Traballo Fin de Máster

Análise e mellora da ferramenta SiZer

Adrián Martínez Rodríguez

Máster en Técnicas Estadísticas

Curso 2017-2018

Proposta de Traballo Fin de Máster

Título en galego: Análise e mellora da ferramenta SiZer

Título en español: Análisis y mejora de la herramienta SiZer

English title: Analysis and improvement of SiZer

Modalidade: Modalidade A

Autor/a: Adrián Martínez Rodríguez, Universidade de Santiago de Compostela

Director/a: Rosa María Crujeiras Casais, Universidade de Santiago de Compostela; Alberto Rodríguez Casal, Universidade de Santiago de Compostela

Breve resumo do traballo:

En 1999, Chaudhuri e Marron propoñen unha ferramenta exploratoria para detectar patróns de multimodalidade en curvas de densidade e regresión, dita ferramenta denominase SiZer. Esta constrúese dende unha perspectiva espazo-escala a partir da estimación tipo núcleo e considerando intervalos de confianza para a derivada. Na práctica, o que se obtén é unha representación nun mapa pixelado onde as cores indican crecementos/decrecementos significativos, ausencia de datos ou rexións con gradiente nulo.

Neste traballo preséntase a ferramenta SiZer aplicada a curvas de densidade, onde se introducen os catro cuantís propostos en Chaudhuri e Marron (1999) para obter os intervalos de confianza. Ademais, tamén se inclúen catro alternativas para obter un dos cuantís (q_1) a través do método bootstrap.

A implementación do SiZer coas diferentes propostas para obter os cuantís é realízase a través do software R, describindo unha implementacin numérica que permite obter o mapa de forma áxil a través da transformada rápida de Fourier.

Por último, lévase a cabo unha estrita comparación dos diferentes cuantís por medio dun estudo de simulación que permite comparar, por unha banda, os mapa SiZers obtidos a partir dos catro cuantís propostos en Chaudhuri e Marron (1999) a través do mapa promedio, o mapa de acerto, o mapa de erro, e unha nova ferramenta denominada MoSiZer; e por outra banda, realízase unha comparación da cobertura dos intervalos de confianza obtidos cos catro cuantís propostos neste traballo co cuantil q_1 introducido no artigo orixinal.

Dona Rosa María Crujeiras Casais, profesora titular da área de Estatística e Investigación Operativa (Dpto. de Estatística, Análise Matemática e Optimización) da Universidade de Santiago de Compostela, don Alberto Rodríguez Casal, profesor titular da área de Estatística e Investigación Operativa (Dpto. de Estatística, Análise Matemática e Optimización) da Universidade de Santiago de Compostela, informan que o Traballo Fin de Máster titulado

Análise e mellora da ferramenta SiZer

foi realizado baixo a súa dirección por don Adrián Martínez Rodríguez para o Máster en Técnicas Estadísticas. Estimando que o traballo está finalizado, dan a súa conformidade para a súa presentación e defensa ante un tribunal.

En Santiago de Compostela, a 3 de Xullo de 2018.

A directora:

Dona Rosa María Crujeiras Casais

Rosa M. Crujeiras Casais

Firmado digitalmente por Rosa M. Crujeiras Casais
Nombre de reconocimiento (DN): cn=Rosa M. Crujeiras Casais, o=Universidade de Santiago de Compostela, ou=Dpto. Estatística, Análise Matemática e Optimización, email=rosa.crujeiras@usc.es, c=ES
Fecha: 2018.07.02 09:48:20 +02'00'

O director:

Don Alberto Rodríguez Casal

RODRIGUEZ CASAL ALBERTO - 34896829X

Firmado digitalmente por RODRIGUEZ CASAL ALBERTO - 34896829X
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-34896829X, givenName=ALBERTO, sn=RODRIGUEZ CASAL, cn=RODRIGUEZ CASAL ALBERTO - 34896829X
Fecha: 2018.07.02 12:15:35 +02'00'

O autor:

Don Adrián Martínez Rodríguez

Agradecementos

En primeiro lugar o meu agradecemento de maneira especial e sincera aos titores que me acompañaron no transcurso deste traballo de fin de Máster, Rosa María Crujeiras e Alberto Rodríguez. Os seus consellos e confianza no meu traballo foron de vital importancia.

Tamén quero expresar o meu agradecemento de forma xeral tanto aos profesores como aos alumnos do MTE, que me acompañaron no transcurso destes dous anos de Máster, por introducirme na atmosfera da estatística.

E por último, mostrar o meu agradecemento aos meus pais, María e Ramón, e á miña irmá Cristina por animarme e apoiarme neste pequeno cambio de dirección que me levou ata a estatística.

Índice xeral

Resumo

Resumen en español

Resumo en galego

A caracterización do comportamento das variables aleatorias pode abordarse a través da estimación, a partir dunha mostra de datos, da súa función de densidade. Con este obxectivo, poden adoitarse alternativas paramétricas (nas que se supón que a densidade pertence a una certa familia, por exemplo, a normal) ou non paramétricas (onde unicamente se requiren certas condicións de regularidade). En ambos os dous enfoques, é posible obter unha aproximación completa da curva de densidade. Sen embargo, en moitas situacións o interesante é pescudar que características (por exemplo, patróns de crecemento e decrecemento significativos) están realmente presentes. Con este propósito, Chaudhuri en Marrron (1999) desenvolveron o SiZer, que se propón como unha ferramenta exploratoria para detectar patróns de multimodalidade en curvas de densidade e tamén de regresión. O SiZer constrúese a partir da estimación tipo núcleo e considerando intervalos de confianza para a derivada da curva suavizada. O que se obtén é unha representación nun mapa pixelado (no eixo horizontal, valores da variable; no eixo vertical, valores do parámetro de suavizado) onde as cores indican crecementos/decrecementos significativos, ausencia de datos ou rexións onde se pode aceptar que o gradiente é nulo.

Neste traballo preséntase a ferramenta SiZer aplicada a curvas de densidade, onde se introducen os catro cuantís propostos en Chaudhuri e Marron (1999) para obter os intervalos de confianza. Ademais, tamén se inclúen catro alternativas para obter un dos cuantís (q_1) a través do método bootstrap.

A implementación do SiZer coas diferentes propostas para obter os cuantís realízase a través do software R, describindo unha implementación numérica que permita obter o mapa de forma áxil a través da transformada rápida de Fourier.

Por último, realízase unha estrita comparación dos diferentes cuantís por medio dun estudo de simulación que permite comparar, por unha banda, os mapa SiZers obtidos a partir dos catro cuantís propostos en Chaudhuri e Marron (1999) a través do mapa promedio, o mapa de acerto, o mapa de erro, e unha nova ferramenta denominada MoSiZer; e por outra banda, realízase unha comparación da cobertura dos intervalos de confianza obtidos cos catro cuantís propostos neste traballo co cuantil q_1 proposto no artigo.

English abstract

The characterization of the behaviour of random variables can be tackled through density estimation, based on a data sample. With this objective, parametric alternatives can be adopted (assuming that the density belongs to a certain family, for instance, the Gaussian family) or non-parametric alternatives (where only some specific regularity conditions are required). In both approaches it is possible to obtain a complete approximation of the density curve. Nevertheless, in many situations it is interesting to identify which features (for example, significant increasing a decreasing patterns) are, in fact, present. For this purpose, Chaudhuri and Marron (1999) developed the SiZer, an exploratory tool used to detect multimodality patterns in density and regression curves. The SiZer is based on the

Kernel estimation considering confidence intervals for the derivative of the smooth curve. This way, it is obtained a representation in a pixelated map (variable values on the horizontal axis; values of the smooth parameter on the vertical axis) where colours indicate significant increases/decreases, absence of data or regions where it can be accepted that the gradient is nule.

In this work, the SiZer tool is presented applied to density curves, where the four quantiles proposed by Chaudhuri and Marron (1999) are introduced in order to get the confidence intervals. Furthermore, four alternatives are included as well, to approximate one of the quantile (q_1) through the bootstrap method.

The implementation of the SiZer to the different proposals in order to get the quantiles is developed using the R software, describing a numeric implementation that allows to obtain the map in a quick way through the Fast Fourier Transform.

Finally, it is carried out a strict comparison of the different quantiles through a simulation study that allows to compare, on the one hand, the SiZers maps obtained from the four quantiles proposed by Chaudhuri and Marron (1999) through the average map, the success map, the mistake map, and a new tool called MoSiZer; and, on the other hand, a comparison of the coverage of confidence intervals is made between the four quantiles proposed in this work and the q_1 quantile proposed in the original paper.

Capítulo 1

Introdución

A función de densidade de probabilidade é un concepto fundamental en estatística. Consideremos unha variable aleatoria X que teña a función de densidade f. Especificar a función f danos una descrición natural da distribución de X e permite que as probabilidades asociadas con X se encontren a partir da relación

$$P(a < X < b) = \int_a^b f(x)dx, \quad \forall a < b.$$

Supoñamos agora, que temos un conxunto de datos observados que se supón que son unha mostra dunha función de densidade descoñecida, é dicir $\boldsymbol{X} = (X_1, \dots, X_n)$ unha m.a.s. de X con función de densidade f descoñecida. Para a estimación da densidade a partir de datos observados, poden adoitarse dous enfoques:

- Un primeiro no que a estimación da densidade é paramétrico, o cal consiste en considerar que a función de densidade que desexamos estimar pertence a unha determinada clase de funcións paramétricas, por exemplo a algunhas das clásicas distribucións: normal, exponencial, Poisson, etc. Dita suposición usualmente baséase en informacións sobre a variable que son externas á mostra, pero cuxa validez pode ser comprobada con posterioridade mediante probas de bondade de axuste. Baixo esta suposición a estimación redúcese a determinar o valor dos parámetros do modelo a partir da mostra. Por exemplo, se asumimos que $\boldsymbol{X} = (X_1, \dots, X_n)$ unha m.a.s. de X con función de densidade f normal con media μ e varianza σ^2 descoñecidos, a densidade f podería entón estimarse obtendo as estimacións de μ e σ^2 a partires dos datos e substituíndo esas estimacións na fórmula da densidade normal. Esta estimación denomínase estimación paramétrica da densidade.

- Na alternativa as suposicións acerca da distribución dos datos observados son menos ríxidas, xa que consisten en non predeterminar a priori ningún modelo para a distribución de probabilidade da variable e deixar que a función de densidade poida adoptar calquera forma, sen máis límites que os impostos polas propiedades que se esixen ás funcións de densidade para ser consideradas como tales, é dicir, que a estimación non sexa nunca negativa e que a súa integral sexa un

  $$\hat{f}(x) \geq 0, \qquad \int_{\mathbb{R}} \hat{f}(x)dx = 1.$$

  Este enfoque, no que se centra a ferramenta SiZer, é o que se denomina estimación non paramétrica da densidade, e ten unha das orixes mais comunmente aceptados nos traballos de Fix e Hodges (1951), sendo o estimador proposto unha versión xeral do que é hoxe coñecido como estimador naïve. De certa maneira o enfoque non paramétrico permite que os datos determinen de forma totalmente libre, sen restricións, a forma da densidade que os ha de representar.

A estimación de curvas suaves é útil para obter información dos datos en estatística, tales como a asimetría ou a multimodalidade. Existe una extensa bibliografía científica que trata o problema da estimación de densidades dende diversos puntos de vista, pero unha das ferramentas máis utilizadas neste campo son os estimadores tipo núcleo (método kernel), propostos por Parzen (1962) e Rosenblatt (1956).

Dada $\boldsymbol{X} = (X_1, \dots, X_n)$ unha m.a.s. de X con función de densidade f descoñecida, definimos a estimación tipo núcleo da densidade con función núcleo K como

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i), \tag{1.1}$$

onde $K_h(u) = \frac{1}{h} K(\frac{u}{h})$, sendo $K(x)$ unha función, denominada función núcleo, que satisfai certas condicións de regularidade, sendo xeralmente unha función de densidade simétrica como por exemplo a distribución normal (a cal será empregada ao longo de todo este traballo), e h é o parámetro de suavizado, denominado comunmente parámetro ventá.

Intuitivamente, o método kernel consiste en colocar un núcleo (unha densidade de probabilidade) sobre cada punto de observación na mostra. A densidade estimada en cada intersección é esencialmente o promedio das densidades de todos os núcleos que se superpoñen a ese punto. As observacións que están cerca dun punto de avaliación contribuirán máis á estimación que aquelas que están lonxe de selo. Polo tanto, a estimación da densidade será alta en áreas con moitas observacións, e baixa en áreas con poucas.

Na Figura 1.1 podemos observar unha estimación tipo núcleo da densidade mostrando os núcleos individuais.

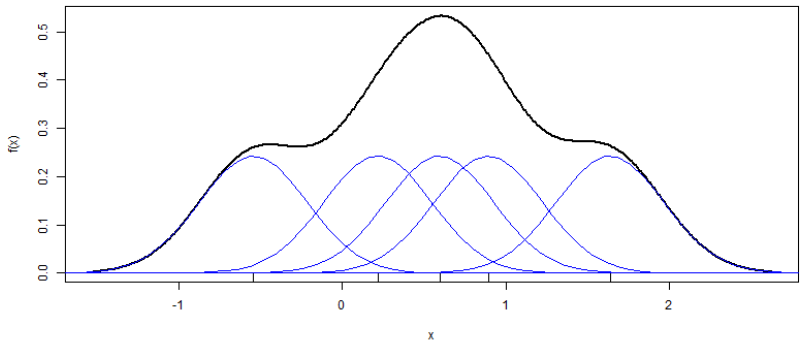

Figura 1.1: Estimación tipo núcleo da función de densidade de $\boldsymbol{X} = (X_1, \dots, X_5)$ unha m.a.s. de X con función de densidade f normal estandarizada.

Moitos traballos foron realizados para escoller de forma idónea o parámetro de suavizado a partir da mostra $\boldsymbol{X}$, e moitas propostas foron realizadas para facer inferencia a partires dos intervalos de confianza de $\hat{f}_h(x)$. O problema principal é que o estimador tipo núcleo da densidade sofre de forma inherente dun nesgo difícil de tratar

$$\mathbb{E}(\hat{f}_h(x)) - f(x) = \frac{1}{2} h^2 \mu_2(K) f''(x) + o(h^2), \tag{1.2}$$

sendo $\mu_2(K) = \int u^2 K(u) du < \infty$. E ademais, ocorre o mesmo coa varianza do estimador

$$Var(\hat{f}_h(x)) = \frac{1}{nh} R(K) f(x) + o((nh)^{-1}), \tag{1.3}$$

con $R(K) = \int K^2(x)dx < \infty$. Das expresións do nesgo e a varianza pódense deducir varios resultados. En primeiro lugar, o nesgo nun punto x depende directamente do valor da segunda derivada da función de densidade $f''(x)$. Isto significa que a estimación tipo núcleo fai un axuste peor nos puntos máximos e mínimos da curva e nas zonas próximas a estes. En segundo lugar, obsérvase como o nesgo está directamente relacionado co valor da ventá h e a varianza, sen embargo, está inversamente relacionada con dito valor. Isto implica que cando se expón calcular h é necesario encontrar un valor de equilibrio, que minimice o nesgo e a varianza de forma conxunta.

Existen diversas estratexias de estimación do parámetro de suavizado h para as cales é necesario dispoñer de mecanismos que midan a bondade de axuste dos mesmos, comunmente denominados criterios de erro. Hai varias posibilidades para medir este axuste, como o Erro Cadrático Integrado (ISE; *Integrated Squared Error*) que se define para un estimador da densidade $\hat{f}$ como

$$ISE(\hat{f}) = \int (\hat{f}(x) - f(x))^2 dx.$$

Este criterio depende da mostra de datos, e polo tanto estase introducindo una certa variabilidade intrínseca á propia mostra e non ao estimador. É empregado polo método da validación cruzada proposto por Bowman (1984).

Outro criterio de erro é o Erro Cadrático Medio Integrado (MISE; *Mean Integrated Squared Error*) que suprime a aleatoriedade procedente de cada mostra individual promediando os resultados obtidos para varias mostras:

$$MISE(\hat{f}) = \int \mathbb{E}(\hat{f}(x) - f(x))^2 dx.$$

O MISE é un criterio de erro global que non depende da mostra empregada. É empregado polo método bootstrap proposto por Taylor (1989). O bootstrap é un procedemento estatístico que serve para aproximar a distribución na mostraxe (normalmente dun estatístico). Para iso procede mediante remostraxe, é dicir, obtendo mostras mediante algún procedemento aleatorio que utilice a mostra orixinal. O autor oficial deste método foi Efron (1979), o cal unificou a potencia do método de Monte Carlo coa resolución de problemas expostos de forma moi xeral.

Normalmente para o estimador tipo núcleo da función de densidade (1.1) non se pode obter una expresión exacta do MISE. Facendo desenvolvementos de Taylor e baixo certas condicións de regularidade sobre f obtense unha aproximación asintótica cuxa expresión ven dada por:

$$AMISE(\hat{f}_{nh}) = \frac{1}{nh} R(K) + \frac{1}{4} h^4 \mu_2(K)^2 R(f''),$$

onde R é unha aplicación que asigna a calquera función a integral do seu cadrado, esto é $R(g) = \int g^2(x)dx$, e AMISE denota o denominado *Asympthotic Mean Integrated Squared Error* que é a parte do MISE que coñecemos de maneira exacta. Este criterio é empregado polo método da regra do pulgar proposto por Silverman (1986) e o método plug-in proposto por Sheather e Jones (1991).

A Figura 1.2 mostra un exemplo da estimación non paramétrica da densidade, onde o obxectivo principal e estimar a densidade f, que supoñeremos a priori descoñecida (no Capítulo 2 mostrarase a procedencia destes datos), que revela a estrutura da mostra $X_1, ..., X_{50}$. A estimación realízase facendo uso dalgún dos métodos mencionados anteriormente para obter o tamaño do parámetro ventá axeitado. En primeiro lugar, o tamaño do parámetro de suavizado obtido a través do método da regra do pulgar é $h = 0.463$, e obsérvase claramente como a estrutura da estimación da función de densidade obtida para este método mostra dúas modas. Posteriormente, a través do método plug-in, para o cal o tamaño de ventá obtido para esta mostra é de $h = 0.270$, obsérvase como a estimación tipo núcleo da densidade revela tres modas. E por último, a través do método da validación cruzada, o cal tende a obter un parámetro de suavizado menor aos demais métodos, obtén un tamaño de ventá de $h = 0.209$, o cal revela unha estrutura da estimación da función de densidade con ata catro modas. A pregunta é, que modas están realmente presentes? Como vemos, incluso con métodos empregados habitualmente na práctica para realizar a estimación tipo núcleo da densidade, obtemos estruturas diferentes que

revelan un número de modas total diferente para cada un dos métodos. Pero unha ferramenta que nos permite responder a esta pregunta é o mapa SiZer (*SIgnificant ZERo crossings*), o cal será obxecto de estudo no presente traballo. Esta metodoloxía consiste en estudar simultaneamente un rango amplo de parámetros de suavizado (parámetros ventá), evitando a clásica necesidade de elexir o parámetro de ventá axeitado. A ferramenta foi proposta por Chaudhuri e Marron (1999), e permite en lugar de superpoñer curvas suavizadas para un rango de valores do parámetro ventá h no mesmo diagrama, representar os resultados empregando o concepto de espazo-escala, usando un mapa de cores.

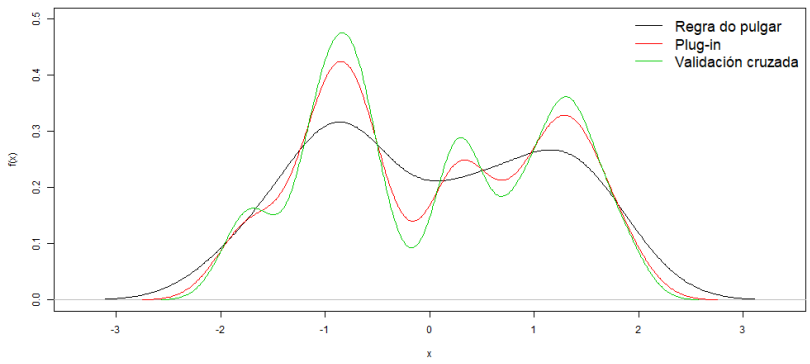

Figura 1.2: Estimación tipo núcleo da función de densidade de $\boldsymbol{X} = (X_1, \ldots, X_{50})$ unha m.a.s. de X con función de densidade f a priori descoñecida.

A ferramenta SiZer baséase en estudar os intervalos de confianza da estimación tipo núcleo da da derivada da función de densidade $\hat{f}_h'(x) = \frac{1}{n}\sum_{i=1}^{n} K_h'(x - X_i)$ para un rango amplo de valores do parámetro de suavizado h. Como veremos, Chaudhuri e Marron (1999) propoñen catro formas de obter o cuantil para estimar os intervalos de confianza: o cuantil independente Gaussiano para cada punto x (q_1), unha aproximación simultánea sobre x do cuantil Gaussiano baseado no número de bloques independentes sobre os datos (q_2), e dous métodos baseados en bootstrap, un deles simultáneo sobre x (q_3) e o outro simultáneo sobre ambos, x e h (q_4). O SiZer e estes cuantís serán expostos con mais detalle no Capítulo 2.

Con isto, o obxectivo deste traballo é facer unha ampla introdución da ferramenta SiZer, onde expoñeremos as principais diferenzas dos cuantís propostos por Chaudhuri e Marron (1999), e ademais, serán propostos catro cuantís máis baseados en bootstrap para tentar mellorar a taxa de cobertura dos intervalos de confianza obtidos co cuantil q_1, baseado no cuantil Gaussiano para cada punto (x, h) do SiZer de forma independente. A implementación do SiZer coas diferentes propostas dos cuantís é levada a cabo sobre o software R, polo que outro dos obxectivos é describir a implementación numérica desta ferramenta a través da transformada rápida de Fourier (FFT; do inglés *Fast Fourier Transform*). Pero o principal propósito deste traballo é levar a cabo unha estrita comparación dos diferentes cuantís por medio dun estudo de simulación que nos permita comparar os mapa SiZers obtidos por cada un dos cuantís propostos por Chaudhuri e Marron (1999) dunha forma obxectiva. Para iso faremos uso de ferramentas propostas neste traballo, como son o mapa SiZer promedio ou o mapa de acerto, os cales nos deixan ver dunha forma clara en que zonas do mapa se están a identificar de forma correcta patróns de crecemento/decrecemento nos SiZers obtidos con cada cuantil; o mapa de erro, co cal veremos a taxa de erro e as zonas onde se está a cometer; e por último a ferramenta MoSiZer (*Mode SiZer*), a cal nos permitirá visualizar dunha forma mais clara a localización e número de modas estimadas polo mapa SiZer. Por último, outro obxectivo deste traballo é facer un estudo de simulación para ver se realmente os cuantís propostos baseados en bootstrap ofrecen unha mellora na taxa de cobertura dos intervalos de confianza con respecto aos intervalos de confianza obtidos por medio do cuantil Gaussiano q_1.

A memoria do traballo fin de Máster organízase da seguinte maneira:

- No Capítulo 2 inicialmente faise unha introdución da ferramenta SiZer, onde se expón a través do exemplo empregado nesta Sección como debe ser a interpretación do mapa. Posteriormente son desenvolvidos os catro cuantís propostos por Chaudhuri e Marron (1999) para a obtención do mapa SiZer. Como dixemos, esta ferramenta fai inferencia sobre os intervalos de confianza da estimación tipo núcleo da derivada da función de densidade para un amplo rango de valores do parámetro de suavizado, e o primeiro dos métodos propostos por Chaudhuri e Marron (1999) obtén os intervalos de confianza, para un nivel de significación dado, a partir dos cuantís dunha densidade normal. Para tamaños de mostra reducidos a utilización destes cuantís pode resultar pouco axeitado, polo que ademais serán introducidos catro novos métodos bootstrap máis (o método percentil, o método percentil simetrizado, o método percentil-t e o método percentil t-simetrizado) que presentan unha posible alternativa deste método. Por último, desenvólvese a implementación numérica que nos permitirá no Capítulo 3 obter o mapa SiZer sobre o software R dun número elevado de mostras nos estudos de simulación levados a cabo.

- No Capítulo 3 lévanse a cabo diferentes estudos de simulación. En primeiro lugar, faise unha comparación dos resultados obtidos polos catro métodos propostos por Chaudhuri e Marron (1999) a través dun mapa promedio do SiZer, dun mapa que permite visualizar o acerto de cada un dos métodos, e dun mapa que permite visualizar o erro. Posteriormente introdúcese unha nova ferramenta, o MoSiZer. Esta permite visualizar de forma mais clara e directa as zonas onde para un número elevado de mostras se están a obter as modas sobre a estimación do mapa SiZer. Ademais, tamén se realiza un estudo de simulación no que se analiza a cobertura dos intervalos de confianza obtidos para todo o mapa SiZer a partires do primeiro dos métodos propostos por Chaudhuri e Marron (1999), onde entran en xogo os os catro novos métodos bootstrap para estudar cal está a ofrecer unha taxa de cobertura mais próxima ao nivel $(1-\alpha)$, sendo α o nivel de significación.

- No Capítulo 4 expóñense as conclusións obtidas neste traballo, centradas especialmente no estudo de simulación levado a cabo.

- No Apéndice A son introducidos os intervalos de confianza convencionais da estimación tipo núcleo da función de densidade, e ademais, desenvólvese a necesidade de empregar a ferramenta SiZer para mostrar as características (modas e vales) da función de densidade f a través dunha mostra $\boldsymbol{X} = X_1, \ldots, X_n$, que poden chegar a ser non visibles a través da estimación tipo núcleo da función de densidade.

- No Apéndice B introdúcense as 15 funcións de densidade descritas por Marron e Wand (1992). Estas densidades están compostas por mestura de normais e convertéronse nun estándar na estimación da densidade. Ademais, xunto a representación da función de densidade de cada unha delas, introdúcese o mapa SiZer da función $f_h = K_h * f$, é dicir, o mapa SiZer da convolución da función núcleo e da función de densidade. Por outra banda, tamén se representa o MoSiZer obtido a través do mapa SiZer agora mencionado.

- Por último, no Apéndice C expóñense resultados do mapa promedio, do mapa de acerto, e do mapa de erro obtidos a partires do estudo de simulación levado a cabo no Capítulo 3 con mostras xeradas coas densidades de Marron e Wand (1992).

Capítulo 2

Ferramenta SiZer

A estimación tipo núcleo da densidade carrexa a problemática da selección da fiestra. Pero en moitos casos, o obxectivo da reconstrución da verdadeira densidade é a identificación de patróns de crecemento/decrecemento, que poden variar dependendo do parámetro de suavizado. Neste contexto, a ferramenta SiZer formúlase dende unha perspectiva diferente, eliminando a idea clásica de buscar un único parámetro de suavizado óptimo dende o que se constrúe todo o proceso inferencial. A motivación básica do SiZer vén da idea de que, cando se elixe un método de suavizado para aproximar una curva en base a un conxunto de datos, quedarse cun só nivel de suavizado supón prescindir de moita información que hai presente nos datos, de feito nas distintas versións suavizadas da curva obxecto de estimación está contida toda a información recollida nos datos. Neste senso Marron e Chung (2001) introducen a idea da familia de suavizadores (*the family approach to smoothing*), como ferramenta exploratoria que permite visualizar os datos dende todos eses niveles de suavizado, xa que moita información útil pode estar dispoñible a diferentes niveles de suavizado. Trátase de analizar o problema inferencial para o conxunto de datos dispoñibles, mirando a todos os membros dunha familia de suavizadores. Os membros da familia corresponden a un tipo particular de suavizadores para todo un rango de parámetros de suavizado. De este modo evítase o problema clásico de elexir o parámetro de suavizado.

A metodoloxía do SiZer está motivada pola idea de localización e escala da visión artificial (Lindeberg 1994). Primeiro, estúdase de forma simultánea un rango amplo de parámetros de suavizado h sobre a estimación tipo núcleo da función de densidade. E en segundo lugar, evítase o problema do nesgo (1.2) e a varianza (1.3) da estimación tipo núcleo da función de densidade (1.1) ao facer inferencia cambiando o enfoque da verdadeira curva subxacente f pola curva suavizada f_h, vista a distintos niveles de resolución como se explica na seguinte Sección.

Isto, permítenos ver resultados de localización e escala a través dun mapa de cores como mostra a Figura 2.1. Un mapa bidimensional no que se representa os valores da localización no eixe x e os valores do parámetro de suavizado no eixe y. Os valores de h son representados en escala logarítmica para obter unha visualización mais precisa. Destaca as zonas relevantes, como son as modas, representando as zonas significativamente crecentes e decrecentes. Cabe mencionar que as modas se atopan onde a derivada ten valor cero entre rexións significativamente crecentes e decrecentes. O mapa mostra en cor azul (vermello) as zonas onde a curva é significativamente crecente (significativamente decrecente), e a cor intermedia púrpura emprégase para representar onde a curva non ten evidencias suficientes para afirmar que é crecente ou decrecente. Por último, a cor gris indica as rexións onde os datos son demasiados escasos para obter conclusións da significación, debido a que hai poucos puntos baixo cada ventá. Con isto, o mapa SiZer deixa claro que estruturas vistas na estimación da densidade son estatísticamente signficativas e cales poden ser descartadas debido a variabilidade natural.

Na Figura 2.1 móstrase o mapa SiZer obtido para a mostra $\boldsymbol{X} = (X_1, \ldots, X_{50})$, para a cal se representou a estimación tipo núcleo da función de densidade a través de diferentes parámetros de suavizado sobre a Figura 1.2. No mapa podemos ver que para niveis grosos de resolución (parámetro de ventá grande), a curva estimada é significativamente crecente (cor azul), para posteriormente ser

significativamente decrecente (cor vermella), dando a entender que estas características están realmente aí para este nivel de resolución. Para parámetros de suavizado mais reducidos (próximos ao nivel $\log_{10}(h) = -1.6$), aparentemente dúas modas son mostradas sendo estatísticamente significativas, xa que a cor do mapa a ese nivel permuta de azul (↑) a vermello (↓), logo de novo a azul (↑), para finalmente retomar de novo a cor vermella (↓).

Con isto, o SiZer danos unha resposta consistente acerca das estimacións realizadas previamente na Figura 1.2. O mapa suxire que a estimación tipo núcleo da función de densidade realizada co parámetro ventá obtido a través do método plug-in está a estimar unha moda ficticia, é dicir, a moda central estimada por medio deste método non é significativa. Por outra banda o número de modas ficticias estimadas a través do método da validación cruzada ascende a dúas. De novo, a mesma moda central estimada a través do método plug-in aparece (é lóxico xa que o parámetro de suavizado é inferior) e polo tanto non é significativa tal e como suxire o mapa SiZer. Ademais, a outra moda que non parece ser significativa tal e como mostra o mapa é a moda situada entorno ao punto $x = -1.8$. O SiZer mostra que unicamente son significativas dúas modas, e estas parecen coincidir coas modas estimadas a través da estimación tipo núcleo da función de densidade co método da regra do pulgar tal e como mostra a Figura 1.2. Por último, cabe mencionar que a cor gris situada na zona inferior do mapa, é debida á escaseza de observacións nesas zonas como mencionamos anteriormente.

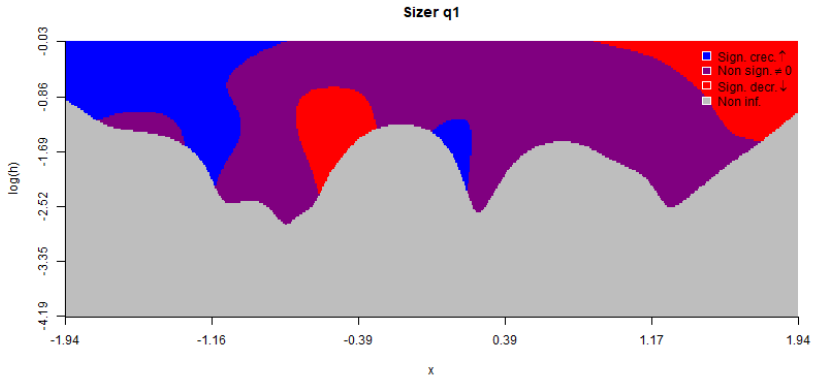

Figura 2.1: Estimación tipo núcleo da función de densidade de $\boldsymbol{X} = (X_1, \ldots, X_{50})$ (empregada na Figura 1.2) unha m.a.s. de X con función de densidade f a priori descoñecida.

A orixe da mostra $\boldsymbol{X} = (X_1, \ldots, X_{50})$ empregada para estimación tipo núcleo da función de densidade e a estimación do mapa SiZer das Figuras 1.2 e 2.1 respectivamente, provén dunha variable aleatoria cuxa función de densidade é a *Bimodal* (#6) proposta por Marron e Wand (1992). Ao longo deste traballo estas densidades serán empregadas como obxecto para realizar o estudo sobre a ferramenta SiZer. Como mencionamos anteriormente, as densidades de Marron e Wand son un estándar no estudo da estimación da función de densidade, e polo tanto resultan idóneas para empregalas na análise do SiZer. No Apéndice B móstranse estas densidades, onde ademais se representa o mapa SiZer obtido a partires da convolución do núcleo e da función de densidade,

$$f_h = K_h * f, \tag{2.1}$$

e por outra banda o MoSiZer, ferramenta que empregaremos no seguinte capítulo e que nos permite ver de forma mais clara a localización das modas obtidas a través do mapa SiZer.

Se observamos a gráfica esquerda da Figura B.6 do Apéndice B podemos ver que a densidade *Bimodal*, como o seu propio nome indica, conta con dúas modas. Isto revela que a estimación realizada polo mapa SiZer obtido na Figura 2.1 a cal apunta que a función de densidade da variable aleatoria X ten dúas modas é correcto. Como vimos na Figura 1.2, a pesar de que os tres métodos empregados para obter o parámetro de suavizado son correctos (e habitualmente empregados na práctica), as estimacións

tipo núcleo da función de densidade mostran un número de modas diferente para cada un deles, sendo unicamente correctas as dúas modas obtidas a través do método da regra do pulgar. Isto revela unha maior consistencia da ferramenta SiZer para a detección e localización das modas que a estimación tipo núcleo da función de densidade.

2.1. Localización e escala

O obxectivo da análise da localización e escala é descubrir as características sobresaíntes de un obxecto de interese que aparecen a diferentes escalas. O obxectivo considerado pode ser, por exemplo, unha serie temporal ou unha imaxe dixital, no que nese caso as características búscanse de forma correspondente en diferentes escalas temporais ou espaciais. A orixe da metodoloxía de localización e escala recae sobre Witkin (1983), pero esta metodoloxía comezou a emerxer no campo da estatística na detección de modas sobre a estimación da función de densidade univariante e bivariante (Minnotte e Scott, 1993).

A familia de estimadores tipo núcleo indexados polo parámetro de suavizado h, son modelos empregados habitualmente na visión artificial. A idea esencial é que parámetros ventá grandes modelan distancias de visión macroscópicas onde só características a gran escala poden ser resoltas, mentres que parámetros de suavizado pequenos modelan resolucións microscópicas de características a pequena escala. En particular, para unha función dada f, unha sinal ruidosa (presente en calquera sistema de visualización), represéntase mediante a convolución $K_h * f$ para diferentes valores do parámetro de suavizado h. De feito, esta familia de convolucións é o centro da análise, coa idea de que esa é toda a información dispoñible dunha cantidade de datos en presenza de ruído (ver mais en Chaudhuri e Marron 1997; Lindeberg 1994). É moi diferente do enfoque estatístico, onde o foco se centra en f.

Exemplos de características na estimación da densidade inclúe modas e vales, que poden ser caracterizados de diferente maneiras, pero o SiZer céntrase no cruce por cero da estimación tipo núcleo da derivada da función densidade. Dicimos que o valor da derivada f'_h e significativamente distinta de cero cando o intervalo de confianza de $\hat{f}'_h$ non contén o valor 0.

Analizando a estimación dos intervalos de confianza da derivada da densidade para un rango amplo de parámetros de suavizado, veuse demostrado que o núcleo Gaussiano $K(x) = (1/\sqrt{2\pi})e^{-x/2}$ ten importantes vantaxes sobre outros núcleos. En particular, o número de cruces por cero é sempre decrecente en función de h (o que non é verdadeiro para outros núcleos usados para a estimación tipo núcleo da densidade). É dicir, só o núcleo Gaussiano respecta a monotonicidade das características da densidade con respecto ao parámetro de suavizado; ver Silverman (1981) e Chaudhuri e Marron (1997). Polo tanto, só o núcleo Gaussino é utilizado no SiZer.

Como diximos, o obxectivo do SiZer é debuxar o mapa de cores como o mostrado na Figura 2.1. Este mapa, que pode ser empregado para a análise exploratoria de datos mostra as rexións de localización e escala (con respecto a os parámetros x e h respectivamente) onde a derivada e significativamente crecente ou decrecente. Como se discute no Apéndice A o enfoque clásico da significación das características baseado nos intervalos de confianza son demasiado conservadores para facer inferencia, ou incluso inválidos debido aos problemas do nesgo. É aquí onde se presenta o aspecto novo do SiZer, adoptando o punto de vista da localización e escala, para este problema do nesgo. En lugar de estudar os intervalos de confianza de f', buscamos os intervalos de confianza de f'_h. Desta forma, o centro dos intervalos estímase de forma correcta, e a varianza estímase de forma eficiente e sinxela como veremos mais tarde. Isto implica, que a significación de calquera característica depende da escala (parámetro ventá h) debe ser interpretada nese sentido. A Figura 2.1 mostra que a estrutura bimodal está presente nalgunhas zonas do mapa, pero esta desaparece para niveis de suavizado altos (Hai só unha moda para parámetros venta grandes).

Por último, cabe mencionar que o enfoque é diferente aos tests de unimodalidade/multimodalidade, onde se trata de contrastar o número de modas, dando unha significación do mesmo. Exemplos disto encóntranse en Silverman (1981), Fisher e Marron (2001) ou Ameijeiras et al. (2016). O SiZer non só indica o número de modas para diferentes niveis de resolución, se non que é unha ferramenta

informativa da localización das mesmas.

2.2. Construción do SiZer

Nesta Sección introdúcese de forma precisa a notación e construción da ferramenta SiZer.

O enfoque para o estudo de características como as modas e vales na familia das curvas estimadas $\{\hat{f}_h(x) : h \in [h_{\text{mín}}, h_{\text{máx}}]\}$ (a selección de $h_{\text{mín}}$ e $h_{\text{máx}}$ serán discutidas posteriormente) baséase no estudo dos intervalos de confianza da estimación tipo núcleo da derivada da función de densidade,

$$\hat{f}'_h(x) = \frac{1}{nh^2}\sum_{i=1}^{n} K'\left(\frac{x - X_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K'_h(x - X_i).$$

Como mencionamos anteriormente, o resultado da estimación sobre x e h preséntase a través dun mapa de cores SiZer, onde a cor azul indica as localización onde $\hat{f}'_h$ é significativamente positivo, a cor vermella onde $\hat{f}'_h$ é significativamente negativo, e a cor púrpura onde non hai significación de que $\hat{f}'_h$ sexa distinto de 0.

Os intervalos de confianza son da seguinte forma

$$\hat{f}'_h(x) \pm q \cdot \widehat{SD}(\hat{f}'_h(x)), \tag{2.2}$$

onde q é o cuantil apropiado e $\widehat{SD}(\hat{f}'_h(x))$ é a raíz cadrada da estimación da varianza de $\hat{f}'_h(x)$,

$$\widehat{var}(\hat{f}'_h(x)) = \widehat{var}\left(\frac{1}{n}\sum_{i=1}^{n} K'_h(x - X_i)\right) = \frac{1}{n}s^2(K'_h(x - X_1), ..., K'_h(x - X_n)).$$

A curva é significativamente crecente, decrecente ou non é significativo de supoñer que é distinto de 0 nun punto (x, h), cando o valor 0 se encontra por debaixo, por arriba, ou dentro dos límites do intervalo de confianza de $\hat{f}'_h(x)$ respectivamente.

2.2.1. Cuantís clásicos

En Chaudhuri e Marron (1999), catro alternativas para determinar q foron propostas: o cuantil independente Gaussiano para cada punto x, unha aproximación simultánea sobre x do cuantil Gaussiano baseado no número de bloques independentes sobre os datos, e dous métodos baseados en bootstrap, un deles simultáneo sobre x e o outro simultáneo sobre ambos, x e h.

Bootstrap

Supoñamos que temos $\boldsymbol{X} = (X_1, ..., X_n)$ unha m.a.s. de X con función de distribución F descoñecida, e desexamos facer inferencia sobre un estatístico $\theta = \theta(F)$. Para isto, gustaríanos coñecer a distribución de $R(\boldsymbol{X}, F)$, certo estatístico función da mostra e da distribución poboacional. Por exemplo $R = R(\boldsymbol{X}, F) = \theta(\hat{F}) - \theta(F)$. Ás veces podemos calcular dita distribución, aínda que soe depender de cantidades poboacionais non coñecidas na práctica, mentres que outras veces só podemos chegar a aproximala cando $n \longrightarrow \infty$.

