Inferencia non paramétrica para big-but-biased data

Ricardo Cao e Laura Borrajo
Grupo MODES, Universidade da Coruña (UDC)
Galicia, España
e-mail: rcao@udc.es, laura.borrajo@udc.es

Machine Learning Workshop Galicia
27 de outubro de 2016

Índice

Motivación

- Estatísticos e informáticos teñen un papel importante na era do Big Data (BD)
- Falsa idea: "con suficientes datos, os números falan por si sós"
- En certas ocasións, a mostra de gran tamaño (BD) non é representativa da nosa poboación

Algúns exemplos de big-but-biased data (B3D)

- Street Bump: http://www.streetbump.org/ (Crawford (2013))

- *Tweets* durante o furacán Sandy: unha base de datos que conta con 20 millóns de *tweets* durante o furacán, 27 de outubro- 1 de novembro, 2012 (Crawford (2013))
- Moitos outros en Hargittai (2015): As enquisas vía web non obedecen unha selección aleatoria, senón que está nesgada como en certas formas que producen mostras que limitan a xeneralización dos resultados

Mostra(s) e tamaño(s)

Consideremos unha poboación con función de distribución acumulada F (densidade f) e unha mostra aleatoria simple,

$$(X_1, \ldots, X_n),$$

de tamaño n. Supoñamos por un intre que non somos capaces de observar esta mostra, pero si observamos, en cambio, outra mostra

$$(Y_1, \ldots, Y_N)$$

dun tamaño moito maior ($N >> n$) procedente dunha distribución nesgada, G (densidade g). Supomos que

$$g(x) = w(x)f(x)$$

para algunha función peso w, tal que $w(x) \geq 0$, $\forall x$.
Consideremos algún parámetro $\theta = \theta(F)$ ...

Preguntas:

- Podemos estimar θ empregando tan só a mostra Y(B3D)?
- Precisamos coñecer a función peso?
- É realista supor que w é coñecida?
- Hai algún xeito de proceder no caso de non coñecer w?
- Ten algo que ver isto coa estimación non paramétrica de curvas?
- Como de grande debe ser o tamaño mostral de Y, N, en relación a n para mellorar os resultados da mostra aleatoria simple?
- Son benvidas máis preguntas (e respostas!!) do público

O caso da mostra aleatoria simple

Centrémonos no problema da estimación da media $\mu = \int x dF(x)$ en un contexto non paramétrico. Se somos quen de observar a mostra X, sinxelo: usamos $\overline{X}$.

- $\overline{X}$ é un estimador innesgado de μ e $Var(\overline{X}) = \frac{\sigma^2}{n}$
- O Teorema Central do Límite é unha boa ferramenta para facer inferencia sobre μ

Pero que ocorre no caso B3D?

Podemos estimar μ no contexto B3D?

Si! Supoñamos que a distribución F é continua, con densidade f. Neste contexto $\overline{Y}$ xa non é un estimador consistente de $\mu = \int x f(x)\, dx$, pero si de

$$\mu_g = \int x g(x)\, dx = \int x w(x) f(x)\, dx.$$

Por outra parte, como $w(x) = \frac{g(x)}{f(x)}$, é fácil comprobar que

$$\begin{aligned} E\left(\frac{Y}{w(Y)}\right) &= \int \frac{y}{w(y)} g(y)\, dy = \int \frac{y}{g(x)/f(x)} g(y)\, dy \\ &= \int y f(y)\, dy = \mu \end{aligned}$$

Isto non é nada novo, xa que hai moitos artigos sobre datos nesgados por lonxitude (incluso cun punto de vista non paramétrico).

Se coñecemos w ...

Se ocorre que w é coñecida (o cal non acontece, na práctica, a miúdo) podemos empregar a anterior esperanza para obter un estimador para μ:

$$\hat{\mu}^{BBBS,w} = \frac{1}{N}\sum_{j=1}^{N}\frac{Y_j}{w\left(Y_j\right)}. \tag{1}$$

Este estimador é unha media mostral (da poboación $Z = Y/w(Y)$) polo que obtemos de forma inmediata o seu nesgo, varianza e distribución asintótica:

$$E\left(\hat{\mu}^{BBBS,w}\right) = \mu \quad , \quad Var\left(\hat{\mu}^{BBBS,w}\right) = \frac{\sigma_Z^2}{N}$$

$$\frac{\sqrt{N}\left(\hat{\mu}^{BBBS,w} - \mu\right)}{\sigma_Z} \quad \rightarrow \quad N(0,1),$$

onde $\sigma_Z^2 = \int y^2 f(y)^2 g(y)^{-1} dy - \mu^2 = \int x^2 f(x) w(x)^{-1} dx - \mu^2$.

Se w é descoñecida, precisamos máis información ...

Cando w non é coñecida (situación moi realista) temos que estimala. Pero a BBBS (mostra Y) non é suficiente para facelo. Podemos pensar en varias situacións:

1. A mostra nesgada pode ser enriquecida con información sobre as frecuencias (ν_i para $i = 1, \ldots, N$) de cada dato Y observado.
2. O mecanismo de mostraxe nesgado pode ser replicado de novo sobre a mostra observada para obter unha submostra "duas veces nesgada" $(Y_{\ell_1}, \ldots, Y_{\ell_k})$ de tamaño k
3. Unha mostra aleatoria simple de tamaño pequeno, n, $(X_1, \ldots, X_n)$, pode extraerse da poboación.

#1 e #2 poden ocorrer no exemplo do Street Bump, e #2 ocorre a miúdo na mostraxe de autoselección (como en datos masivos de redes sociais). Nesta charla consideramos as situacións #2 e #3.

Estimador núcleo para #3

O estimador de Parzen-Rosenblatt pode ser empregado para estimar $f(x)$ e $g(x)$:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$$

$$\hat{g}_b(x) = \frac{1}{N}\sum_{i=1}^{N} K_b(x - Y_i)$$

Agora, a función w pode estimarse de forma sinxela:

$$\hat{w}_{h,b}(x) = \frac{\hat{g}_b(x)}{\hat{f}_h(x)}$$

e substituíndo este estimador en (1) obtemos

$$\hat{\mu}^{BBBS,\hat{w}_{h,b}} = \frac{1}{N}\sum_{j=1}^{N} \frac{Y_j}{\hat{w}_{h,b}(Y_j)} = \frac{1}{N}\sum_{j=1}^{N} \frac{Y_j \hat{f}_h(Y_j)}{\hat{g}_b(Y_j)} \tag{2}$$

Un estimador fácil de estudar

Para estudar as propiedades asintóticas, desprazámonos cara unha versión aproximada de (2) que supón que g é coñecida. Situación natural no contexto B3D xa que $N >> n$ (i.e., $\lim_{n\to\infty} N/n = \infty$):

$$\hat{\mu}^{BBBS,\hat{w}_h} = \frac{1}{N}\sum_{j=1}^{N} \frac{Y_j \hat{f}_h\left(Y_j\right)}{g\left(Y_j\right)} \tag{3}$$

Este "estimador" depende dun só parámetro: h.

Nesgo e varianza asintótica

É sinxelo comprobar que $\hat{\mu}^{BBBS,\hat{w}_h}$ é un estimador innesgado:

$$\begin{aligned} E\left(\hat{\mu}^{BBBS,\hat{w}_h}\right) &= E\left(\frac{Y_1 \hat{f}_h\left(Y_1\right)}{g\left(Y_1\right)}\right) = E\left(\frac{Y_1 \cdot K_h * f\left(Y_1\right)}{g\left(Y_1\right)}\right) \\ &= \int y \cdot K_h * f\left(y\right) dy = \int y \cdot f\left(y\right) dy = \mu. \end{aligned}$$

A súa varianza asintótica ($h \to 0$, $nh \to \infty$ e $N/n \to \infty$) é:

$$\begin{aligned} MSE\left(\hat{\mu}^{BBBS,\hat{w}_h}\right) &= Var\left(\hat{\mu}^{BBBS,\hat{w}_h}\right) = \frac{C_1}{n} + \frac{C_2 h^2}{n} + \frac{C_3}{Nnh} + \frac{C_4}{N} \\ &+ O\left(\frac{h^4}{n}\right) + O\left(\frac{h^2}{N}\right) + O\left(\frac{1}{Nn}\right), \end{aligned}$$

Varianza asintótica

onde

$$
\begin{aligned}
C_1 &= \int x^2 f(x)dx - \mu^2 \geq 0, \\
C_2 &= \mu_2(K)\left[\frac{1}{2}\int x^2 f''(x)dx + \int x f'(x)dx\right], \\
C_3 &= R(K)\int x^2 \frac{f(x)}{g(x)}dx \geq 0, \\
C_4 &= \int x^2 \frac{f(x)^2}{g(x)}dx - \mu^2 \geq 0,
\end{aligned}
$$

$$
R(K) = \int K(u)^2 du, \qquad \mu_2(K) = \int u^2 K(u)du.
$$

AMSE e ventá óptima

O erro cadrático medio asintótico, en función do parámetro de suavizado, h, resulta

$$AMSE\,(h) = \frac{C_1}{n} + \frac{C_2 h^2}{n} + \frac{C_3}{Nnh} + \frac{C_4}{N}.$$

Se $C_2 > 0$, minimizando a expresión anterior en h, obtemos:

$$h_{AMSE} \quad = \quad \left(\frac{C_3}{2C_2}\right)^{1/3} N^{-1/3}.$$

Mais neste caso, todos os termos da expresión do $AMSE(h)$ son positivos, co que

$$AMSE(h_{AMSE}) > MSE(\overline{X})$$

Modelo 1

Consideremos unha poboación con distribución $N(3,1)$ (densidade f).

