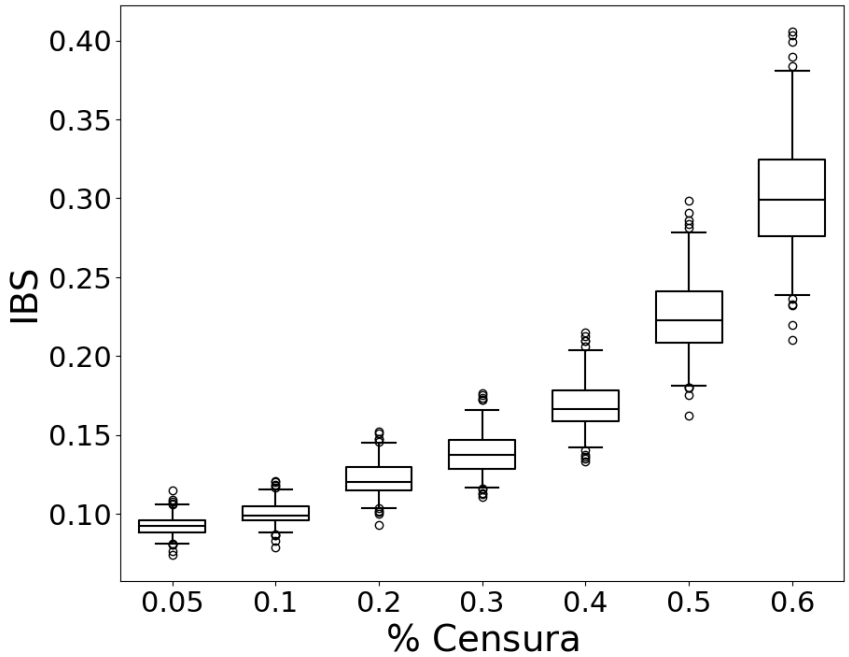
% Censura	w_1	w_2	$\sigma(t_m)$
0.05	78.90	21.10	6.18 (5.12, 7.15)
0.1	76.43	23.57	6.44 (5.37, 7.50)
0.2	70.83	29.17	7.21 (6.08, 8.31)
0.3	72.91	27.09	8.43 (7.25, 9.48)
0.4	70.93	29.07	8.60 (7.57, 9.63)
0.5	70.47	29.53	8.96 (8.00, 9.85)
0.6	73.46	26.54	8.60 (7.60, 9.38)

Tabela: Pesos estimados w_1, w_2 e desvio padrão dos tempos de falha estimados $\sigma(t_m)$ para amostras com diferentes taxas de censura.

- A menor taxa de censura possui ao menor desvio padrão dos tempos de falha estimados. Por outro lado, o maior desvio padrão ocorre quando metade da amostra é composta por censuras.

Figura: Métricas de performance do estimador (CI e IBS) para amostras com diferentes taxas de censura.





UNIVERSIDADE FEDERAL DO CEARÁ

Base de dados SUPPORT

- Devido ao tamanho e complexidade da base de SUPPORT, tivemos limitações de tempo computacional.
- Constatamos que o SBKM é competitivo em relação aos demais modelos avaliados nesta base também.
- Ao testar diferentes tamanhos de amostra, observamos que o desempenho do estimador permaneceu semelhante em ambas as métricas de avaliação para todos os tamanhos avaliados, sendo o menor $n = 100$ e o maior $n = 4968$.
- Ao variar a taxa de censura da base de treinamento, notamos que apenas o IBS foi significativamente prejudicado à medida que a porcentagem de censuras aumentou.

Conclusões

- O estimador de Kaplan-Meier baseado em similaridade apresenta desempenho competitivo nas métricas avaliadas;
- Diferentemente dos métodos estatísticos, o estimador proposto não assume distribuições de probabilidade ou riscos proporcionais;
- Comparado aos algoritmos de aprendizado de máquina, o estimador oferece uma interpretação mais simples dos parâmetros estimados e evitamos problemas de sobreajuste com maior facilidade.
- Ao utilizar covariáveis categóricas, o SBKM apresenta a vantagem de ser adaptável e mais parcimonioso em comparação aos demais modelos avaliados.

Trabalhos futuros

- Comparar o estimador de Kaplan-Meier baseado em similaridade (SBKM) com diferentes adaptações do estimador de Kaplan-Meier (KM) discutidas na literatura;
- Explorar propriedades mais específicas do estimador utilizando dados de sobrevivência simulados;
- Investigar estratégias para otimizar o tempo computacional do estimador.

Obrigada pela Atenção!

Contato:

isabelcastro@alu.ufc.br

UNIVERSIDADE FEDERAL DO CEARÁ