O análogo bootstrap consiste en substituír a distribución poboacional (descoñecida) F por unha estimación $\hat{F}$ da mesma. Para iso, obtemos condicionalmente á mostra observada, remostras tal que $\boldsymbol{X}^* = (X_1^*, ... X_n)$ con distribución $\hat{F}$. A partir das remostras obtemos a distribución na remostraxe $R^* = R(\boldsymbol{X}^*, \hat{F})$, chamada distribución bootstrap. E isto permítenos aproximar a distribución na mostraxe de R pola distribución bootstrap R^*. En raras ocasións a distribución bootstrap é calculable directamente, pero sempre pode ser aproximada mediante o método de Monte Carlo.

Habitualmente empréganse catro métodos para obter as remostras bootstrap:

- O primeiro deles é o bootstrap uniforme, tamén chamado bootstrap naïve, e é aquel no que se substitúe a distribución poboacional (descoñecida) F pola distribución empírica

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}. \tag{2.3}$$

  É dicir, $\hat{F} = F_n$, e polo tanto $R^* = R(\boldsymbol{X}^*, F_n)$.

- Supoñamos que sabemos que a función de distribución poboacional pertence a unha certa familia paramétrica. É dicir $F = F_\theta$ para algún vector d-dimensional $\theta \in \Theta$. Neste caso, estimaríase θ a partir da mostra, e con iso as remostras serían obtidas a partir da estimación da función de distribución $F_{\hat{\theta}}$ e non de F_n. Este método é o bootstrap paramétrico.

  Polo tanto, $\hat{F} = F_{\hat{\theta}}$, e polo tanto $R^* = R(\boldsymbol{X}^*, F_{\hat{\theta}})$.

- Por outro lado, pode ocorrer que coñezamos que a función de distribución poboacional é simétrica entorno a certo valor, polo que sería aconsellable empregar o bootstrap simetrizado. Iso significa que existe un valor c tal que $F(c-a) = 1-F(c+a)$ para todo $a > 0$. Pode demostrarse que dito centro de simetría c ha de ser a media da distribución. Así, para estimar a función de distribución poboacional F, é razoable utilizar unha versión simetrizada da distribución empírica F_n^{sim}. Ese estimador empírico outorga igual masa de probabilidade a unha mostra artificialmente construída simetrizando a mostra orixinal:

$$Y_i = \{X_1, \ldots, X_n, 2\overline{X} - X_1, \ldots, 2\overline{X} - X_n\},$$

  onde $\overline{X}$ é a estimación da media μ. Obtendo desta forma a estimación da función de distribución:

$$F_n^{sim}(x) = \frac{1}{2n}\sum_{i=1}^{2n} \mathbf{1}\{Y_i \leq x\}. \tag{2.4}$$

  Polo tanto, $\hat{F} = F_n^{sim}$, e polo tanto $R^* = R(\boldsymbol{X}^*, F_n^{sim})$.

- Por último, cando a distribución poboacional F é continua é lóxico incorporar dita información ao bootstrap, por medio do método comunmente chamado bootstrap suavizado. Iso significa que a función de distribución ten unha función de densidade asociada tal que $f(x) = F'(x)$. Para isto, debemos utilizar un método bootstrap que faga a remostraxe a partires dun universo bootstrap continuo, mais concretamente empregando un estimador da función de densidade e facendo a remostraxe a partir del. É dicir, nesta ocasión as remostras serán obtidas a partires da estimación tipo núcleo da función de densidade con $\hat{f}(x)$ (1.1).

Estes métodos poden ser combinados. Por exemplo se coñecemos que a distribución poboacional é continua e simétrica, sería lóxico incorporar ambos métodos ao proceso de remostraxe.

Para todos os métodos bootstrap utilizados para obter os cuantís deste traballo emprégase unicamente o bootstrap uniforme no proceso de remostraxe. A xustificación é sinxela: por un lado resulta imposible empregar o bootstrap paramétrico ou simetrizado, xa que un dos obxectivos do SiZer é obter información da densidade da variable aleatoria que estamos estudando, polo tanto en ningún caso coñeceremos se esta pertence a unha familia de distribucións, algún parámetro, ou se esta é simétrica. E por outra banda, un dos obxectivos do SiZer é evitar a problemática que carrexa utilizar un parámetro de suavizado, polo que carece de sentido empregar un bootstrap suavizado.

Cuantil q_1

Na proposta de Chaudhuri e Marron (1999), o primeiro cuantil que consideran é o cuantil Gaussiano, q_1, que é unha aproximación a través da densidade normal que se obtén como:

$$q_1(h) = q_1 = \Phi^{-1}[1 - (\alpha/2)], \tag{2.5}$$

sendo Φ a densidade Gaussiana estandarizada, é dicir, a densidade normal con $\mu = 0$ e $\sigma = 1$, sendo μ e σ a media e desviación típica respectivamente da función de densidade. Por outra banda, α representa o nivel de significación utilizado.

Cuantil q_2

O cuantil q_2 baséase na idea de facer inferencia simultánea sobre x, tentando determinar o número de bloques independentes sobre a mostra. Para isto fundaméntase no feito de que cando x e x' están o suficientemente separados, o que implica que o núcleo centrado en x e x' son "disxuntos", a estimación de $\hat{f}'_h(x)$ e $\hat{f}'_h(x')$ son independentes. Pero cando x e x' están o suficientemente cerca, a estimación de ambas ten unha correlación alta. Con isto, o problema do límite dos intervalos de confianza simultáneos aproxímase por m problemas de intervalos de confianza independentes, onde m reflexa o número de bloques independentes. Estimamos m a partir do tamaño efectivo de mostra (ESS), definido para cada (x, h) como

$$ESS(x, h) = \frac{\sum_{i=1}^{n} K_h(x - X_i)}{K_h(0)}.$$

Cando K é uniforme, ESS é o número de observacións baixo a ventá centrada en x. Para outras formas do núcleo, asígnase un peso a cada punto en función da forma do núcleo. Posteriormente, obtemos m para ser esencialmente o número de bloques independentes sobre a mostras de tamaño n,

$$m(h) = \frac{n}{\text{avg}_x ESS(x, h)}.$$

Agora, asumindo independencia deses $m(h)$ bloques de data, a aproximación do cuantil é

$$q_2 = q_2(h) = \Phi^{-1}\left(\frac{1 + (1 - \alpha)^{1/m}}{2}\right).$$

Cabe resaltar, que ESS é útil para destacar as rexións onde a aproximación (2.2) podería non ser axeitada. As rexións onde $ESS(x, h) < n_0$ (onde é escollido por Chaudhuri e Marron (1999) como $n_0 = 5$) son pintadas de cor gris, para evitar estimar características inexistentes, debido a escaseza de puntos baixo o núcleo nesas rexións. Ademais, isto implica que o tamaño de bloques independentes $m(h)$ é modificado para evitar problemas con pequenos ESS,

$$m(h) = \frac{n}{\text{avg}_{x \in D_h} ESS(x, h)},$$

onde D_h é o conxunto de localizacións x onde a mostra é densa,

$$D_h = \{x : ESS(x, h) \geq n_0\}.$$

Cuantil q_3

O cuantil q_3 é o primeiro dos métodos baseados en bootstrap, o cal busca ter en conta a posible dependencia dos valores da variable x. Neste caso obtén un cuantil para cada h a partires da distribución de

$$Z(x, h) = \frac{\hat{f}'_h(x) - f'_h(x)}{SD(f'_h(x))} \tag{2.6}$$

Para cada h, o cuantil bootstrap $q_3 = q_3(h)$ é o cuantil empírico $\text{máx}_{x\in D_h} |Z^*(x,h)|$ calculado sobre as réplicas bootstrap. Este método baséase no método percentil (Efron 1993; Davison e Hinkley 1997) de forma simetrizada. A característica deste estatístico é que non está estandarizado a partires da desviación típica estimada $\widehat{SD}(f'_h(x))$, se non a partir da desviación típica da función teórica $SD(f'_h(x))$. O procedemento para obter o cuantil vese reflectido no seguinte algoritmo:

Algoritmo 1 Cuantil q_3

Obtemos o cuantil a través do intervalo de confianza empregando o método percentil simetrizado maximizando en x, sendo

$$R(h) = \max_{x\in D_h} |Z(x,h)|$$

Con isto:

1: Para cada $i = 1, ..., n$ arroxar X_i^* a partir de F_n.
2: Obter $\boldsymbol{X}^* = (X_1^*, ..., X_n^*)$.
3: Calcular $R(h)^* = \text{máx}_{x\in D_h} |\frac{\hat{f}_h'^*(x) - \hat{f}'_h(x)}{\widehat{SD}(f'_h(x))}|$.
4: Repetir B veces os pasos 1-3 para obter as réplicas bootstrap $R(h)^{*(1)}, ..., R(h)^{*(B)}$.
5: Empregar os resultados obtidos en 3 para obter o cuantil que deixa á dereita unha proporción de α remostras.
6: O cuantil superior será o cuantil obtido en 5, e o cuantil inferior será o mesmo co signo oposto.

Cuantil q_4

O procedemento para obter o cuantil q_4 é semellante ao procedemento para obter o cuantil q_3. Nesta ocasión o cuantil ten en conta a posible dependencia dos valores tanto da variable x como da variable h.

Empregando de novo a distribución de (2.6), nesta ocasión obtense un único cuantil para todo o mapa SiZer, empregando para este o cuantil empírico $\text{máx}_h \text{máx}_{x\in D_h} |Z^*(x,h)|$ calculado sobre as réplicas bootstrap. O procedemento para obter este cuantil reflíctese no seguinte algoritmo:

Algoritmo 2 Cuantil q_4

Obtemos o cuantil a través do intervalo de confianza empregando o método percentil simetrizado maximizando en x e h, sendo

$$R = \max_h \max_{x\in D_h} |Z(x,h)|$$

Con isto:

1: Para cada $i = 1, ..., n$ arroxar X_i^* a partir de F_n.
2: Obter $\boldsymbol{X}^* = (X_1^*, ..., X_n^*)$.
3: Calcular $R^* = \text{máx}_h \text{máx}_{x\in D_h} |\frac{\hat{f}_h'^*(x) - \hat{f}'_h(x)}{\widehat{SD}(f'_h(x))}|$.
4: Repetir B veces os pasos 1-3 para obter as réplicas bootstrap $R^{*(1)}, ..., R^{*(B)}$.
5: Empregar os resultados obtidos no paso 3 para obter o cuantíl que deixa á dereita unha proporción de α remostras.
6: O cuantil superior será o cuantil obtido no paso 5, e o cuantil inferior será o mesmo co signo oposto.

2.2.2. Comparativa de Chaudhuri e Marron

Aínda que no Capítulo 3 se presentan estudos de simulación que comparan a obtención do mapa SiZer a partir dos diferentes cuantís propostos, Chaudhuri e Marron narran que aínda que a aproxi-

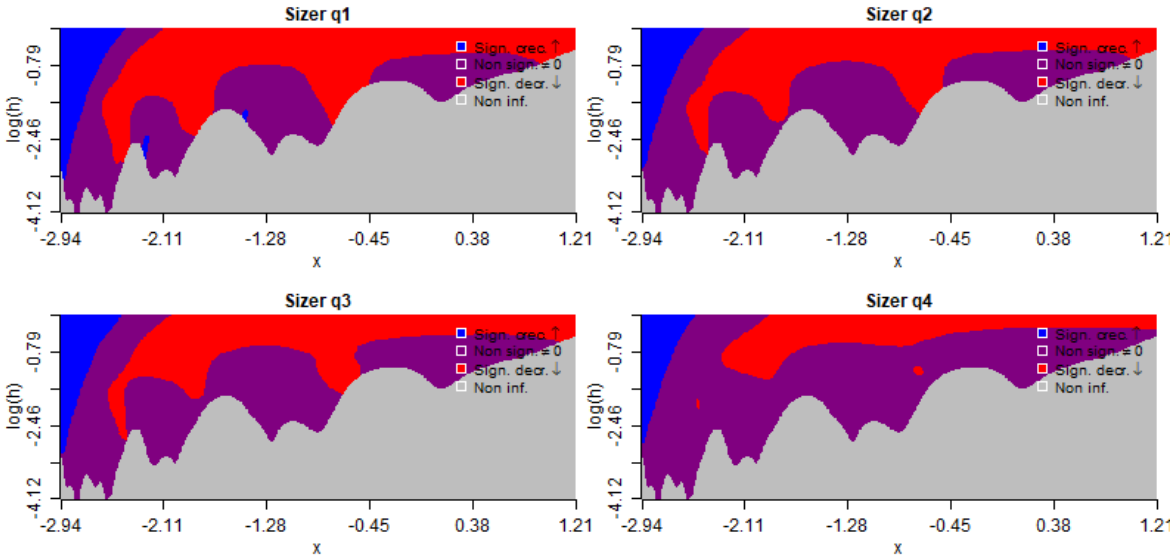

Figura 2.2: Estimación do SiZer con q_1, q_2, q_3 e q_4 para unha mostra xerada a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

mación Gaussiana do cuantil (q_1) funciona bastante ben, esta non é recomendada, xa que a versión deste mapa SiZer suxire que demasiadas características (modas ou vales) son significativas.

Na a Figura 2.2 represéntase o mapa SiZer de $\boldsymbol{X} = (X_1, \ldots, X_{100})$, unha m.a.s. de X con función de densidade *Strongly Skewed* (#3) proposta por Marron e Wand (1992), a partir dos catro métodos para obter o cuantil que acabamos de mencionar, empregando para todos eles un nivel de significación de $\alpha = 0.05$. Ademais o tamaño da grella de puntos sobre o eixe x é de $n(x) = 512$ puntos, e o número de valores do parámetro de suavizado empregado foi dun total de $n(h) = 151$ puntos. Esta densidade móstrase na gráfica esquerda da Figura B.3 do Apéndice B, a cal se pode ver xunto o mapa SiZer (gráfica do centro) obtido a partir da convolución (2.1) con núcleo Gaussiano. A densidade está formada pola mestura de oito distribucións normais e intenta reflectir a estrutura da distribución lognormal (un pico moi alto cunha longa cola á dereita).

Na Figura 2.3 móstrase a estimación tipo núcleo da función de densidade (1.1) para tres valores distintos do parámetro de suavizado ($h = \{0.10, 0.22, 0.50\}$) xunto a curva orixinal f. Apréciase claramente como esta densidade é difícil de estimar, xa que para zonas próximas ao pico un parámetro de suavizado pequeno é máis axeitado para evitar o sobresuavizado da curva, mentres que un parámetro de suavizado mais grande é mais axeitado para a zona da cola, onde é importante non estimar modas inexistentes. Como mostra a Figura 2.2, o mapa SiZer obtido a partir da aproximación do cuantil Gaussiano, presenta de forma incorrecta que algunhas destas modas obtidas sobre a cola son significativas. Por exemplo, os picos mostrados en -2.1 ou en -1.2. O problema é entendido baixo a clásica interpretación frecuentista; obtendo o intervalo de confianza dun amplo número de valores, unha proporción aproximada α non cubre o verdadeiro valor. Unha solución a este problema é axustar a lonxitude dos intervalos para facer inferencia simultánea, o cal é o obxectivo dos demais cuantís mencionados anteriormente. Volvendo á Figura 2.2, vemos como o mapa SiZer obtido a partires dos cuantís q_2, q_3 e q_4 non presentan características estimadas de forma errónea, e mostran correctamente a única moda presente na función de densidade *Strongly Skewed*.

Por outra banda, Chaudhuri e Marron indican que os mapas SiZer baseados en q_2, q_3 e q_4 mostran que en moitos casos hai pouca diferencia entre os cantís q_2 e o cuantil bootstrap simultáneo sobre x

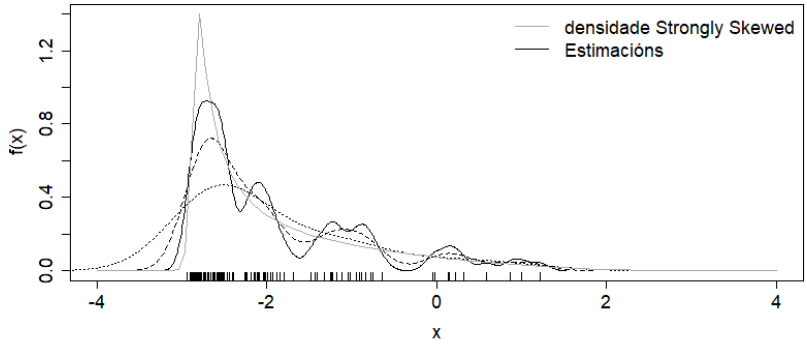

Figura 2.3: Densidade *Strongly Skewed* (#3) proposta por Marron e Wand (1992) e estimación da densidade con diferentes parámetros de suavizado dunha mostra xerada.

(q_3). Ademais, xeralmente aparecen algunhas características menos como significativas no cuantil q_4, aínda que sorprendentemente (a pesar de ser un cuantil simultáneo sobre x e h) case sempre é moi similar aos mapas SiZer obtidos con q_2 e q_3.

A pesar disto, o cuantil q_4 é un tanto conservador. Na a Figura 2.4 represéntase o mapa SiZer de $\boldsymbol{X} = (X_1, \dots, X_{500})$, unha m.a.s. de X con función de densidade *Bimodal* (#6) proposta por Marron e Wand (1992), de novo, a partir dos catro métodos para obter o cuantil, e con un nivel de significación de $\alpha = 0.05$. Ademais o tamaño da grella de puntos sobre o eixe x é de $n(x) = 512$ puntos, e o número de valores do parámetro de suavizado empregado foi dun total de $n(h) = 151$ puntos. Esta densidade móstrase na gráfica esquerda da Figura B.6 do Apéndice B, a cal está formada por dúas modas próximas entre elas, e pódese ver xunto o mapa SiZer (gráfica do centro) obtido a partir da convolución (2.1). Na Figura 2.5 móstrase a estimación tipo núcleo da función de densidade (1.1) para tres valores distintos do parámetro de suavizado ($h = \{0.10, 0.22, 0.50\}$) xunto a curva orixinal f.

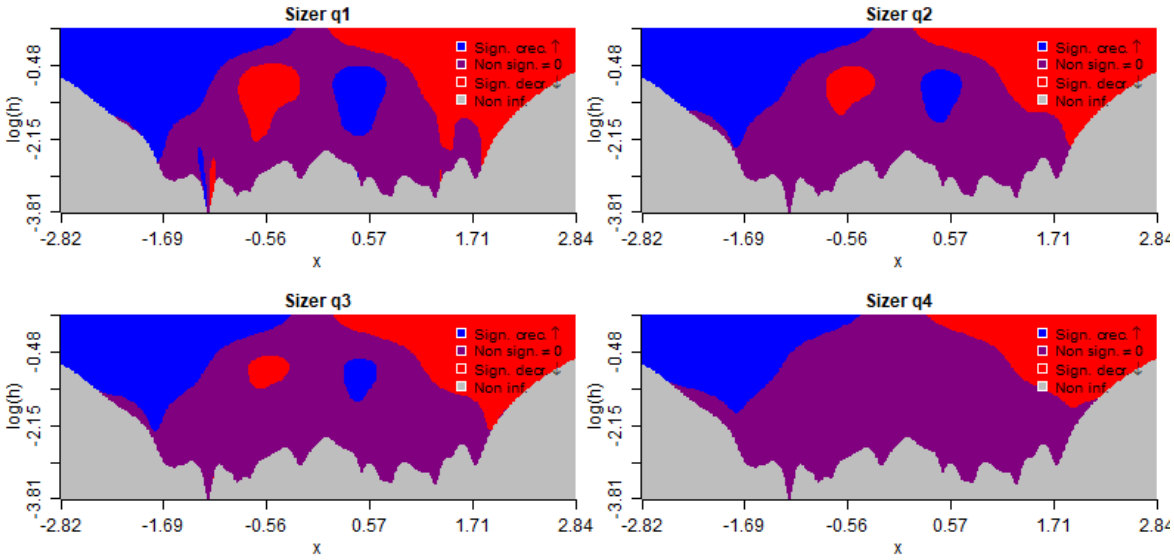

Figura 2.4: Estimación do SiZer con q_1, q_2, q_3 e q_4 para unha mostra xerada a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand.

Obsérvase como para valores pequenos do parámetro de suavizado o estimador tipo núcleo da función de densidade pode chegar a ter problemas para detectar as dúas modas. Volvendo á Figura 2.4, o mapa SiZer obtido para q_1, q_2 e q_3 mostra a presenza das dúas modas existentes na distribución, mentres que estas desaparecen para o mapa SiZer co cuantil q_4. No seu lugar obtén unha única moda central inexistente en $f(x)$, o que mostra a problemática do cuantil q_4 para representar de forma significativa características existentes e visibles polos demais cuantís.

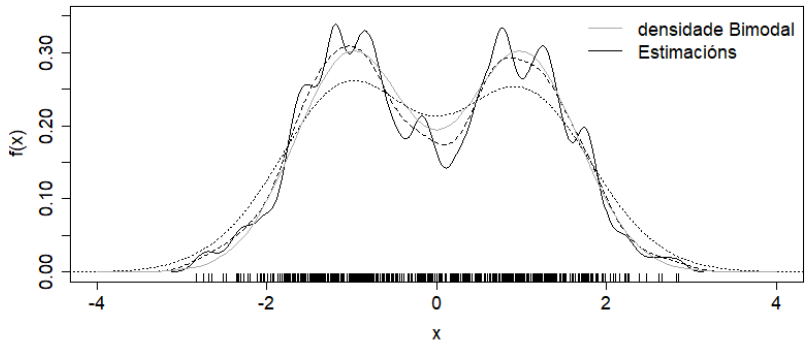

Figura 2.5: Densidade *Bimodal* (#6) proposta por Marron e Wand e estimación da densidade con diferentes parámetros de suavizado dunha mostra xerada.

Chaudhuri e Marron (1999) suxerían usar o cuantil q_2 como primeira alternativa debido á alta carga computacional que supoñía o método bootstrap naquel momento. Pero no caso de haber dúbidas acerca das conclusións obtidas a partires do mapa SiZer obtido co cuantil q_2, os cuantís q_3 e q_4 servirían como ferramenta para verificar. Aínda que o método q_4 é o único procedemento que brinda unha proba rigorosa da importancia das características, tamén é en xeral conservador, polo que recomendaban que

as características encontradas no mapa SiZer a través dos cuantís q_2 e q_3 que non aparecen na versión do SiZer co cuantil q_4 deberían ser investigadas independentemente por medio de un test de modas.

Unha comparativa mais extensa dos resultados obtidos por cada un dos cuantís propostos por Chaudhuri e Marron na ferramenta SiZer realizarase no seguinte Capítulo. Nel levaremos a cabo un estudo de simulación que nos permita ver os resultados obtidos por cada un dos cuantís, observando a partir de diferentes ferramentas a porcentaxe de erro nas modas estimadas ou a porcentaxe na que un cuantil non é capaz de detectar as características presentes na función de densidade xeradora das mostras empregadas no estudo de simulación.

2.3. Cuantís bootstrap

Na teoría, o cuantil q_1 ofrece unha taxa de cobertura dos intervalos de confianza da estimación tipo núcleo da derivada da función de densidade $\hat{f}'_h(x)$ en cada punto (x, h) entorno ao nivel $(1-\alpha)$. Tal e como veremos na Sección 3.3 isto vai ser así independentemente do valor do parámetro de suavizado h e do punto da grella x. Pero para tamaños de mostra menores (nos que consideramos tamaños de mostra tal que $n \leq 20$) veremos que a cobertura non vai ser tan acertada.

Tal e como se menciona en Hall (1988), facendo un desenvolvemento de Edgeworth, o cuantil q_1 Gaussiano ten unha orde de erro de cobertura do intervalo de confianza bilateral de $O(n^{-1})$, polo que nesta Sección imos propoñer métodos bootstrap que poidan ofrecer unha taxa de cobertura similar ou maior.

Con isto, propoñeremos catros formas máis de obter o cuantil, tendo en conta cada punto x de forma independente, e o mesmo para cada valor do parámetro de suavizado h, tentando desta forma mellorar os resultados obtidos por q_1:

- q_5: Bootstrap percentil: Baséase en obter o cuantil superior $(1-\alpha)$ e inferior α para cada punto (x, h) do mapa SiZer de forma independente, aproximando a distribución mediante bootstrap de (2.6).

- q_6: Bootstrap percentil simetrizado: Baséase en obter o cuantil superior $(1-\alpha)$ para cada punto (x, h) do mapa SiZer de forma independente, aproximando o valor absoluto da distribución mediante bootstrap de (2.6).

- q_7: Bootstrap percentil-t: Baséase en obter o cuantil superior $(1-\alpha)$ e inferior α para cada punto (x, h) do mapa SiZer de forma independente, aproximando a distribución mediante bootstrap de:

$$Z_2(x,h) = \frac{\hat{f}'_h(x) - f'_h(x)}{SD(\hat{f}'_h(x))}. \tag{2.7}$$

- q_8: Bootstrap percentil-t simetrizado: Baséase en obter o cuantil superior $(1-\alpha)$ para cada punto (x, h) do mapa SiZer de forma independente, aproximando o valor absoluto da distribución mediante bootstrap de (2.7).

2.3.1. Cuantil q_5

Para cada punto (x, h), o cuantil bootstrap $q_5 = q_5(x, h)$ é o cuantil empírico superior e inferior de $Z^*(x, h)$ calculado sobre as réplicas bootstrap. O procedemento para obter o cuantil vese reflectido no seguinte algoritmo:

Algoritmo 3 Cuantil q_5: Bootstrap percentil

Obtemos o cuantil a través do intervalo de confianza empregando o método percentil, sendo

$$R(x,h) = Z(x,h).$$

Con isto:

1: Para cada $i = 1, ..., n$ arroxar X_i^* a partir de F_n.
2: Obter $\boldsymbol{X}^* = (X_1^*, ..., X_n^*)$.
3: Calcular $R(x,h)^* = \frac{\hat{f}_h'^*(x) - \hat{f}_h'(x)}{\widehat{SD}(f_h'(x))}$.
4: Repetir B veces os pasos 1-3 para obter as réplicas bootstrap $R(x,h)^{*(1)}, ..., R(x,h)^{*(B)}$.
5: Empregar os resultados obtidos no paso 3 para obter os cuantís superior e inferior de modo que deixen unha proporción de $\alpha/2$ remostras a cada unha das dúas bandas.

Tal e como se menciona en Hall (1988), facendo un desenvolvemento de Edgeworth, o cuantil q_5 ten unha orde de erro de cobertura do intervalo de confianza bilateral de $O(n^{-1/2})$.

2.3.2. Cuantil q_6

De novo, para cada punto (x,h), o cuantil bootstrap $q_6 = q_6(x,h)$ é o cuantil empírico de $|Z^*(x,h)|$ calculado sobre as réplicas bootstrap. O procedemento para obter o cuantil vese reflectido no seguinte algoritmo:

Algoritmo 4 Cuantil q_6: Bootstrap percentil simetrizado

Obtemos o cuantil a través do intervalo de confianza empregando o método percentil simetrizado, sendo

$$R(x,h) = Z(x,h).$$

Con isto:

1: Para cada $i = 1, ..., n$ arroxar X_i^* a partir de F_n.
2: Obter $\boldsymbol{X}^* = (X_1^*, ..., X_n^*)$.
3: Calcular $R(x,h)^* = \frac{\hat{f}_h'^*(x) - \hat{f}_h'(x)}{\widehat{SD}(f_h'(x))}$.
4: Repetir B veces os pasos 1-3 para obter as réplicas bootstrap $R(x,h)^{*(1)}, ..., R(x,h)^{*(B)}$.
5: Empregar o valor absoluto dos resultados obtidos no paso 3 para obter o cuantíl que deixa á dereita unha proporción de α remostras.
6: O cuantil superior será o cuantil obtido no paso 5, e o cuantil inferior será o mesmo co signo oposto.

A orde do erro da cobertura do intervalos de confianza bilateral do cuantil q_6 é descoñecida.

2.3.3. Cuantil q_7

Para cada punto (x,h), o cuantil bootstrap $q_7 = q_7(x,h)$ é o cuantil empírico superior e inferior de $Z_2^*(x,h)$ calculado sobre as réplicas bootstrap. Este método baséase no método percentil-t (Efron 1993; Davison e Hinkley 1997). A característica deste estatístico é que non está estandarizado a partires da desviación típica da función teórica $SD(f_h'(x))$, se non a partir da desviación típica estimada $\widehat{SD}(f_h'(x))$.

O procedemento para obter o cuantil vese reflectido no seguinte algoritmo:

Algoritmo 5 Cuantil q_7: Bootstrap percentil-t

Obtemos o cuantil a través do intervalo de confianza empregando o método percentil-t (Efron 1993; Davison e Hinkley 1997), sendo

$$R(x,h) = Z_2(x,h).$$

Con isto:

1: Para cada $i = 1, ..., n$ arroxar X_i^* a partir de F_n.
2: Obter $\boldsymbol{X}^* = (X_1^*, ..., X_n^*)$.
3: Calcular $R(x,h)^* = \frac{\hat{f}_h'^*(x) - \hat{f}_h'(x)}{\widehat{SD}(f_h'(x))}$.
4: Repetir B veces os pasos 1-3 para obter as réplicas bootstrap $R(x,h)^{*(1)}, ..., R(x,h)^{*(B)}$.
5: Empregar os resultados obtidos no paso 3 para obter os cuantís superior e inferior de modo que deixen unha proporción de $\alpha/2$ remostras a cada unha das dúas bandas.

De novo, seguindo as indicacións presentes en Hall (1988), facendo un desenvolvemento de Edgeworth, o cuantil q_7 ten unha orde de erro de cobertura do intervalo de confianza bilateral de $O(n^{-1})$.

2.3.4. Cuantil q_8

De novo, para cada punto (x,h), o cuantil bootstrap $q_8 = q_8(x,h)$ é o cuantil empírico de $|Z_2^*(x,h)|$ calculado sobre as réplicas bootstrap. O procedemento para obter o cuantil vese reflectido no seguinte algoritmo:

Algoritmo 6 Cuantil q_8: Bootstrap percentil-t simetrizado

Obtemos o cuantil a través do intervalo de confianza empregando o método percentil-t simetrizado, sendo

$$R(x,h) = Z_2(x,h).$$

Con isto:

1: Para cada $i = 1, ..., n$ arroxar X_i^* a partir de F_n.
2: Obter $\boldsymbol{X}^* = (X_1^*, ..., X_n^*)$.
3: Calcular $R(x,h)^* = \frac{\hat{f}_h'^*(x) - \hat{f}_h'(x)}{\widehat{SD}(f_h'(x))}$.
4: Repetir B veces os pasos 1-3 para obter as réplicas bootstrap $R(x,h)^{*(1)}, ..., R(x,h)^{*(B)}$.
5: Empregar o valor absoluto dos resultados obtidos no paso 3 para obter o cuantíl que deixa á dereita unha proporción de α remostras.
6: O cuantil superior será o cuantil obtido no paso 5, e o cuantil inferior será o mesmo co signo oposto.

Facendo un desenvolvemento de Edgeworth, o cuantil q_8 ten unha orde de erro de cobertura do intervalo de confianza bilateral de $O(n^{-3/2})$.

Na Sección 3.3 veremos que este cuantil ofrece unha taxa de cobertura superior a dos cuantís q_1, q_5, q_6 e q_7 para tamaños de mostra reducidos.

2.4. Implementación numérica

O rango de parámetros ventá $[h_{\text{mín}}, h_{\text{máx}}]$ pode ser escollido de moitas maneiras. Un enfoque consiste en coller unha gama ampla de parámetros de suavizado os cales deberían capturar as características mais interesantes, como propoñen Marron e Chung (2001). Neste caso, Chaudhuri e Marron propoñen

coller un rango amplo que sexa determinado mais polo axuste da estimación da curva que pola mostra. Nos exemplos do artigo Chaudhuri e Marron (1999) toman como $h_{\text{mín}}$ o parámetro ventá mais pequeno para o cal non hai distorsión na estimación da curva, tal que $h_{\text{mín}} = 2 \cdot (\text{binwidth})$. É dicir, o dobre da distancia entre dous puntos consecutivos da grella x, sendo a distancia entre dous puntos un valor fixo. Por outra banda $h_{\text{máx}}$ é tomado como a distancia que hai no rango da mostra.

2.4.1. Uso da tranformada rápida de Fourier na estimación tipo núcleo

A implementación da estimación tipo núcleo da derivada da densidade $\hat{f}'_h(x)$ foi levada a través da transformada de Fourier.

En xeral, a transformada discreta de Fourier (DFT) dun vector $c = (c_1, \dots, c_M) \in \mathbb{R}^M$ ven dada pola expresión $\text{DFT}(\boldsymbol{c}) = \boldsymbol{d}$, sendo $\boldsymbol{d} = (d_1, \dots, d_M) \in \mathbb{R}^M$ con

$$d_l = \sum_{k=1}^{M} c_k e^{2\pi i(l-1)(k-1)/M}. \quad l = 1, \dots, M$$

Para calcular a DFT de modo rápido, empregase a transformada rápida de Fourier (FFT), que en `R` está implementada na función fft. En concreto,

$$DFT(\boldsymbol{c}) = \texttt{fft(c, inverse = TRUE)}.$$

A opción `inverse = FALSE` proporciona a DFT co expoñente cambiado de signo, que en moitos lugares é a definición estándar para a DFT. Este método FFT funciona mellor se $M = 2^m$ para algún $m \in \mathbb{N}$.

Por outra banda, para obter o intervalo de confianza de $\hat{f}'_h(x)$ a través da ecuación (2.2) é necesario obter a estimación da desviación típica de $\hat{f}'_h(x)$, tal que

$$\widehat{SD}(\hat{f}'_h(x)) = \sqrt{\widehat{var}(\hat{f}'_h(x))},$$

sendo $\widehat{var}(\hat{f}'_h(x))$ a varianza da derivada da función de densidade, a cal podemos obter sen mais que

$$\widehat{var}(\hat{f}'_h(x)) = \frac{1}{n}\left((\hat{f}'_h)^2(x) - (\hat{f}'_h(x))^2\right).$$

Polo que necesitaremos obter tamén a estimación tipo núcleo da derivada da densidade ao cadrado. A implementación desta estimación tamén foi levada a cabo a través da transformada de Fourier, a cal veremos a continuación.

Estimación tipo núcleo da derivada da densidade e da derivada da densidade ao cadrado

Se definimos $\varphi_L(t) = \int e^{itx} L(x)dx$ para cada función $L \in L_1$, pódese escribir a función característica da estimación tipo núcleo da derivada da densidade da seguinte maneira:

$$\begin{aligned}\varphi_{f'_{nh}}(t) &= \int_{-\infty}^{\infty} e^{itx} f'_{nh}(x)dx = \int_{-\infty}^{\infty} e^{itx} \frac{1}{nh^2} \sum_{j=1}^{n} K'\left(\frac{x - X_j}{h}\right) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{nh} \sum_{j=1}^{n} e^{it(yh+X_j)} K'(y)dy = \frac{1}{nh} \sum_{j=1}^{n} e^{itX_j} \int_{-\infty}^{\infty} e^{ityh} K'(y)dy = \frac{1}{h}\varphi_n(t)\varphi_{K'}(th),\end{aligned}$$

onde $\varphi_n(t) = \frac{1}{n}\sum_{j=1}^{n} e^{itX_j}$ denota a función característica empírica, $\varphi_{K'}$ a función característica da derivada do núcleo e $y = \frac{x - X_j}{h}$, polo tanto $dy = \frac{dx}{h}$. Mentres que a estimación tipo núcleo da derivada da densidade ao cadrado queda da seguinte forma:

$$\begin{aligned}\varphi_{(f'_{nh})^2}(t) &= \int_{-\infty}^{\infty} e^{itx}(f'_{nh})^2(x)dx = \int_{-\infty}^{\infty} e^{itx}\frac{1}{n}\sum_{j=1}^{n}\left(\frac{K'\left(\frac{x-X_j}{h}\right)}{h^2}\right)^2 dx \\ &= \int_{-\infty}^{\infty} e^{itx}\frac{1}{nh^4}\sum_{j=1}^{n}\left(K'\left(\frac{x-X_j}{h}\right)\right)^2 dx = \int_{-\infty}^{\infty}\frac{1}{nh^3}\sum_{j=1}^{n} e^{it(yh+X_j)}(K')^2(y)dy \\ &= \frac{1}{nh^3}\sum_{j=1}^{n} e^{itX_j}\int_{-\infty}^{\infty} e^{ityh}(K')^2(y)dy = \frac{1}{h^3}\varphi_n(t)\varphi_{(K')^2}(th),\end{aligned}$$

onde $\varphi_{(K')^2}$ denota a función característica da derivada do núcleo ao cadrado e $y = \frac{x-X_j}{h}$, polo tanto $dy = \frac{dx}{h}$. Ademais, con esta factorización, a única parte que depende da mostra é $\varphi_n(t)$.

O noso propósito é pintar os estimadores de f'_{nh} e $(f'_{nh})^2$ para $x \in [a,b]$, e para iso, necesitaremos coñecer o valor de f'_{nh} e $(f'_{nh})^2$ nunha grella de puntos de $[a,b]$; e dicir, dados M puntos $a = r_1 < r_2 < ... < r_m = b$, o noso obxectivo é calcular $f'_{nh}(r_k)$ e $(f'_{nh})^2(r_k)$, $k = 1, ..., M$. Para isto, calcularemos $\varphi_{f'_{nh}}$ e $\varphi_{(f'_{nh})^2}$ nunha grella adecuada e logo aplicaremos a transformada inversa de Fourier, xa que:

$$f'_{nh}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi_{f'_{nh}},$$

$$(f'_{nh})^2(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi_{(f'_{nh})^2}.$$

Este procedemento ten dous problemas que haberá que resolver: primeiro, como avaliar de maneira rápida $\varphi_n(t)$ nunha grella de puntos para logo obter $\varphi_{f'_{nh}}$ e $\varphi_{(f'_{nh})^2}$, e segundo, como calcular a transformada inversa, tamén de forma rápida.

Función característica

Como vemos, para a estimación tipo núcleo da derivada da densidade e da derivada da densidade ao cadrado serán necesarias as funcións características da derivada do núcleo e da derivada do núcleo ao cadrado respectivamente. De seguido, son mostrados os resultados obtidos do desenvolvemento de $\varphi_{K'}$:

$$\begin{aligned}\varphi_{K'}(t) &= \int_{-\infty}^{\infty} e^{itx}K'(x)dx = \int_{-\infty}^{\infty} e^{itx}\frac{-x}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}dx = e^{\frac{-t^2}{2}}\int_{-\infty}^{\infty}\frac{-x}{\sqrt{2\pi}}e^{\frac{-(x-it)^2}{2}}dx \\ &= e^{\frac{-t^2}{2}}\int_{-\infty-it}^{\infty-it}\frac{-y-it}{\sqrt{2\pi}}e^{\frac{-y^2}{2}}dy = e^{\frac{-t^2}{2}}\left(\int_{-\infty-it}^{\infty-it}\frac{-y}{\sqrt{2\pi}}e^{\frac{-y^2}{2}}dy + \int_{-\infty-it}^{\infty-it}\frac{-it}{\sqrt{2\pi}}e^{\frac{-y^2}{2}}dy\right) \\ &= (0-it)e^{\frac{-t^2}{2}} = -ite^{\frac{-t^2}{2}},\end{aligned}$$

sendo $y = x - it$, e polo tanto $dy = dx$. Onde ademais o resultado do primeiro sumando das integrais se obtivo da seguinte maneira:

$$\int_{-\infty-it}^{\infty-it}\frac{-y}{\sqrt{2\pi}}e^{\frac{-y^2}{2}}dy = \frac{1}{\sqrt{2\pi}}\int_{-\infty-it}^{\infty-it} -ye^{\frac{-y^2}{2}}dy = \frac{1}{\sqrt{2\pi}}\left[e^{\frac{-y^2}{2}}\right]_{-\infty-it}^{\infty-it} = \frac{1}{\sqrt{2\pi}}\left[e^{\frac{-(x-it)^2}{2}}\right]_{-\infty}^{\infty} = 0.$$

Mentres que o resultado do segundo sumando das integrais foi obtido da seguinte forma:

$$\int_{-\infty-it}^{\infty-it}\frac{-it}{\sqrt{2\pi}}e^{\frac{-y^2}{2}}dy = -it\int_{-\infty-it}^{\infty-it}\frac{1}{\sqrt{2\pi}}e^{\frac{-y^2}{2}}dy = -it\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{\frac{-(x-it)^2}{2}}dx = -it.$$

E por outra banda, os resultados obtidos dos desenvolvemento de $\varphi_{(K')^2}$:

$$\begin{aligned}
\varphi_{(K')^2}(t) &= \int_{-\infty}^{\infty} e^{itx} K'(x)^2 dx = \int_{-\infty}^{\infty} e^{itx} \left(\frac{-x}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \right)^2 dx = \int_{-\infty}^{\infty} e^{itx} \frac{x^2}{2\pi} e^{-x^2} dx \\
&= \frac{1}{2\pi} e^{\frac{-t^2}{4}} \int_{-\infty}^{\infty} x^2 e^{-(x-it)^2} dx = \frac{1}{2\pi} e^{\frac{-t^2}{4}} \int_{-\infty-\frac{it}{2}}^{\infty-\frac{it}{2}} \left(y + \frac{it}{2} \right)^2 e^{-y^2} dy \\
&= \frac{1}{2\pi} e^{\frac{-t^2}{4}} \int_{-\infty-\frac{it}{2}}^{\infty-\frac{it}{2}} \left(y^2 + ity - \frac{t^2}{4} \right) e^{-y^2} dy \\
&= \frac{1}{2\pi} e^{\frac{-t^2}{4}} \left(\int_{-\infty-\frac{it}{2}}^{\infty-\frac{it}{2}} y^2 e^{-y^2} dy + \int_{-\infty-\frac{it}{2}}^{\infty-\frac{it}{2}} ity e^{-y^2} dy + \int_{-\infty-\frac{it}{2}}^{\infty-\frac{it}{2}} -\frac{t^2}{4} e^{-y^2} dy \right) \\
&= \frac{1}{2\pi} e^{\frac{-t^2}{4}} \left(\frac{\sqrt{\pi}}{2} + 0 + \frac{-t^2\sqrt{\pi}}{4} \right) = \frac{1 - t^2/2}{4\sqrt{\pi}} e^{\frac{-t^2}{4}},
\end{aligned}$$

sendo $y = x - \frac{it}{2}$, e polo tanto $dy = dx$. Onde o primeiro sumando das integrais se obtivo da seguinte maneira:

$$\int_{-\infty-\frac{it}{2}}^{\infty-\frac{it}{2}} y^2 e^{-y^2} dy = \frac{\sqrt{\pi}}{2}.$$

O resultado do segundo sumando:

$$\int_{-\infty-\frac{it}{2}}^{\infty-\frac{it}{2}} ity e^{-y^2} dy = \frac{it}{-2} \int_{-\infty-\frac{it}{2}}^{\infty-\frac{it}{2}} -2y e^{-y^2} dy = \frac{-it}{2} \left[e^{-y^2} \right]_{-\infty-\frac{it}{2}}^{\infty-\frac{it}{2}} = \frac{-it}{2} \left[e^{-(x-\frac{it}{2})^2} \right]_{-\infty}^{\infty} = 0.$$

Mentres que o resultado do terceiro sumando obtívose da seguinte maneira:

$$\begin{aligned}
\int_{-\infty-\frac{it}{2}}^{\infty-\frac{it}{2}} -\frac{t^2}{4} e^{-y^2} dy &= -\frac{t^2}{4} \int_{\sqrt{2}(-\infty-\frac{it}{2})}^{\sqrt{2}(\infty-\frac{it}{2})} \frac{1}{\sqrt{2}} e^{\frac{-u^2}{2}} du \\
&= -\frac{t^2\sqrt{\pi}}{4} \int_{\sqrt{2}(-\infty-\frac{it}{2})}^{\sqrt{2}(\infty-\frac{it}{2})} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du = -\frac{t^2\sqrt{\pi}}{4},
\end{aligned}$$

sendo $u = \sqrt{2}y$, e polo tanto $du = \sqrt{2}dy$.