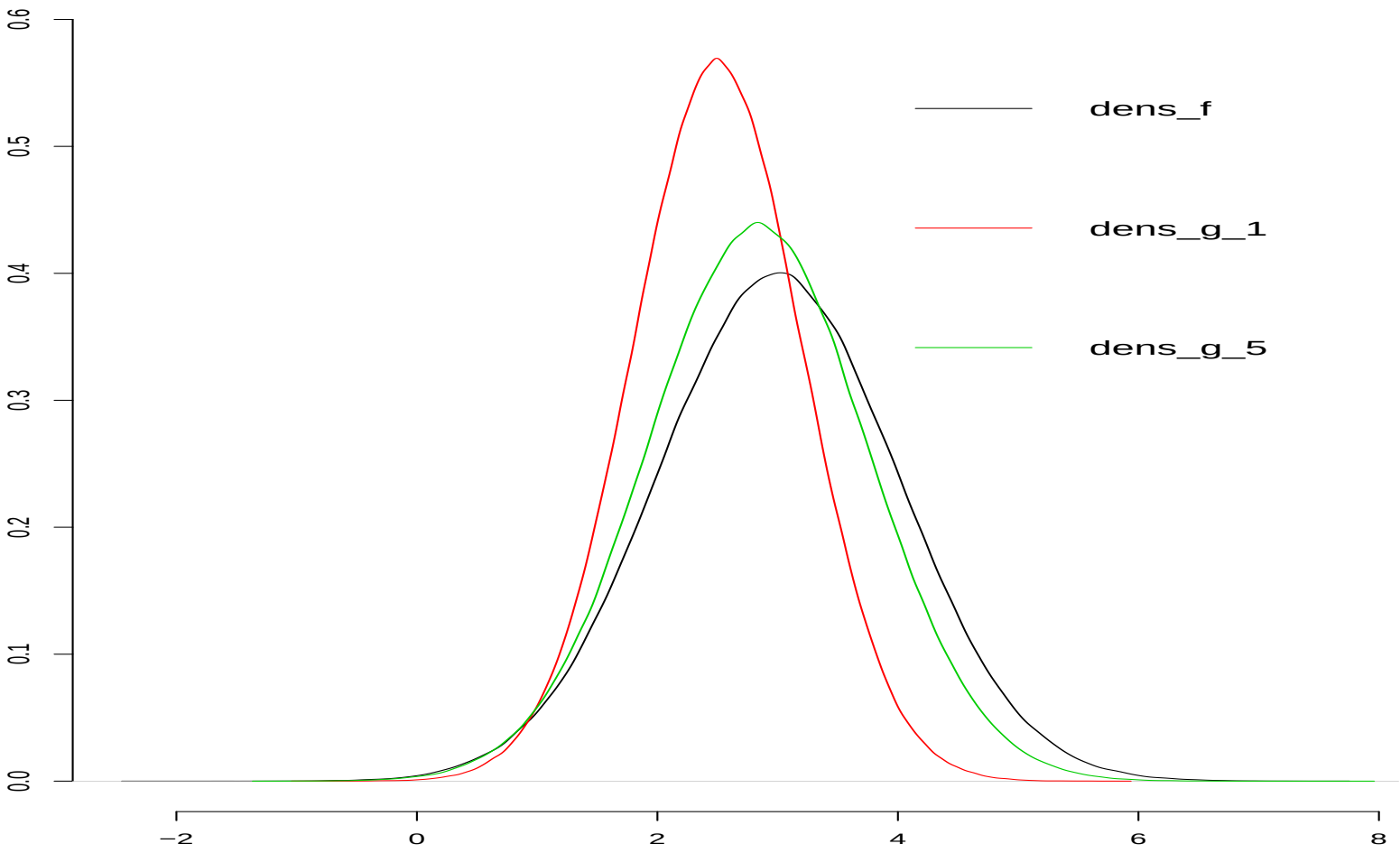
Consideramos a seguinte clase de funcións peso

$$w(x) = \frac{1}{\sigma_w \sqrt{2\pi}} e^{\left(-\frac{(x-2)^2}{2\sigma_w^2}\right)},$$

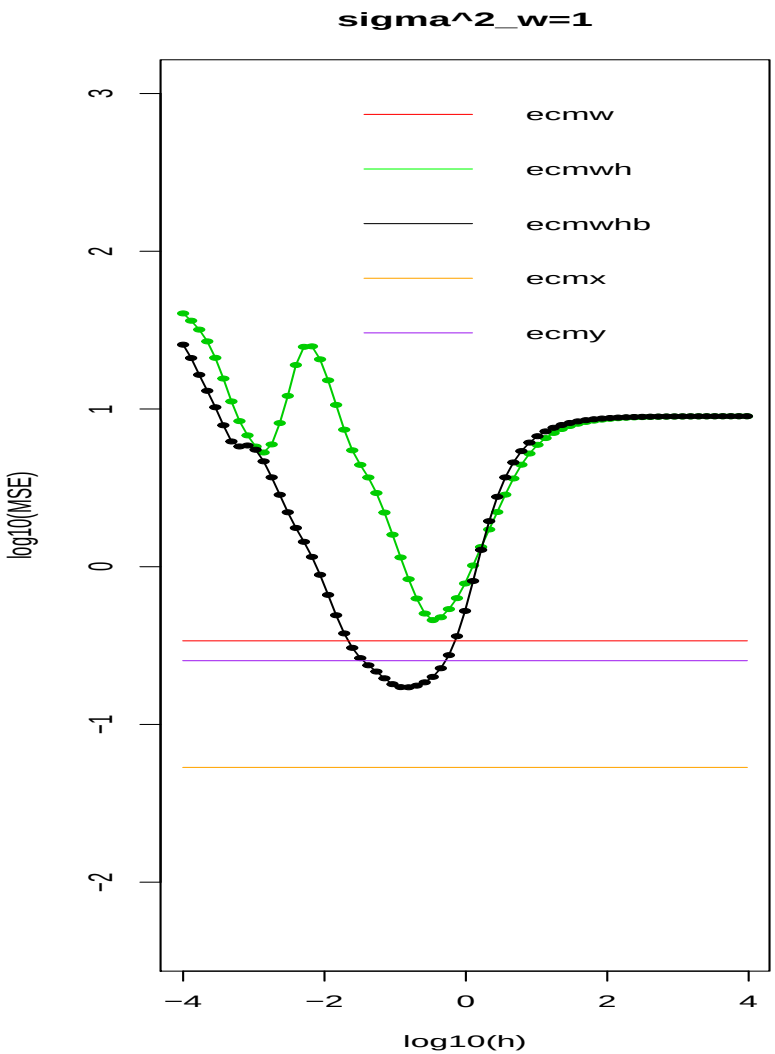
con dúas eleccións: $\sigma_w^2 = 1$ e $\sigma_w^2 = 5$, sendo a densidade nesgada