Grella de puntos

O primeiro paso a seguir é o que se coñece como discretización dos datos (en inglés, *binning*). A discretización consiste en pasar dos puntos $X_1, \ldots, X_n$ que teñen peso $1/n$ cada un, aos puntos da grella $r_1, \ldots, r_M$ cargados con certos pesos $\xi_1, \ldots, \xi_M$ de modo que ξ_k reflicte dalgunha maneira a cantidade de puntos X_i que están cerca de r_k. Imos supoñer que a grella na que imos avaliar $f'_{nh}(x)$ e $(f'_{nh})^2(x)$ é equidistante, e dicir, que se tomamos $\delta = (b-a)/(M-1)$ entón $r_k = a + (k-1)\delta$ para $k = 1, \ldots, M$.

Existen diversas maneiras de asignar peso ξ_k ao punto r_k da grella. A opción máis sinxela, coñecida como *simple binning* (Silverman, 1982), consiste en engadir o peso $1/n$ que inicialmente corresponde ao dato X_i ao punto r_k da grella que teña máis próximo. Deste modo, o peso final que terá o punto r_k da grella será:

$$\xi_{k,simple} = \frac{1}{n} \sum_{i=1}^{n} I_{\{|X_i - r_k| = \text{mín}_{j=1,\ldots,M} |X_i - r_j|\}} = \frac{1}{n} \sum_{i=1}^{n} I_{\{|X_i - r_k| < \delta/2\}}.$$

No que segue, imos empregar a asignación que se coñece como *linear binning*, que Jones e Lotwick (1983) proban que é máis preciso que o *simple binning*. Co *linear binning* o peso que cada observación

X_i aporta ao punto r_k da grella vén dado por:

$$\text{Peso que } X_i \text{ aporta a } \xi_k = \begin{cases} 0 & \text{si } |X_i - r_k| \geq \delta \\ (\delta - |X_i - r_k|)/(n\delta) & \text{si } |X_i - r_k| < \delta \end{cases}$$

e dicir, para calcular o peso ξ_k asignado a r_k só nos fixamos nos datos que están nos intervalos que rodean r_k, que son $(r_{k-1}, r_k]$ e $[r_k, r_{k+1})$, e o peso que aporta a cada un deses datos a ξ_k é maior canto máis cerca está dito dato d punto r_k. Con isto,

$$\xi_{k,linear} = \frac{1}{n}\sum_{i=1}^{n}(1 - |X_i - r_k|/\delta)I_{\{|X_i - r_k| < \delta\}},$$

e pódese comprobar que $\sum_{k=1}^{M} \xi_k = 1$.

Estimacións

Ao cambiar a mostra $X_1, \dots, X_n$ con pesos $1/n$ cada un, polos puntos da grella $r_1, \dots, r_M$ con pesos $\xi_1, \dots, \xi_M$, en realidade estamos aproximando a medida empírica usual $\mu_n(A) = \frac{1}{n}\sum_{j=1}^{n} I_{\{X_j \in A\}}$ pola discretizada $\nu_M(A) = \sum_{k=1}^{M} \xi_k I_{\{r_k \in A\}}$. Entón, para cada $t \in \mathbb{R}$, podemos aproximar a función característica empírica mediante

$$\varphi_n(t) = \sum_{j=1}^{n} \frac{1}{n} e^{it_l X_j} \approx \sum_{k=1}^{M} \xi_k e^{itr_k} = e^{ita} \sum_{k=1}^{M} \xi_k e^{it(k-1)\delta}.$$

Esa última suma ten a forma dunha DFT se tomamos un valor de $t = t_l$ axeitado. En concreto se

$$t = t_l = 2\pi(l-1)/(\delta M), \quad l = 1, \dots, M$$

entón o vector $\boldsymbol{Y} = (\varphi_n(t_1), \dots, \varphi_n(t_M))$ está directamente relacionado coa DFT do vector de pesos $\boldsymbol{\xi} = (\xi_1, \dots, \xi_M)$, en concreto

$$Y_l = \varphi_n(t_l) \approx e^{ita} \sum_{k=1}^{M} \xi_k e^{2\pi i(l-1)(k-1)/M}, \quad l = 1, \dots, M \tag{2.8}$$

e dicir, Y_l pódese aproximar pola DFT$(\boldsymbol{\xi})_l$, a coordenada l-ésima da DFT de $\boldsymbol{\xi}$, (que se pode calcular de forma rápida empregando a FFT), multiplicada por $e^{it_l a}$.

Ademais, os puntos $\{t_l\}_{l=1}^{M}$ forman unha grella equidistante no intervalo $\left[0, \frac{2\pi(M-1)^2}{M(b-a)}\right] \approx \left[0, \frac{2\pi M}{(b-a)}\right]$, intervalo que se aproxima a $[0, \infty)$ cando a grella orixinal $\{r_k\}_{k=1}^{M}$ do intervalo $[a, b]$ se afina (e dicir, cando $M \to \infty$). Con isto conseguimos o primeiro dos propósitos que buscábamos.

Para o segundo, no caso da estimación da derivada da densidade, empregando a aproximación rápida de $\varphi_n(t_l)$ pódese calcular

$$\zeta_l = \varphi_n(t_l)\varphi_{K'}(t_l h), \quad l = 1, \dots, M$$

polo que $\frac{1}{h}\zeta_l$ é o mesmo que avaliar $\varphi_{f'_{nh}}(t_l)$ nunha grella de M puntos equidistantes, con distancia entre eles $\Delta = t_2 - t_1 = 2\pi/(\delta M)$. Ademais, tendo en conta que $t_1 = 0$ e que $\varphi_{f'_{nh}}(-t) = \overline{\varphi_{f'_{nh}}(t)}$, resulta que $\frac{1}{h}(\overline{\zeta_M}, \dots, \overline{\zeta_2}, \zeta_1, \dots, \zeta_M)$ corresponde a avaliar $\varphi_{f'_{nh}}(t)$ nos $2M-1$ puntos equidistantes $-t_M, \dots, -t_2, t_1, \dots, t_M$. Con isto, dado que $t_l r_k = t_l a + 2\pi(l-1)(k-1)/M$, utilizando a aproximación

integral por rectángulos, séguese que

$$\begin{aligned} f'_{nh}(r_k) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itr_k}\varphi_{f'_{nh}}(t)dt \approx \frac{1}{2\pi h}\sum_{l=1}^{M}\Delta\zeta_l e^{-it_l r_k} + \frac{1}{2\pi h}\sum_{l=2}^{M}\Delta\overline{\zeta_l}e^{-it_l r_k} \\ &= \frac{1}{2\pi h}\sum_{l=1}^{M}\Delta\zeta_l e^{-it_l r_k} + \frac{1}{2\pi h}\sum_{l=1}^{M}\Delta\overline{\zeta_l}e^{-it_l r_k} = \frac{2}{\delta M h}\mathrm{Re}\left\{\sum_{l=1}^{M}\zeta_l e^{-it_l r_k}\right\} \\ &= \frac{2}{\delta M h}\mathrm{Re}\left\{\sum_{l=1}^{M}[\zeta_l e^{-it_l a}]e^{-2\pi i(l-1)(k-1)/M}\right\}. \end{aligned}$$

Se chamamos $\boldsymbol{Z}$ ao vector cuxas coordenadas son as de $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_M)$ multiplicadas polo valor de $e^{-it_l a}$ correspondente, entón o sumando da expresión anterior non é mais que $2/(\delta M h)$ pola parte real da DFT inversa de $\boldsymbol{Z}$, e dita DFT inversa pódese calcular eficientemente mediante a FFT. Ademais, cabe destacar que en realidade non é necesario facer a multiplicación polo valor $e^{-it_l a}$, xa que como $\zeta_l = e^{it_l a}\mathrm{DFT}(\boldsymbol{\xi})_l\varphi_{K'}(t_l h)$, resulta que $Z_l = \mathrm{DFT}(\boldsymbol{\xi})_l\varphi_{K'}(t_l h)$.

No caso da estimación da derivada da densidade ao cadrado, empregando de novo a aproximación rápida de $\varphi_n(t_l)$, pódese calcular

$$\gamma_l = \varphi_{(f'_{nh})^2}(t_l) = \varphi_n(t_l)\varphi_{(K')^2}(t_l h), \quad l = 1, \dots, M$$

polo que $\frac{1}{h^3}\gamma_l$ é o mesmo que avaliar $\varphi_{(f'_{nh})^2}(t_l)$ nunha grella de M puntos equidistantes como mencionamos no anterior caso. Cumpríndose as mesmas propiedades que na estimación da derivada da densidade, e utilizando a aproximación da integral por rectángulos, séguese que

$$\begin{aligned} (f'_{nh})^2(r_k) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itr_k}\varphi_{f'_{nh}}(t)dt \approx \frac{1}{2\pi h^3}\sum_{l=1}^{M}\Delta\gamma_l e^{-it_l r_k} + \frac{1}{2\pi h^3}\sum_{l=2}^{M}\Delta\overline{\gamma_l}e^{-it_l r_k} \\ &= \frac{1}{2\pi h^3}\sum_{l=1}^{M}\Delta\gamma_l e^{-it_l r_k} + \frac{1}{2\pi h^3}\sum_{l=1}^{M}\Delta\overline{\gamma_l}e^{-it_l r_k} - \frac{1}{4\sqrt{\pi}\delta M h^3} \\ &= \frac{2}{\delta M h^3}\mathrm{Re}\left\{\sum_{l=1}^{M}\gamma_l e^{-it_l r_k}\right\} - \frac{1}{4\sqrt{\pi}\delta M h^3} \\ &= \frac{2}{\delta M h^3}\mathrm{Re}\left\{\sum_{l=1}^{M}[\gamma_l e^{-it_l a}]e^{-2\pi i(l-1)(k-1)/M}\right\} - \frac{1}{4\sqrt{\pi}\delta M h^3}. \end{aligned}$$

Seguindo o mesmo procedemento que no anterior caso, e tomando $\boldsymbol{Y}$ o vector cuxas coordenadas son as de $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_M)$ multiplicadas polo valor de $e^{-it_l a}$ correspondente, chegamos a que o primeiro sumando da expresión anterior non é mais que $2/(\delta M h^3)$ pola parte real da DFT inversa de $\boldsymbol{Y}$.

Capítulo 3

Estudo de simulación

Neste capítulo imos realizar un estudo de simulación no que imos comparar os resultados ofrecidos polos diferentes cuantís propostos por Chaudhuri e Marron (1999). Ademais faremos unha análise da taxa de cobertura ofrecida polo cuantil q_1 (o cuantil Gaussiano), en comparación coas novas propostas presentes na Sección 2.3, as cales, da mesma maneira, obteñen un cuantil sen ter en conta a dependencia entres os diferentes valores de x e h, pero por medio do procedemento bootstrap. Para iso simularemos mostras xeradas a partires das densidades de Marron e Wand (1992) (presentes no Apéndice B), habitualmente utilizadas para estudos de simulación na análise da estimación da función de densidade. Realizaremos o estudo de simulación para diferentes tamaños de mostra n, para diferentes funcións de densidade, e simulando un total de N mostras para cada caso particular.

En primeiro lugar obteremos un mapa SiZer promedio e un mapa de acerto que nos permitirán ver en que zonas do mapa cada un dos cuantís propostos por Chaudhuri e Marron (1999) está a identificar de forma correcta patróns de crecemento/decrecemento. Tamén obteremos un mapa de erro que nos indicará que cuantil está a cometer un maior erro e en que zonas do mapa. E Ademais, para comparar estes catro cuantís, por último estudaremos os resultados obtidos a través do MoSiZer, ferramenta que nos permitirá ver dunha forma clara que modas está a estimar cada un dos cuantís sobre o mapa SiZer.

Para levar a cabo a comparación da taxa de cobertura entre os cuantís q_1 e os cuantís bootstrap propostos (q_5, q_6, q_7 e q_8) analizaremos a cobertura ofrecida por cada cuantil en función do valor do parámetro de suavizado h. Isto permitiranos ver se o cuantil ofrece unha taxa acertada entorno ao valor $(1-\alpha)$, independentemente do valor de h.

3.1. Taxa de acerto e taxa de erro do SiZer

O primeiro estudo de simulación realizado está centrado en analizar a taxa de acerto e de erro dos diferentes cuantís expostos sobre as cores do mapa SiZer. Basicamente, estudaremos as zonas significativas obtidas do mapa, analizando se a forma da curva se está a estimar de xeito correcto. Mais concretamente imos obter un mapa SiZer promedio que represente a media de N SiZers a partires de N mostras simuladas. Sobre esas mesmas mostras obteremos un mapa de acerto, que será representativo das zonas nas que cada cuantil está a identificar de forma correcta patróns de crecemento/decrecemento, e por último un mapa de erro o cal nos dará a información do erro que está a cometer cada cuantil nas cores do SiZer.

3.1.1. Implementación

O centro da análise do SiZer é a familia de convolucións f_h (2.1) para diferentes valores do parámetro de suavizado h. Neste estudo de simulación comparamos a cor do mapa SiZer obtido sobre unha mostra xerada a partires dunha función de densidade f, coa cor do mapa SiZer obtido directamente a partir da función f_h.

Por exemplo, dada a función de densidade *Bimodal* (#6) proposta por Marron e Wand (1992) mostrada na Figura B.6 (esquerda) do Apéndice B, o mapa SiZer desa función de densidade obtido a partir da convolución seguindo a ecuación (2.1), para un rango discreto de valores para h, é o mostrado no centro da mesma Figura. Nesta ocasión, en lugar de facer inferencia a partires dos intervalos de confianza de $\hat{f}'_h$, representando de cor azul (vermella) aquelas zonas onde a curva é significativamente crecente (decrecente), represéntase directamente en cor azul ou vermella, aquelas zonas onde f'_h ten valor positivo ou negativo respectivamente.

O obxecto do estudo é comparar o mapa SiZer obtido para unha mostra calquera co mapa SiZer obtido a partir de f_h. Neste analizaranse mostras xeradas a partir de diferentes funcións de densidade de Marron e Wand, en cada caso tomando distintos tamaños de mostra n, e xerando para cada un deles un total de N mostras.

SiZer promedio

En primeiro lugar obteremos para cada caso (para unha función de densidade e tamaño de mostra fixo) un mapa SiZer que represente a media dos N SiZers obtidos. Traballaremos cunha escala `rgb` que nos permita representar este mapa, tomando unha gama de cores que vai dende a cor azul (crecente) á cor vermella (decrecente).

A descrición `rgb` (do inglés *Red, Green, Blue*; "vermello, verde, azul") dunha cor fai referencia a súa composición da intensidade das cores primarias coas que se forma: o vermello, o verde e o azul. Este é un modelo de cor baseado no que se coñece como síntese aditiva, co que é posible representar unha cor pola mestura por adición das tres cores de luz primarios. Para indicar con que proporción se mestura cada cor, asignámoslle un valor a cada unha das cores primarias, e así, por exemplo, o valor 0 significa que non intervén na mestura, e na medida que ese valor aumenta, aportará máis intensidade á mestura. O que coñecemos como píxel é en realidade un conxunto de tres puntos, un vermello, un verde e un azul, cada un dos cales brilla con una determinada intensidade.

A función do mapa SiZer que implementamos en `R` está a empregar tres cores para representar onde a curva $\hat{f}_h$ é significativamente crecente, significativamente decrecente ou onde non é significativamente distinta de cero. Para representar onde $\hat{f}_h$ é signicativamente crecente empregase a cor azul, a cal nunha escala `rgb` de cero a un, se está a empregar a cor azul pura `rgb(0,0,1)`. Mentres que por outra banda, para representar as zonas onde a curva $\hat{f}_h$ é significativamente decrecente se emprega a cor vermella pura `rgb(1,0,0)`. Finalmente, as zonas onde a curva non é significativa de que $\hat{f}'_h$ sexa distinto de cero son representadas de cor púrpura. Un púrpura o cal representa perfectamente o punto intermedio entre a cor azul e a cor vermella nunha escala `rgb`, tomando o valor `rgb(0.5,0,0.5)`. Isto será de vital importancia xa que nos permite traballar nunha escala lineal con esa gama de cores para poder representar o SiZer promedio.

No SiZer obtido para cada mostra faremos un rexistro global da cor obtida en cada punto (x, h). Como aquelas zonas onde se está a facer inferencia poden ser representadas de tres cores distintas (azul, púrpura ou vermello) o rexistro será feito nun *array* A de $3 \times n(x) \times n(h)$, sendo $n(x)$ o número de valores distintos da grella x e $n(h)$ o número de valores distintos do parámetro de suavizado h, xa que este *array* será empregado para obter tanto o mapa promedio, como o mapa de acerto e o mapa de erro. Isto leva a que traballemos con tres matrices distintas, onde serán rexistradas as tres cores por separado seguindo o seguinte algoritmo:

Algoritmo 7 *Array A*

O rexistro das cores de cada punto (x, h) do mapa SiZer é realizado nun *array* A de dimensións $3 \times n(x) \times n(h)$. É dicir, 3 matrices nas que se realizarán os rexistros das cores azul, púrpura e vermella respectivamente. Con isto:

1: Obter o mapa SiZer de $\boldsymbol{X} = (X_1, \ldots, X_n)$ unha m.a.s. de X con función de densidade f descoñecida. Para cada punto (x, h) do mapa SiZer:

- Se o punto (x, h) do mapa SiZer é azul: $A(1, x, h) = A(1, x, h) + 1$.
- Se o punto (x, h) do mapa SiZer é púrpura: $A(2, x, h) = A(2, x, h) + 1$.
- Se o punto (x, h) do mapa SiZer é vermello: $A(3, x, h) = A(3, x, h) + 1$.

2: Repetir o punto 1 un total de N veces coas N mostras.

Con isto, empregando o *array* A, obtemos o SiZer promedio seguindo o seguinte algoritmo:

Algoritmo 8 SiZer promedio

1: Definir a cor de cada punto (x, h) utilizando o valor do punto (x, h) das tres matrices de A.

2: Asignar valores de cor aos puntos (x, h) e empregando a escala `rgb` como segue:

- Cor vermello: $\texttt{r}(x, h) = \frac{A(3,x,h)+0.5 \cdot A(2,x,h)}{A(1,x,h)+A(2,x,h)+A(3,x,h)}$
- Cor verde: $\texttt{g}(x, h) = 0$
- Cor azul: $\texttt{b}(x, h) = \frac{A(1,x,h)+0.5 \cdot A(2,x,h)}{A(1,x,h)+A(2,x,h)+A(3,x,h)}$

Polo tanto:

- Se $A(1, x, h) + A(2, x, h) + A(3, x, h) < m$ (cor gris):

$$\text{cor}(x, h) = \texttt{rgb}(0.5, 0.5, 0.5)$$

- Se $A(1, x, h) + A(2, x, h) + A(3, x, h) \geq m$:

$$\text{cor}(x, h) = \texttt{rgb}(\texttt{r}(x, h), \texttt{g}(x, h), \texttt{b}(x, h))$$

Onde m é o parámetro que limita as zonas onde o SiZer promedio vai ser pintado de cor gris en referencia ao valor n_0 empregado por Chaudhuri e Marron (1999).

O que vai a ocorrer naqueles puntos (x, h) onde o SiZer tende a estimar unha cor vermella (azul), e que o SiZer promedio vai a tender a pintalos de cor vermella (azul), xa que estamos aumentando a intensidade da cor primaria vermella (azul) neses puntos. Por outra banda, vai a ocorrer que a cor tenda a ser púrpura naquelas zonas onde se estima cunha proporción semellante a cor azul e vermella do SiZer, ou onde a estimación do SiZer sexa púrpura, xa que neses puntos aumentaremos con igual proporción a intensidade da cor primaria vermella e azul. Mentres que se representarán de cor gris aqueles puntos onde non se chegou a realizar inferencia un número de veces superior ou igual ao valor m, debido ao valor de n_0 na estimación do SiZer.

Mapa de acerto

En segundo lugar, obteremos un mapa de acerto no cal imos representar utilizando a mesma gama de cores as zonas onde se está a estimar de forma correcta a cor do SiZer. Empregando o mesmo *array* A, imos obter a porcentaxe de veces que a cor de cada punto (x, h) coincide coa verdadeira cor do SiZer de f_h.

As simulacións van a ser feitas en todos os casos empregando as densidades propostas por Marron e Wand, é dicir, mestura de normais, provocando que en todo momento a forma da curva f_h sexa crecente ou decrecente, sen atopar en ningún momento zonas onde a curva sexa plana. Isto vai a implicar que o SiZer de f_h estea composto en todo momento unicamente por puntos azuis e vermellos, provocando que o algoritmo empregado para obter o mapa de acerto sexa o seguinte:

Algoritmo 9 Mapa de acerto

1: Definir a cor de cada punto (x, h) utilizando o valor do punto (x, h) das tres matrices de A.

2: Asignar valores de cor aos puntos (x, h) e empregando a escala `rgb` como segue:

- Se o punto (x, h) do SiZer de f_h é azul:
  - Cor vermello: $\mathtt{r}(x, h) = \frac{A(3,x,h)+0.5\cdot A(2,x,h)}{A(1,x,h)+A(2,x,h)+A(3,x,h)}$
  - Cor verde: $\mathtt{g}(x, h) = 0$
  - Cor azul: $\mathtt{b}(x, h) = \frac{A(1,x,h)+0.5\cdot A(2,x,h)}{A(1,x,h)+A(2,x,h)+A(3,x,h)}$
- Se o punto (x, h) do SiZer de f_h é vermello:
  - Cor vermello: $\mathtt{r}(x, h) = \frac{A(1,x,h)+0.5\cdot A(2,x,h)}{A(1,x,h)+A(2,x,h)+A(3,x,h)}$
  - Cor verde: $\mathtt{g}(x, h) = 0$
  - Cor azul: $\mathtt{b}(x, h) = \frac{A(3,x,h)+0.5\cdot A(2,x,h)}{A(1,x,h)+A(2,x,h)+A(3,x,h)}$

Polo tanto:

- Se $A(1, x, h) + A(2, x, h) + A(3, x, h) < m$ (cor gris):

$$\text{cor}(x, h) = \mathtt{rgb}(0.5, 0.5, 0.5)$$

- Se $A(1, x, h) + A(2, x, h) + A(3, x, h) \geq m$:

$$\text{cor}(x, h) = \mathtt{rgb}(\mathtt{r}(x, h), \mathtt{g}(x, h), \mathtt{b}(x, h))$$

Onde m é o parámetro que limita as zonas onde o mapa de acerto vai ser pintado de cor gris.

Neste ocasión, como se pode apreciar a través do algoritmo, o acerto vai ser representado coa cor azul, xa que estamos a aumentar a intensidade desta cor primaria do punto (x, h) cando a cor dese punto do SiZer de f_h coincide coa da estimación do SiZer. O erro vai ser representado coa cor vermella, xa que aumenta a intensidade desta cor primaria cando o SiZer de f_h e a estimación do SiZer nun punto teñen cores distintas, mentres que a cor púrpura representará un punto intermedio onde na maioría dos casos será indicativo de que $\hat{f}_h'$ non é significativamente distinto de cero nese punto.

Mapa de erro

O mapa de acerto non será esclarecedor acerca do erro que está a cometer cada cuantil nas cores do mapa SiZer. Cabe destacar que en ningún caso un erro está definido por unha cor púrpura do mapa SiZer estimado en calquera punto (x, h) cando a verdadeira cor do mapa SiZer de f_h nese mesmo punto é azul ou vermella, se non por puntos onde se está a obter a cor azul na estimación do mapa SiZer cando a verdadeira cor do SiZer de f_h é vermella, ou viceversa. A cor púrpura é representativa de que non hai suficiente información para concluír que en calquera punto (x, h) de $\hat{f}_h$ a curva sexa crecente ou decrecente, polo tanto non é indicativo de que se estea a cometer un erro na estimación da forma da curva.

Tendo en conta que no peor dos casos (cuantil q_1), imos obter intervalos de confianza cun nivel de significación α, se estes intervalos son obtidos de forma correcta, van a tender a conter o verdadeiro valor nunha proporción do nivel $(1 - \alpha)$. Por exemplo, supoñendo que estamos a estimar o intervalo de confianza dun punto (x, h) de $\hat{f}'_h$ onde a curva f_h é crecente (é dicir, f'_h positivo), este intervalo de confianza vai a conter o verdadeiro valor de f'_h nunha proporción do nivel $(1-\alpha)$. Provocando que nese mesmo nivel de proporción ese punto no SiZer sexa pintado coa cor verdadeira de f'_h (azul), ou que o intervalo de confianza conteña tamén o valor cero (púrpura). Por outra banda, existe a posibilidade de que o intervalo de confianza aínda non contendo o verdadeiro valor de f'_h, se atope por encima dese valor, provocando desta maneira, que a cor do mapa SiZer nese punto siga sendo correcta. Ademais, supoñendo o caso oposto onde intervalo de confianza se atopa por debaixo do verdadeiro valor, este ou ben pode seguir contendo unicamente valores positivo, provocando que a cor siga sendo azul, ou pode conter o valor cero, dando lugar a non obter información significativa e polo tanto pintando o mapa SiZer de cor púrpura nese punto.

Ao que nos leva todo isto, é que na maioría dos casos a proporción de erros na cor dun punto calquera (x, h) do SiZer vai ser inferior, no peor dos casos (cuantil q_1), ao nivel de significación α. Tendo en conta unha taxa tan baixa de erro en cada punto, o mapa de acerto vai a estar unicamente representando por cores azuis, indicativo de onde se están a identificar patróns de crecemento/decrecemento de forma correcta, e por cores púrpuras, indicativo de onde non se están a identificar.

Para representar as zonas do mapa onde se está a cometer erro imos obter un mapa de erro que nos permita visualizar isto dunha forma mais clara. Sabedores de que o erro da cor do SiZer cometido en cada punto (x, h) vai ser inferior ao nivel de significación, e tomando este cun valor $\alpha = 0.05$, imos supoñer que na maioría dos puntos o erro da cor estimada non é maior a unha proporción do 0.04. Nesta ocasión, a gama de cores empregada para representar a porcentaxe de erro vai a ir dende a cor verde pura, representativa de que a proporción de erro e igual a cero nun punto dado, ata a cor vermella pura, representando que a proporción de erro nun punto e maior ou igual ao nivel 0.04.

Desta forma, sabendo que o mapa SiZer de f_h só vai a conter cores azuis e vermellos, o mapa de erro, empregando de novo o *array* A, vai ser obtido segundo o seguinte algoritmo:

Algoritmo 10 Mapa de erro

1: Sendo $p_{erro}(x,h)$ a proporción de erro en cada punto:

- Se o punto (x,h) do SiZer de f_h é azul:

$$p_{\text{erro}}(x,h) = \frac{A(3,x,h)}{N} \tag{3.1}$$

- Mentres que se o punto (x,h) do SiZer de f_h é vermello:

$$p_{\text{erro}}(x,h) = \frac{A(1,x,h)}{N} \tag{3.2}$$

2: Obtendo a cor de cada punto do mapa de erro da seguinte maneira:

- Cor vermello: $\mathtt{r}(x,h) = \frac{p_{\text{erro}}(x,h)}{0.04}$
- Cor verde: $\mathtt{g}(x,h) = \frac{0.04 - p_{\text{erro}}(x,h)}{0.04}$
- Cor azul: $\mathtt{b}(x,h) = 0$

Polo tanto:

$$\text{cor}(x,h) = \mathtt{rgb}(\mathtt{r}(x,h), \mathtt{g}(x,h), \mathtt{b}(x,h))$$

Desta maneira, estamos a medir a proporción de veces que nun punto dado (x,h) a estimación da cor SiZer sobre unha mostra é diferente á cor do SiZer de f_h. A cor tenderá a ser vermella cando os valores da proporción sexan próximos ao valor 0.04 (tendo en conta que seleccionaremos un nivel de significación $\alpha = 0.05$), mentres que tenderá a ter a cor verde cando a proporción de erro se aproxime a 0.

3.1.2. Resultados

A simulación foi levada a cabo utilizando mostras simuladas a partir das densidades *Strongly Skewed* (#3), *Bimodal* (#6) e *Sep. Bimodal* (#7) de Marron e Wand (1992), as cales podemos ver sobre as Figura B.3, B.6 e B.7 do Apéndice B xunta o mapa SiZer e MoSiZer (o cal veremos na seguinte Sección) de f_h. Os tamaños de mostra empregados foron $n = \{50, 200, 500\}$, e simulamos un total de $N = 1000$ mostras para cada caso. O nivel de significación empregado é de $\alpha = 0.05$, e ademais, sobre cada un dos mapa SiZer a grella de puntos x estivo comprendida dentro do intervalo $[-4, 4]$, onde se utilizou un total de $n(x) = 512$ puntos. Por outra banda, o número de valores do parámetro de suavizado empregado foi dun total de $n(h) = 151$ puntos, e o parámetro m que limita a zona gris do mapa foi do valor $m = 50$.

Debido á similitude dos resultados obtidos entre as diferentes funcións de densidades propostas, centrarémonos en expoñer as conclusións obtidas acerca das simulacións levadas a cabo coa densidade *Strongly Skewed* (#3) de Marron e Wand (1992), a cal vemos reflectida na Figura B.3 (esquerda) do Apéndice B, e que conta tan só con unha moda.

Dirixíndonos directamente aos resultados obtidos para o SiZer promedio, podemos observar a través da Figura 3.1, onde a simulación foi realizada para un tamaño de mostra de $n = 50$, como a taxa de cobertura das zonas de crecemento e de decrecemento sobre as cores do mapa SiZer é maior para o cuantil q_1 que para o resto de cuantís, no cal os intervalos de confianza son mais curtos, debido a que os demais cuantís contemplan a dependencia entre os puntos da grella x (e no caso do cuantil q_4 tamén a dependencia dos diferentes valores do parámetro de suavizado), contendo o valor cero un número

inferior de veces. Como antes mencionamos, no peor dos casos (cuantil q_1) o intervalo de confianza vai tender a conter o verdadeiro valor (sempre e cando os intervalos de confianza estean obtidos de forma axeitada) nun 95 % das ocasións, o que provoca que o SiZer promedio na súa medida tenda a aproximarse ao SiZer de f_h. Ademais, aínda que difícil de apreciar, a taxa de cobertura da verdadeira cor do mapa SiZer de f_h é maior para o cuantil q_2 que para o cuantil q_3. Isto vese dunha forma mais clara na Figura 3.4, onde se está a ver as zonas onde a cor do mapa SiZer cubre a verdadeira cor. Por último, o cuantil q_4 cubre dunha forma menos eficiente a verdadeira cor do SiZer de f_h.

Por outra banda, podemos ver reflectido tanto a través das Figuras do SiZer promedio 3.2 e 3.3 como dos mapas de acerto das Figuras 3.5 e 3.6 para as simulacións realizadas cos tamaños de mostra $n = 200$ e $n = 500$ respectivamente, como a medida que o tamaño de mostra medra, todos os cuantís tenden a cubrir en maior medida a verdadeira cor do mapa SiZer de f_h.

Estas Figuras parecen dar a entender que o cuantil q_1 tende a mostrar dunha maneira mais clara as características que podemos chegar a observar nun mapa SiZer. Observando a Figura 3.7 podemos ver como para todos os cuantís, a maioría dos erros cometidos estimando a verdadeira cor do mapa SiZer se producen na fronteira do cambio de cor vermella a azul ou viceversa. Isto é debido a que nesas zonas o valor de f'_h está próximo ao valor cero, e polo tanto, cando os intervalos de confianza estimados para eses puntos non conteñen o verdadeiro valor de f'_h pódese dar o caso, en maior medida, de que o intervalos de confianza conteñan tan só valores do signo oposto. Ademais, pódese ver como os cuantís q_2, q_3 e q_4 apenas cometen erros fóra desa fronteira. Por outra banda, o cuantil q_1 si parece mostrar unha porcentaxe mais alta de erros fóra da fronteira, contradicindo o que parecían mostrar tanto o mapa do SiZer promedio como o mapa de acerto acerca da boa cobertura das cores do cuantil q_1.

Observando as Figuras 3.8 e 3.9 podemos ver como a medida que o tamaño de mostra medra ($n = 200$ e $n = 500$ respectivamente) o erro das cores no SiZer cometido polos cuantís q_2, q_3 e q_4 tende a diminuír. Non obstante, non parece ocorrer o mesmo para o cuantíl q_1, onde se pode apreciar perfectamente unha taxa de erro maior nas zonas inferiores do mapa SiZer. Como mencionamos anteriormente, En Chaudhuri e Marron (1999) suxiren que o cuantil Gaussiano podería ser pouco axeitado para zonas do mapa onde o parámetro ESS é demasiado baixo ($ESS(x, h) < n_0 = 5$), e aínda que neste estudo de simulación non se está a facer inferencia para aqueles puntos (x, h) onde $ESS(x, h) < 5$, a taxa de erro para os tamaños de mostra $n = 200$ e $n = 500$ en zonas baixas do mapa parece ser mais elevada. Na seguinte Sección faremos un estudo de simulación onde discutiremos acerca do papel que segue o parámetro n_0.

Aínda que a simulación foi levada a cabo empregando un maior número de densidades propostas por Marron e Wand, todas elas parecen mostrar os mesmos resultados. É dicir, unha maior taxa de cobertura da verdadeira cor do SiZer sobre os puntos (x, h) do cuantil q_1; a taxa de erro de cada cuantil é maior en zonas próximas á fronteira do cambio de cor (onde o valor de f'_h é próximo a cero); e o mais importante, que a taxa de erro para o cuantil q_1 é maior en zonas inferiores do mapa SiZer (e dicir, onde o parámetro de suavizado h é menor), a medida que o tamaño de mostra crece, dando lugar a dúbidas da correcta aproximación do parámetro n_0. Os resultados obtidos para o estudo de simulación coas mostras xeradas a partir das densidades *Bimodal* (#6) e *Sep. Bimodal* (#7) móstranse nas Figuras do Apéndice C.

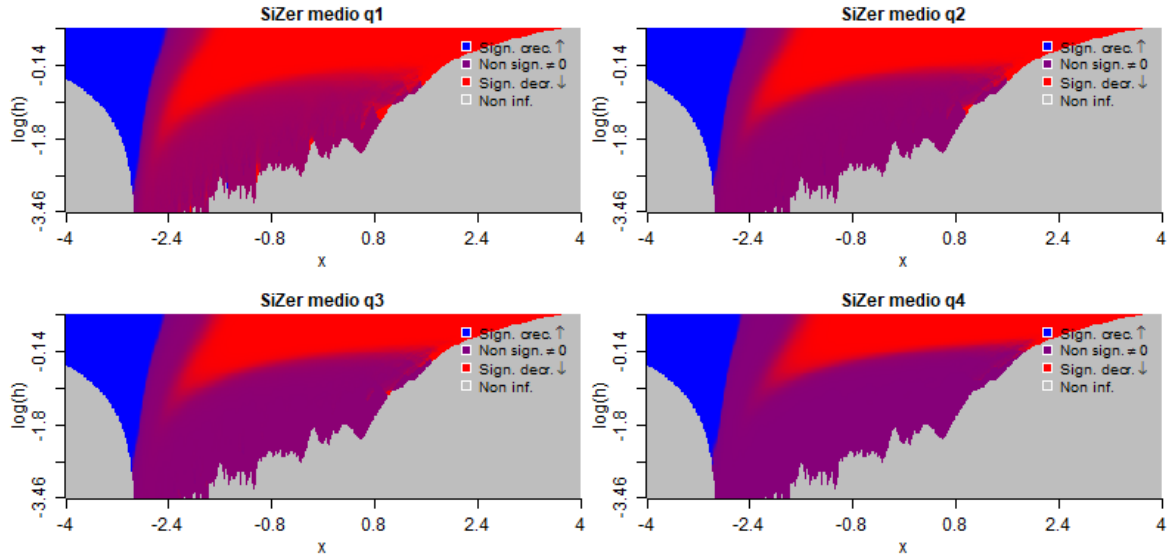

Figura 3.1: Estimación do SiZer promedio para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 50$, xeradas a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

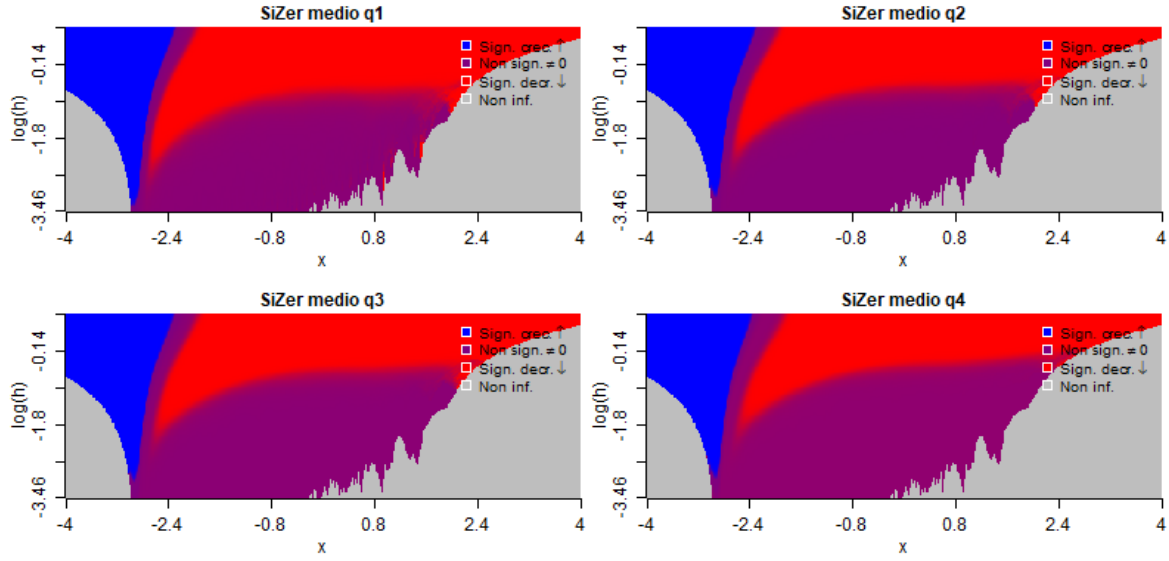

Figura 3.2: Estimación do SiZer promedio para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 200$, xeradas a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

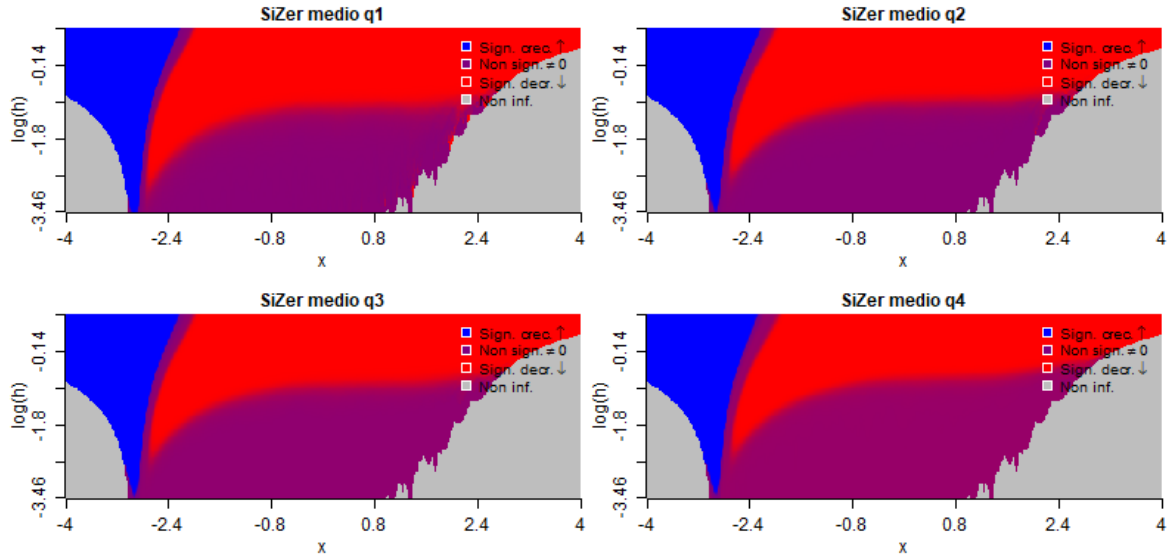

Figura 3.3: Estimación do SiZer promedio para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 500$, xeradas a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

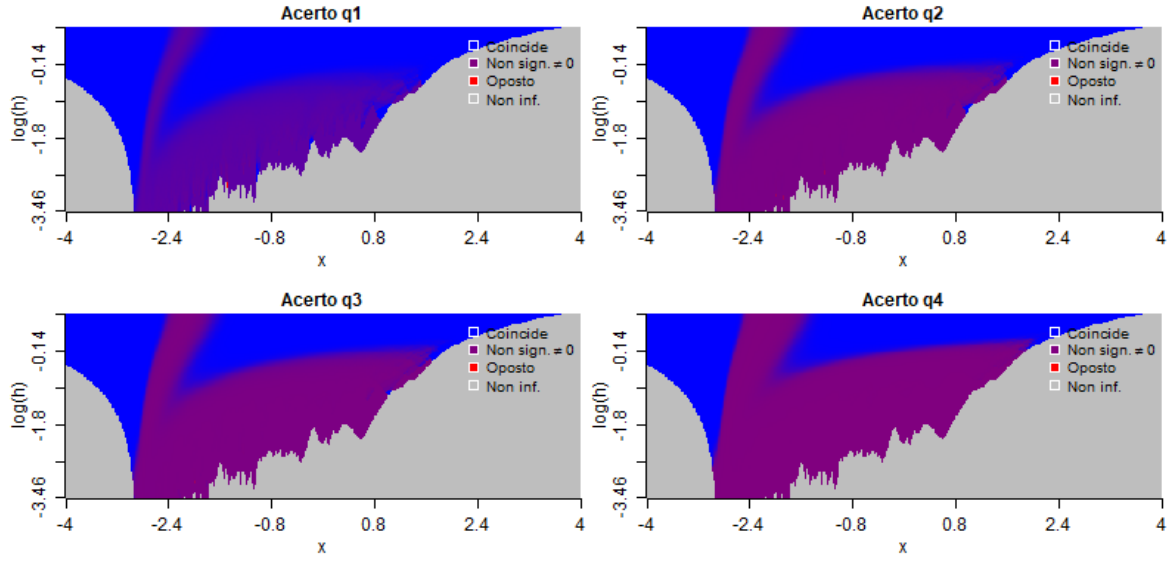

Figura 3.4: Mapa de acerto para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 50$, xeradas a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

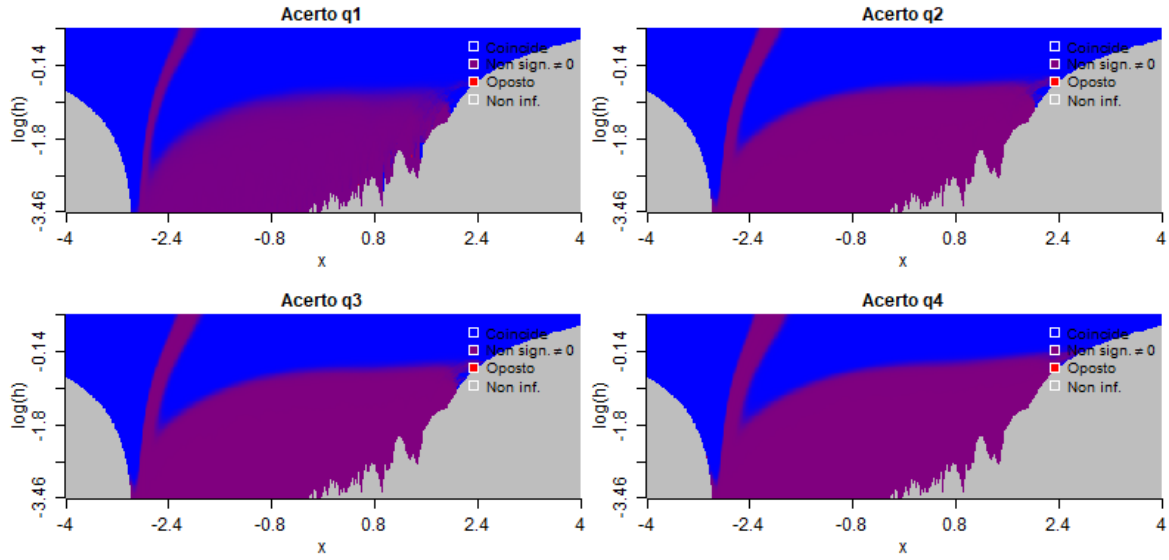

Figura 3.5: Mapa de acerto para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 200$, xeradas a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

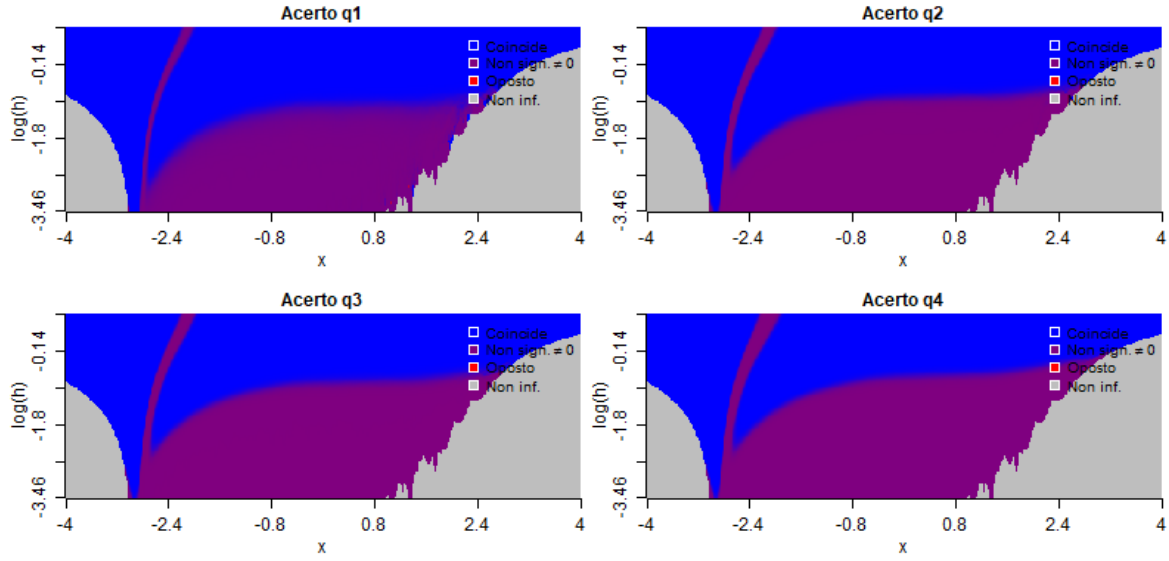

Figura 3.6: Mapa de acerto para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 500$, xeradas a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

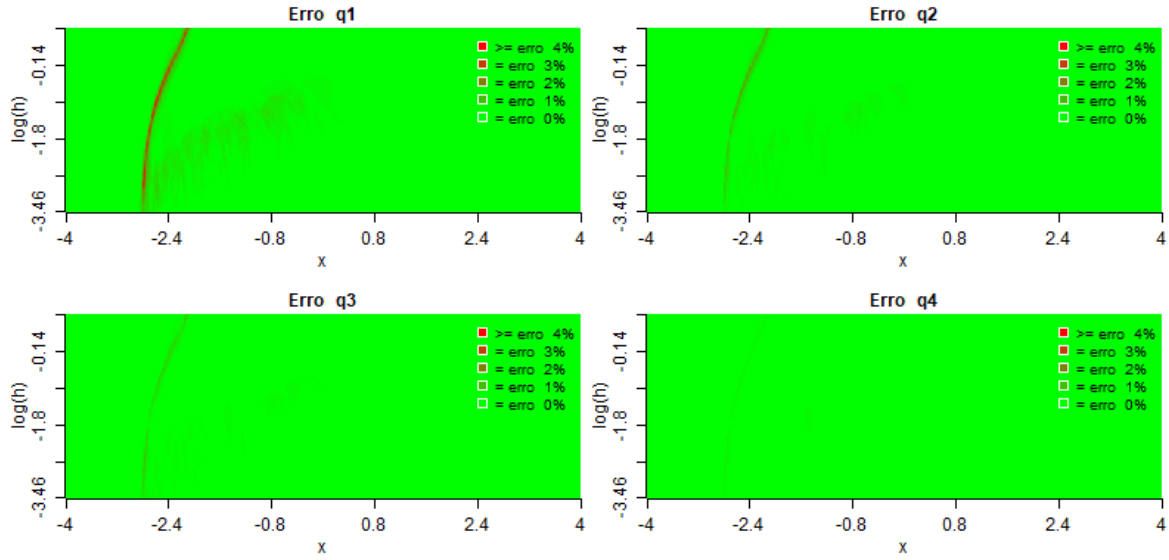

Figura 3.7: Mapa de erro para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 50$, xeradas a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

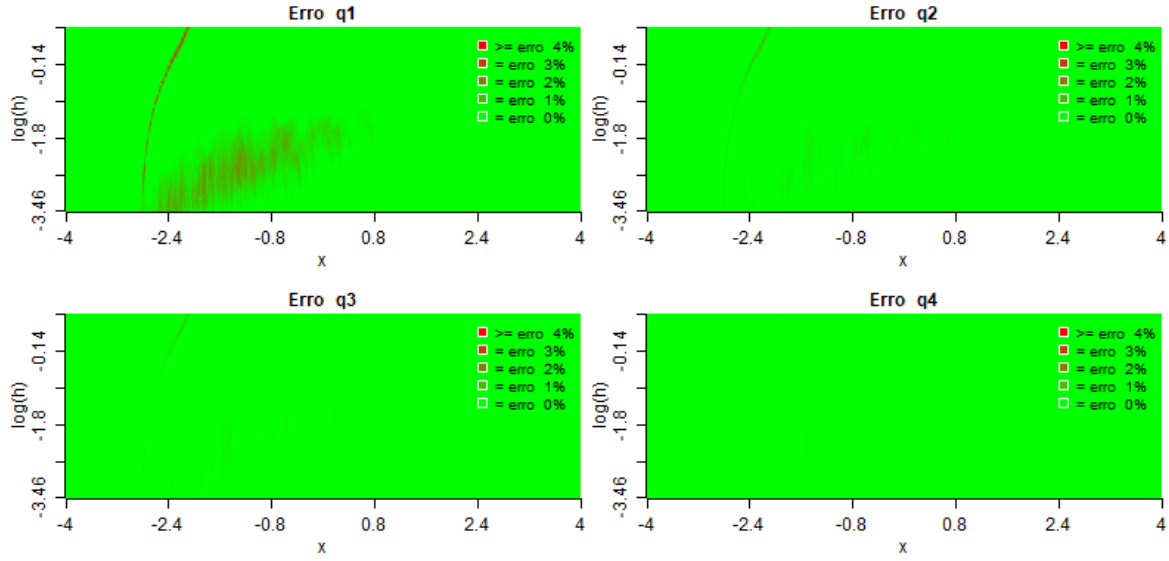

Figura 3.8: Mapa de erro para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 200$, xeradas a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

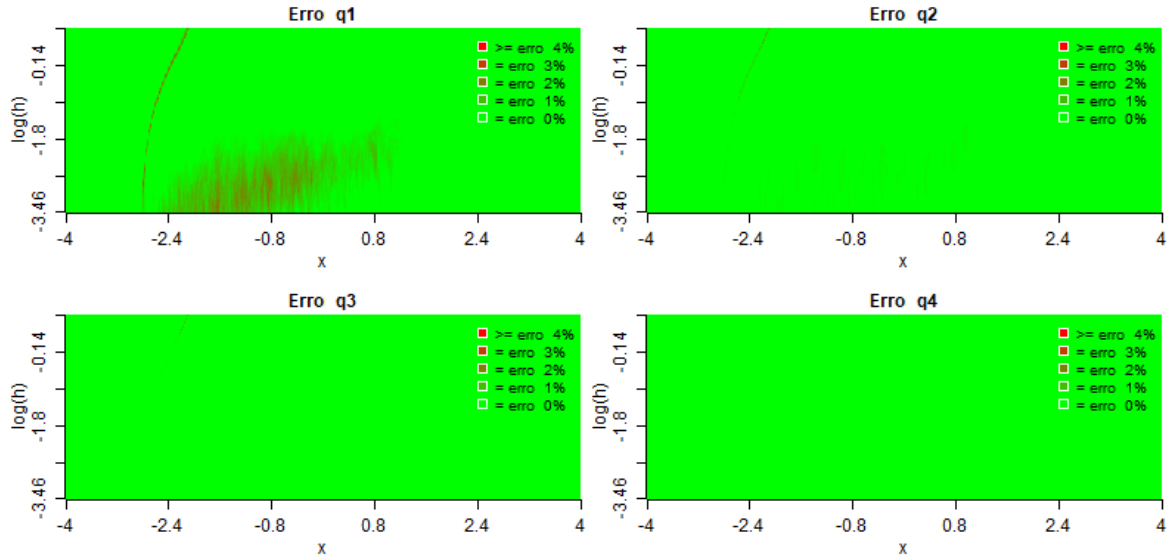

Figura 3.9: Mapa de erro para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 500$, xeradas a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

3.2. Análise das modas

O obxectivo do SiZer e facer inferencia acerca da forma que segue a función f_h, e máis concretamente informar da localización das modas, polo que estudar o erro que está a cometer cada SiZer na estimación da cor, non nos permite obter unha conclusión clara acerca de que cuantil está a obter mellores resultados. Nesta Sección imos realizar un estudo de simulación que nos permita analizar como os cuantís propostos por Chaudhuri e Marron (1999) son capaces de detectar as modas dunha función de densidade a través dun mapa que nos permita visualizar como cada cuantil está a estimar as modas. A ferramenta empregada para isto será o MoSiZer, a cal nos permitirá visualizar dunha forma mais clara a localización e o número de modas estimadas por cada SiZer.

Ademais, discutiremos acerca da influencia do parámetro n_0, que como veremos, vai afectar no número de modas estimadas segundo o valor escollido.

3.2.1. Implementación

De novo, estudaremos diferentes casos que nos permita ver as principais diferenzas entre os catro cuantís propostos por Chaudhuri e Marron (1999). Imos simular mostras a partir das densidades propostas por Marron e Wand (1992), con diferentes tamaños de mostra n, e simulando un total de N mostras para cada caso particular.

O obxectivo vai ser estudar como cada cuantil identifica as modas a través da ferramenta SiZer, debuxando un mapa de modas (MoSiZer). Por exemplo, tomando nesta ocasión a densidade *Trimodal* (#9) proposta por Marron e Wand (1992) mostrada na Figura B.9 (esquerda) do Apéndice C, o SiZer obtido a través de f_h é o mostrado no centro da mesma Figura, mentres que o MoSiZer se atopa na parte dereita da Figura. Neste mapa vemos identificado de forma clara onde se están a estimar as modas con respecto ao SiZer, onde ademais, se empregan diferentes cores para mostrar o número de modas que son detectadas con respecto ao parámetro de suavizado h sobre o eixe y. Con isto, o obxectivo do estudo vai ser xerar un MoSiZer como o mostrado na Figura B.9 (dereita), obtido a través das N mostras xeradas para cada caso particular.

Dado un mapa SiZer, e analizando cada fila do mapa por separado (é dicir, as cores obtidas a través dos intervalos de confianza de $\hat{f}'_h$ con h fixo), imos entender que a localización dunha moda se encontra no punto intermedio que hai entre o último punto dunha zona de puntos azuis, e o primeiro punto dunha zona de putos vermellos. Existen catro posibilidades: A primeira (pouco probable), que un punto azul veña seguido dun punto vermello. Nese caso podemos interpretar que a moda se encontra sobre calquera dos dous puntos, pero por defecto imos entender que se encontra sobre o punto vermello. Como segundo caso pode ocorrer que un punto azul veña seguido por unha zona púrpura e posteriormente un punto vermello, polo que imos interpretar que a moda se encontra no punto medio da zona púrpura. Ademais, pódese dar a posibilidade de que un punto azul veña seguido dunha zona gris e posteriormente un punto vermello, no que tamén imos interpretar que a moda se encontra no punto medio da zona gris. E por último, que un punto azul veña seguido dunha zona con puntos púrpuras e grises, seguido de novo por un punto vermello, no que seguiremos co mesmo procedemento de tomar o punto intermedio da zona onde non se están a obter conclusións.

Posto que imos traballar cun número de mostras N grande, o MoSiZer onde se vai a rexistrar a localización das modas pode resultar difícil de interpretar debido ao solapamento que se pode producir entre os resultados das diferentes mostras. Pois ben, para obter unha interpretación máis clara, imos a resaltar as zonas onde se están a estimar un número maior de modas. Con isto, o rexistro da localización das modas sobre o MoSiZer vai ser realizado nunha matriz A coas mesmas dimensións que as do mapa SiZer, obtendo o mapa segundo o seguinte algoritmo:

Algoritmo 11 MoSiZer

O rexistro das modas obtidas a partires da estimación do mapa SiZer é realizado nunha matriz A de dimensións $n(x)\times n(h)$ cuxo valor inicial é 0 en todos os seus puntos. No bucle tamén se empregará unha matriz B coas mesmas dimensións e con valor inicial 0. Con isto:

1: Obter o mapa SiZer de $\boldsymbol{X} = (X_1, \dots, X_n)$ unha m.a.s. de X con función de densidade f descoñecida.

2: Para cada fila do SiZer con h fixo:

- Obter as modas da fila, interpretando como unha moda o punto intermedio entre o último punto dunha zona de puntos azuis e o primeiro punto dunha zona de puntos vermellos. Posteriormente, contaremos o número de modas, e cada unha delas terá un identificador cun enteiro igual ao número de modas total, que rexistraremos na matriz B. Por exemplo, se nunha fila detectamos dúas modas nos puntos (x_1, h) e (x_2, h), esas modas van a ter un identificador co número enteiro 2 sobre a matriz B neses mesmos puntos.

3: Copiar os valores da matriz B sobre a matriz A da seguinte maneira:

- Se $A(x,h) < B(x,h)$:
$$A(x,h) = B(x,h)$$
- Se $A(x,h) \geq B(x,h)$:
$$A(x,h) = A(x,h)$$

4: Repetir os pasos anteriores N veces.

Coa matriz A obtida no anterior algoritmo obtemos o MoSiZer desexado no exercicio de simulación, o cal se verá representado sobre un mapa sen mais que empregando unha cor distinta para cada identificador enteiro da matriz. Debido á lóxica empregada de conservar os números enteiros con maior valor sobre a matriz, lograremos resaltar aquelas zonas onde se está a detectar un maior número de modas.

Aínda que o MoSiZer nos da unha idea de como cada cuantil está a detectar as modas, debido ao solapamento entre as diferentes mostras, é difícil identificar cal é o número máximo de modas detectadas sobre cada mostra. Por iso, ademais de facer a representación das modas que estamos a detectar, imos cuantificar cal é o número máximo de modas que estamos a detectar sobre cada mostra e expoñer unha porcentaxe dos resultados sobre as N mostras. Iso daranos unha idea mais aproximada do que realmente esta a ocorrer.

3.2.2. Resultados

Para este primeiro caso, onde comparamos os diferentes resultados obtidos polos catro cuantís propostos por Chaudhuri e Marron (1999), a simulación foi levada a cabo utilizando mostras simuladas a partir das densidades *Strongly Skewed* (#3), *Bimodal* (#6) e *Sep. Bimodal* (#7) de Marron e Wand (1992), as cales podemos ver sobre as Figura B.3, B.6 e B.7 do Apéndice B xunta o mapa SiZer e MoSiZer de f_h. Os tamaños de mostra empregados foron $n = \{50, 200, 500\}$, e simulando un total de $N = 1000$ mostras para cada caso. O nivel de significación empregado é de $\alpha = 0.05$, e de novo, sobre cada un dos mapa SiZer a grella de puntos x estivo comprendida dentro do intervalo $[-4, 4]$, onde se utilizou un total de $n(x) = 512$ puntos. Ademais, o número de valores do parámetro de suavizado empregado foi dun total de $n(h) = 151$ puntos.

En primeiro lugar, imos expoñer os resultados obtidos para a densidade #3 de Marron e Wand. Tal e como mostra a Figura B.3 o SiZer (centro) e MoSiZer (dereita) obtido para f_h desta densidade ten unha única moda para todo h.

Observando os resultados obtidos para o tamaño de mostra $n = 50$ na Figura 3.10 podemos ver como q_2, q_3 e q_4 estiman de forma bastante correcta a localización da moda. Ademais, os tres cuantís estiman cunha proporción moi alta de mostras a existencia dunha única moda sobre todo o SiZer, chegando incluso o cuantil q_4 a facelo na súa totalidade. Sen embargo, o cuantil q_1 comete algún erro, estimando erroneamente nunha proporción do 0.091 das mostras a existencia de dúas modas sobre o mapa SiZer. Ademais, analizando os resultados obtidos para os tamaños de mostra $n = 200$ e $n = 500$ nas Figuras 3.11 e 3.12 respectivamente, podemos ver como esa proporción de erro se ve incrementada considerablemente a medida que o tamaño de mostra medra, chegando incluso a detectar 5 modas nalgunha mostra. Nestes dous exemplos podemos ver claramente a problemática de q_1, o cal a medida que o tamaño de mostra medra perde as poucas posibilidades que ten de competir co resto dos cuantís. Para as mostras con $n = 200$ unicamente é capaz de detectar como máximo unha moda en todo SiZer nunha proporción de 0.479 das mostras, mentres que esa proporción se ve aínda mais reducida para $n = 500$, chegando estimar correctamente a existencia dunha única moda no 0.253 dos casos. Por outra banda, os cuantís q_2, q_3 e q_4 son capaces de estimar correctamente a existencia dunha moda case no 100 % das mostras, e loxicamente, a medida que o tamaño de mostra medra a proporción de acerto en todos eles vese incrementada.

Tal e como mencionamos no anterior exemplo de simulación, q_1 está a cometer unha porcentaxe de erro maior en zonas inferiores do SiZer onde o parámetro de suavizado ten un tamaño menor. Agora, no MoSiZer, vemos como efectivamente eses erros se ven traducidos en estimacións de modas inexistentes en f_h. Que as modas sexan detectas na zona baixa do SiZer fainos dubidar acerca da correcta aproximación do parámetro n_0, para o cal trataremos de ver a súa influencia nos seguintes pasos.

Agora, imos expoñer os resultados obtidos para a densidade *Bimodal* (#6) de Marron e Wand (1992), a cal vemos reflectida na Figura B.6 (esquerda) do Apéndice B. Ademais, no mapa do centro e da dereita, podemos ver a través do SiZer e MoSiZer de f_h respectivamente, como para parámetros ventá h grandes unicamente hai unha moda en f_h, mentres que a medida que o parámetro de suavizado vai diminuíndo comezan a aparecer dúas modas que se van separando ata chegar a unha posición fixa.

Analizando en primeiro lugar os resultados obtidos para o tamaño de mostra $n = 50$, podemos

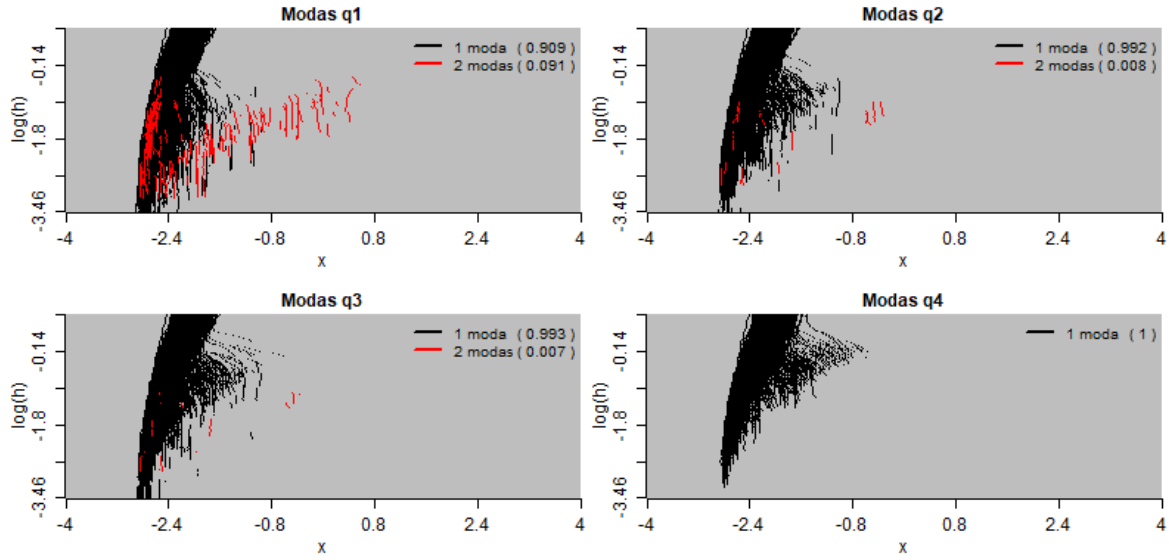

Figura 3.10: MoSiZer para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 50$, xeradas a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

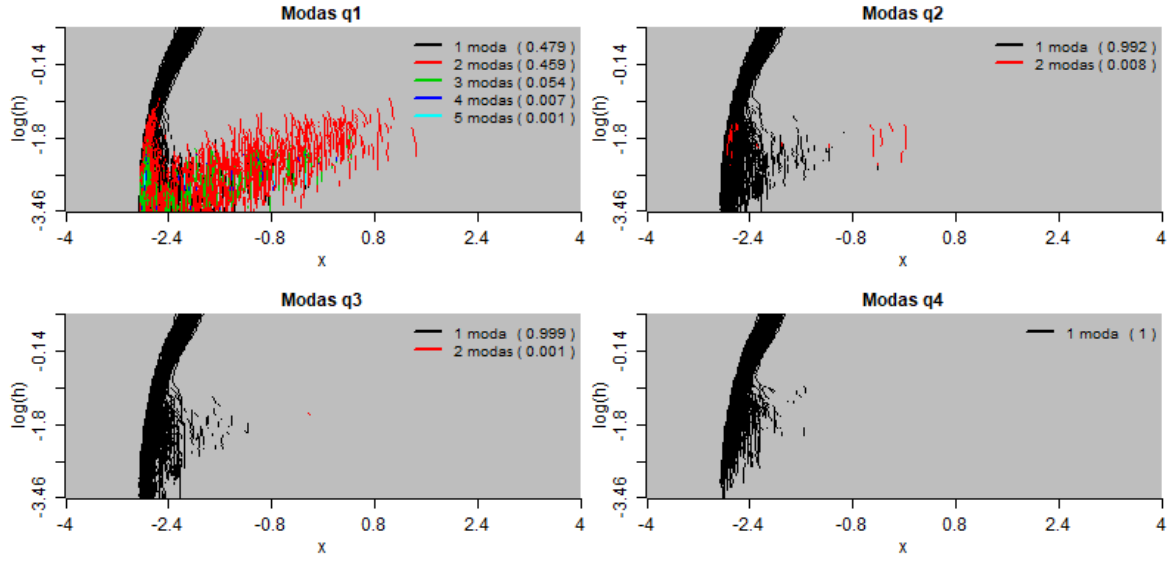

Figura 3.11: MoSiZer para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 200$, xeradas a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

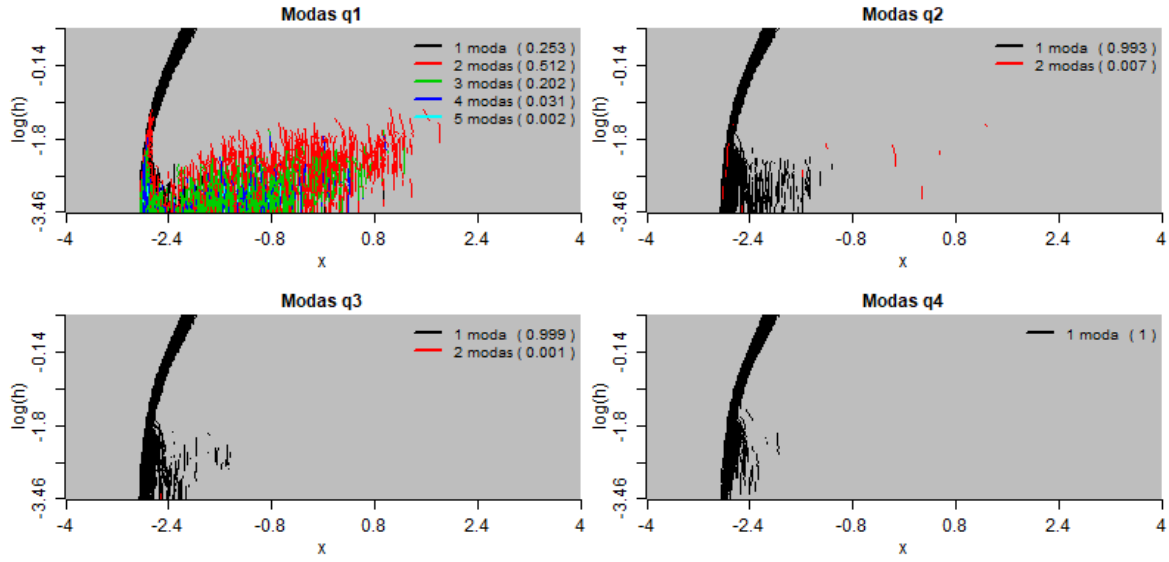

Figura 3.12: MoSiZer para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 500$, xeradas a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

ver sobre a Figura 3.13 como todos os cuantís teñen bastantes dificultades para detectar as dúas modas existentes. Mentres que q_2, q_3 e q_4 son incapaces de mostrar as dúas modas presentes en zonas baixas do SiZer de f_h, o cuantil q_1 a pesar de obter malos resultados, si é capaz de detectalas nunha proporción do 0.195 das mostras. Estes resultados reflicten o conservador que son os demais cuantís. Cunha probabilidade alta as características mostradas por q_2, q_3 e q_4 son correctas, pero vense limitados a non mostrar moitas outras que si están presentes para o cuantil q_1.

Por outra banda, observando os resultados para o tamaño de mostra $n = 200$ sobre a Figura 3.14 podemos ver como de novo para o cuantil q_1, a medida que o tamaño de mostra medra a inferencia é realizada en zonas baixas do SiZer, provocando que haxa un maior erro nesas zonas e dando lugar a modas inexistentes. Aínda con isto, o cuantil q_1, para as mostras xeradas a partires desta función de densidade, parece ser capaz de detectar a existencia de dúas modas en f_h como máximo de mellor forma que o resto dos cuantís. Neste caso, o cuantil q_1 chega a unha proporción do 0.574 das mostras nos que si é capaz de estimar correctamente a presenza de dúas modas en f_h, pero como no anterior caso, comezan a aparecer mostras nas que o cuantil q_1 comete o erro de estimar en zonas baixas do SiZer a presenza de tres, catro ou cinco modas. Co que respecta aos cuantís q_2, q_3 e q_4, aínda cun tamaño de mostra mais elevado, estes son incapaces de detectar a presenza de dúas modas, chegando tan só a proporcións do 0.108, 0.047 e 0.006 respectivamente.

Para o tamaño de mostra de $n = 500$ podemos ver sobre a Figura 3.15 como o erro cometido polo cuantil q_1 se ve incrementado. Este cuantil parece capaz de detectar o número de modas, con mostras desta densidade, de forma mais correcta que o resto dos cuantís en zonas altas do mapa SiZer, pero a medida que o tamaño do parámetro de suavizado se ve reducido, a estimación comeza a ser fatídica e dá lugar a modas inexistente en f_h. Neste caso o cuantil q_1 detecta nunha proporción do 0.491 das mostras a presenza de dúas modas, proporción que se ve diminuída con respecto á porcentaxe obtida para o tamaño de mostra $n = 200$. Vese de forma clara, como nas zonas baixas do mapa, o SiZer comeza a detectar ata seis modas nalgunha ocasión, chegando a ter unha proporción preocupante do 0.333 das mostras nos que está a estimar a presenza de tres modas. Por outra banda, para este tamaño

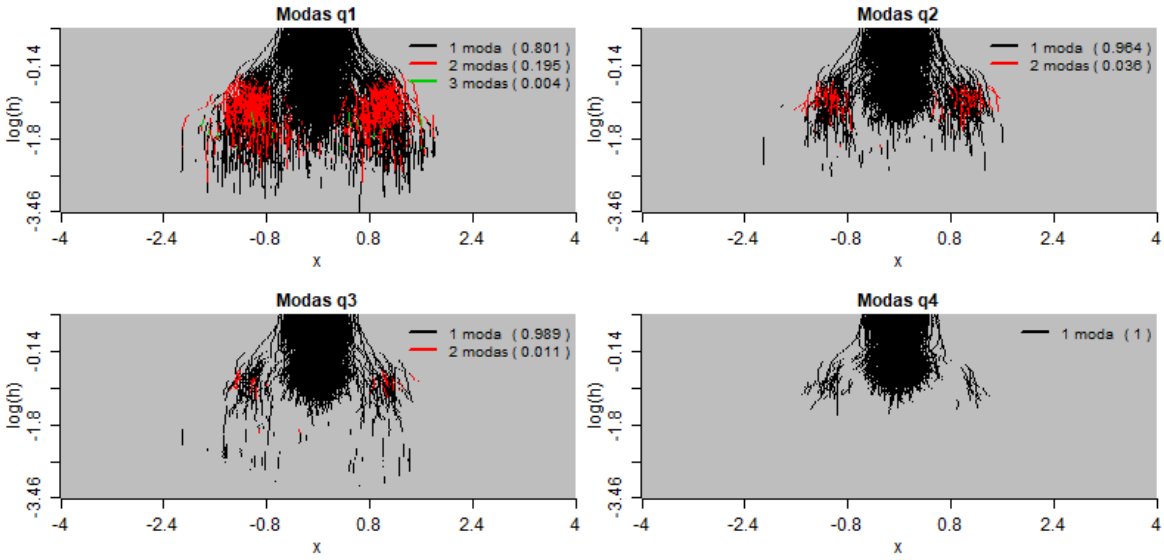

Figura 3.13: MoSiZer para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 50$, xeradas a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

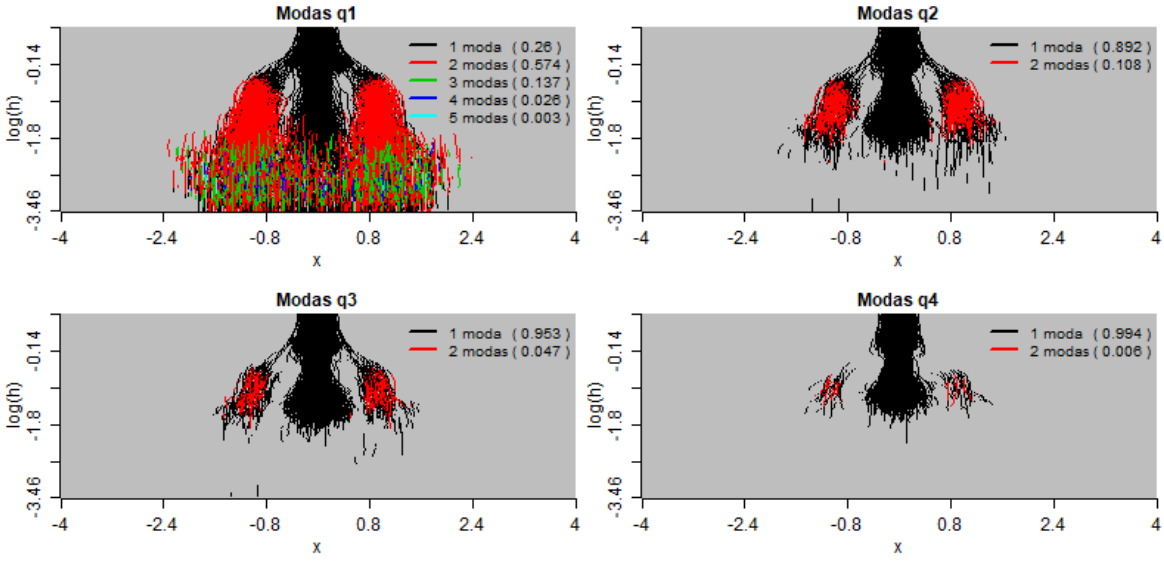

Figura 3.14: MoSiZer para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 200$, xeradas a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

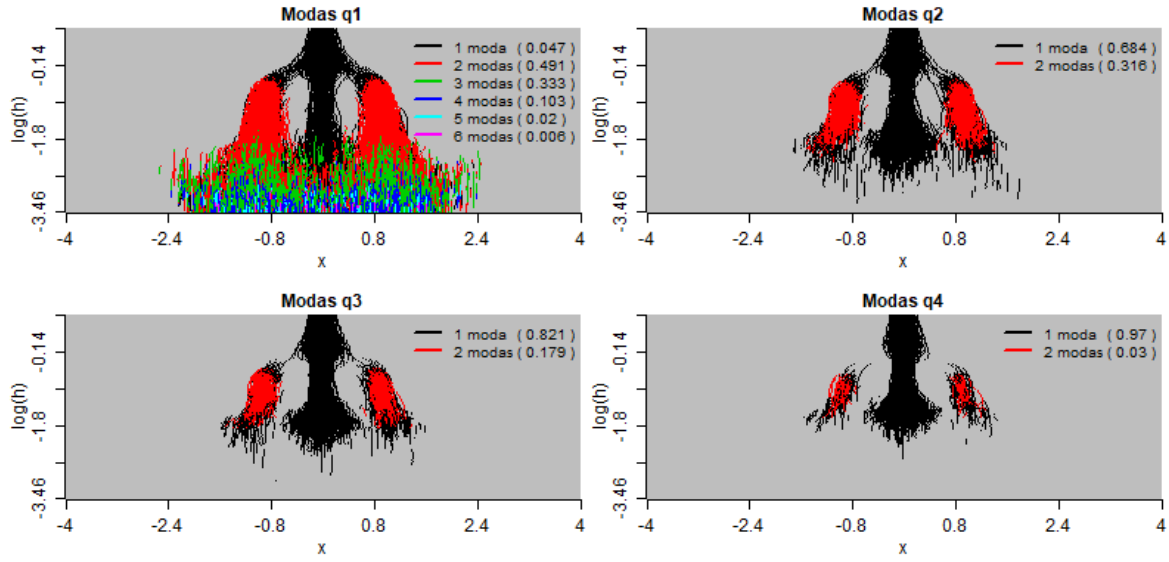

Figura 3.15: MoSiZer para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 500$, xeradas a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

de mostra tamén se ve de forma clara que o cuantil q_3 é mais conservador que o cuantil q_2, o cal, a pesar de ser un tamaño de mostra elevado, tan só detecta de forma correcta a presenza de dúas modas nunha proporción do 0.179 das mostras fronte á proporción 0.316 do cuantil q_2. Por último, podemos ver como o cuantil q_4 segue sendo demasiado conservador, dando lugar a que a pesar de contar cun tamaño de mostra elevado ($n = 500$), tan só sexa capaz de detectar nunha proporción do 0.03 das mostras a presenza de dúas modas.