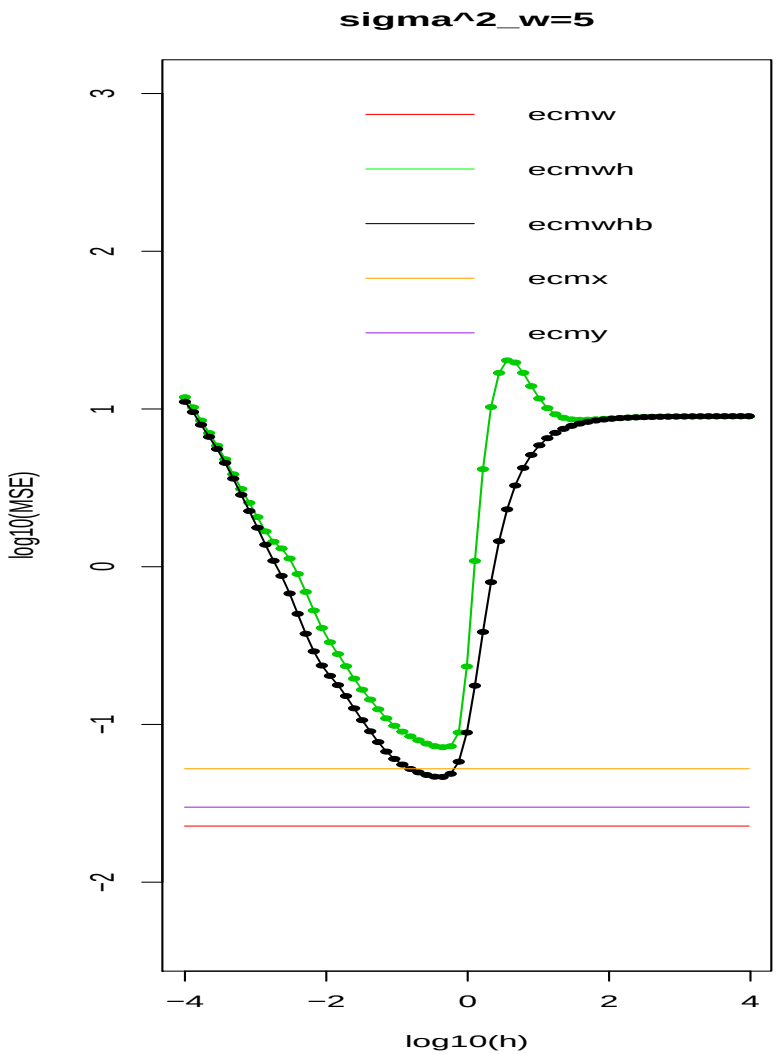
$$g(x) = \begin{cases} \frac{e^{\frac{1}{4}}}{\sqrt{\pi}} e^{\left(-\frac{(x-2)^2+(x-3)^2}{2}\right)} & \text{se} \quad \sigma_w^2 = 1 \\ \frac{\sqrt{3}e^{\frac{1}{12}}}{\sqrt{5\pi}} e^{\left(-\frac{(x-2)^2+5(x-3)^2}{10}\right)} & \text{se} \quad \sigma_w^2 = 5 \end{cases}$$

Modelo 1: densidade e densidades nesgadas

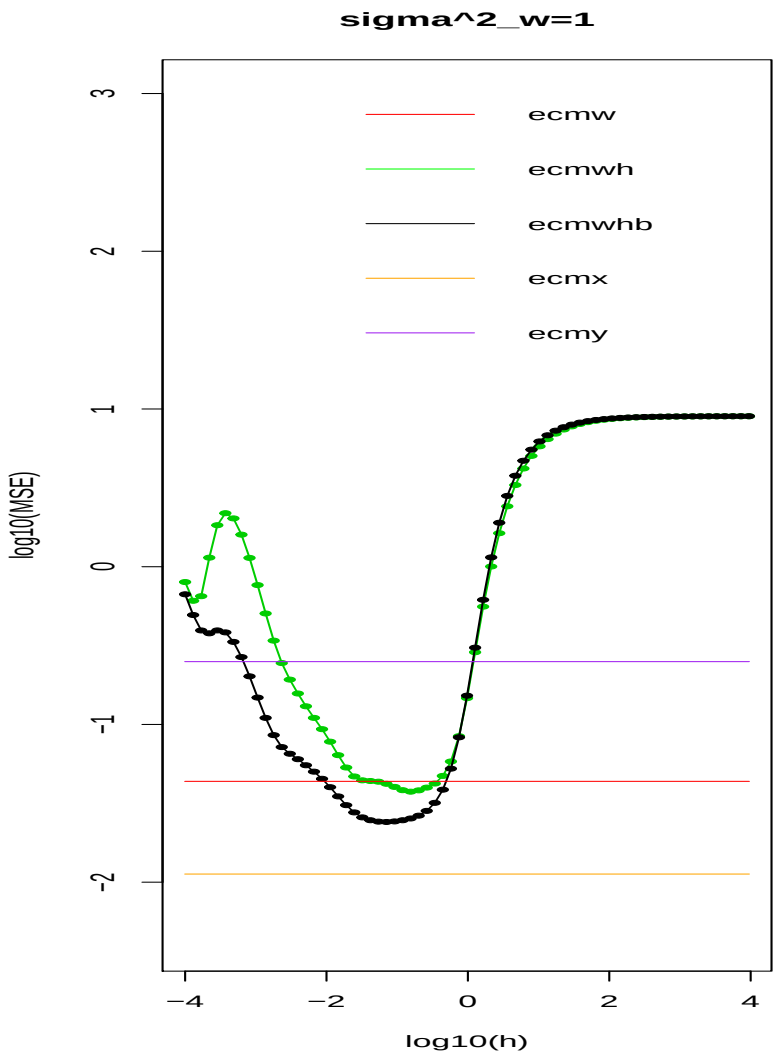


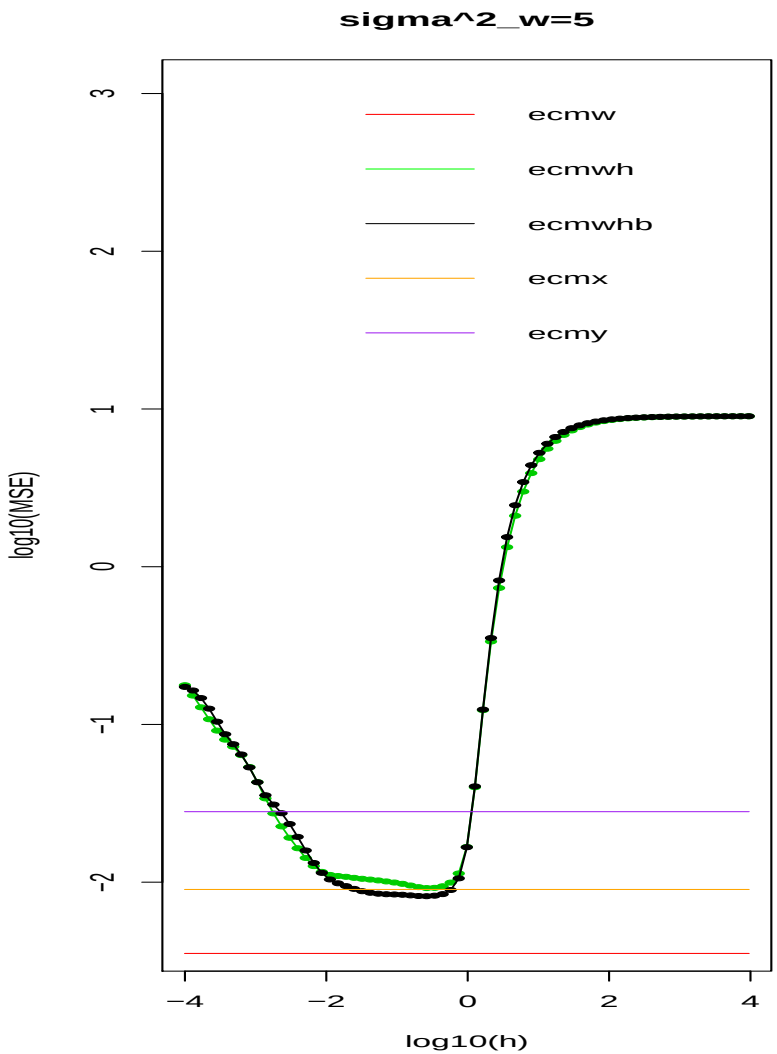
MSE Modelo 1 ($n = 20$, $N = n^2 = 400$, trials=100, b_{ROT})