De seguido imos ver os resultados obtidos para a simulación da densidade *Sep. Bimodal* (#7) de Marron e Wand (1992) mostrada sobre a Figura B.7 (esquerda) do Apéndice B xunto o SiZer (centro) e MoSiZer (dereita) de f_h. Veremos que a pesar de contar con dúas modas como a densidade *Bimodal* (#6) (Figura B.6) que acabamos de presentar, a localización destas modas vai a influír drasticamente nos resultados obtidos para cada cuantil.

Tal e como mostra a Figura B.7, podemos ver como as modas presentes na función de densidade se encontran mais distantes entre elas que no caso da función de densidade *Bimodal* (#6) de Marron e Wand (1992), e ademais, o vale presente entre elas é moito mais profundo. Isto vai a facilitar aos cuantís a estimar en proporcións mais altas a presenza das dúas modas.

Observando en primeiro lugar os resultados obtidos para o tamaño de mostra $n = 50$, podemos ver como aínda os resultados obtidos polo cuantil q_1 son superiores aos demais cuantís. Todos eles detectan de forma adecuada a presenza de dúas modas en f_h con proporcións moi próximas a 1 (a excepción de q_4 no que a proporción tan só chega ao 0.81), polo que podemos ver de forma clara que non só inflúe o número de modas, se non tamén a localización das mesmas.

Con respecto aos resultados obtidos para o tamaño de mostra $n = 200$ mostrados na Figura 3.17 podemos ver como o escenario e completamente diferente o da densidade #6 de Marron e Wand. Na Figura 3.14 vimos como a pesar de que o cuantil q_1 cometía erros nas zonas onde o parámetro de suavizado era mais pequeno, este era capaz de detectar dunha forma mais eficiente que o resto de cuantís a presenza de dúas modas en f_h. Sen embargo, na Figura 3.17 vemos como nesta ocasión os cuantís q_2, q_3 e q_4 detectan na totalidade das mostras as dúas modas presentes en f_h. Con respecto ao

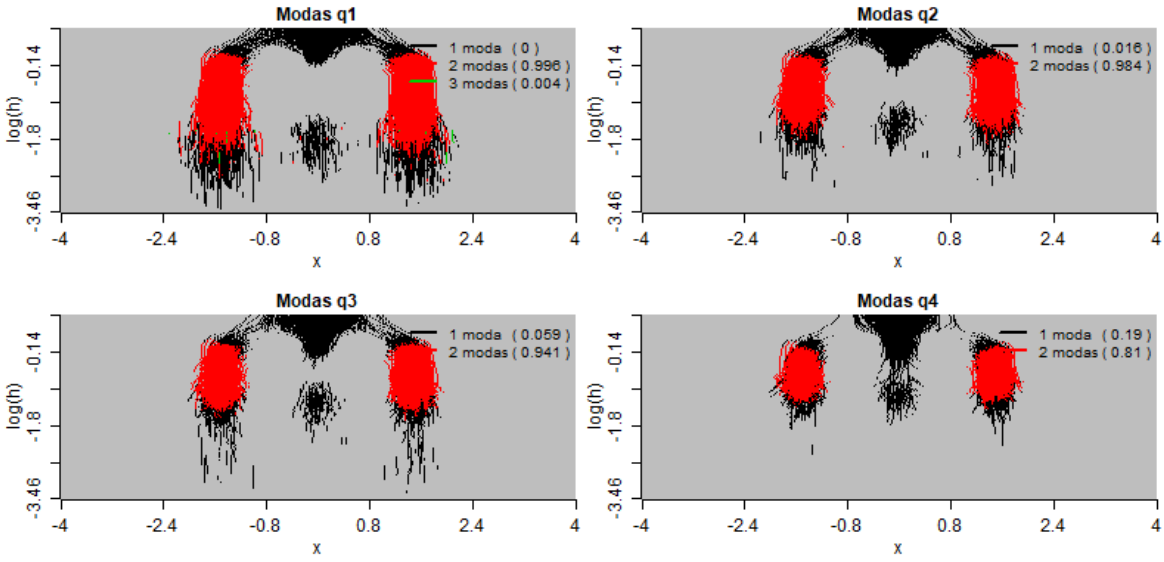

Figura 3.16: MoSiZer para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 50$, xeradas a partir dunha variable aleatoria con función de densidade *Sep. Bimodal* (#7) de Marron e Wand (1992).

cuantil q_1 de novo comete o erro de estimar nas zonas mais baixas do SiZer a presenza de mais modas, reducíndose a proporción da correcta estimación de dúas modas ata o 0.833.

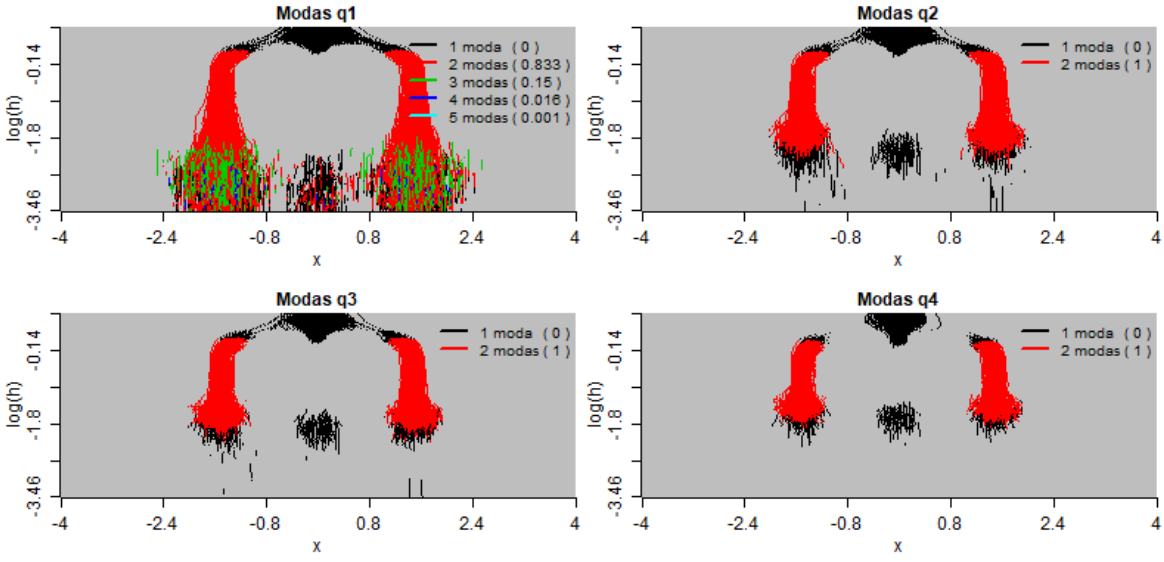

Figura 3.17: MoSiZer para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 200$, xeradas a partir dunha variable aleatoria con función de densidade *Sep. Bimodal* (#7) de Marron e Wand (1992).

Por último, os resultados obtidos para o tamaño de mostra $n = 500$ parecen mostrar o mesmo

patrón seguido polo tamaño de mostra $n = 200$. E dicir, os cuantis q_2, q_3 e q_4 estiman na totalidade das mostras as dúas modas presentes en f_h, mentres que a proporción de veces que o cuantil q_1 estima como máximo dúas modas segue diminuíndo. Nesta ocasión a proporción é de tan só 0.634, e chega nalgunha ocasión a estimar ata seis ou sete modas sobre algunha mostra.

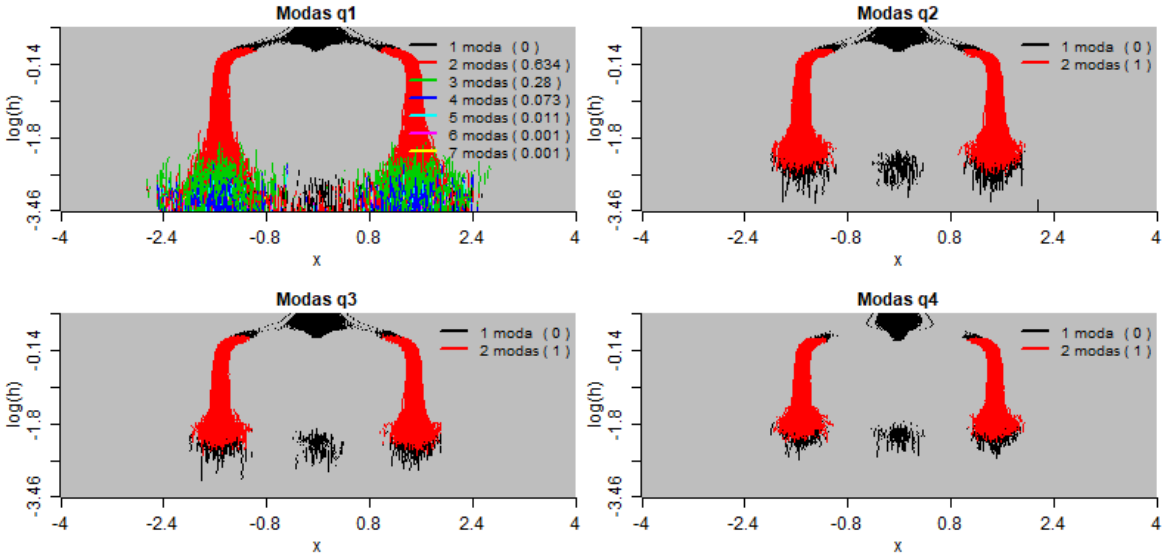

Figura 3.18: MoSiZer para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 500$, xeradas a partir dunha variable aleatoria con función de densidade *Sep. Bimodal* (#7) de Marron e Wand (1992).

Os resultados obtidos parecen mostrar de forma evidente que en ocasións os cuantís q_2, q_3 e q_4 poden chegar a ser un tanto conservadores. Estes estiman de forma moi acertada a presenza de características moi definidas sobre a función de densidade orixinal, como é o caso da única moda presente na densidade *Strongly Skewed* (#3) ou as dúas modas presentes na densidade *Sep. Bimodal* (#7), pero cando o tamaño de mostra non é tan elevado e as modas non están tan acentuadas teñen a dificultade de mostrar a verdadeira forma de f_h como ocorre para a densidade *Bimodal* (#6) de Marron e Wand (1992). Por outra banda, o cuantil q_1 é capaz de estimar con maior facilidade a presenza das modas que se encontran en f_h, pero coa adversidade de cometer un maior erro que finalmente se ve traducido en modas inexistentes sobre a función orixinal.

Variación do parámetro n_0

Como vimos en todos os casos anteriores, o maior erro cometido por q_1 está presente en zonas do SiZer onde o parámetro de suavizado é mais pequeno. A medida que o tamaño de mostra medra, a ferramenta SiZer comeza a pintar con cores azuis, púrpuras e vermellas aquelas zonas onde para tamaños de mostra mais pequenos eran de cor gris. Nesas zonas o SiZer cunha probabilidade mais alta comeza a mostrar modas que non están presentes en f_h, o que a priori, nos fai dubidar acerca da correcta elección do parámetro n_0. Tal e como mostran os resultados obtidos para o anterior exemplo de simulación, en ocasións este parámetro parece depender do tamaño de mostra, polo que neste apartado faremos de novo un estudo de simulación que nos deixe ver a posible influencia do parámetro n_0 sobre as estimacións.

Neste estudo de simulación repetimos o proceso realizado na anterior Sección para as funcións de densidade *Bimodal* (#6) (Figura B.6 do Apéndice B) e *Sep. Bimodal* (#7) (Figura B.7 do Apéndice

B) de Marron e Wand (1992), pero para unha simulación na que analizamos tan só os resultados obtidos polo cuantil q_1 con diferentes valores do parámetro n_0. Nesta ocasión simulamos un total de $N = 1000$ mostras para os tamaños de mostra $n = \{50, 200, 500\}$ das dúas densidades, onde de novos utilizamos unha grella de $n(x) = 512$ puntos comprendidos no intervalo $x = [-4, 4]$. Tamén empregamos $n(h) = 151$ puntos, utilizando un nivel de significación de $\alpha = 0.05$, e os valores escollidos para o parámetro n_0 foron $n_0 = \{5, 6, 7, 8\}$.

Observando en primeiro lugar os resultados obtidos para a densidade *Bimodal* (#6) de Marron e Wand (1992) cun tamaño de mostra $n = 50$, podemos ver a través da Figura 3.19 como o SiZer que está a obter mellores resultados é o SiZer con $n_0 = 5$. Este está a estimar de forma correcta a presenza de dúas modas en tan só o 0.178 das mostras, pero que aínda así mellora os resultados obtidos polos demais SiZers con valores de n_0 maiores.

Por outra banda, centrándonos na Figura 3.20, podemos ver que cun tamaño de mostra maior ($n = 200$) o cuantil q_1 con $n_0 = 5$ aumenta considerablemente a probabilidade de erro de estimar un número de modas maior ao presente en f_h. Este cuantil con $n_0 = 5$ continua a ofrecer mellores resultados que o resto, pero podemos ver como os demais resultados (con $n_0 = \{6, 7, 8\}$) van aproximando a súa proporción de acerto á de $n_0 = 5$.

Xa para o tamaño de mostra $n = 500$, podemos ver a través da Figura 3.21 como o cuantil q_1 con $n_0 = 5$ presenta unha proporción de erro moi alta co que respecta a estimar un número de modas maior a dous. Mentres que para este obtemos unha proporción de acerto de que en f_h hai un número máximo de dúas modas de 0.517, podemos ver que para $n_0 = \{6, 7, 8\}$ obtemos proporcións do 0.554, 0.572 e 0.599 respectivamente. O que nos mostra que para esta densidade con este tamaño de mostra parece mais axeitado empregar o cuantil q_1 cun parámetro $n_0 = 8$ ou superior.

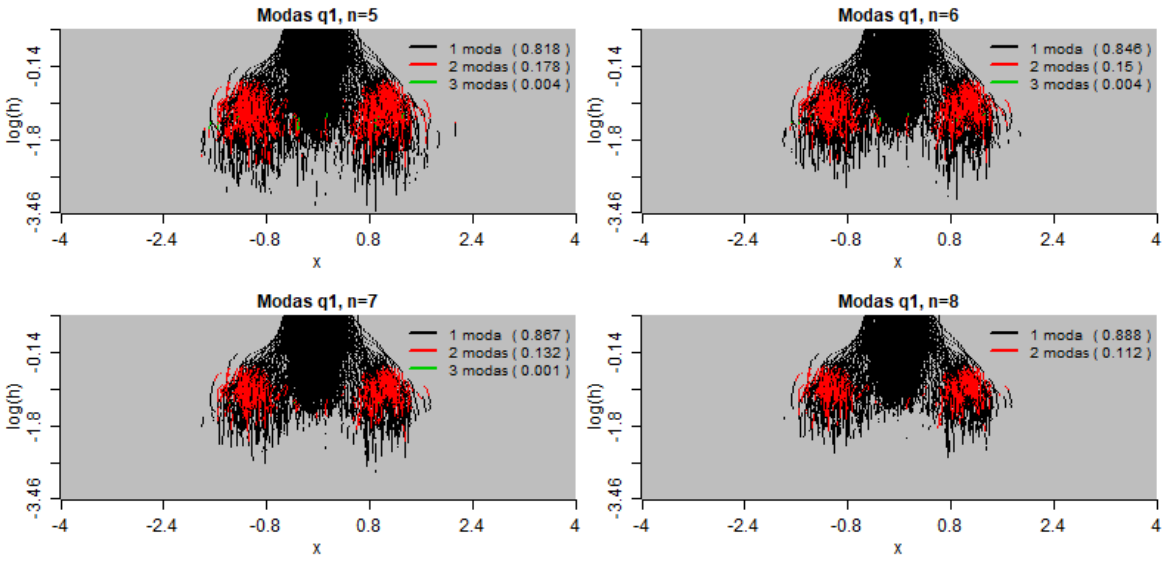

Figura 3.19: MoSiZer para o cuantil q_1 con $n_0 = \{5, 6, 7, 8\}$, para $N = 1000$ mostras, con tamaño de mostra $n = 50$, xeradas a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

Os resultados anteriores parecen dar a entender que o parámetro n_0 poda chegar a depender do tamaño de mostra, pero veremos a continuación sobre os resultados obtidos coa densidade #7 de Marron e Wand que non só iso vai a influír na correcta elección do parámetro.

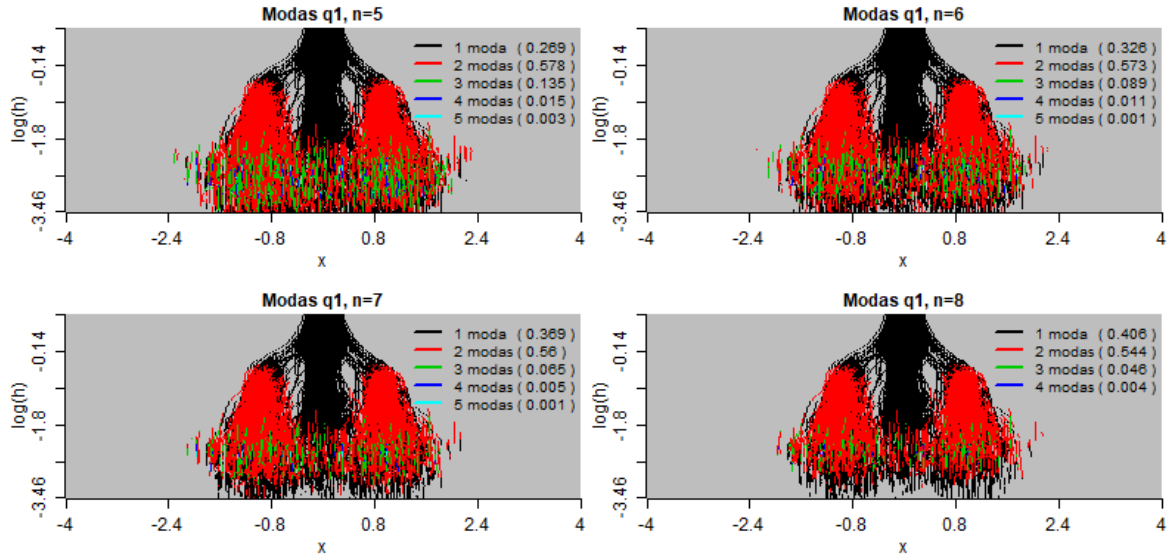

Figura 3.20: MoSiZer para o cuantil q_1 con $n_0 = \{5, 6, 7, 8\}$, para $N = 1000$ mostras, con tamaño de mostra $n = 200$, xeradas a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

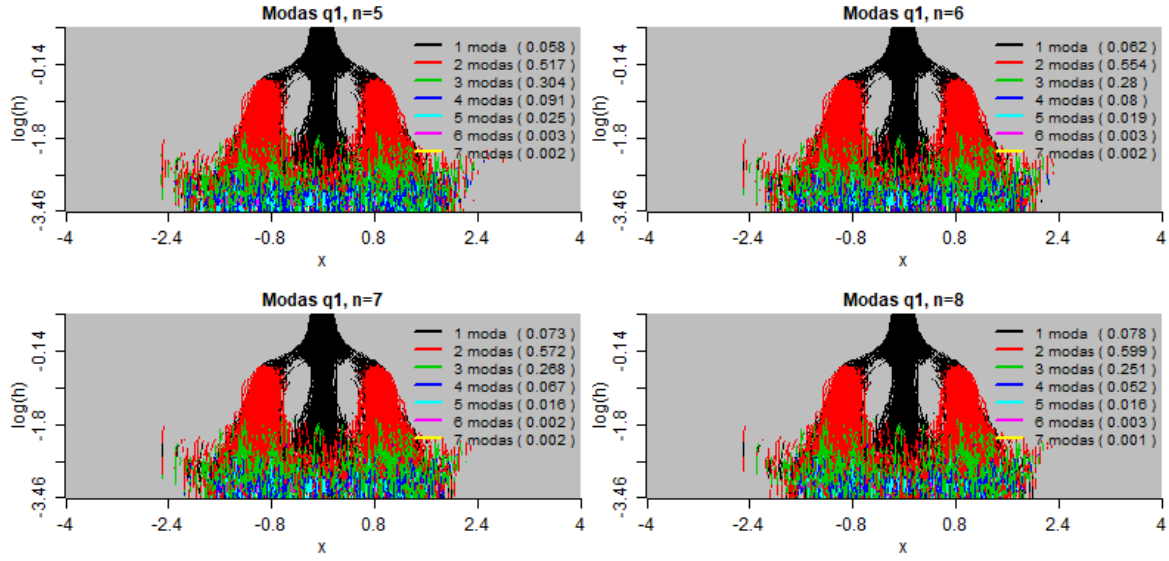

Figura 3.21: MoSiZer para o cuantil q_1 con $n_0 = \{5, 6, 7, 8\}$, para $N = 1000$ mostras, con tamaño de mostra $n = 500$, xeradas a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

Como mencionamos, a densidade *Sep. Bimodal* (#7) de Marron e Wand (mostrada na Figura B.7) está formada por dúas modas como a densidade *Bimodal* (#6) (Figura B.6) pero cunha maior separación e un vale mais profundo entre elas. Con isto, analizando os resultados mostrados para o tamaño de mostra $n = 50$ sobre a Figura 3.22 podemos ver como nesta ocasión os mellores resultados son os ofrecidos polo cuantil q_1 cun parámetro $n_0 = 7$ e $n_0 = 8$. Estes obteñen proporcións mais altas na estimación dun número máximo de dúas modas que o cuantil q_1 con $n_0 = \{5, 6\}$.

Por outra banda, observando os resultados obtidos para o tamaño de mostra $n = 200$ e $n = 500$ sobre as Figuras 3.23 e 3.24 respectivamente, podemos ver como claramente as estimacións realizadas polo cuantil q_1 con $n_0 = 8$ teñen unha probabilidade mais elevada da correcta estimación dun número máximo de dúas modas, cunha proporción de 0.942 e 0.713 respectivamente, fronte ás proporcións para un parámetro $n_0 = 5$ de 0.858 e 0.611 para os tamaños de mostra $n = 200$ e $n = 500$ respectivamente.

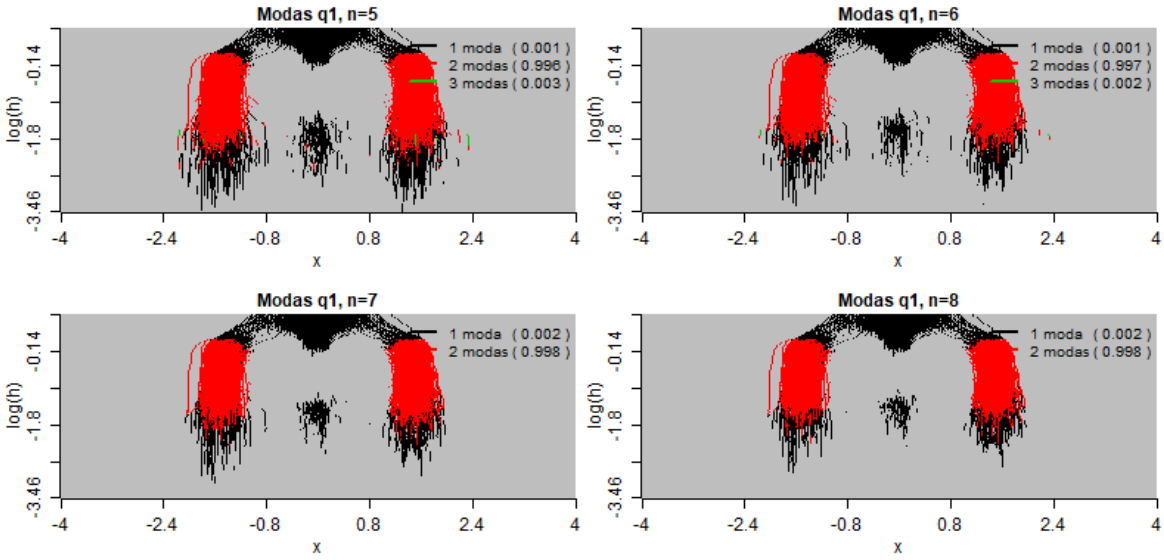

Figura 3.22: MoSiZer para o cuantil q_1 con $n_0 = \{5, 6, 7, 8\}$, para $N = 1000$ mostras, con tamaño de mostra $n = 50$, xeradas a partir dunha variable aleatoria con función de densidade *Sep. Bimodal* (#7) de Marron e Wand (1992).

Os resultados obtidos sobre a densidade *Bimodal* (#6) de Marron e Wand (1992) parecían mostrar que o parámetro n_0 podía chegar a depender directamente do tamaño de mostra, onde para un tamaño de mostra de $n = 50$ obtiñamos mellores resultados co parámetro $n_0 = 5$, o mesmo para o tamaño de mostra de $n = 200$ e finalmente para $n = 500$ obtiñamos mellores resultados con $n_0 = 8$. Posteriormente, analizando os resultados obtidos para a densidade *Sep. Bimodal* (#7) de Marron e Wand (1992) vimos que para todos os tamaños de mostra estudados ($n = \{50, 200, 500\}$) o parámetro mais axeitado n_0 para o cuantil q_1 era o parámetro $n_0 = 8$. Isto lévanos a supoñer que o parámetro n_0 non só probablemente dependa do tamaño de mostra, se non que tamén da forma orixinal da función de densidade coa que estamos a xerar as mostras. Isto é unha información que imos descoñecer para facer inferencia coa ferramenta SiZer, pero deixa aberto posibles futuras vías de investigación que leven á posibilidade de aproximar o valor do parámetro n_0 sen a necesidade de coñecer a función de densidade da variable aleatoria xeradora da mostra.

Outra dúbida que nos xera os resultados obtidos, e a razón pola cal o erro cometido se está a producir maiormente nas zonas do SiZer onde o parámetro de suavizado é mais pequeno. En Chaudhuri e Marron (1999) mencionan que o cuantil Gaussiano non é adecuada para obter os intervalos de confianza onde

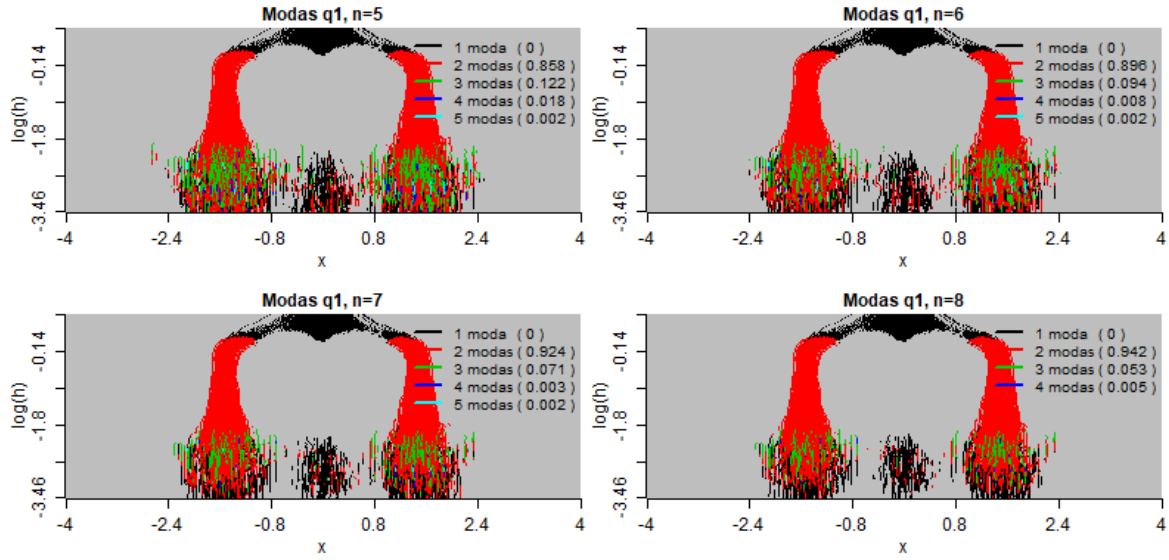

Figura 3.23: MoSiZer para o cuantil q_1 con $n_0 = \{5, 6, 7, 8\}$, para $N = 1000$ mostras, con tamaño de mostra $n = 200$, xeradas a partir dunha variable aleatoria con función de densidade *Sep. Bimodal* (#7) de Marron e Wand (1992).

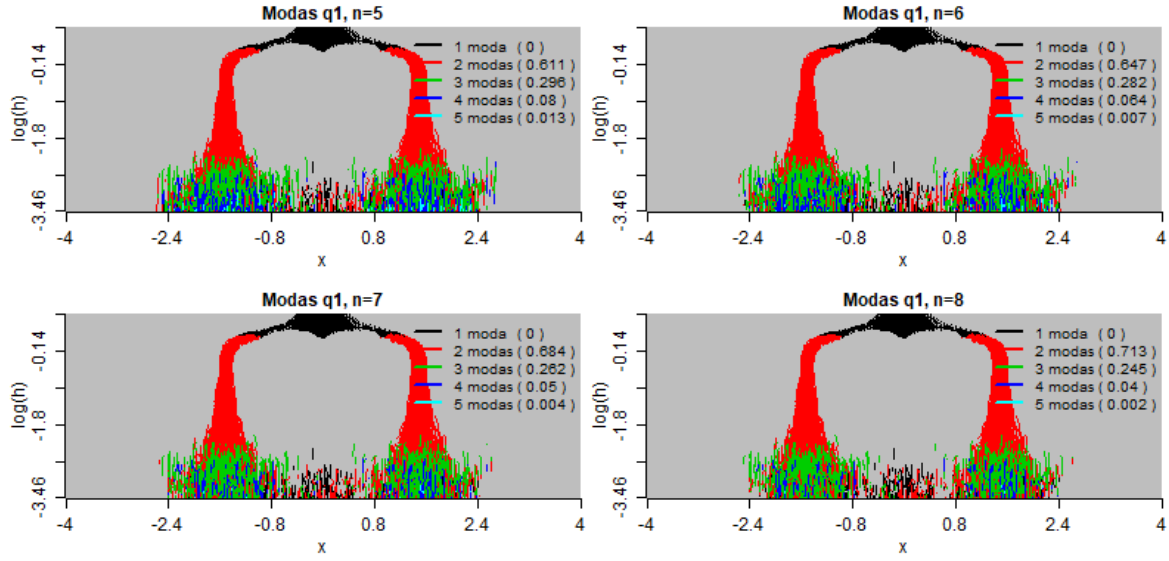

Figura 3.24: MoSiZer para o cuantil q_1 con $n_0 = \{5, 6, 7, 8\}$, para $N = 1000$ mostras, con tamaño de mostra $n = 500$, xeradas a partir dunha variable aleatoria con función de densidade *Sep. Bimodal* (#7) de Marron e Wand (1992).

$ESS > n_0 = 5$, o que nos leva a na seguinte Sección facer un estudo de simulación para ver se realmente é iso o que está a producir unha taxa de erro maior nas zonas mais baixas do SiZer. Nela faremos unha análise para ver se a cobertura dos intervalos de confianza de $\hat{f}'_h$ sobre o cuantil q_1 é próxima ao nivel $(1-\alpha)$.

3.3. Taxa de cobertura dos intervalos de confianza

Como vimos nos exemplos de simulación levados a cabo, o cuantil q_1 parece cometer unha maior taxa de erro en zonas do SiZer onde o parámetro de suavizado ten un valor pequeno. En Chaudhuri e Marron (1999) os autores comentan que o cuantil normal non é adecuado para obter os intervalos de confianza da estimación da derivada da densidade onde $ESS > n_0$. Ademais, suxiren que un valor axeitado para este parámetro é $n_0 = 5$.

Tal e como vimos na Sección 3.1, a pesar de obter o mapa SiZer das mostras empregando o parámetro $n_0 = 5$, os resultados mostraban que a taxa de erro para o cuantil q_1 era igualmente maior en zonas do mapa SiZer onde h tiña un valor pequeno. Ademais, na Sección 3.2 vimos como eses erros se acababan traducindo en estimación de modas inexistentes.

Nesta Sección vamos a estudar que está a ocorrer cos intervalos de confianza do cuantil q_1 obtidos para a estimación da derivada da función de densidade, analizando se estes teñen a cobertura adecuada ou realmente a afirmación que suxiren Chaudhuri e Marron é certa. Ademais, veremos os resultados ofrecidos polas diferentes versións bootstrap do cuantil q_1 (q_5, q_6, q_7 e q_8), naqueles modelos nos que q_1 non parece obter resultados adecuados.

3.3.1. Implementación

Como mencionamos, imos tratar de estudar a cobertura dos intervalos de confianza obtidos para o mapa SiZer. Trataremos de estudalo de dúas formas diferentes, en primeiro lugar analizando os intervalos de confianza de toda a función $\hat{f}'_h$ para h fixo, e posteriormente analizando a cobertura dos intervalos de confianza de todo o mapa SiZer punto por punto, é dicir, de cada punto (x, h). De novo, na simulación levada a cabo traballaremos con funcións de densidade propostas por Marron e Wand (1992), con diferentes tamaños de mostra, e simulando para caso un total de N mostras. Ademais, para ver a influencia real do parámetro n_0 imos facer o exercicio de simulación con diferentes valores para este parámetro.

Para obter os resultados precisos imos traballar con dúas matrices coa mesma dimensión que a matriz obtida para o mapa SiZer. Nunha delas, a matriz A, imos facer o rexistro da cantidade de veces que o intervalo de confianza de $\hat{f}'_h$ conten o verdadeiro valor de f'_h sobre cada punto (x, h) do mapa SiZer. Mentres que na outra matriz, a matriz B, imos facer o rexistro da cantidade de veces que se fixo inferencia sobre cada punto do SiZer, é dicir, a cantidade de veces que en cada punto (x, h) o valor ESS e maior que n_0. Desta maneira, obtemos as matrices A e B a partires das N mostras de cada caso segundo o seguinte algoritmo:

Algoritmo 12 Matrices A e B

1: Obter os intervalos de confianza superior e inferior de $\hat{f}'_h$ a partires do cuantil q_1 para todos os puntos (x, h) do SiZer, e a matriz cos valores ESS de cada punto (x, h).
2: Sumar unha unidade a aqueles puntos da matriz $A(x, h)$ nos que o intervalos de confianza de $\hat{f}'_h$ conteña o verdadeiro valor de f'_h, pero que o valor nese punto de ESS sexa tal que $ESS(x, h) > n_0$.
3: Sumar unha unidade a aqueles puntos da matriz $B(x, h)$ nos que no mapa SiZer $ESS(x, h) > n_0$.
4: Repetir os pasos anteriores N veces.

Unha vez obtidos os resultados sobre as matrices A e B, imos obter a cobertura de cada f_h para h fixo segundo o seguinte algoritmo:

Algoritmo 13 Cobertura de f_h para h fixo

1: Obter as matrices A e B a partires do Algoritmo 12.
2: Obter un vector a coa suma dos valores de cada fila de A, é dicir, $a(h) = \sum_x A(x,h)$.
3: Obter un vector b coa suma dos valores de cada fila de B, é dicir, $b(h) = \sum_x B(x,h)$.
4: Se $b(h) \geq N$:

$$\text{Cob}(h) = a(h)/b(h).$$

Se $b(h) < N$:

$$\text{Cob}(h) = \text{NULL}.$$

Con isto imos obter a cobertura de f'_h só naqueles f'_h nos que como mínimo, en media, se fixo inferencia nun punto x de f'_h nas N mostras. Posteriormente a cobertura Cob(h) será representada frente a h, onde cada curva referente a cada valor de n_0 será representada cunha cor.

Por outra banda, o procedemento para obter a cobertura de cada punto (x,h) de todo o SiZer a partir das matrices A e B vai ser mais sinxelo segundo o seguinte algoritmo:

Algoritmo 14 Cobertura de f_h para cada punto (x,h)

1: Obter as matrices A e B a partires do Algoritmo 12.
2: Se $B(x,h) \geq 100$:

$$\text{Cob}(x,h) = A(x,h)/B(x,h)$$

Se $B(x,h) < 100$:

$$\text{Cob}(x,h) = \text{NULL}$$

Polo que imos obter a cobertura daqueles puntos nos que se fixo inferencia un mínimo de 100 ocasións, supoñendo que o valor de N sexa o suficientemente grande. Posteriormente será representada a cobertura nun mapa, onde a cor púrpura será indicativa de que o intervalo de confianza ten unha cobertura próxima ao nivel $(1-\alpha)$, a cor azul que a cobertura tende a encontrarse por debaixo dese nivel, e unha cor púrpura avermellada será indicativo de que a cobertura é maior a $(1-\alpha)$.

As limitacións propostas de que se faga un mínimo de inferencia sobre cada f_h ou sobre cada punto x, sérvennos para non obter conclusións naquelas zonas onde non hai suficiente información coma para facelo.

3.3.2. Resultados

De novo, a simulación foi levada a cabo empregando as funcións de densidade *Strongly Skewed* (#3) e *Bimodal* (#6) de Marron e Wand (1992), onde empregamos os tamaños de mostra $n = \{10, 20, 50, 200, 500\}$, un nivel de significación $\alpha = 0.05$, un total de $n(x) = 512$ puntos sobre a grella x no intervalo $x = [-4, 4]$, e $n(h) = 151$ parámetros de suavizado. Ademais, para cada caso simulamos un total de $N = 1000$ mostras, onde os parámetros n_0 empregados foron $n_0 = \{1, 2, 3, 4, 5, 6, 7\}$. Cabe destacar que o parámetro $n_0 = 0$ non foi tido en conta debido a que iso provocaría que fixeramos inferencia en zonas onde non temos datos na maioría das mostras, e polo tanto obviamente a cobertura dos intervalos de confianza obtidos neses puntos non sería certeira.

Ademais, levamos acabo o mesmo exemplo de simulación pero para tamaños de mostra mais reducidos $n = \{10, 20\}$, onde comparamos os resultados obtidos polos cuantís q_1, q_5, q_6, q_7 e q_8, para ver que cuantil está a ofrecer mellores resultados de cobertura. Neste caso, a simulación foi levada a cabo unicamente para o valor $n_0 = 5$.

Analizando en primeiro lugar os resultados obtidos para a densidade número *Strongly Skewed* (#3) de Marron e Wand (1992), para os tamaños de mostra $n = \{50, 200, 500\}$, podemos ver a través da

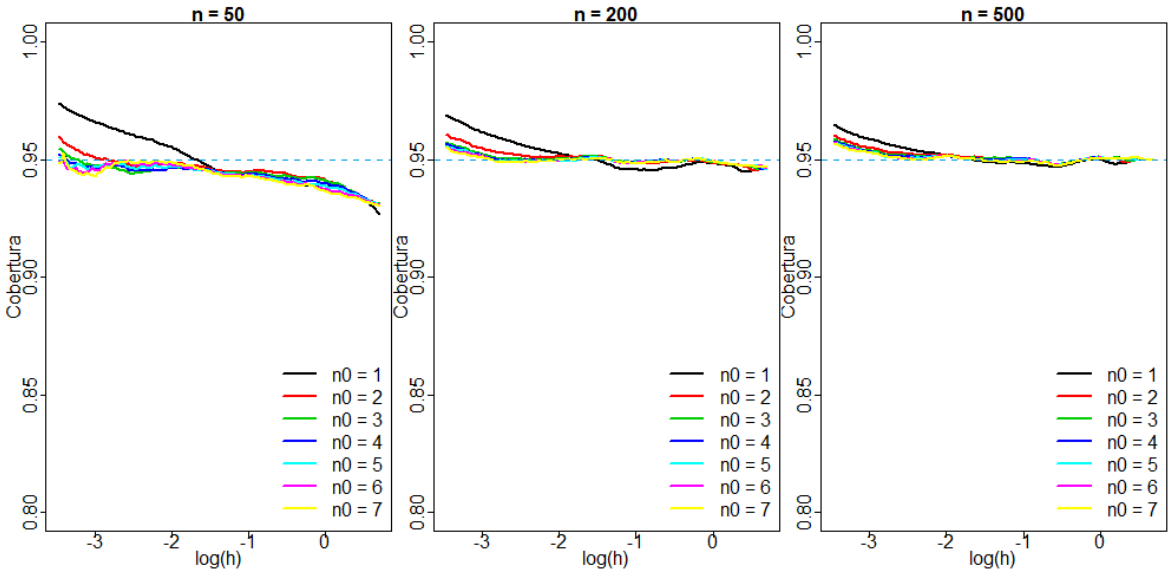

Figura 3.25: Taxa de cobertura do cuantil q_1 con respecto a h, con $n_0 = \{1, 2, 3, 4, 5, 6, 7\}$, para $N = 1000$ mostras, con tamaño de mostra $n = \{50, 200, 500\}$, xeradas a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

Figura 3.25 como independentemente do parámetro n_0 seleccionado, a cobertura dos intervalos de confianza se encontra sempre entorno ao valor $(1 - \alpha) = 0.95$. Como vemos, ocorre para calquera tamaño de mostra dos considerados, e incluso, podemos ver como para o parámetro mais pequeno $n_0 = 1$ a cobertura é un pouco superior para zonas onde o parámetro de suavizado é menor. Todas as coberturas, independentemente do parámetro n_0 ou tamaño de mostra n parecen mostrar uns valores de cobertura por encima do nivel $(1 - \alpha)$ en zonas con pequenos valores de h, para posteriormente estabilizarse no nivel $(1 - \alpha)$. Xusto o oposto aos resultados obtidos ata o momento.

Como vemos, obtemos estes mesmos resultados analizando cada punto (x, h) a través da Figura 3.26 (contemplando unicamente os resultados obtidos para $n = 50$), onde todos os puntos teñen unha cobertura próxima ao nivel $(1 - \alpha)$ independentemente do parámetro n_0 seleccionado. Nesa mesma Figura, na primeira das gráficas tamén vemos reflectido o anteriormente mencionado de que para un parámetro $n_0 = 1$, nas zonas do mapa onde o parámetro de suavizado h é mais pequeno, a cobertura dos intervalos de confianza parece ser maior que $(1-\alpha)$ debido ao color púrpura avermellado que parece mostrar. Obviamente tamén vemos reflectido que canto maior é o parámetro n_0 imos facer inferencia en menos puntos (x, h) do mapa SiZer.

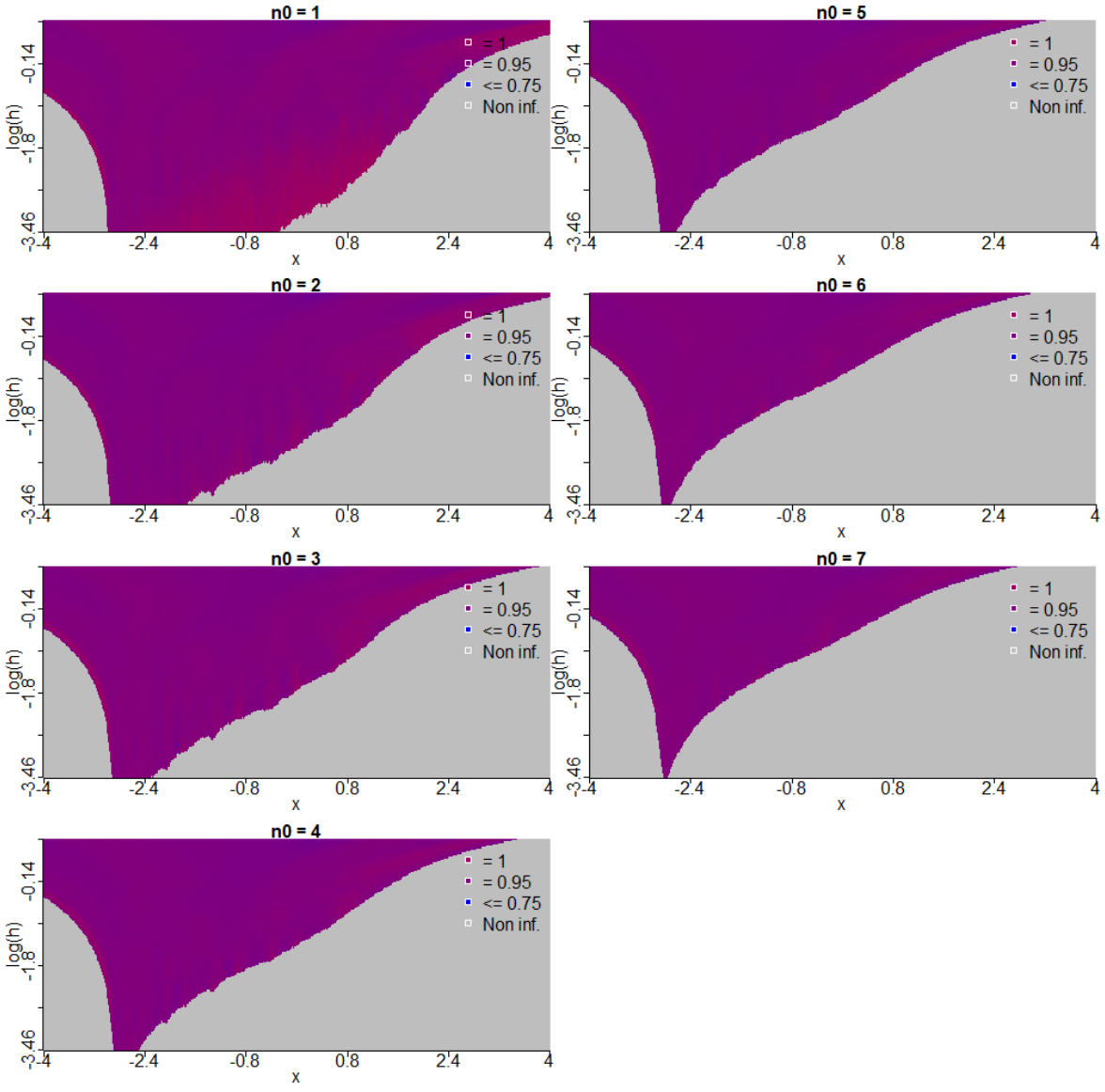

Figura 3.26: Taxa de cobertura do cuantil q_1 con respecto a (x, h), con $n_0 = \{1, 2, 3, 4, 5, 6, 7\}$, para $N = 1000$ mostras, con tamaño de mostra $n = 50$, xeradas a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

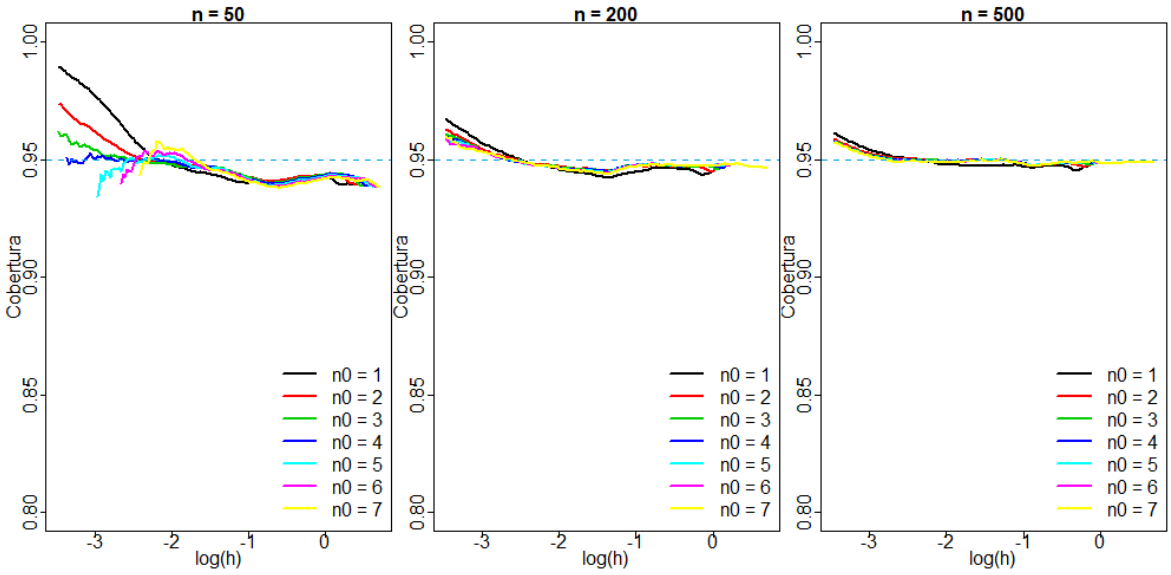

Figura 3.27: Taxa de cobertura do cuantil q_1 con respecto a h, con $n_0 = \{1, 2, 3, 4, 5, 6, 7\}$, para $N = 1000$ mostras, con tamaño de mostra $n = \{50, 200, 500\}$, xeradas a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

Por outro lado, analizando os resultados obtidos para a densidade *Bimodal* (#6) de Marron e Wand (1992) sobre as Figuras 3.27 e 3.28 podemos ver como ocorre exactamente o mesmo. A cobertura dos intervalos de confianza xa sexa para todo f_h ou para $f_h(x)$ encóntrase entorno ao valor $(1 - \alpha)$ independentemente do parámetro n_0. Ademais, nestas Figuras vese un pouco mais claro que non só obtemos cobertura mais altas para o valor $n_0 = 1$ en zonas do mapa onde o parámetro de suavizado h é menor, se non que a cobertura parece ser mais grande nesas zonas conforme o parámetro n_0 é mais pequeno.

Os resultados obtidos parecen mostrar que a cobertura dos intervalos de confianza de f'_h para todo h considerado, en parte é independente do parámetro n_0, ou polo menos podemos afirmar que en ningún caso parece ser inferior ao nivel $(1 - \alpha)$ considerado.

Nas simulacións levadas a cabo nas Seccións 3.1 e 3.2 vimos que a maior taxa de erro no SiZer se producía en zonas do mapa onde o parámetro de suavizado h ten un valor mais pequeno, é dicir, o oposto ao obtido na análise de cobertura dos intervalos de confianza de f_h. Que está a ocorrer realmente nesas zonas do mapa SiZer? Mostrémolo a través dun exemplo.

No seguinte caso, simulamos unha mostra de $n = 200$ a partir da función de densidade *Bimodal* (#6) de Marron e Wand (1992). Na Figura 3.29 estamos a facer a estimación do intervalo de confianza de $\hat{f}'_h$, cun nivel de significación $\alpha = 0.05$, e cun parámetro de suavizado relativamente grande tal que $\log(h) = -1$. A curva f'_h, a cal se ve representada en cor negro, é unha curva que debido ao parámetro de suavizado empregado non presenta valores demasiado grandes xa sexan positivos ou negativos. Ao obter a estimación de $\hat{f}'_h$, debido a que estamos a empregar un parámetro h grande, baixo cada ventá temos unha cantidade de observacións relativamente constante que provoca que a estimación de $\hat{f}'_h$ non presente unha variabilidade grande. A pesar diso, debido á lentitude no crecemento e decrecemento da curva $\hat{f}_h$, a estimación de $\hat{f}'_h$ vai a conter valores próximos ao cero, provocando ou ben que os intervalos de confianza de $\hat{f}'_h$ conteñan o verdadeiro valor, ou no caso de non contelo na maioría dos casos vai a conter o valor cero, provocando que no SiZer obteñamos puntos púrpuras e non erráticos. Na Figura 3.29, podemos ver como neste caso nas zonas onde se está a facer inferencia

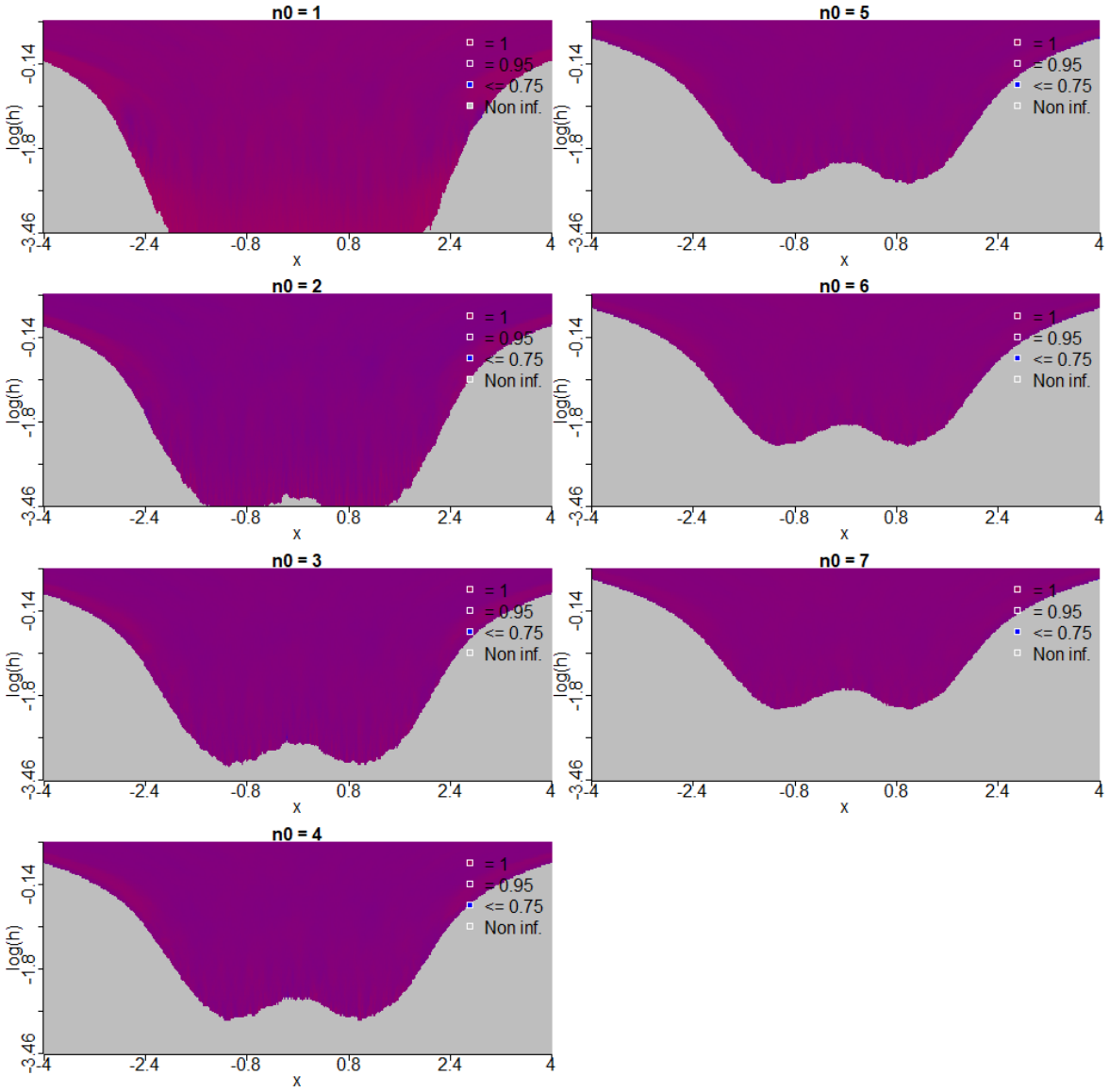

Figura 3.28: Taxa de cobertura do cuantil q_1 con respecto a (x, h), con $n_0 = \{1, 2, 3, 4, 5, 6, 7\}$, para $N = 1000$ mostras, con tamaño de mostra $n = 50$, xeradas a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

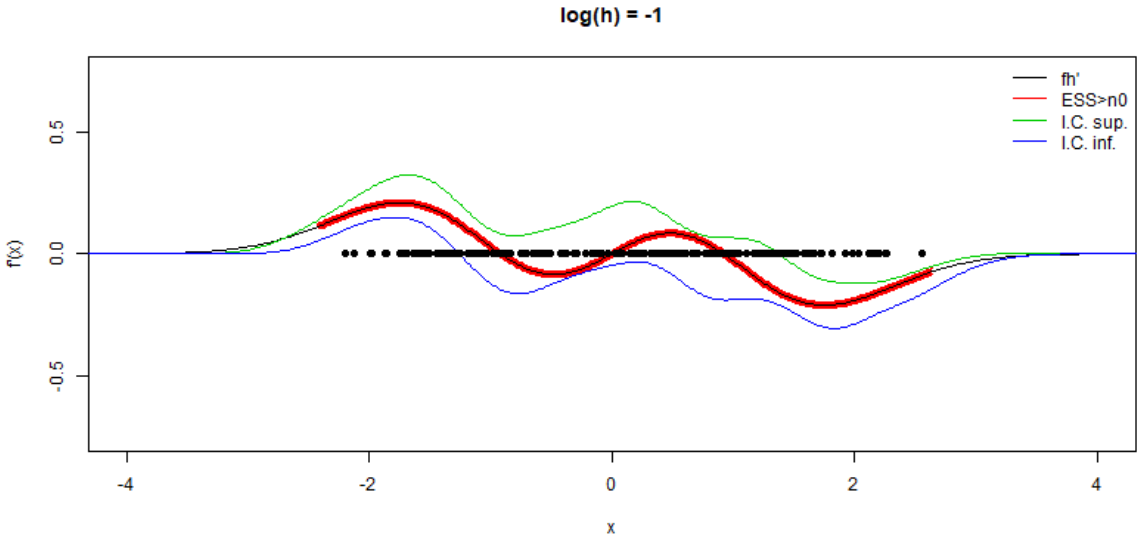

Figura 3.29: Intervalo de confianza de $\hat{f}'_h(x)$ con $\log(h) = -1$ para unha mostra de tamaño $n = 200$ xerada a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

no SiZer (é dicir, onde $ESS > 5$, representadas de cor vermello), os intervalos de confianza (límite superior en verde, e límite inferior en azul), conteñen sempre o verdadeiro valor de f'_h. O mais lóxico sería que conterá o verdadeiro valor de f'_h no $(1 - \alpha) = 0.95$ dos puntos, pero debido á mostra obtida (representada con puntos negros) a estimación parece mellorar a cobertura esperada. Podemos ver como no vale (entorno ao punto $x = -0.5$), ou na segunda das modas (entorno ao punto $x = 0.5$) o intervalo de confianza está a punto de deixar o verdadeiro valor de f'_h fóra do intervalo de confianza, pero aínda así tendendo a conter o valor cero, o que non provocaría unha estimación errática na cor do SiZer.

Por outra banda, temos a estimación dos intervalos de confianza de f'_h na Figura 3.30, onde se utilizou un parámetro de suavizado menor tal que $\log(h) = -2.66$. A elección dun parámetro de suavizado pequeno implica que a variabilidade de $\hat{f}'_h$ sexa maior, provocando por un lado que a estimación de $\hat{f}'_h$ obteña valores afastados do verdadeiro valor de f'_h, e por outro lado que os intervalos de confianza sexan mais amplos, provocando polo tanto que conteñan o verdadeiro valor nunha taxa do nivel $(1 - \alpha)$. Isto vese claramente reflectido sobre a Figura 3.30, onde os intervalos de confianza son moito mais amplos que no caso de $\log(h) = -1$ (Figura 3.29), pero debido á forma dos intervalos de confianza presentes na gráfica, a estimación de $\hat{f}_h$ parece ser moi distinta á verdadeira curva f'_h. Se nos centramos entorno aos puntos comprendidos no intervalo $x = [-0.5, 0]$, podemos ver como existen dúas zonas onde os intervalos de confianza non conteñen o verdadeiro valor de f'_h. Vendo dunha forma mais clara esta zona sobre a Figura 3.31 podemos ver que no medio desas dúas zonas onde o intervalo de confianza de $\hat{f}'_h$ non cubre o verdadeiro valor de f'_h se encontra un cúmulo de observacións da mostra. Esta parte da mostra está provocando, debido ao parámetro de suavizado pequeno seleccionado, que a estimación $\hat{f}_h$ sexa extremadamente crecente antes do cúmulo e extremadamente decrecente despois dó cúmulo, orixinando unha moda inexistente sobre a verdadeira curva f_h. Como nesa zona estamos a obter unha moda bastante alta que non existe na curva f_h, os intervalos de confianza de $\hat{f}'_h$, aínda coa gran variabilidade da curva nesa zona, non son capaces de conter o verdadeiro valor de f'_h ou o valor cero, provocando que erros habituais como este, cando o parámetro de suavizado h é pequeno, se

traduzan en modas inexistentes sobre o mapa SiZer. Aínda que os intervalos de confianza de $\hat{f}'_h$ conteñen o verdadeiro valor de f'_h nunha taxa próxima ao nivel $1-\alpha$, a estimación de f'_h ten unha gran variabilidade que en ocasións provoca que modas inexistentes en f_h sexan significativas na estimación de $\hat{f}_h$.

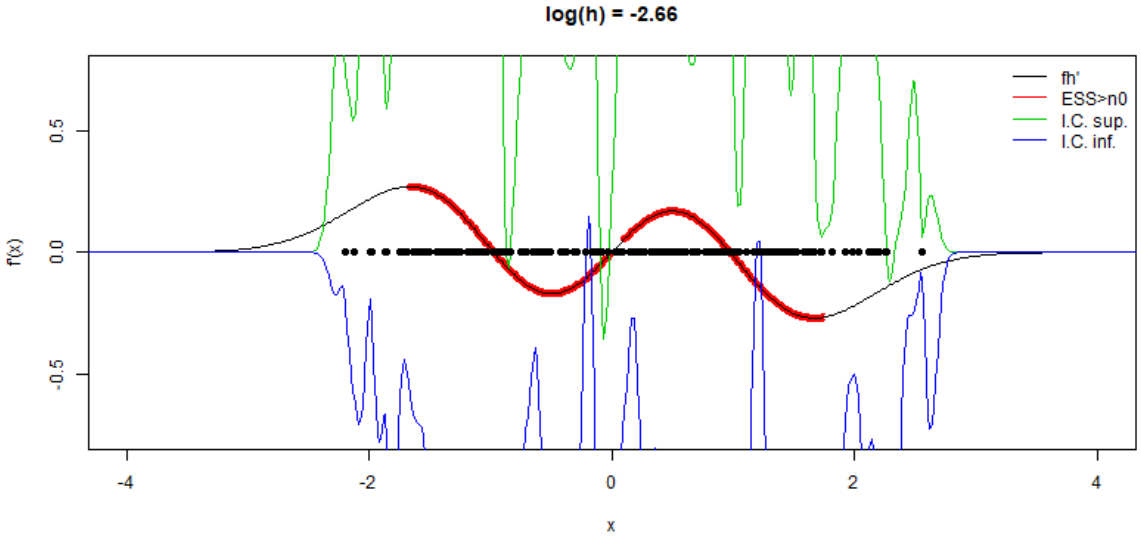

Figura 3.30: Intervalo de confianza de $\hat{f}'_h(x)$ con $\log(h) = -2.66$ para unha mostra de tamaño $n = 200$ xerada a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

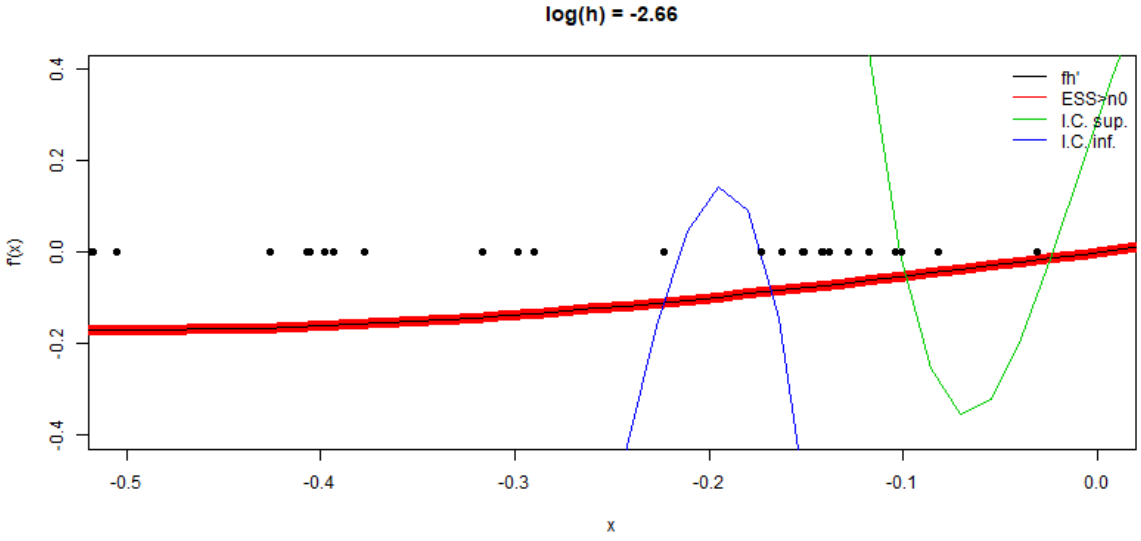

Figura 3.31: Intervalo de confianza de $\hat{f}'_h(x)$ no rango $x = [-0.5, 0]$, con $\log(h) = -1$ para unha mostra de tamaño $n = 200$ xerada a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

Por último, podemos ver sobre a Figura 3.32 como efectivamente o que acabamos de mencionar ocorre na estimación do mapa SiZer para o cuantil q_1, presentando unha moda inexistente no intervalo $x = [-0.5, 0]$ do mapa onde o parámetro de suavizado h é menor, unha moda que non está presente na función de densidade *Bimodal* (#6) como podemos apreciar a través da Figura B.6 do Apéndice B.

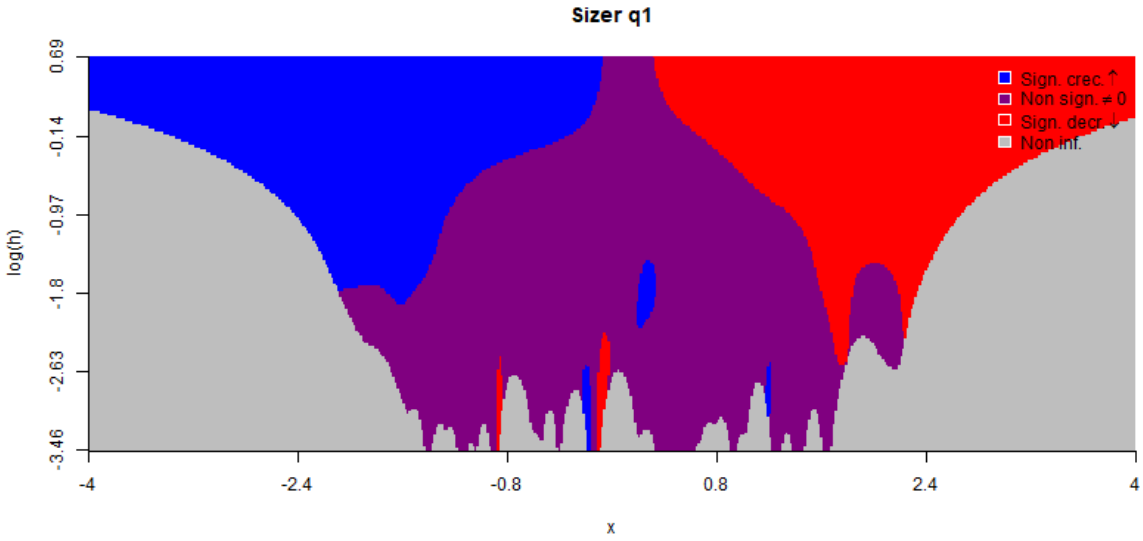

Figura 3.32: SiZer dunha mostra de tamaño $n = 200$ xerada a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

Tamaño de mostra reducido

Sen embargo, a cobertura dos intervalos de confianza non parece ser tan axeitada para tamaños de mostra máis reducidos.

Analizando en primeiro lugar os resultados obtidos para a densidade *Strongly Skewed* (#3) de Marron e Wand (1992), podemos ver a través da Figura 3.33 como case en todo momento, para calquera valor de n_0, a cobertura dos intervalos de confianza parece encontrarse por debaixo do valor $(1 - \alpha) = 0.95$. Os resultados non parecen ser tan obvios para o tamaño de mostra $n = 20$ (gráfica dereita da Fígura 3.33), pero observando os resultados obtidos para $n = 10$ (gráfica esquerda), podemos ver a que cobertura dos intervalos de confianza obtidos co cuantil q_1 ronda o valor 0.90, moi por debaixo do valor correcto, $(1 - \alpha) = 0.95$.

Por outra banda, no mesmo exemplo de simulación levado a cabo coa densidade *Bimodal* (#6) de Marron e Wand (1992), podemos ver a través da Figura 3.34 como os resultados obtidos son semellantes aos da Figura 3.33. Para o tamaño de mostra $n = 10$, de novo a cobertura dos intervalos de confianza parece rondar o valor 0.90, mentres que neste caso para o tamaño de mostra $n = 20$ parece verse de forma mais clara a baixa cobertura dos intervalos de confianza do cuantil q_1 para este tamaño de mostra.

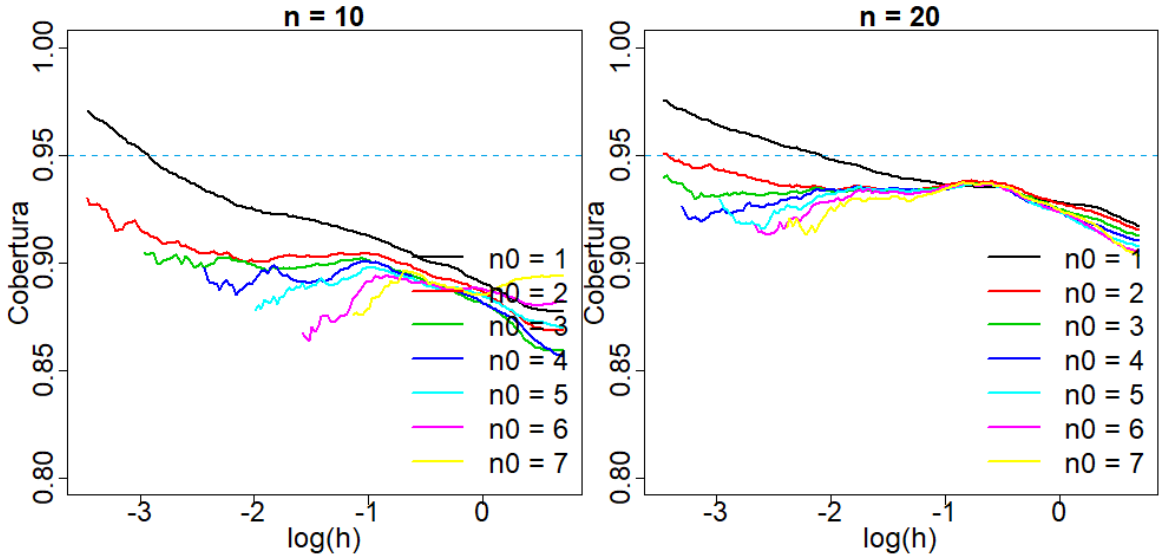

Figura 3.33: Taxa de cobertura do cuantil q_1 con respecto a h, con $n_0 = \{1, 2, 3, 4, 5, 6, 7\}$, para $N = 1000$ mostras, con tamaño de mostra $n = \{10, 20\}$, xeradas a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

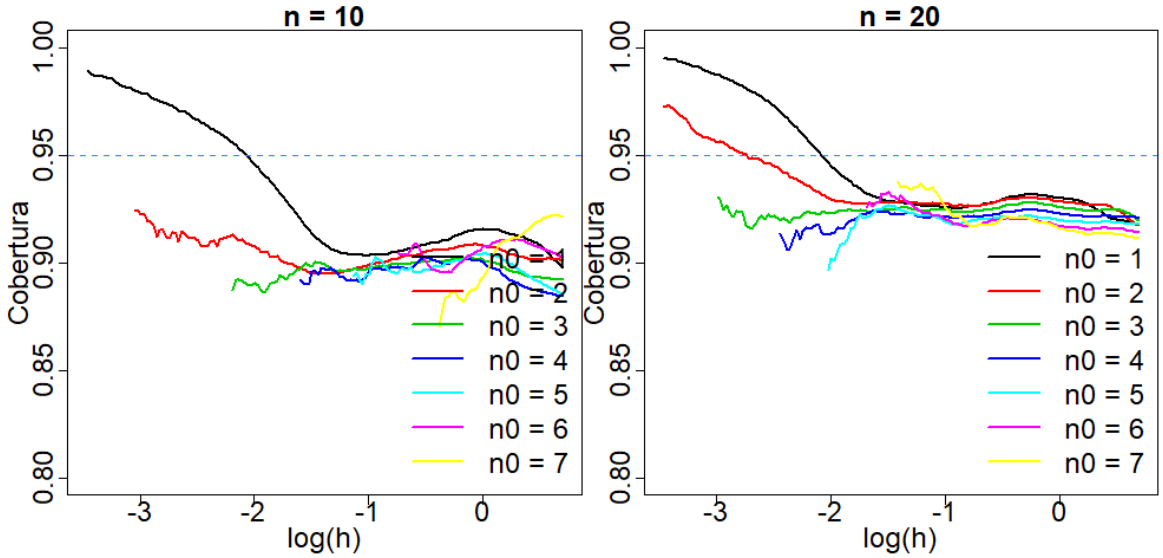

Figura 3.34: Taxa de cobertura do cuantil q_1 con respecto a h, con $n_0 = \{1, 2, 3, 4, 5, 6, 7\}$, para $N = 1000$ mostras, con tamaño de mostra $n = \{10, 20\}$, xeradas a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

O cuantil Gaussiano non parece ser axeitado para obter taxas de cobertura do nivel $(1 - \alpha)$ cando o tamaño de mostra é pequeno. Cando o tamaño de mostra medra, os cuantís converxen aos dunha normal, pero para tamaños mais reducidos parece obvio empregar outra ferramenta que nos permita obter unha converxencia do erro mais rápida ao valor cero.

Para tratar de mellorar isto, como mencionamos anteriormente, tamén levamos a cabo este exercicio de simulación no que comparamos o cuantil q_1 coas versións bootstrap do mesmo: o cuantil percentil (q_5), o cuantil percentil simetrizado (q_6), o cuantil percentil-t (q_7), e o cuantil percentil-t simetrizado (q_8). Neste caso a simulación foi realizada para o valor $n_0 = 5$, e para os tamaños de mostra $n = \{10, 20, 50\}$.

Observando en primeiro lugar os resultados obtidos para as mostras simuladas a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992), podemos ver a través da Figura 3.35, como claramente para o tamaño de mostra $n = 10$ a taxa de cobertura do cuantil q_1 se encontra moi por debaixo do valor correcto $(1 - \alpha)$. A converxencia do erro deste cuantil ao valor cero é da orde $O(n^{-1})$, a cal non é suficiente para obter unha boa taxa de cobertura para este tamaño de mostra. Por outra banda, podemos ver como os cuantís q_5, q_6 e q_7 obteñen uns resultados moi semellantes aos do cuantil q_1. A converxencia do erro destes cuantís é da orde $O(n^{-1/2})$ para o cuantil q_5, da orde $O(n^{-1})$ para o cuantil q_7, e descoñecida para o cuantil q_6, e a pesar de ter ordes diferentes, podemos ver como os resultados dos catro cuantís son moi similares. Por último fixándonos nos resultados obtidos polo cuantil q_8, podemos ver como a taxa de cobertura deste cuantil é moi superior á dos demais cuantís. A velocidade converxencia do erro deste cuantil ao valor cero é maior a dos demais cuantís, da orde de $O(n^{-3/2})$, e podemos observar claramente como en todo momento a taxa de cobertura ronda o valor $(1 - \alpha) = 0.95$.

Por outra banda, observando os resultados obtidos para o tamaño de mostra $n = 20$, podemos ver como os cuantís q_1, q_5, q_6 e q_7 aínda ofrecen unha taxa de cobertura inferior a do nivel $(1 - \alpha)$. O cuantil q_1 parece ofrecer unha taxa de cobertura considerablemente estable para todo h inferior ao nivel de significación $(1 - \alpha)$, descendendo levemente esa taxa para valores do parámetro de suavizado h maiores. O mesmo parece ocorrer para os cuantís q_5 e q_6 pero ofrecendo unha taxa de cobertura un pouco inferior a do cuantil q_1. Por outra banda, se nos fixamos no cuantil q_7 podemos ver como este ofrece unha taxa de cobertura superior a do nivel $(1 - \alpha)$ para valores de h pequenos, pero descende rapidamente a valores inferiores ao do nivel $(1 - \alpha)$ a medida que o parámetro ventá h medra. De novo, estes catro cuantís non parecen ser axeitados para ofrecer unha boa taxa de cobertura para este tamaño de mostra, pero observando os resultados obtidos para o cuantil q_8, podemos ver como este ofrece unha taxa de cobertura bastante estable entorno ao valor $(1 - \alpha)$ para todo h, mostrando unha vez mais que ofrece os mellores resultados para un tamaño de mostra reducido.

Por último, observando os resultados obtidos para o tamaño de mostra $n = 50$, vese claramente como tanto o cuantil Gaussiano q_1, e o cuantil bootstrap percentil-t simetrizado q_8 ofrecen os mellores resultados, dando lugar a valores da taxa de cobertura entorno ao nivel $(1 - \alpha)$ para todo h. A medida que o tamaño de mostra medra, os cuantís óptimos tenden a aproximarse aos cuantís dunha densidade normal, ofrecendo polo tanto este tan bos resultados como o cuantil q_8. Por outro lado, podemos observar como tanto os cuantís q_5 como o cuantil q_6 aínda ofrecen unha taxa de cobertura inferior ao nivel $(1 - \alpha)$ para todo h. Ademais, o cuantil q_7 de novo ofrece unha cobertura por encima do nivel $1 - \alpha$ para valores de h pequenos, mentres que a taxa vai diminuíndo ata valores inferiores ao nivel $(1 - \alpha)$ a medida que o parámetro de suavizado h medra.

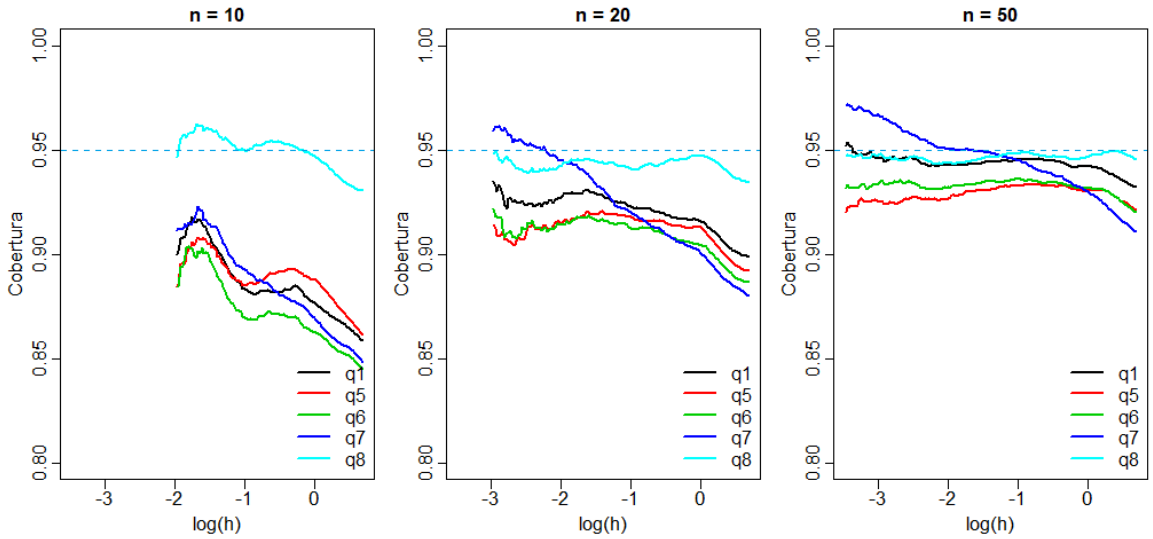

Figura 3.35: Taxa de cobertura do cuantil q_1, q_5, q_6, q_7 e q_8 con respecto a h, con $n_0 = 5$, para $N = 1000$ mostras, con tamaño de mostra $n = \{10, 20, 50\}$, xeradas a partir dunha variable aleatoria con función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

Analizando agora os resultados obtidos para as mostras xeradas a partir dunha variable aleatoria coa densidade *Bimodal* (#6) de Marron e Wand (1992), podemos observar a través da Figura 3.36, como de novo para o tamaño de mostra $n = 10$, o único cuantil que é capaz de obter unha taxa de cobertura entorno ao valor $(1 - \alpha)$ e o cuantil q_8. Nesta ocasión, o cuantil q_8 parece ofrecer valores de cobertura menos estables entorno a ese valor con respecto a h, pero en ningún momento ofrece unha taxa de cobertura inferior ao nivel 0.92, ou superior ao nivel 0.97. Por outra banda, todos os demais cuantís parecen ofrecer os mesmos resultados. Presentan unha gran variabilidade da taxa de cobertura con respecto a h, pero ofrecendo para a maioría dos casos taxas de cobertura inferiores ao nivel 0.90.

Para o tamaño de mostra $n = 20$, de novo, os resultados ofrecidos son moi similares aos obtidos para a función de densidade *Strongly Skewed* (#3) de Marron e Wand (1992). É dicir, unha maior estabilidade da taxa de cobertura do cuantil q_8 entorno ao valor $(1 - \alpha)$ con respecto ao parámetro de suavizado h, e o mesmo para os cuantís q_1, q_5 e q_6 pero obtendo taxas de cobertura inferiores ao nivel $(1 - \alpha)$. Ademais, de novo o cuantil q_7 comeza ofrecendo taxas de cobertura superiores ao nivel $(1 - \alpha)$ para parámetros ventá pequenos, e vai diminuíndo esta taxa a medida que o parámetro h medra ata situarse en valores inferiores a 0.95.

E por último, podemos ver para o tamaño de mostra $n = 50$ como de novo o cuantil q_1 ofrece unha taxa de cobertura estable con respecto a h entorno ao nivel $(1 - \alpha)$, mostrando unha vez mais que os cuantís tenden a aproximarse aos dunha densidade normal a medida que o tamaño de mostra medra. Ademais, obtemos resultados similares para o cuantil q_8, e de novo, os cuantís q_5 e q_6 ofrecen taxas de cobertura inferiores a $(1 - \alpha)$. Por outra banda, podemos ver como de novo o cuantil q_7 volve a seguir o mesmo patrón, ofrecendo taxas de cobertura moi altas para valores de h pequenos, e taxas de cobertura moi baixas para valores de h grandes.