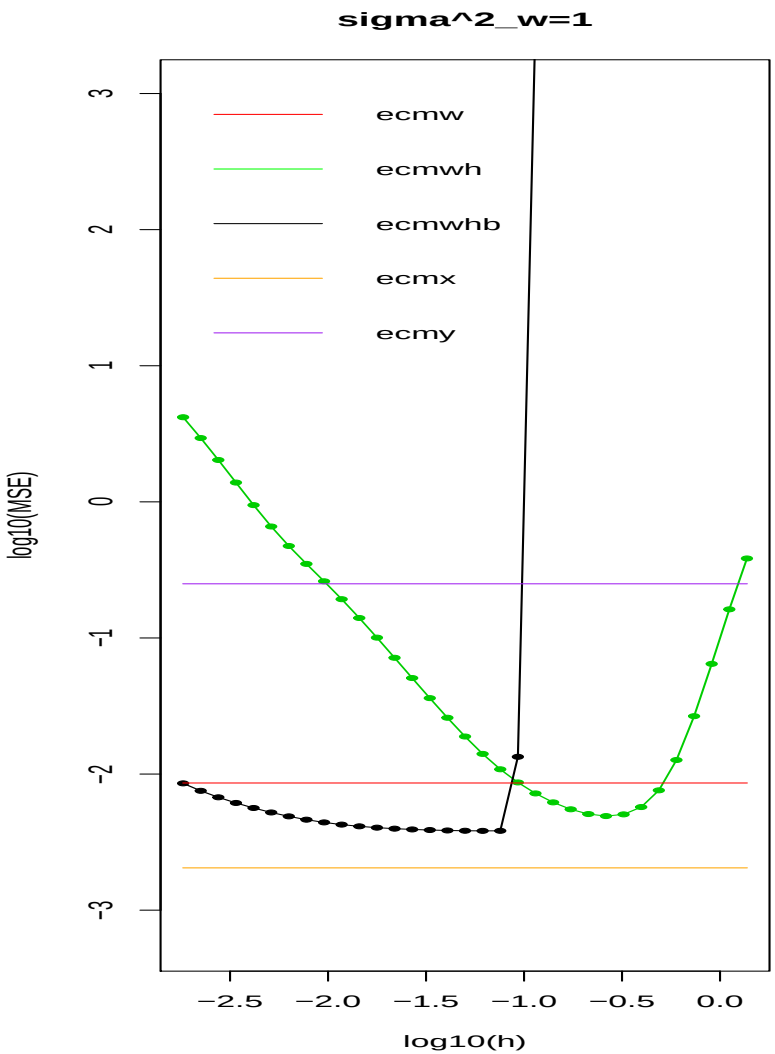


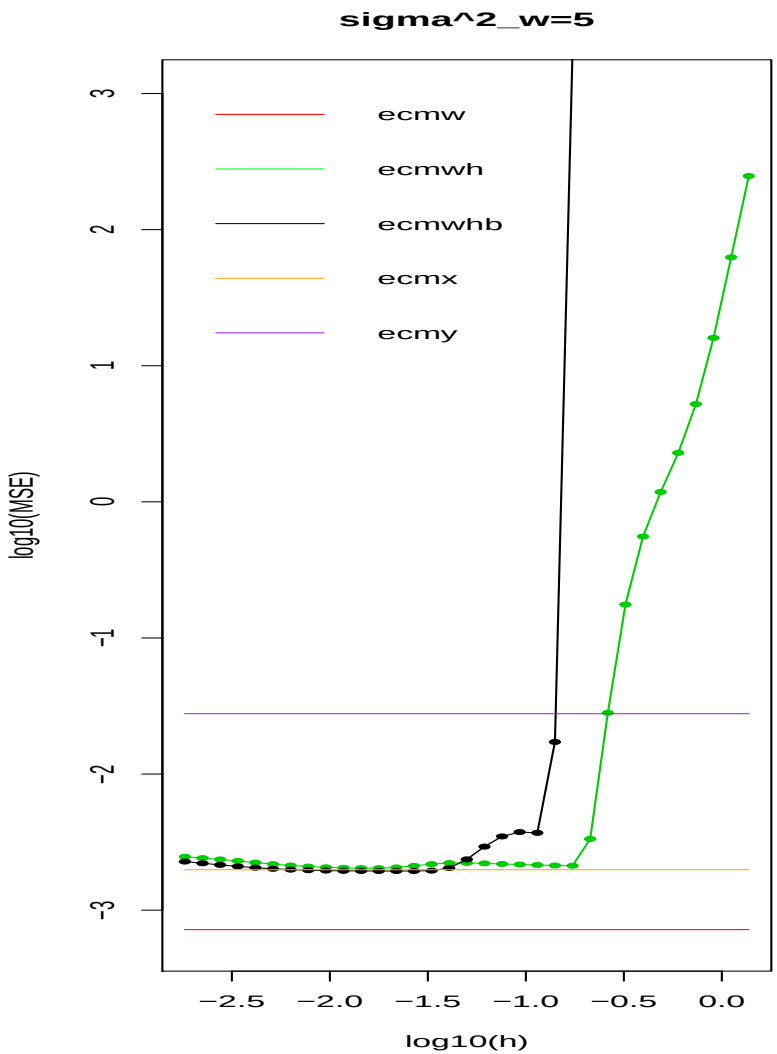
MSE Modelo 1 ($n = 100$, $N = 10000$, trials=100, b_{ROT})





MSE Modelo 1 ($n = 500$, $N = 250000$, trials=1000, b_{ROT})





Se non coñecemos X ...

Trasladémonos agora ao escenario #2. Supoñamos que non somos quen de observar a mostra X, pero si observamos a mostra

$$(Y_1, \ldots, Y_N)$$

de tamaño N procedente dunha distribución nesgada, G (densidade g) e unha mostra

$$(Z_1, \ldots, Z_n)$$

de tamaño $n << N$ procedente dunha distribución dúas veces nesgada, M (densidade m). É dicir

$$m(x) = w(x)g(x)$$

para a mesma función peso w, tal que $w(x) \geq 0$, $\forall x$ que cumpre $g(x) = w(x)f(x)$.

Estimadores para #2

- w coñecida: $w(x) = \frac{m(x)}{g(x)}$

$$\hat{\mu}^{BBBS_2,w} = \frac{1}{N}\sum_{j=1}^{N}\frac{Y_j}{w\left(Y_j\right)}$$

- w descoñecida: $\hat{w}_{h,b}(x) = \frac{\hat{m}_b(x)}{\hat{g}_h(x)}$

$$\hat{g}_h\left(x\right) = \frac{1}{N}\sum_{i=1}^{N}K_h\left(x - Y_i\right)$$

$$\hat{m}_b\left(x\right) = \frac{1}{n}\sum_{i=1}^{n}K_b\left(x - Z_i\right)$$

$$\hat{\mu}^{BBBS_2,\hat{w}_{h,b}} = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i}{\hat{w}_{h,b}\left(Y_i\right)} = \frac{1}{N}\sum_{i=1}^{N}Y_i\frac{\hat{g}_h\left(Y_i\right)}{\hat{m}_b\left(Y_i\right)} \tag{4}$$

Estimadores para #2

Unha versión aproximada de (4) que supón que g é coñecida:

$$\hat{\mu}^{BBBS_2,\hat{w}_b} = \frac{1}{N}\sum_{i=1}^{N} Y_i \frac{g\left(Y_i\right)}{\hat{m}_b\left(Y_i\right)}$$

Este "estimador" depende dun só parámetro: b.

Modelo 6

Consideremos unha poboación con función de densidade uniforme descoñecida

$$f(x) = \begin{cases} \frac{1}{2} & \text{se} \quad x \in [0,2] \\ 0 & \text{se} \quad x \notin [0,2] \end{cases}$$

Consideremos a función peso seguinte

$$w(x) = \begin{cases} \frac{1}{2}x & \text{se} \quad x \in [0,2] \\ 0 & \text{se} \quad x \notin [0,2] \end{cases}$$