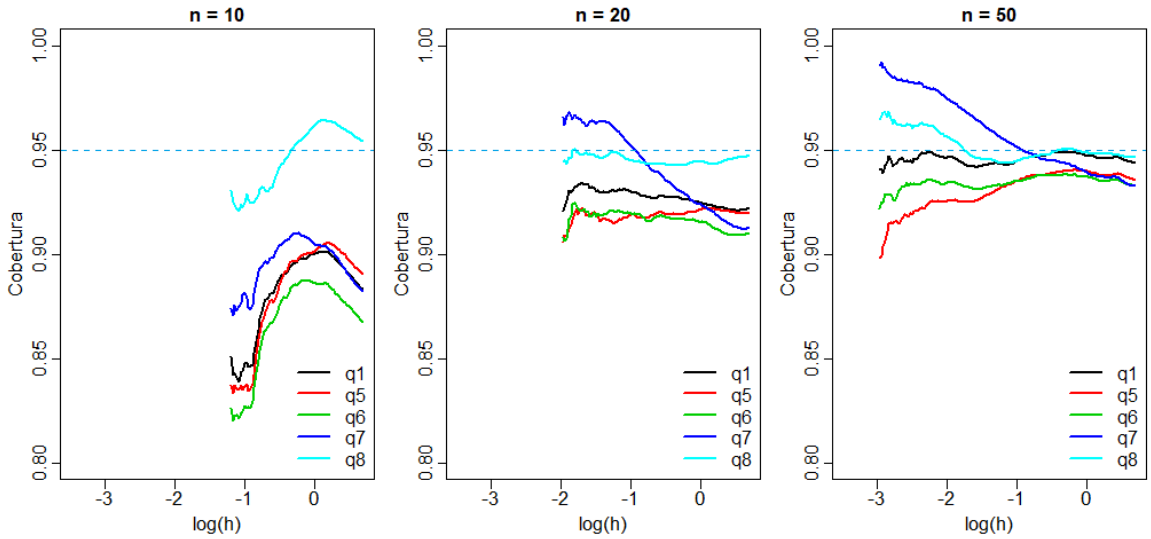

Figura 3.36: Taxa de cobertura do cuantil q_1, q_5, q_6, q_7 e q_8 con respecto a h, con $n_0 = 5$, para $N = 1000$ mostras, con tamaño de mostra $n = \{10, 20, 50\}$, xeradas a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

Como vimos, o cuantil q_1 ofrece taxas de cobertura do intervalo de confianza de $\hat{f}'_h$ entorno ao nivel $(1 - \alpha)$ independentemente do valor do parámetro de suavizado h cando o tamaño de mostra é relativamente grande (nas simulacións levadas a cabo mostra resultados satisfactorios para $n \geq 50$). Esta taxa de cobertura deixa de ser correcta cando o tamaño de mostra diminúe, e como vimos a través das simulacións coas densidades *Strongly Skewed* (#3) e *Bimodal* (#6) de Marron e Wand (1992), a taxa é inferior ao nivel $(1 - \alpha) = 0.95$, situándose incluso por debaixo do valor 0.90 para un tamaño de mostra $n = 10$.

Unha alternativa ao cuantil Gaussiano q_1 para aqueles tamaños de mostra onde o cuantil non ten boas taxas de cobertura, parece ser o cuantil percentil-t simetrizado q_8. Este cuantil, a diferencia do cuantil q_1, mostra nas simulacións obter unha taxa de cobertura entorno ao valor $(1 - \alpha)$ independentemente do parámetro de suavizado h cando o tamaño de mostra é reducido. O obxectivo principal do cuantil q_1 é ofrecer taxas de cobertura do intervalo de confianza de $\hat{f}'_h$ do valor $(1 - \alpha)$, este non é capaz de lograr esas taxas cando o tamaño de mostra é reducido, polo tanto, o cuantil q_8 supón unha alternativa mellor cando n é reducido.

Capítulo 4

Conclusións

No contexto da estimación non paramétrica da densidade, se ben esta permite recuperar a forma global da curva de interese, un problema crucial é a selección do parámetro de suavizado. É ben sabido que valores altos do parámetro levarán aparelladas estimacións sobresuavizadas, mentres que valores baixos tenderán a infrasuavizar. Así, a selección óptima deste parámetro é un dos ámbitos de traballo máis relevantes no contexto da estatística non paramétrica. Sen embargo, e como xa se ten indicado, en moitas ocasións a cuestión de interese pode respostarse identificando as características significativas da curva (é dicir, os patróns de crecemento e decrecemento). Cando este problema se aborda dende unha perspectiva non paramétrica, teñen cabida as ideas de localización e escala da visión artificial. É aquí onde xorde a ferramenta SiZer, a cal estuda de forma simultánea un rango amplo de parámetros de suavizado h sobre a estimación tipo núcleo da derivada da función de densidade. Baséase en estudar os intervalos de confianza da función, onde Chaudhuri e Marron (1999) propoñen catro formas de obter o cuantil: o cuantil independente Gaussiano para cada punto x (q_1), unha aproximación simultánea sobre x do cuantil Gaussiano baseado no número de bloques independentes sobre os datos (q_2), e dous métodos baseados en bootstrap, un deles simultáneo sobre x (q_3) e o outro simultáneo sobre ambos, x e h (q_4). Ademais, neste traballo foron propostos catro cuantís mais para tratar de mellorar a cobertura ofrecida polos intervalos de confianza da derivada da función de densidade a través do cuantil q_1: o bootstrap percentil (q_5), bootstrap percentil simetrizado (q_6), bootstrap percentil-t (q_7) e o bootstrap percentil-t simetrizado (q_8).

O obxectivo principal do traballo foi levar a cabo un estudo de simulación que nos permitira por unha banda realizar unha comparación exhaustiva dos resultados ofrecidos polos diferentes SiZers obtidos a partir dos cuantís propostos en Chaudhuri e Marron (1999), e por outra banda, unha comparacións da cobertura ofrecida polos intervalos de confianza dos SiZers a través do cuantil q_1 e os novos métodos propostos. Debido á alta carga computacional, levouse a cabo unha implementación numérica da estimación tipo núcleo da derivada da función de densidade, así como da súa varianza a través transformada rápida de Fourier (FFT), o cal logrou mellorar de forma sobresaínte, a través da súa implementación en R, os tempos de execución do estudo de simulación levado a cabo.

Co primeiro exemplo de simulación tratamos de ver por medio do SiZer promedio, o mapa de acerto e o mapa de erro en que zonas do mapa SiZer, a través dos catro cuantís propostos por Chaudhuri e Marron (1999), se estaban a identificar de forma correcta patróns de crecemento/decremento (zonas azuis/vermellas). Os resultados mostraron, como se supoñía a priori, que en media as cores do SiZer obtido co cuantil q_1 tenden a parecerse en maior medida ás cores do SiZer da curva suavizada f_h. Pero por outra banda, vimos que o SiZer obtido a través do cuantil q_1 asumía unha maior taxa de erro no que se refire a obter conclusións de crecemento (decrecemento) en zonas da curva f_h nos que realmente a curva e decrecente (crecente). Con respecto ao cuantil q_4 ocorre xustamente o oposto: o cuantil tende a devolver un mapa SiZer moito mais conservador, provocando que as cores do mapa sexan en menor medida iguais ás do mapa SiZer de f_h, pero asumindo unha taxa de erro ínfima estimando a cor azul (vermella) nun punto (x, h), cando a verdadeira cor é vermella (azul).

No segundo exemplo de simulación introducimos a ferramenta MoSiZer, a cal nos permitiu ver dunha forma obxectiva en que zonas do mapa o SiZer está a cometer un maior erro para mostrar modas presentes ou non sobre a curva f_h. Os resultados mostraron que en ocasións os cuantís q_2, q_3 e q_4 poden chegar a ser moi conservadores. Estiman de forma acertada a presencia de modas moi definidas sobre a función f_h, pero en casos nos que o tamaño de mostra é máis reducido e as modas non están tan acentuadas, teñen dificultades para mostrar as modas presentes sobre a curva f_h. Por outra banda o cuantil q_1 ten maior facilidade para mostrar modas presentes en f_h, pero cometendo un maior erro que finalmente se ve traducido en modas inexistentes sobre a curva orixinal. Unha particularidade foi que o maior erro cometido polo cuantil q_1 para detectar modas, a pesar de que os intervalos de confianza para a curva $f'_h(x)$ en todos os puntos (x, h) do mapa SiZer ofrecen unha taxa de cobertura similar, se cometía en zonas inferiores do mapa onde o parámetro de suavizado é inferior, a medida que o tamaño de mostra medraba. Isto levounos á segunda parte deste exemplo de simulación onde puidemos ver a través do mesmo modelo co cuantil q_1 e coa variación do parámetro n_0, como a correcta elección deste último parece depender non só do tamaño de mostra, se non que tamén da forma da curva orixinal f. Isto deixa aberta a posibilidade de futuros estudos que estean centrados na correcta elección deste parámetro tratando de evitar a problemática de non coñecer a función de densidade f.

O último exemplo de simulación estivo centrado en estudar a taxa de cobertura do cuantil q_1 e os cuantís bootstrap propostos q_5, q_6, q_7 e q_8 como posibles alternativas a este primeiro. Para tamaños de mostra elevados a taxa de cobertura dos intervalos de confianza obtidos a través do cuantil q_1 tenden a aproximarse ao nivel $(1 - \alpha)$, pero non ocorre o mesmo cando o tamaño de mostra se ve reducido, ofrecendo taxas de cobertura inferiores e dando lugar a un maior erro na estimación do mapa SiZer. Os resultados ofrecidos polo exemplo mostran que conseguimos ofrecer unha taxa de cobertura máis próxima ao nivel $(1 - \alpha)$ co cuantil bootstrap percentil-t simetrizado (q_8) cando o tamaño de mostra é pequeno. A pesar da maior carga computacional debido ao procedemento bootstrap, este presenta unha boa e mellor alternativa ao cuantil q_1 para obter unha taxa de cobertura acertada, e polo tanto, a posibilidade de obter resultados mais acertados a través da ferramenta SiZer.

Apéndice A

Intervalos de confianza convencionais

Na estatística paramétrica, un enfoque tradicional para mostrar a variabilidade dos estimadores dos parámetros de interese son os intervalos de confianza. Moitos intentos foron realizados para estender esa idea á estimación non paramétrica da función de densidade. Pero hai dous obstáculos para o uso efectivo desta técnica:

- En lugar de estimar o intervalo de confianza dun só parámetro, son obtidos os intervalos de confianza dunha curva completa. Ademais a inferencia sobre característica involucra aspectos da inferencia simultánea.

- Ao contrario que a estimación paramétrica, a estimación de curvas involucra unha importante parte do nesgo.

Se o obxectivo é facer inferencia sobre a verdadeira curva f, entón o enfoque clásico dos intervalos de confianza non é axeitado debido a que ignora o nesgo. E ademais, a natureza puntual dos intervalos fainos demasiado curtos para facer inferencia.

Un enfoque tradicional para manexar o nesgo é facer que sexa insignificante sen mais que infrasuavizando a curva, é dicir, empregando un parámetro de suavizado moi pequeno. Isto é feito simplemente asumindo que asintoticamente cando o tamaño de mostra medra, o parámetro ventá tende mais rápido a cero que o óptimo, facendo que o nesgo tenda a cero mais rápido. Isto deixa aberto, a forma de como debería ser seleccionado o tamaño do parámetro ventá, e o feito de que para unha mostra fixa calquera parámetro de suavizado terá polo menos un pouco de nesgo. Pero incluso ignorando eses problemas, os intervalos de confianza baseados en parámetros de suavizado non son atractivos, xa que poden chegar a ser demasiado longos e polo tanto, características significativas poden non chegar a ser encontradas.

Outra aproximación consiste en intentar estimar o nesgo e axustalo en consecuencia. Un intento disto foi presentado en Hardle e Marron (1991) o cal foi asintoticamente exitoso, pero proporciona unha cobertura incorrecta nas simulacións.

Coñecedor do fracaso nos métodos da corrección do nesgo, Hall (presentado mais detalladamente en Hall 1991) propuxo elexir o parámetro de suavizado para facer que as probabilidades de cobertura tan próximas como sexa posible aos valores desexados. Teoría asintótica foi desenvolvida para obter o parámetro ventá seguindo este criterio, e mostra que cando parámetros de suavizado óptimos son usados obtemos intervalos de confianza mais curtos que no caso de empregar calquera tipo de corrección do nesgo.

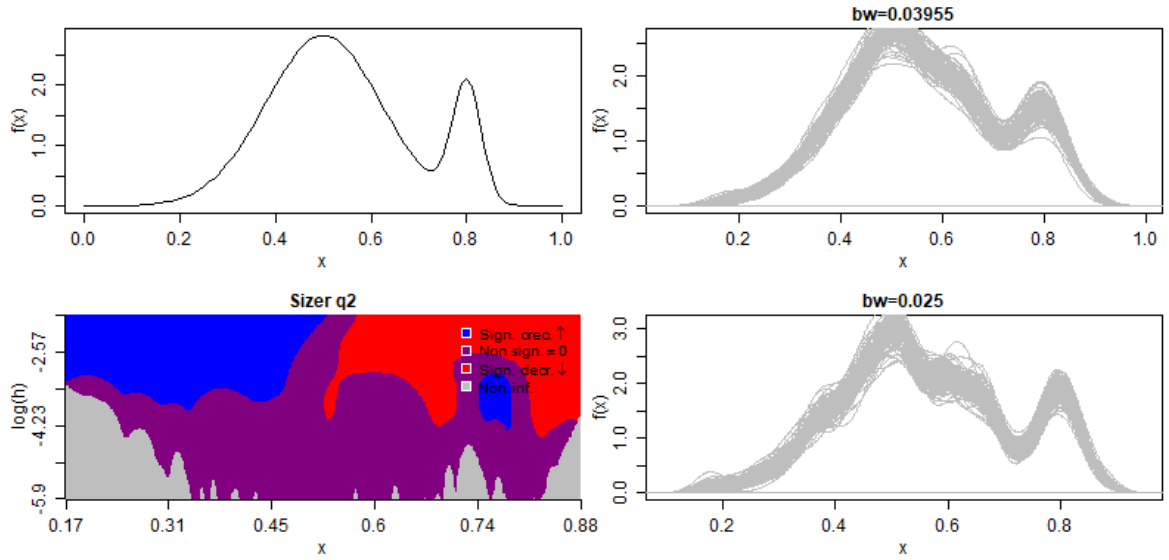

Figura A.1: Función de densidade f (arriba-esquerda), SiZer de $\boldsymbol{X} = (X_1, \ldots, X_{500})$ m.a.s. de X con densidade f (abaixo-esquerda), estimación tipo núcleo con $h = 0.03955$ de 100 réplicas bootstrap de $\boldsymbol{X}$ (arriba-dereita) e estimación tipo núcleo con $h = 0.0250$ de 100 réplicas bootstrap de $\boldsymbol{X}$ (abaixo-dereita).

Analizando mais detalladamente os parámetros de ventá óptimos a pregunta é: como é a cobertura destes intervalos? A Figura A.1 mostra un exemplo centrado neste punto.

A gráfica superior esquerda mostra a función de densidade, a cal é unha mestura de normais tal que

$$0.85 \cdot N(0.5, 0.0144) + 0.15 \cdot N(0.8, 0.0009)$$

Aquí imos estudar a estimación a partir dunha mostra con $n = 500$, poñendo o foco na moda situada no punto $x = 0.8$. A efectos prácticos a cobertura do intervalo de confianza para un parámetro de suavizado óptimo, seguindo a regra do pulgar de Silverman, é mostrado na gráfica superior dereita desa mesma Figura, onde superpoñemos a estimación tipo núcleo con ese mesmo parámetro de suavizado de 100 réplicas bootstrap independentes. Neste caso, a envoltura das curvas suxire que, con este parámetro ventá, non hai suficiente información na mostra para que a moda situada en $x = 0.8$ sexa significativa, xa que envoltura superior cerca do vale situado en $x = 0.72$ está mais arriba que a envoltura inferior da moda en $x = 0.8$. Por outra banda, a gráfica inferior dereita da mesma Figura, onde se está a empregar un parámetro de suavizado menor $h = 0.025$, suxire que hai suficiente información na mostra para que a segunda moda sexa significativa. A envoltura aí obtida mostra que a este nivel de resolución hai moita información e que a segunda moda debería ser unha característica significativa. O mapa SiZer da gráfica inferior esquerda da mesma figura encontra ambas modas, e polo tanto está utilizando a información dispoñible na mostra mais eficientemente que os intervalos de confianza obtidos co parámetro de suavizado óptimo.

Cabe destacar que incluso se fora posible obter de forma efectiva os intervalos de confianza clásicos (sen a problemática do nesgo), o SiZer sería aínda unha ferramenta analítica mais efectiva. Os intervalos de confianza necesitan centrarse en un único parámetro de suavizado, o cal (incluso cando pode ser escollido de forma correcta a partir da mostra) pode non encontrar características que si aparecen a outros niveis de resolución.

Apéndice B

Densidades de Marron e Wand

Neste Apéndice expóñense as 15 funcións de densidade descritas por Marron e Wand (1992). Estas densidades están compostas por mestura de normais e convertéronse nun estándar na estimación da densidade. Ademais, xunto a representación da función de densidade de cada unha delas, introdúcese o mapa SiZer da función f_h e o MoSiZer obtido a través do SiZer.

Normal (*Gaussian*)

$$\#1 : N(0, 1)$$

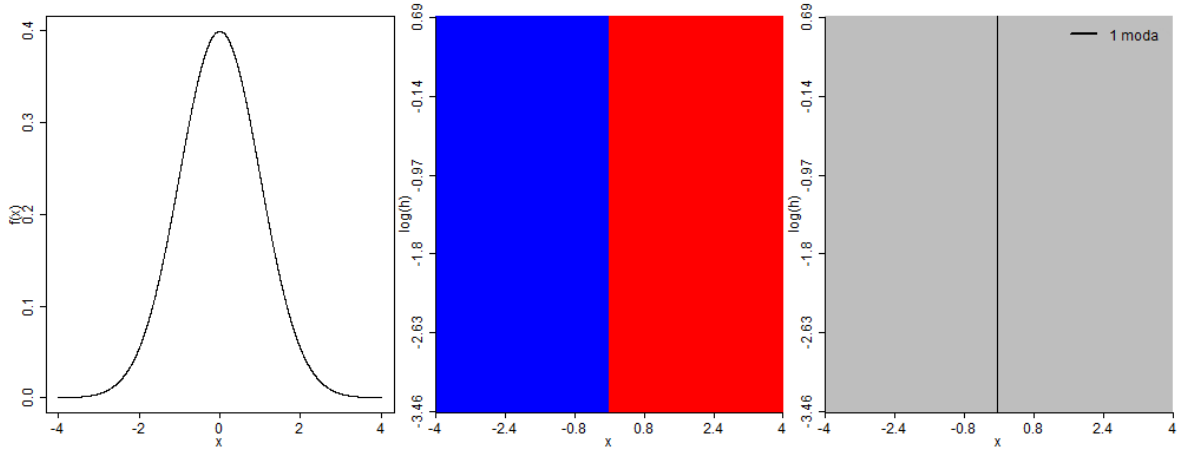

Figura B.1: Densidade f (esquerda), SiZer de f_h (centro) e MoSiZer de f_h (dereita) da densidade *Gaussian* (#1) de Marron e Wand (1992).

Unimodal asimétrica (*Skewed Unimodal*)

$$\#2 : \frac{1}{5} \cdot N(0,1) + \frac{1}{5} \cdot N\left(\frac{1}{2}, \left(\frac{2}{3}\right)^2\right) + \frac{3}{5} \cdot N\left(\frac{13}{12}, \left(\frac{5}{9}\right)^2\right)$$

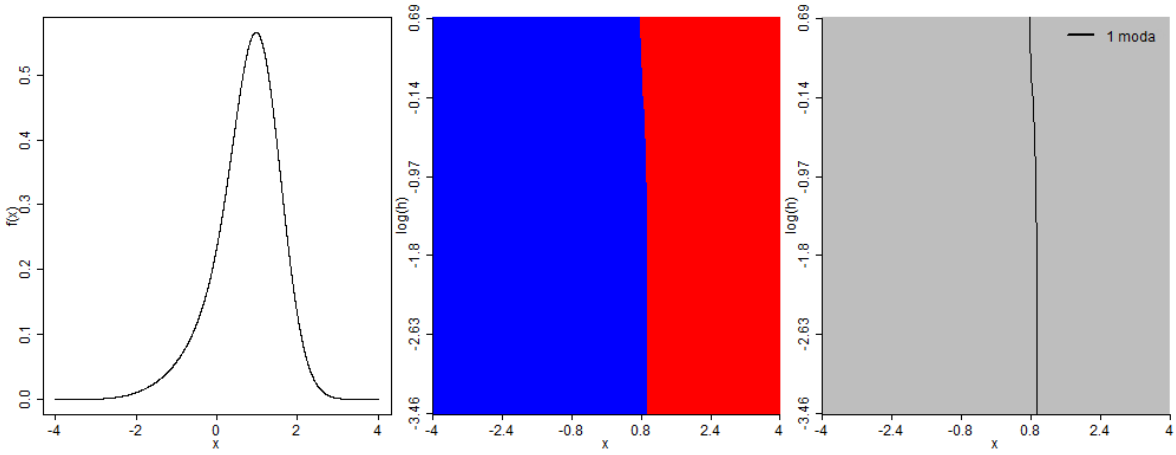

Figura B.2: Densidade f (esquerda), SiZer de f_h (centro) e MoSiZer de f_h (dereita) da densidade *Skewed Unimodal* (#2) de Marron e Wand (1992).

Fortemente asimétrica (*Strongly Skewed*)

$$\#3 : \sum_{l=0}^{7} \frac{1}{8} \cdot N\left(3\left\{(\frac{2}{3})^t - 1\right\}, \left(\frac{2}{3}\right)^{2l}\right)$$

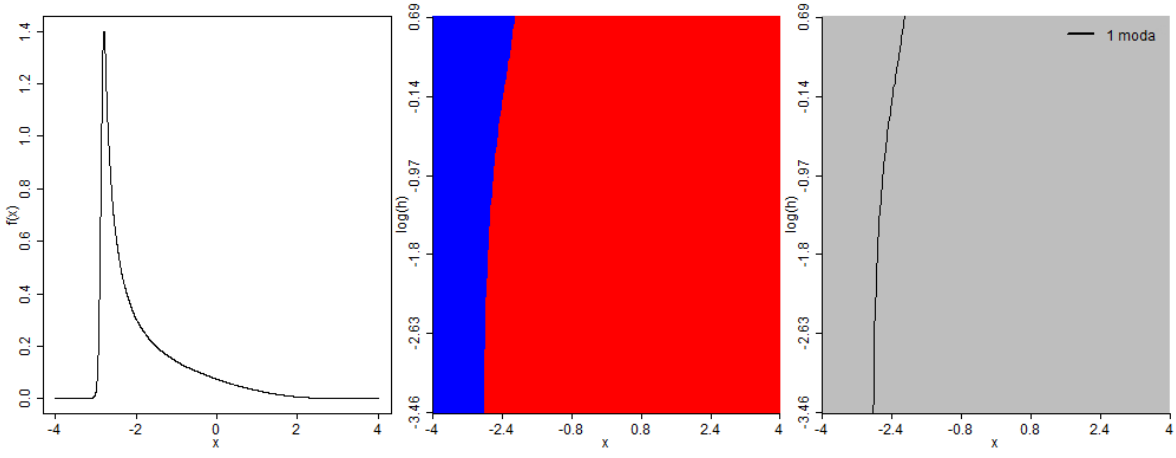

Figura B.3: Densidade f (esquerda), SiZer de f_h (centro) e MoSiZer de f_h (dereita) da densidade *Strongly Skewed* (#3) de Marron e Wand (1992).

Con alta curtose (*Kurtotic*)

$$\#4 : \frac{2}{3} \cdot N(0,1) + \frac{1}{3} \cdot N\left(0, \left(\frac{1}{10}\right)^2\right)$$

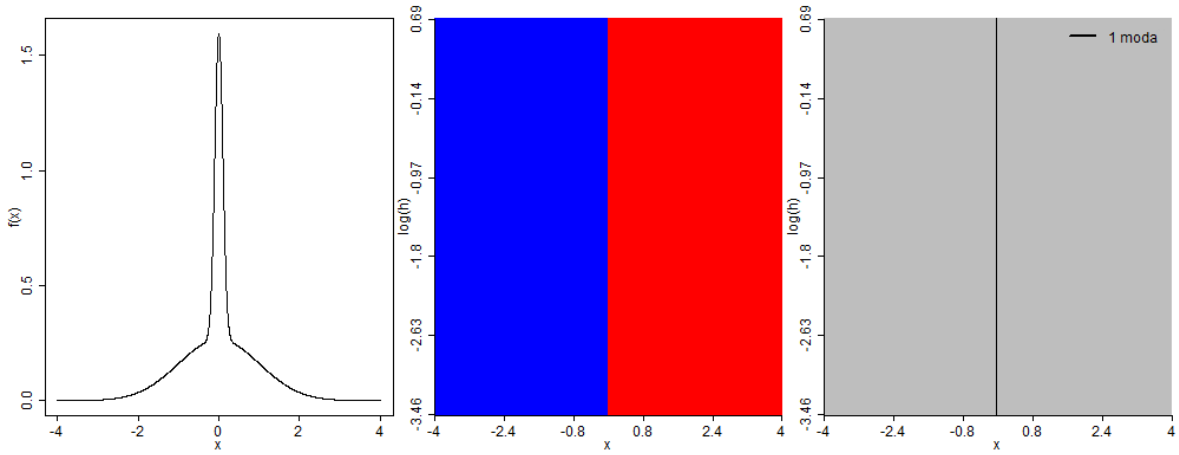

Figura B.4: Densidade f (esquerda), SiZer de f_h (centro) e MoSiZer de f_h (dereita) da densidade *Kurtotic* (#4) de Marron e Wand (1992).

Con atípico (*Outlier*)

$$\#5 : \frac{1}{10} \cdot N(0,1) + \frac{9}{10} \cdot N\left(0, \left(\frac{1}{10}\right)^2\right)$$

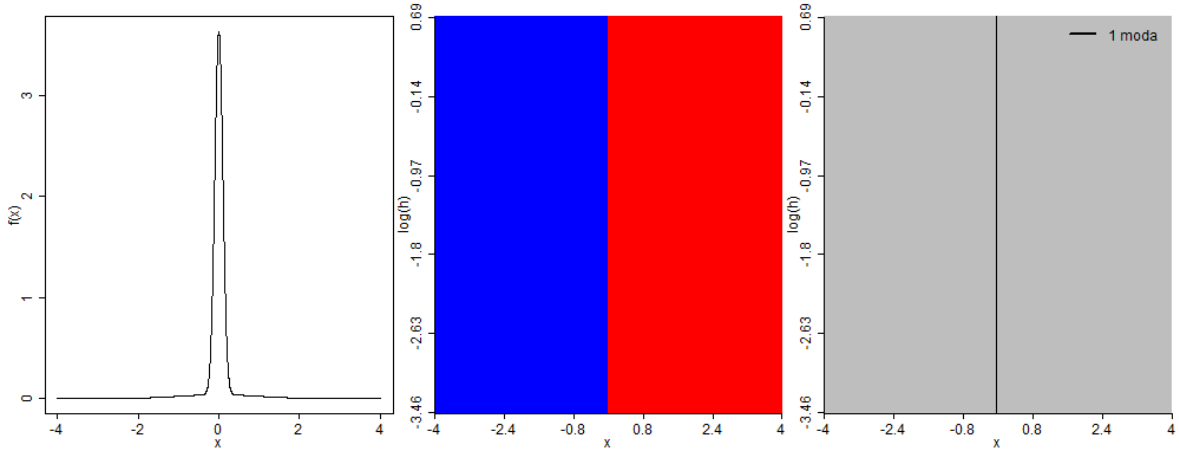

Figura B.5: Densidade f (esquerda), SiZer de f_h (centro) e MoSiZer de f_h (dereita) da densidade *Outlier* (#5) de Marron e Wand (1992).

Bimodal (*Bimodal*)

$$\#6 : \frac{1}{2} \cdot N\left(-1, \left(\frac{2}{3}\right)^2\right) + \frac{1}{2} \cdot N\left(1, \left(\frac{2}{3}\right)^2\right)$$

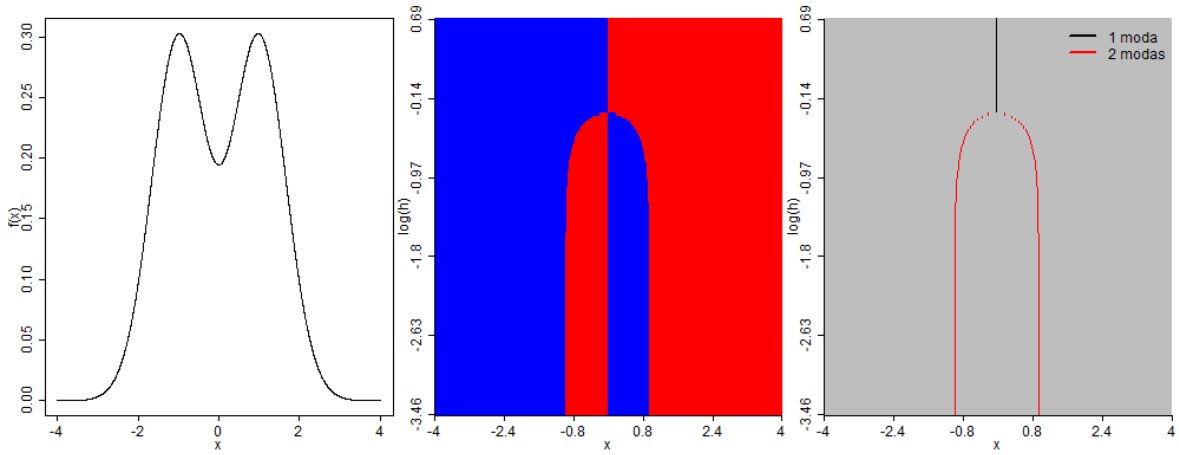

Figura B.6: Densidade f (esquerda), SiZer de f_h (centro) e MoSiZer de f_h (dereita) da densidade *Bimodal* (#6) de Marron e Wand (1992).

Bimodal con separación (*Separated Bimodal*)

$$\#7 : \frac{1}{2} \cdot N\left(-\frac{3}{2}, \left(\frac{1}{2}\right)^2\right) + \frac{1}{2} \cdot N\left(\frac{3}{2}, \left(\frac{1}{2}\right)^2\right)$$

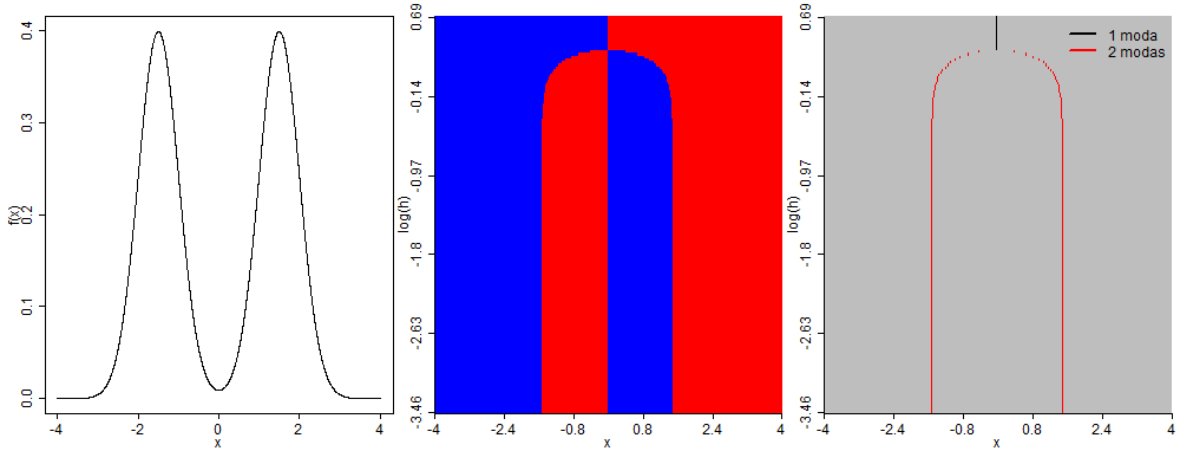

Figura B.7: Densidade f (esquerda), SiZer de f_h (centro) e MoSiZer de f_h (dereita) da densidade *Sep. Bimodal* (#7) de Marron e Wand (1992).

Bimodal asimétrica (*Asymetric Bimodal*)

$$\#8 : \frac{3}{4} \cdot N(0,1) + \frac{1}{4} \cdot N\left(\frac{3}{2}, \left(\frac{1}{3}\right)^2\right)$$

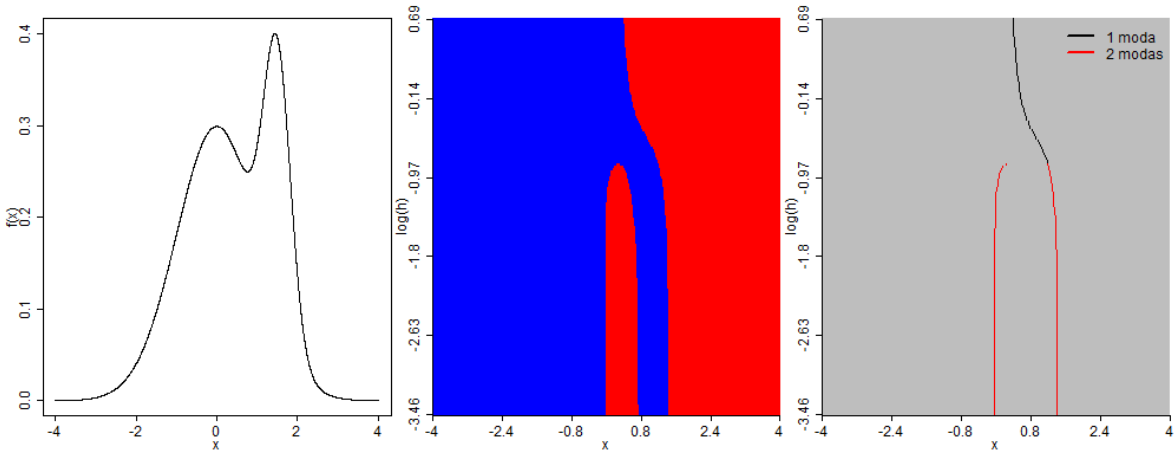

Figura B.8: Densidade f (esquerda), SiZer de f_h (centro) e MoSiZer de f_h (dereita) da densidade *Asym. Bimodal* (#8) de Marron e Wand (1992).

Trimodal (*Trimodal*)

$$\#9 : \frac{9}{20} \cdot N\left(-\frac{6}{5}, \left(\frac{3}{5}\right)^2\right) + \frac{9}{20} \cdot N\left(\frac{6}{5}, \left(\frac{3}{5}\right)^2\right) + \frac{1}{10} \cdot N\left(0, \left(\frac{1}{4}\right)^2\right)$$

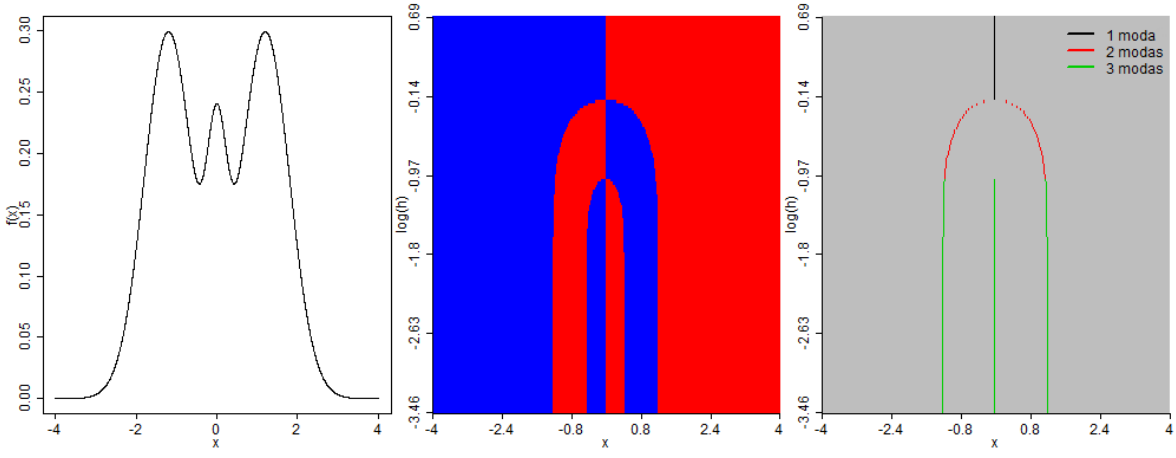

Figura B.9: Densidade f (esquerda), SiZer de f_h (centro) e MoSiZer de f_h (dereita) da densidade *Trimodal* (#9) de Marron e Wand (1992).

Gadoupa (*Claw*)

$$\#10 : \frac{1}{2} \cdot N(0,1) + \sum_{l=0}^{4} \frac{1}{10} \cdot N\left(l/2 - 1, \left(\frac{1}{10}\right)^2\right)$$

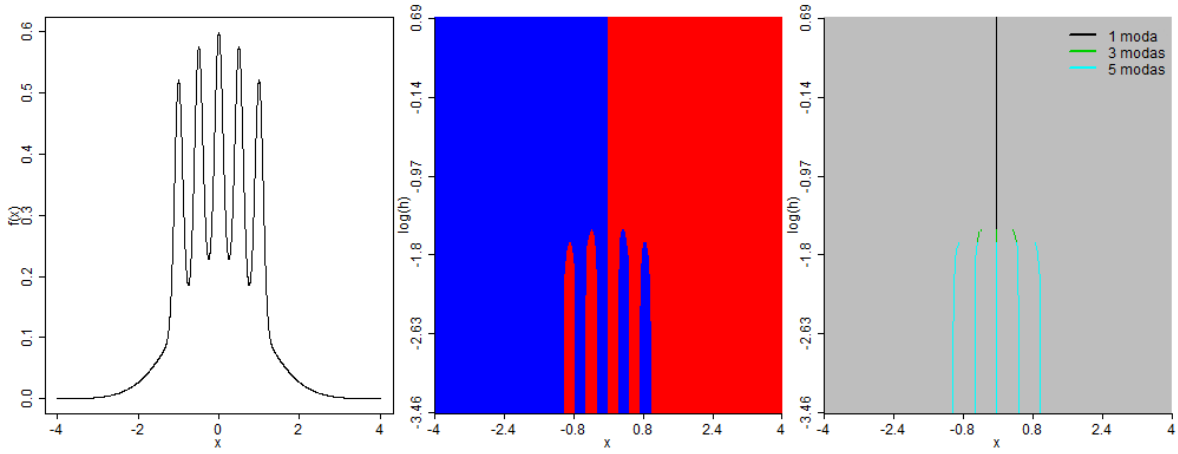

Figura B.10: Densidade f (esquerda), SiZer de f_h (centro) e MoSiZer de f_h (dereita) da densidade *Claw* (#10) de Marron e Wand (1992).

Dobre gadoupa (*Double Claw*)

$$\#11 : \frac{49}{100} \cdot N\left(-1, \left(\frac{2}{3}\right)^2\right) + \frac{49}{100} \cdot N\left(1, \left(\frac{2}{3}\right)^2\right) + \sum_{l=0}^{6} \frac{1}{350} \cdot N\left((l-3)/2, \left(\frac{1}{100}\right)^2\right)$$

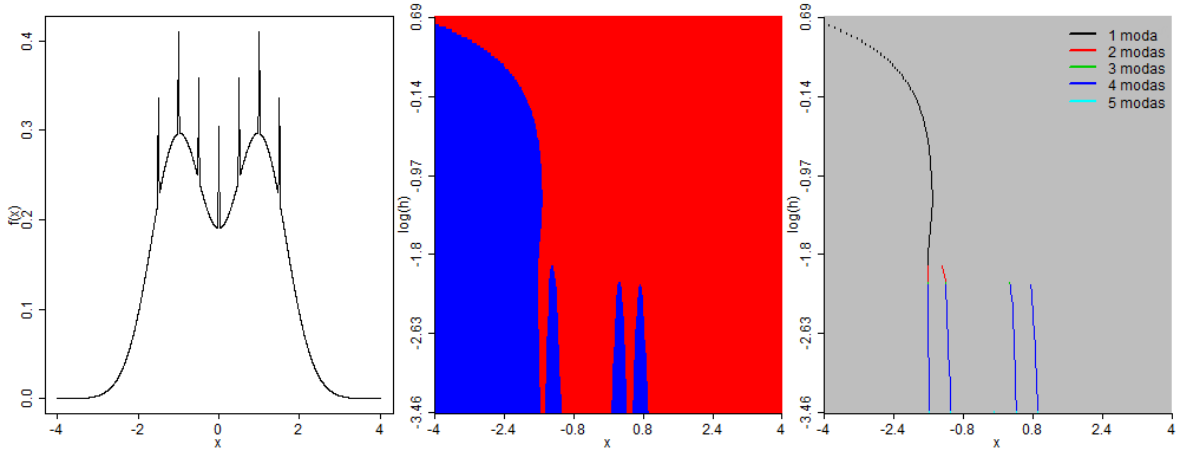

Figura B.11: Densidade f (esquerda), SiZer de f_h (centro) e MoSiZer de f_h (dereita) da densidade *Double Claw* (#11) de Marron e Wand (1992).