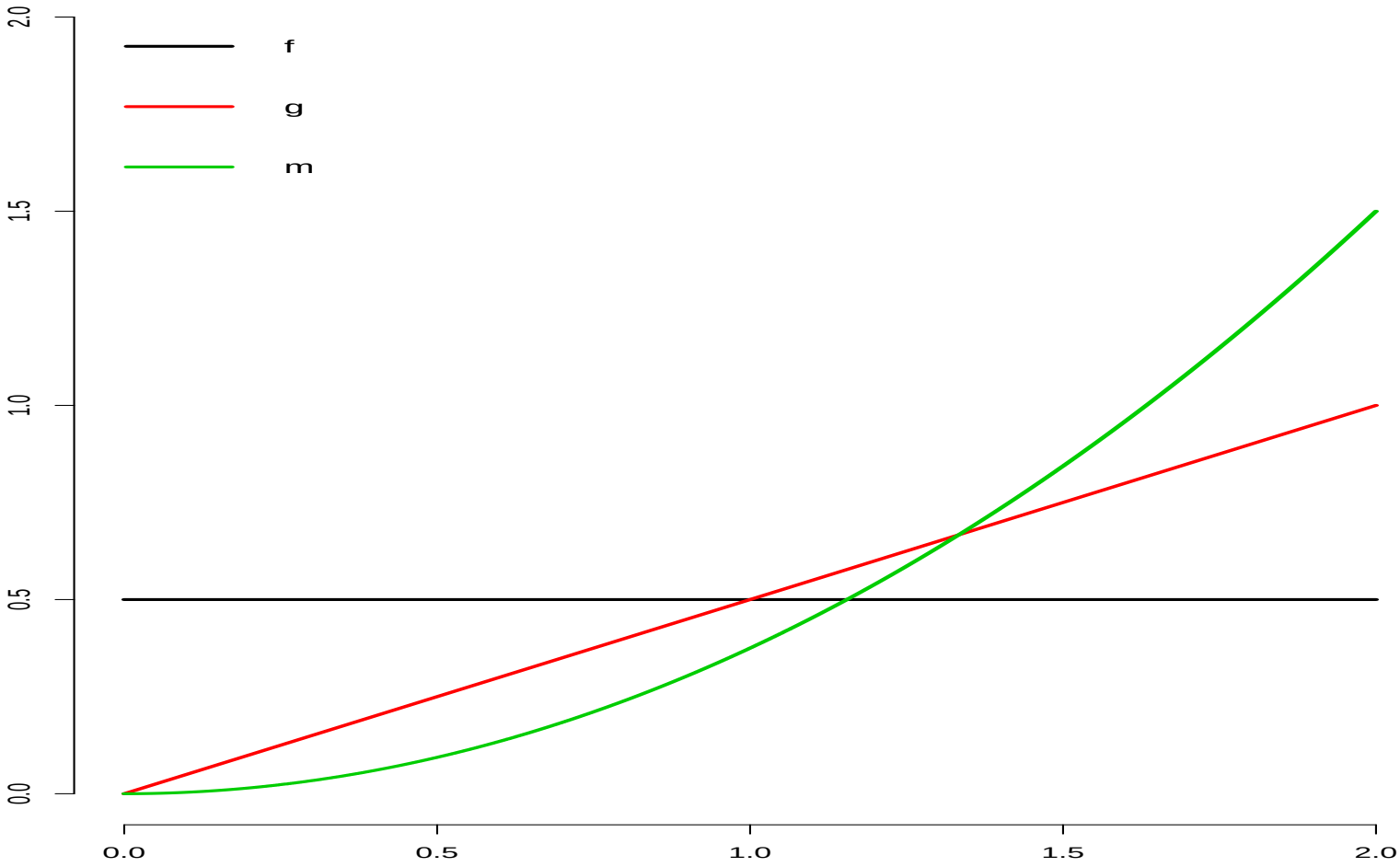
Supoñamos que coñecemos a función de densidade nesgada

$$g(x) = \begin{cases} \frac{1}{2}x & \text{se} \quad x \in [0,2] \\ 0 & \text{se} \quad x \notin [0,2] \end{cases}$$

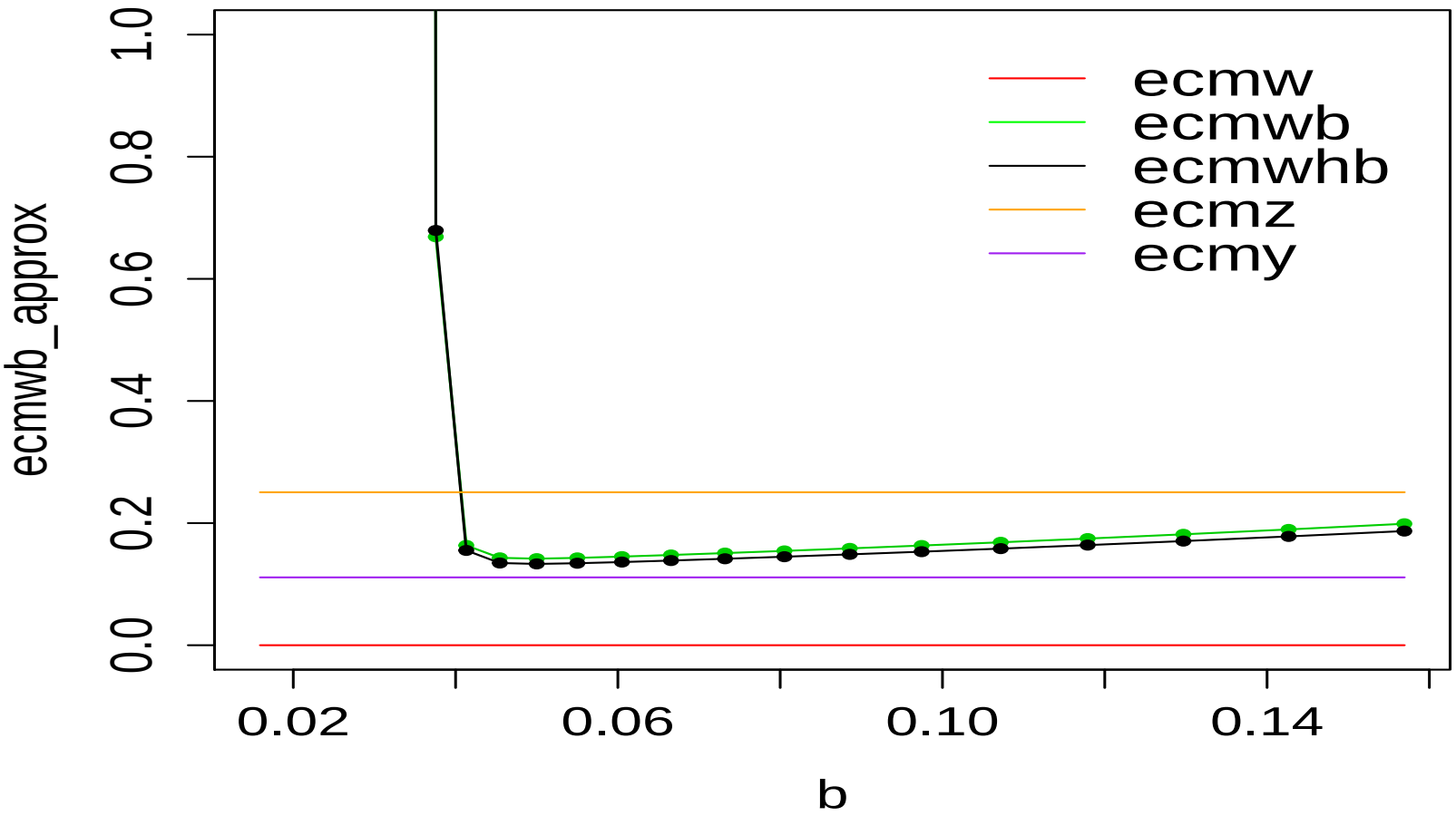
e a función de densidade dúas veces nesgada

$$m(x) = \begin{cases} \frac{3}{8}x^2 & \text{se} \quad x \in [0,2] \\ 0 & \text{se} \quad x \notin [0,2] \end{cases}$$

Modelo 6: densidade e densidades nesgadas



MSE Modelo 6 ($n = 8000$, $N = 64000000$, trials=100)



Modelo 8

Consideremos unha poboación con función de densidade expoñencial descoñecida ($X = \exp(\lambda)$).
Consideremos a función peso seguinte

$$w(x) = \begin{cases} e^{-\tau x} & \text{se} \quad x \in [0, \infty] \\ 0 & \text{se} \quad x \notin [0, \infty] \end{cases}$$

Supoñamos que coñecemos a función de densidade nesgada expoñencial

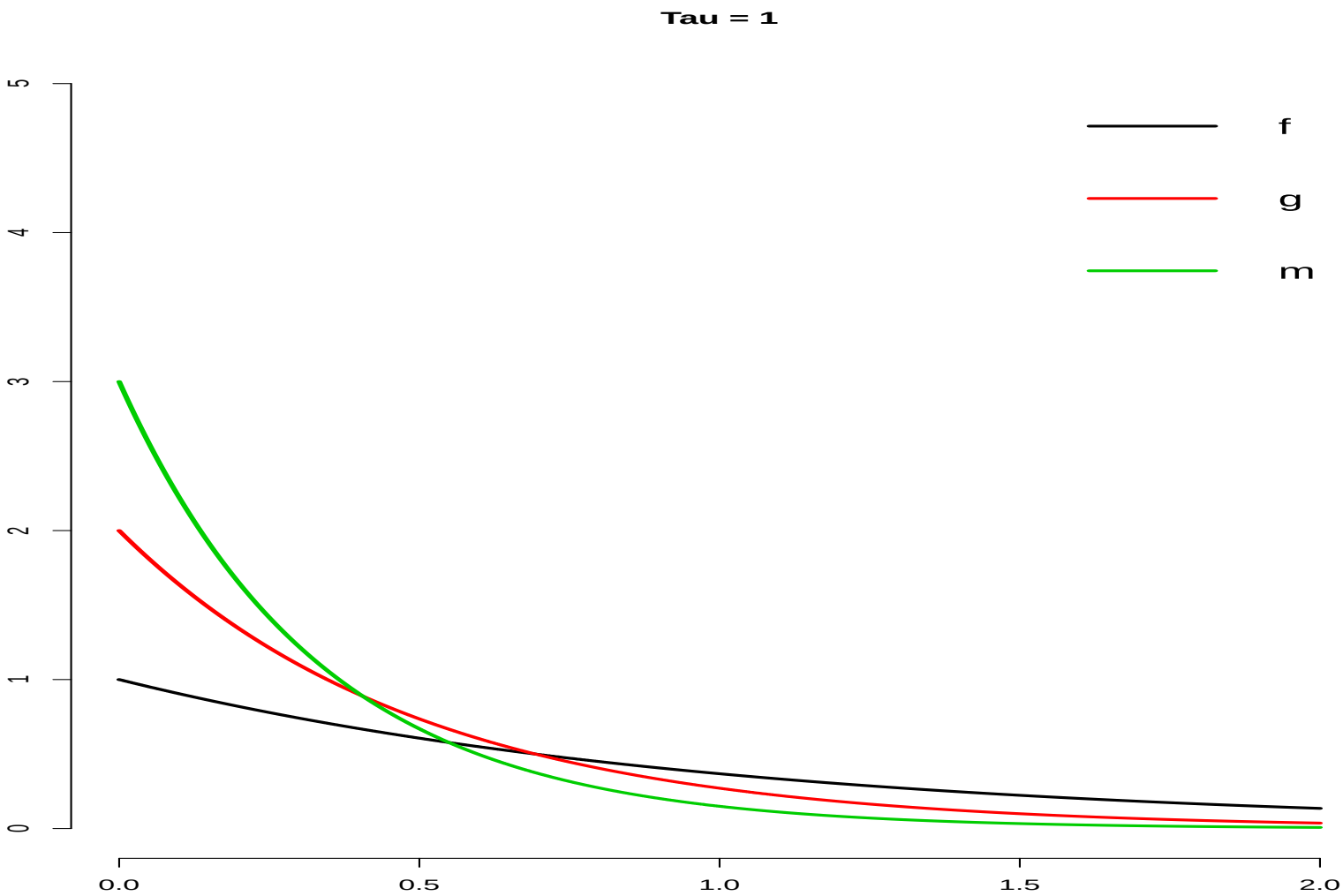
$$g(x) = \begin{cases} (\lambda + \tau) e^{-(\lambda+\tau)x} & \text{se} \quad x \in [0, \infty] \\ 0 & \text{se} \quad x \notin [0, \infty] \end{cases}$$

e a función de densidade expoñencial dúas veces nesgada

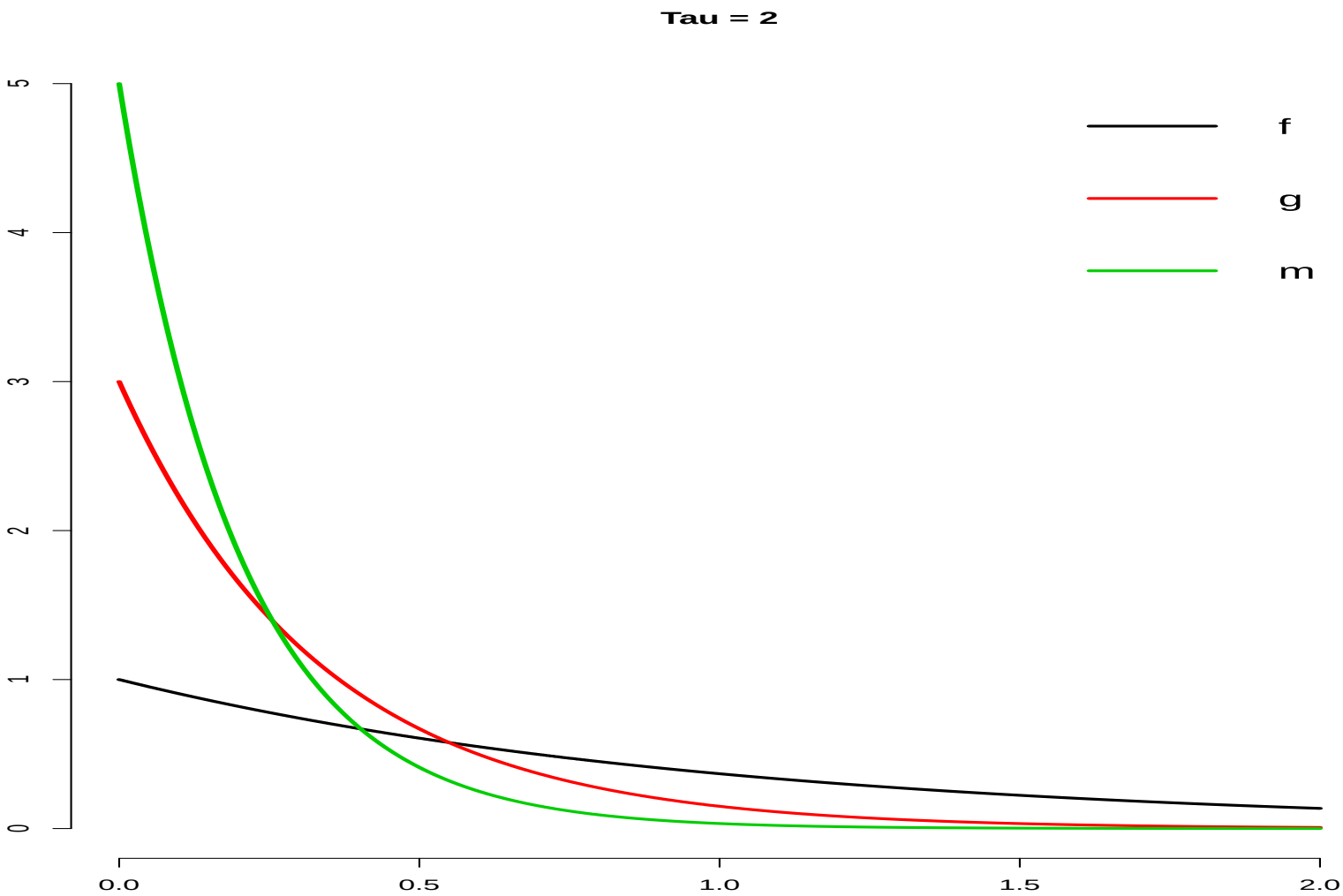
$$m(x) = \begin{cases} (\lambda + 2\tau) e^{-(\lambda+2\tau)x} & \text{se} \quad x \in [0, \infty] \\ 0 & \text{se} \quad x \notin [0, \infty] \end{cases}$$

En particular, analizaremos os casos $\lambda = 1$ e $\tau = 1, 2, 4, 8$.

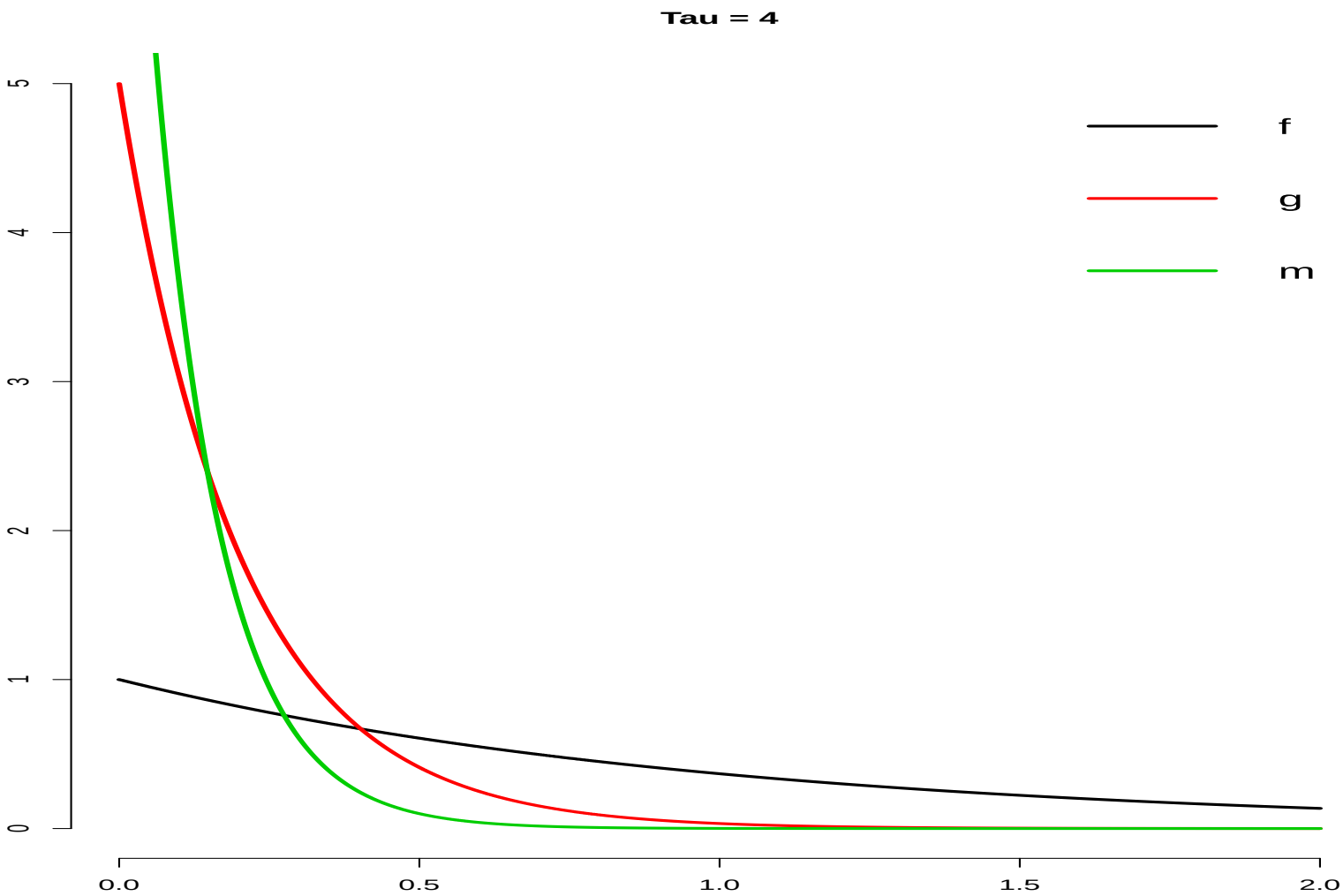
Modelo 8: densidade e densidades nesgadas



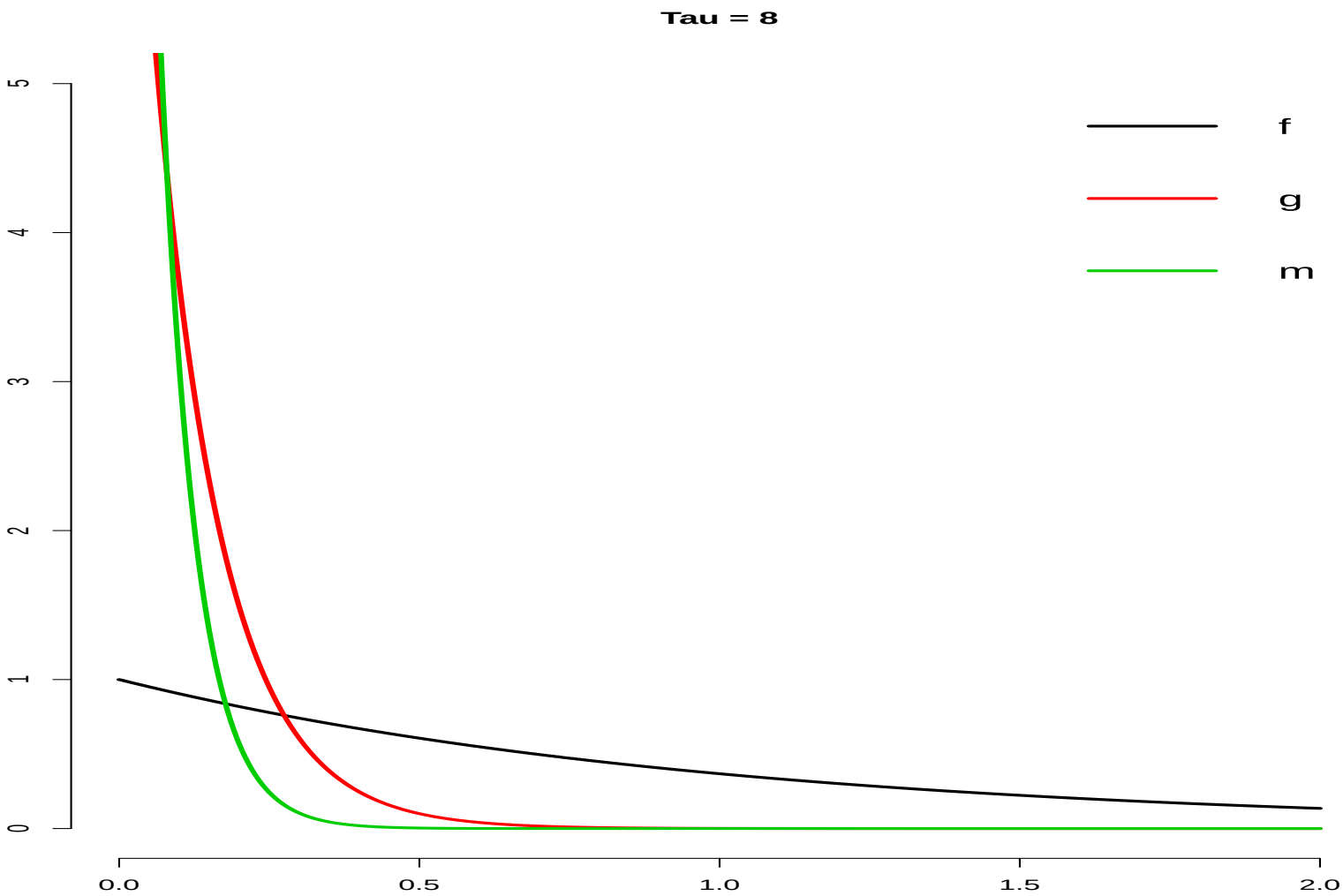
Modelo 8: densidade e densidades nesgadas



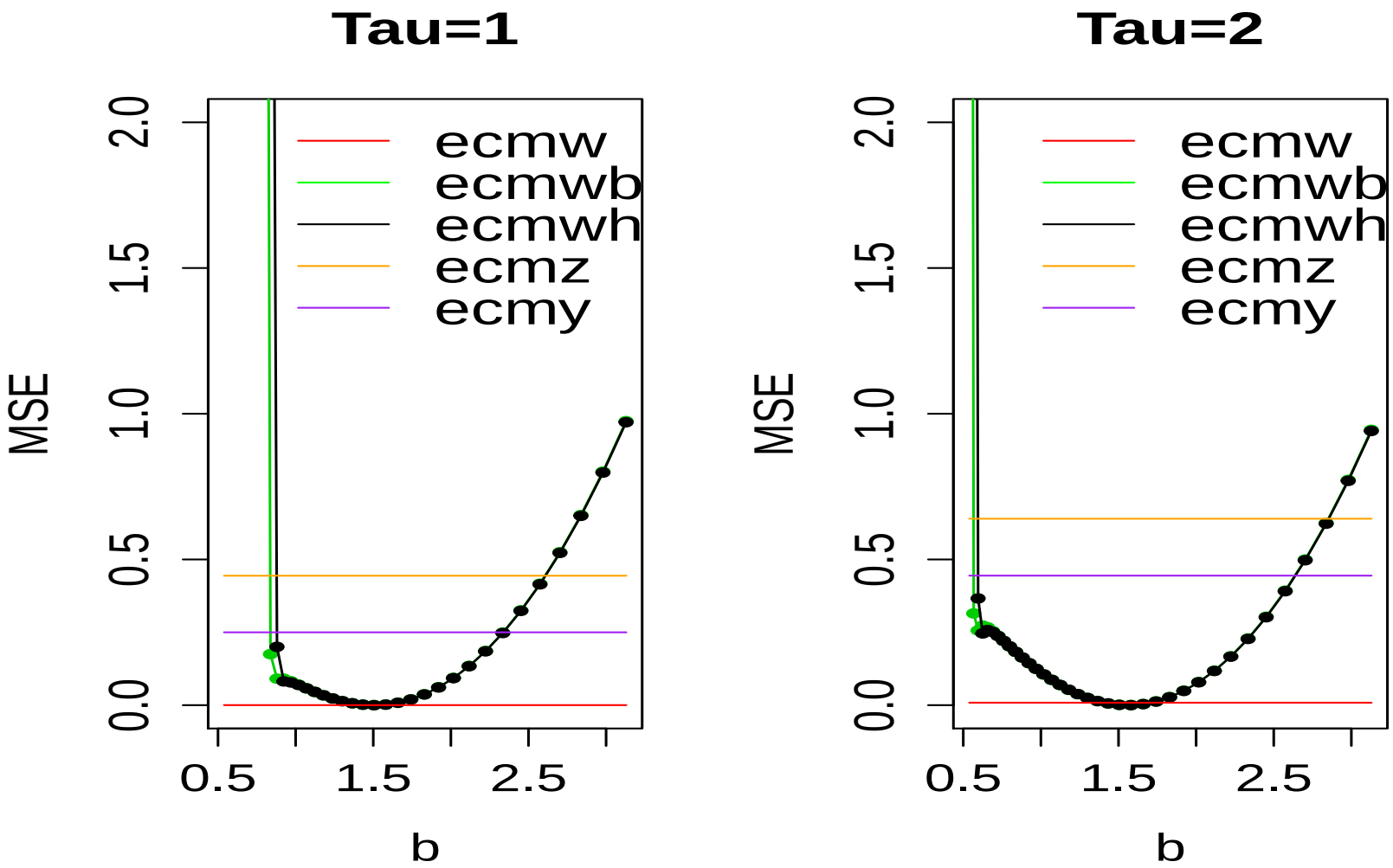
Modelo 8: densidade e densidades nesgadas



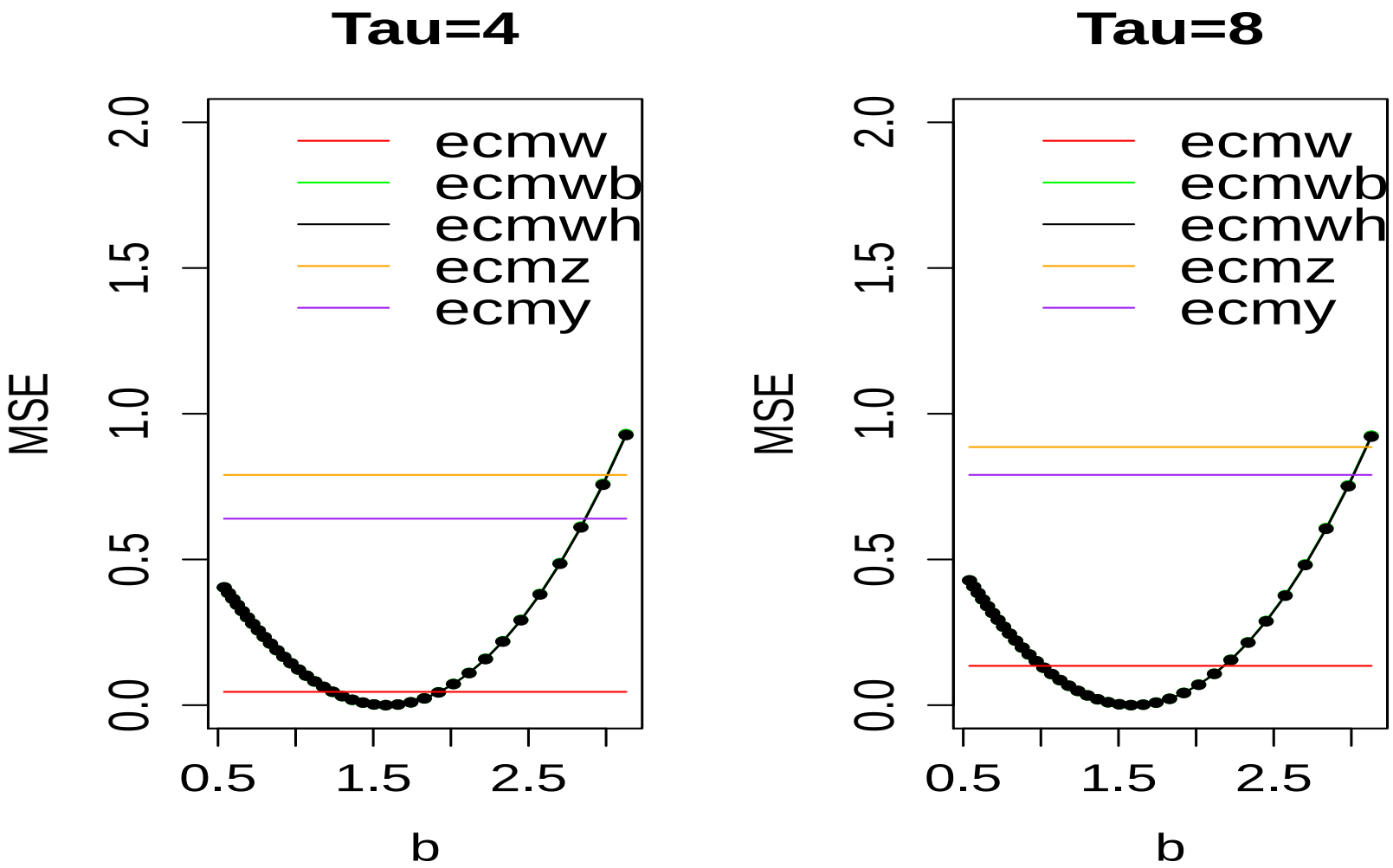
Modelo 8: densidade e densidades nesgadas



MSE Modelo 8 ($n = 1000$, $N = 1000000$, trials=1000)



MSE Modelo 8 ($n = 1000$, $N = 1000000$, trials=1000)



Necesidades computacionais

M	Proc.	RAM	n	N	Trials	$\#h$'s	$\#b$'s	t	t.total
1	$i5$	4GB	20	400	100	70	1	2h	4h
	$i7$	24GB	100	10^4	100	70	1	2h	4h
			100	10^4	10^4	40	1	86h	86h
			500	$2.5{\cdot}10^5$	10^3	30	1	14h	28h
6	$i7$	32GB	$8{\cdot}10^3$	$6.4{\cdot}10^7$	100	1	25	22h	22h
8	$i7$	32GB	10^3	10^6	10^3	1	35	14h	56h

1. Intel(R) Core(TM) i5-3337U CPU @ 1.80GHz 1.80GHz. RAM: 4GB
2. Intel(R) Core(TM) i7 CPU 950 @ 3.07GHz 3.06GHz. RAM: 24GB
3. Intel(R) Core(TM) i7-3770 CPU @ 3.40GHz 3.40GHz. RAM: 32GB

Próximos pasos...

- Método de contraste sobre a existencia de nesgo en BD
- Nun contexto B3D, que criterio seguir para decantarnos por un estimador ou outro?
- Aproximación por Bootstrap do h e do MSE cando $C_2 < 0$ nos escenarios #2 e #3
- Propiedades asintóticas no escenario #2
- Futura aplicación a datos reais dunha entidade financeira

Conclusións principais

- En BD preséntanse retos interesantes e novos obxectivos estatísticos
- O nesgo aparece con maior probabilidade en BD (B3D): e.g. mostraxe de autoselección
- En caso de posuír unha mostra aleatoria simple de tamaño pequeno as opcións oscilan entre o uso desta ou da mostra nesgada
- O nesgo pode ser correxido con algunha información adicional (e.g. mostra dúas veces nesgada)
- Resultados sobre datos nesgados por lonxitude e datos nesgados en xeral teñen cabida no contexto B3D
- As técnicas de suavizado non paramétricas poden ter un papel importante na corrección do nesgo en B3D
- Contrastar se estamos nun contexto B3D é un problema importante en Big Data
- Máis preguntas e moitas máis respostas son necesarias en B3D

Referencias


Cao, R. (2015). Inferencia estadística con datos de gran volumen. La Gaceta de la RSME, 18, 393-417.


Crawford, K. (2013). The hidden biases in big data. Harvard Business Review 2013, april 1st. Available at https://hbr.org/2013/04/the-hidden-biases-in-big-data


Hargittai, E. (2015). Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. The Annals of the American Academy of Political and Social Science, 659, 63-76.


Lloyd, C.J., Jones, M.C. (2000). Nonparametric Density Estimation from Biased Data with Unknown Biasing Function. Journal of the American Statistical Association, 95, 865-876.

Agradecementos

Esta publicación científica foi financiada polas axudas de apoio á etapa predoutoral nas universidades do Sistema universitario galego, nos organismos públicos de investigación de Galicia e noutras entidades do Sistema galego de I+D+i, cuxo financiamento procede do Fondo Social Europeo nun 80% e no 20% restante da Secretaría Xeral de Universidades, pertencente á Consellería de Cultura, Educación e Ordenación Universitaria da Xunta de Galicia, que patrocinan a investigación da segunda autora con referencia ED481A-2016/367. Ambos autores recoñecen o apoio da subvención MINECO MTM2014-52876-R (incluído o apoio do FEDER da UE).

Información de contacto

Grazas pola súa atención!

Contacto: rcao@udc.es
laura.borrajo@udc.es

2017
International Workshop
on Functional
and Operational Statistics

IWFOS