Gadoupa asimétrica (*Asymetric Claw*)

$$\#12 : \frac{1}{2} \cdot N(0,1) + \sum_{l=-2}^{2} (2^{1-l}/31) \cdot N\left(l + \frac{1}{2}, (2^{-l}/10)^2\right)$$

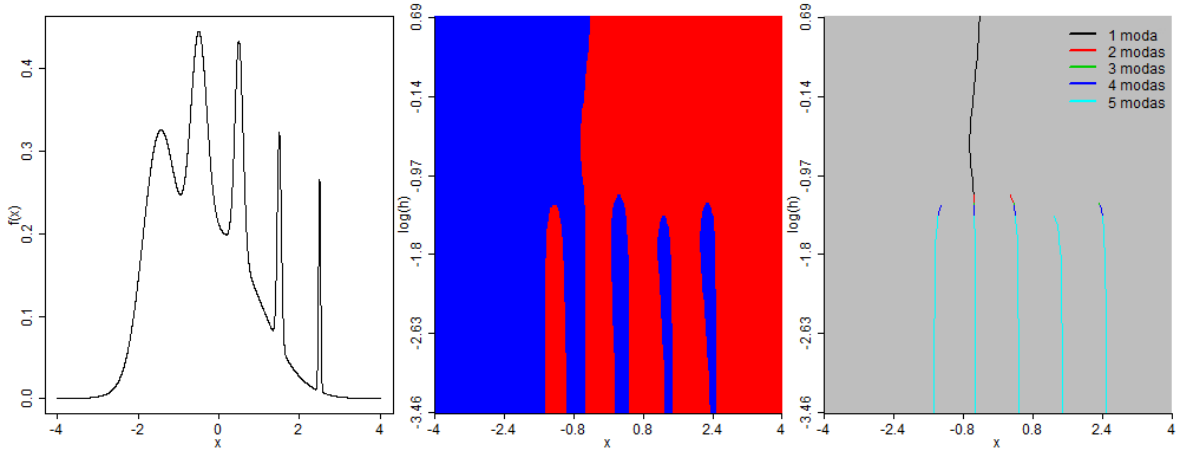

Figura B.12: Densidade f (esquerda), SiZer de f_h (centro) e MoSiZer de f_h (dereita) da densidade *Asym. Claw* (#12) de Marron e Wand (1992).

Dobre gadoupa asimétrica (*Asymetric Double Claw*)

$$\#13 : \sum_{l=0}^{1} \frac{46}{100} \cdot N\left(2l - 1, \left(\frac{2}{3}\right)^2\right) + \sum_{l=1}^{3} \frac{1}{300} \cdot N\left(-l/2, \left(\frac{1}{100}\right)^2\right) + \sum_{l=1}^{3} \frac{7}{300} \cdot N\left(l/2, \left(\frac{7}{100}\right)^2\right)$$

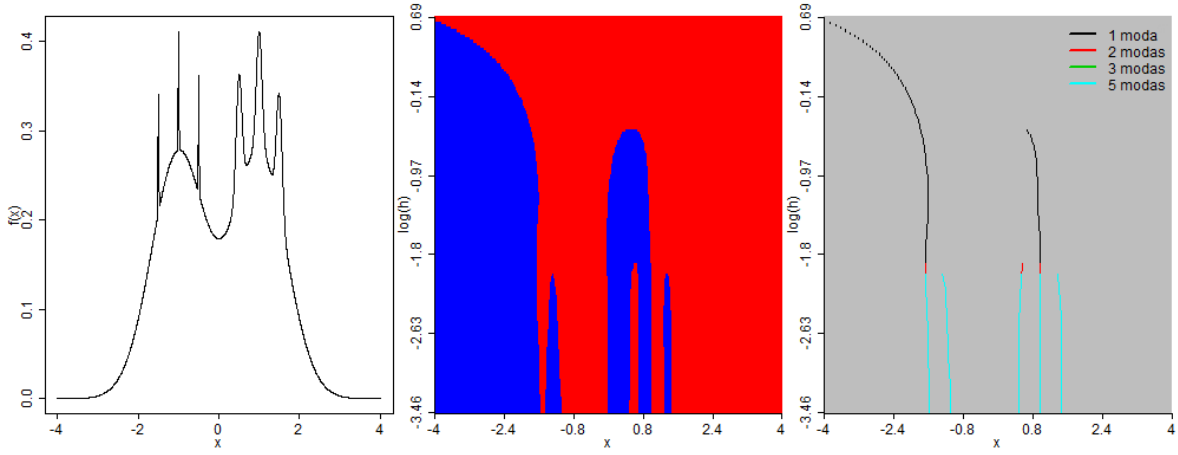

Figura B.13: Densidade f (esquerda), SiZer de f_h (centro) e MoSiZer de f_h (dereita) da densidade *Asym. Db. Claw* (#13) de Marron e Wand (1992).

Peite suave (*Smooth Comb*)

$$\#14 : \sum_{l=0}^{5} (2^{5-l}/63) \cdot N\left(65 - 96\left(\frac{1}{2}\right)^{l}/21, \left(\frac{32}{63}\right)^{2}/2^{2l}\right)$$

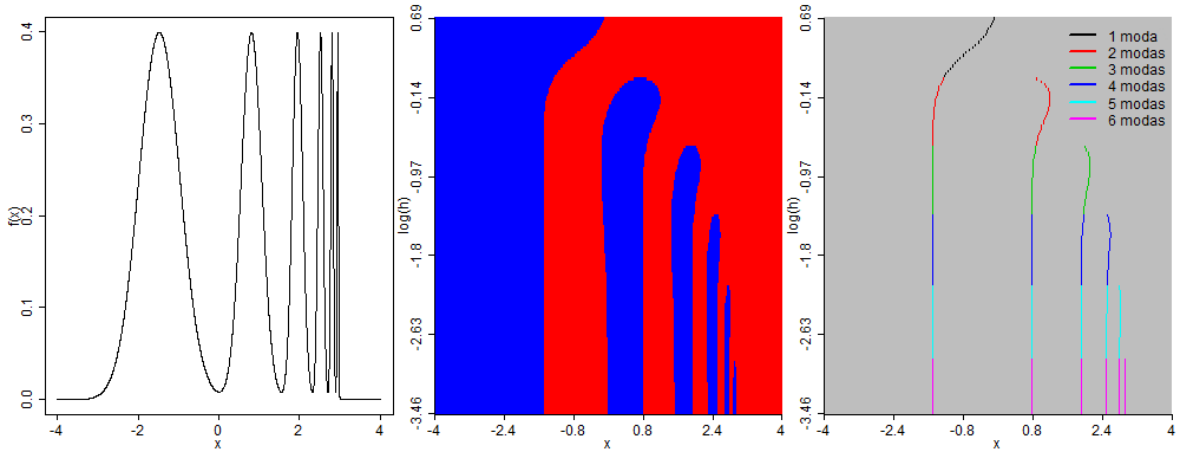

Figura B.14: Densidade f (esquerda), SiZer de f_h (centro) e MoSiZer de f_h (dereita) da densidade *Smooth Comb* (#14) de Marron e Wand (1992).

Peite discreto (*Discrete Comb*)

$$\#15 : \sum_{l=0}^{2} \frac{2}{7} \cdot N\left((12l - 15)/7, \left(\frac{2}{7}\right)^{2}\right) + \sum_{l=8}^{10} \frac{1}{21} \cdot N\left(2l/7, \left(\frac{1}{21}\right)^{2}\right)$$

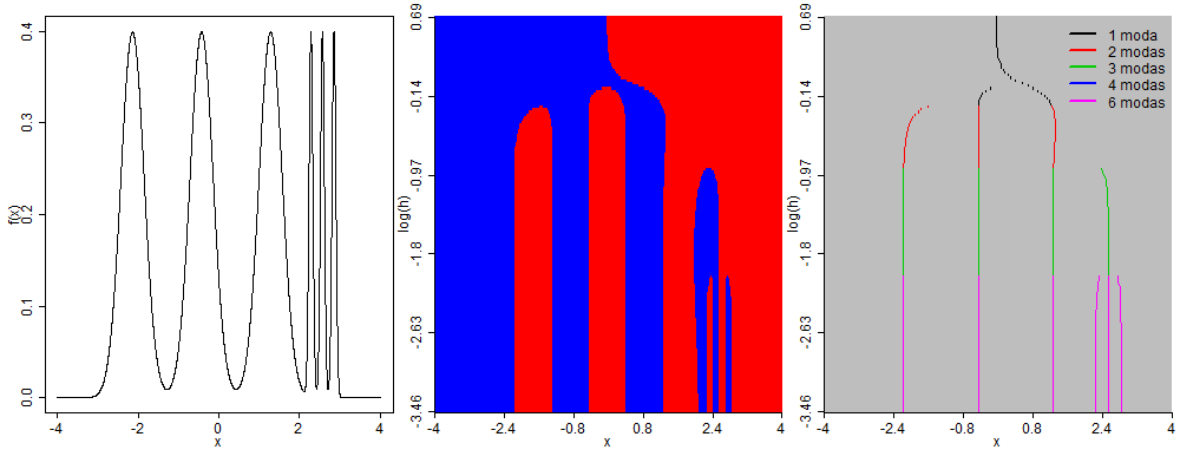

Figura B.15: Densidade f (esquerda), SiZer de f_h (centro) e MoSiZer de f_h (dereita) da densidade *Discrete Comb* (#15) de Marron e Wand (1992).

Apéndice C

Resultados da taxa de acerto e taxa de erro do SiZer

Neste Apéndice son presentados os resultados da simulación levada a cabo co SiZer promedio, o mapa de acerto e o mapa de erro para as mostras simuladas a partir das densidades *Bimodal* (#6) e *Sep. Bimodal* (#7) de Marron e Wand (1992), onde os tamaños de mostra empregados foron $n = \{50, 200, 500\}$, e simulamos un total de $N = 1000$ mostras para cada caso. O nivel de significación empregado é de $\alpha = 0.05$, e ademais, sobre cada un dos mapa SiZer a grella de puntos x estivo comprendida dentro do intervalo $[-4, 4]$, onde se utilizou un total de $n(x) = 512$ puntos. Por outra banda, o número de valores do parámetro de suavizado empregado foi dun total de $n(h) = 151$ puntos.

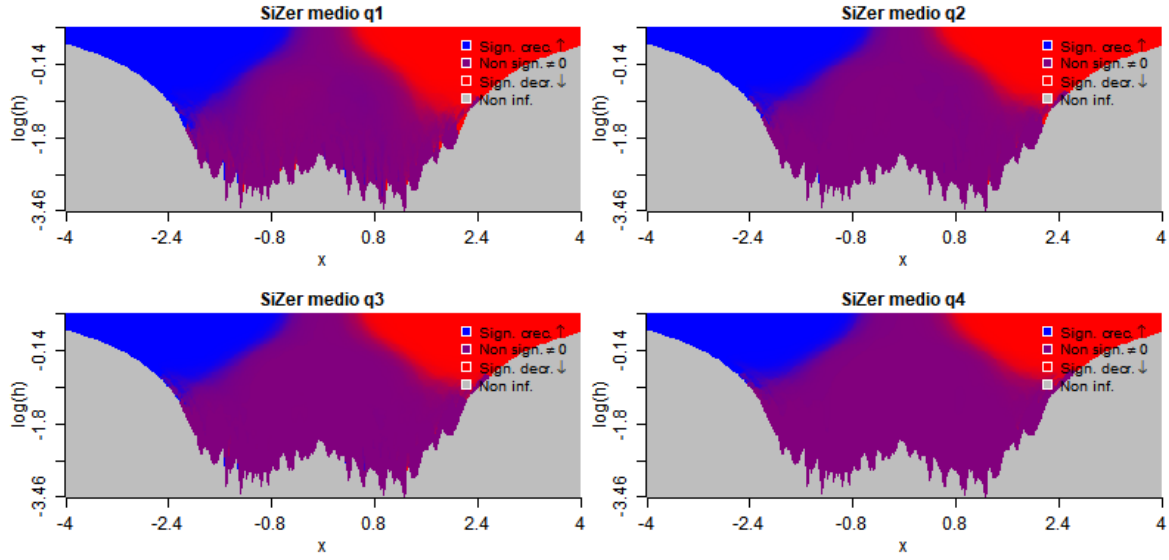

Figura C.1: Estimación do SiZer promedio para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 50$, xeradas a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

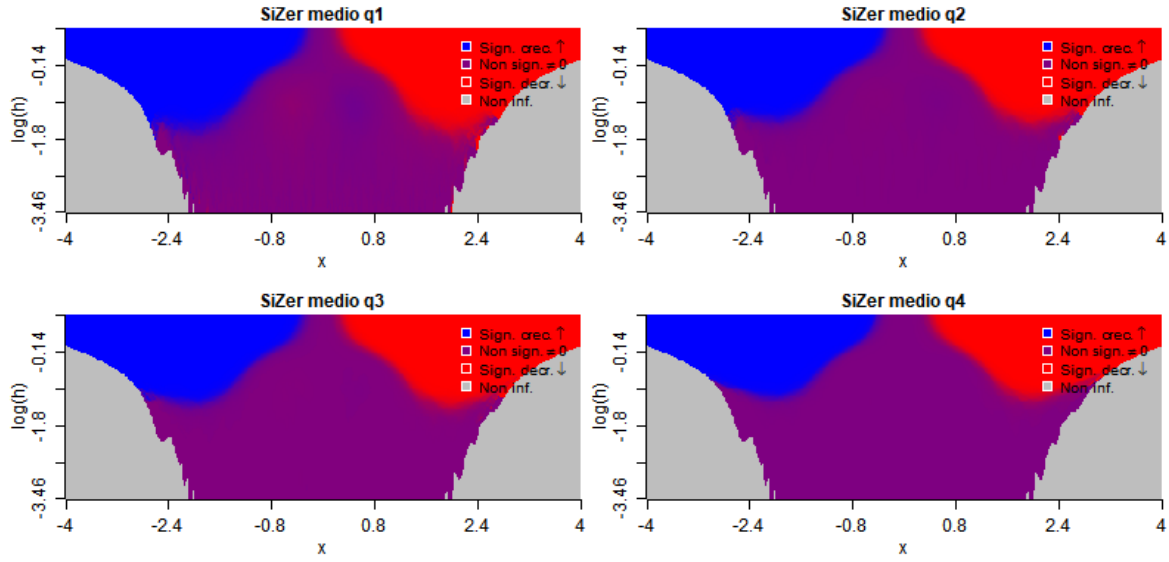

Figura C.2: Estimación do SiZer promedio para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 200$, xeradas a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

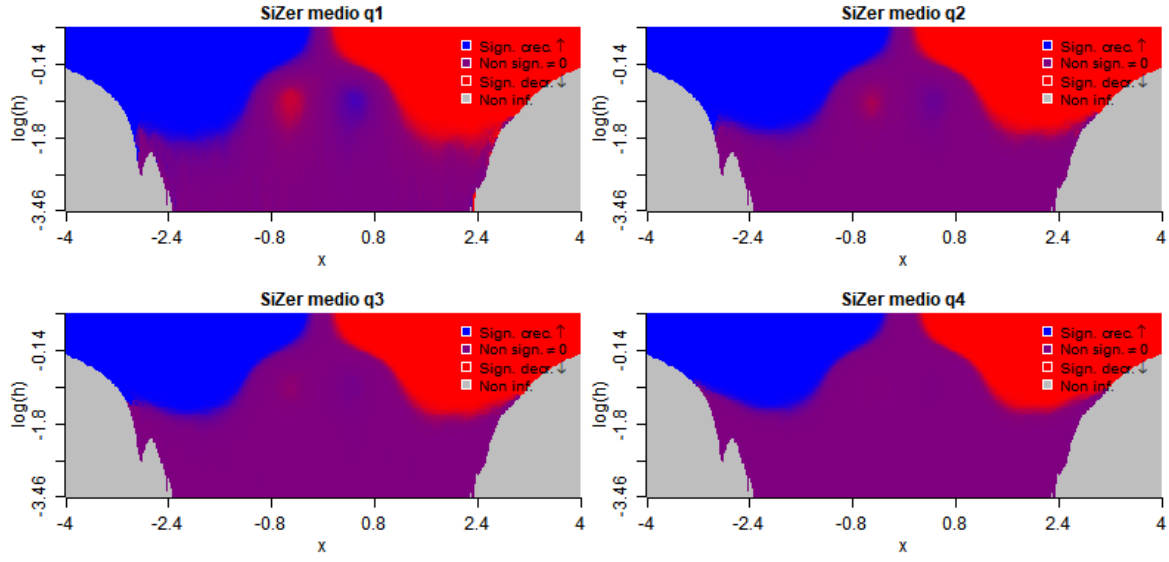

Figura C.3: Estimación do SiZer promedio para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 500$, xeradas a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

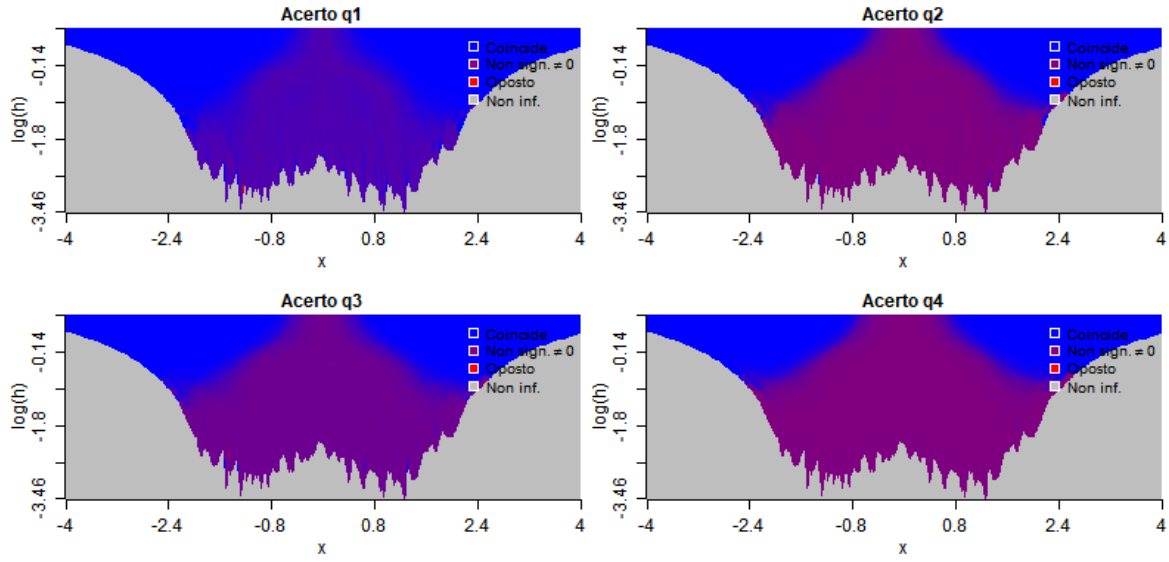

Figura C.4: Mapa de acerto para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 50$, xeradas a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

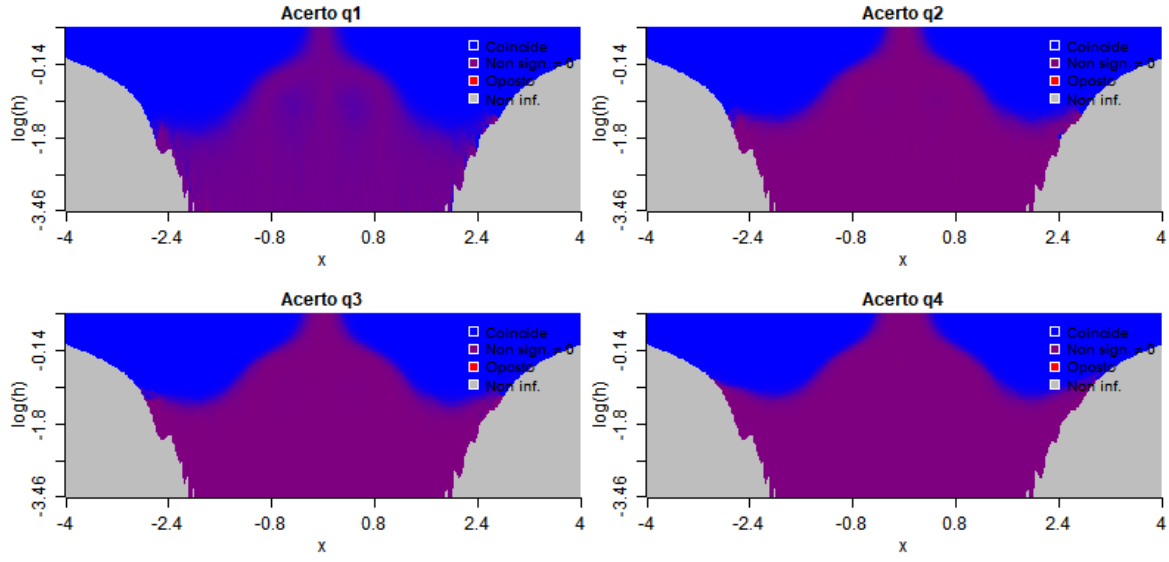

Figura C.5: Mapa de acerto para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 200$, xeradas a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

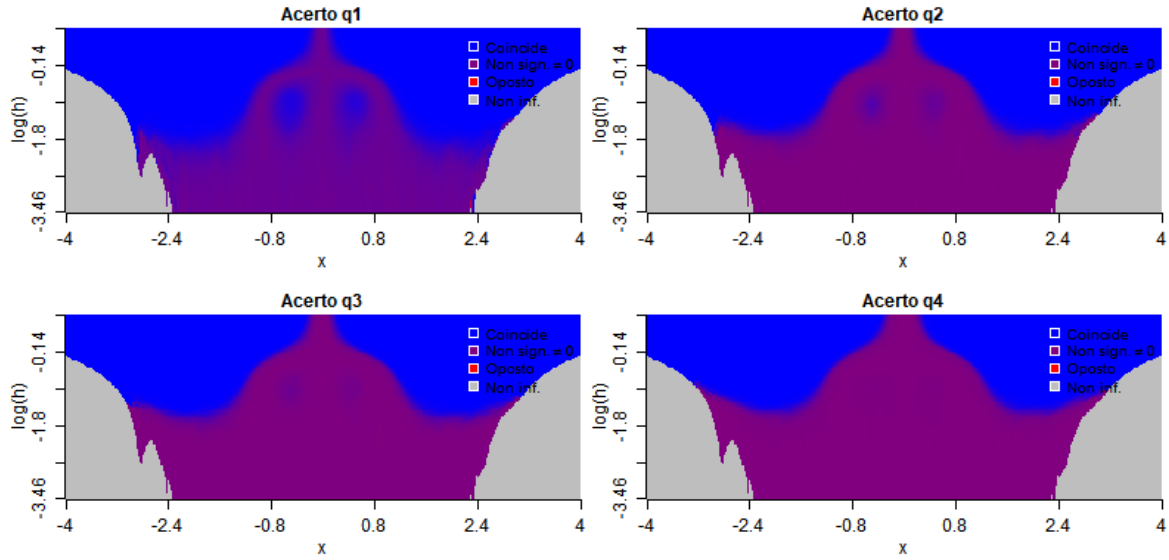

Figura C.6: Mapa de acerto para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 500$, xeradas a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

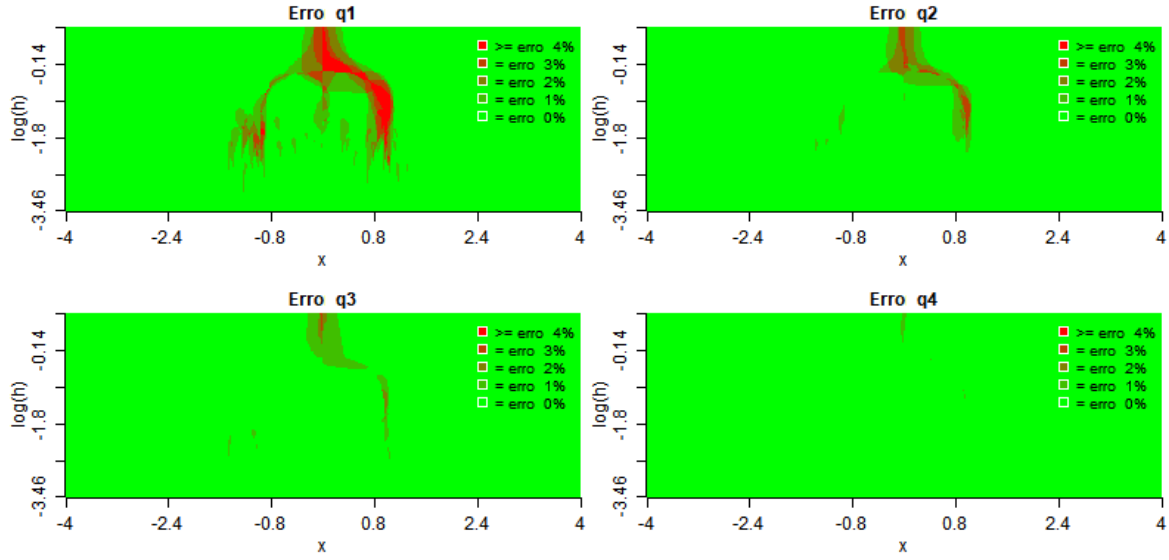

Figura C.7: Mapa de erro para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 50$, xeradas a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

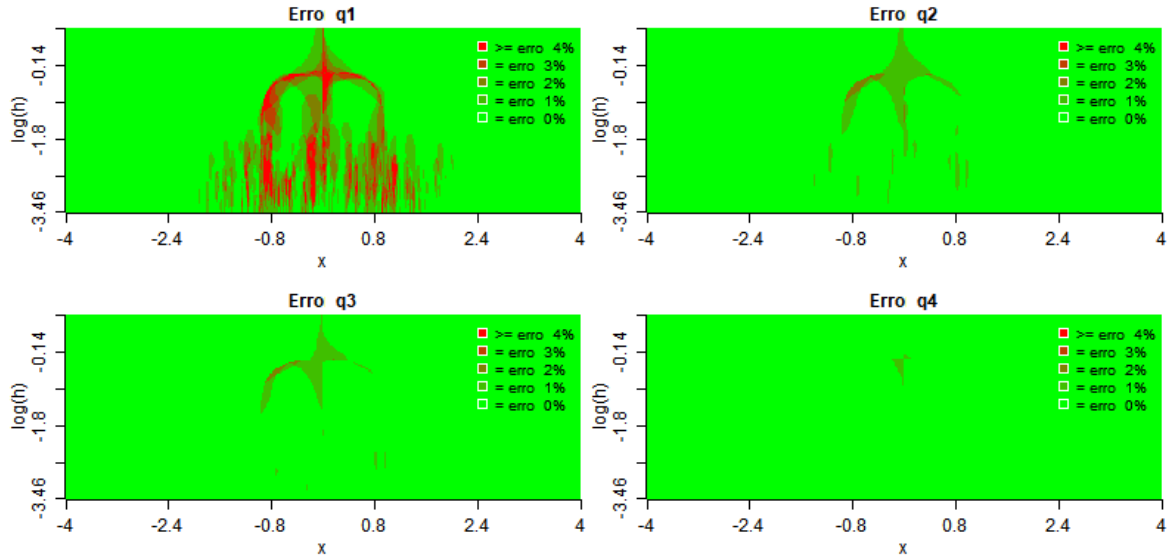

Figura C.8: Mapa de erro para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 200$, xeradas a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

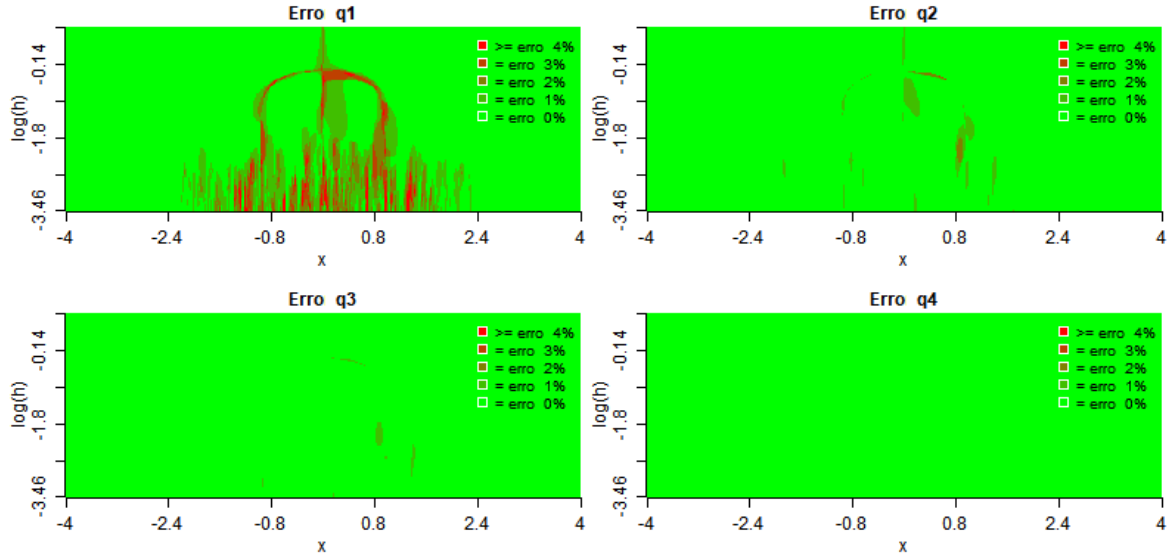

Figura C.9: Mapa de erro para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 500$, xeradas a partir dunha variable aleatoria con función de densidade *Bimodal* (#6) de Marron e Wand (1992).

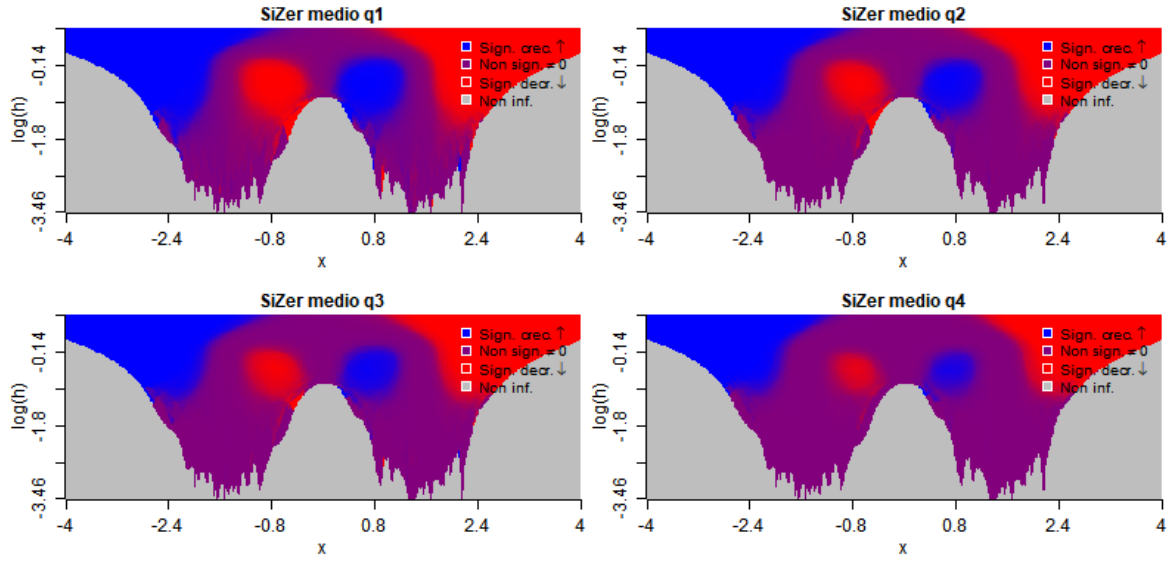

Figura C.10: Estimación do SiZer promedio para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 50$, xeradas a partir dunha variable aleatoria con función de densidade *Sep. Bimodal* (#7) de Marron e Wand (1992).

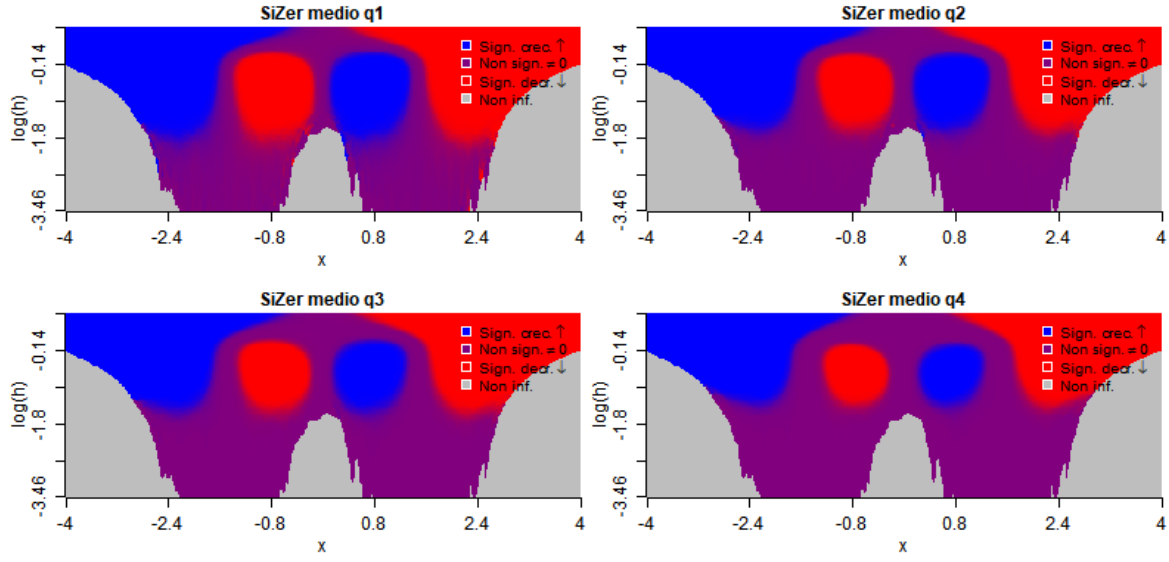

Figura C.11: Estimación do SiZer promedio para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 200$, xeradas a partir dunha variable aleatoria con función de densidade *Sep. Bimodal* (#7) de Marron e Wand (1992).

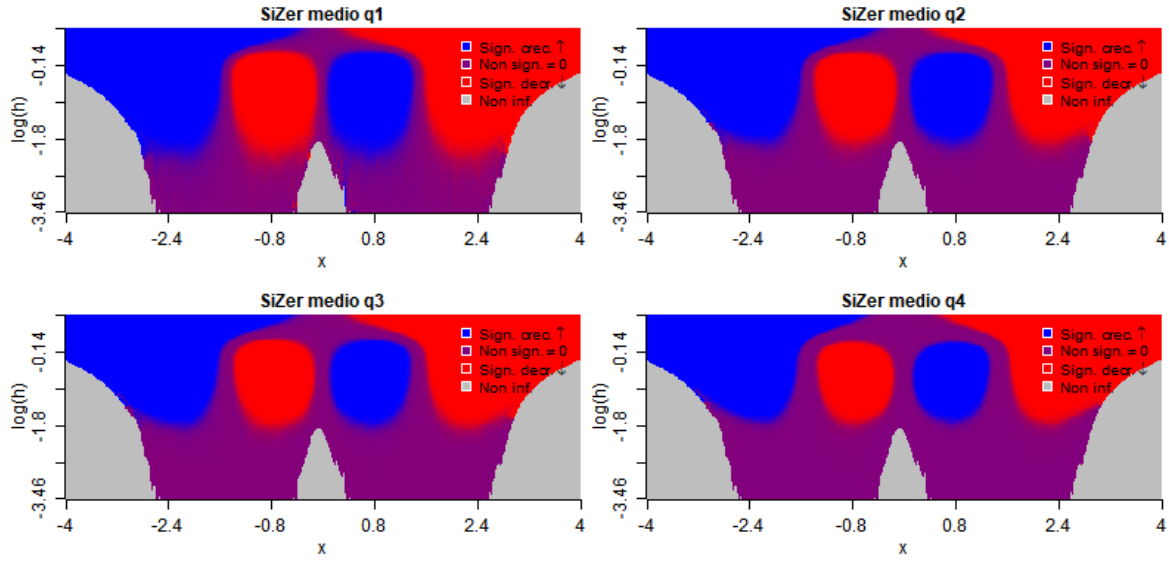

Figura C.12: Estimación do SiZer promedio para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 500$, xeradas a partir dunha variable aleatoria con función de densidade *Sep. Bimodal* (#7) de Marron e Wand (1992).

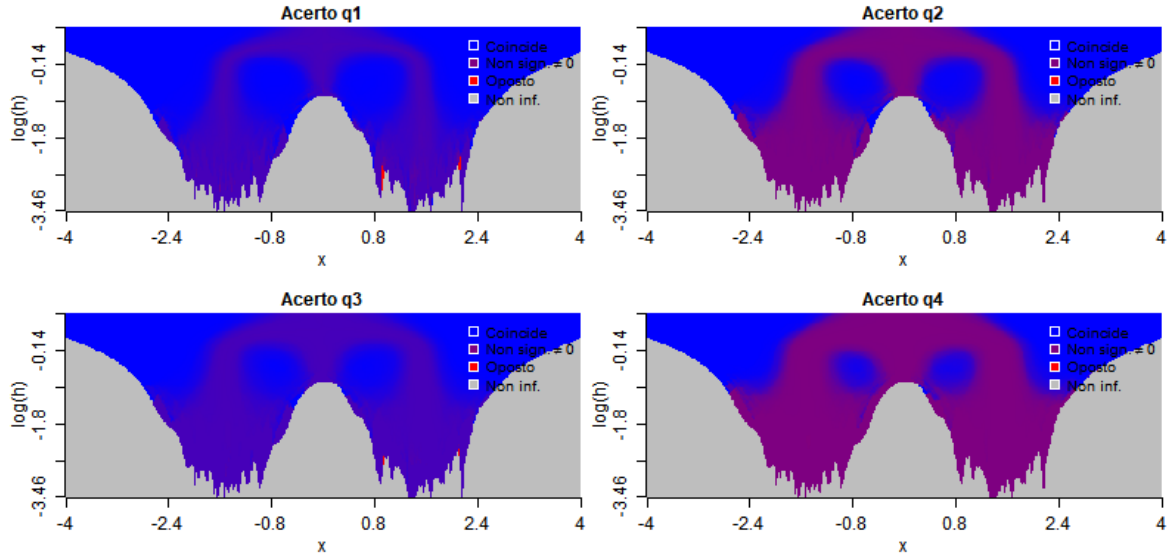

Figura C.13: Mapa de acerto para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 50$, xeradas a partir dunha variable aleatoria con función de densidade *Sep. Bimodal* (#7) de Marron e Wand (1992).

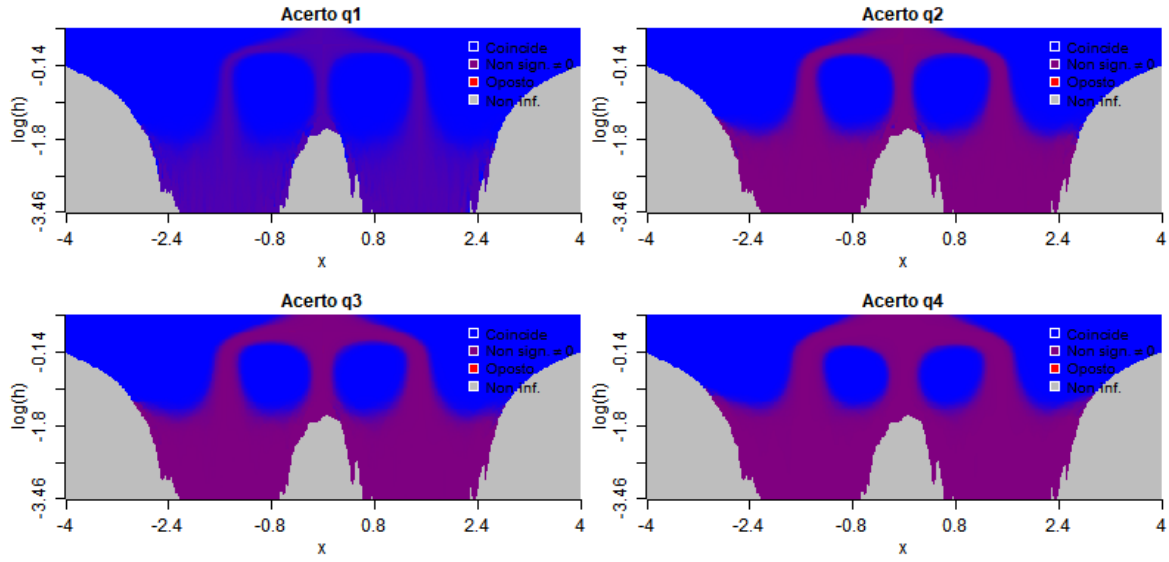

Figura C.14: Mapa de acerto para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 200$, xeradas a partir dunha variable aleatoria con función de densidade *Sep. Bimodal* (#7) de Marron e Wand (1992).

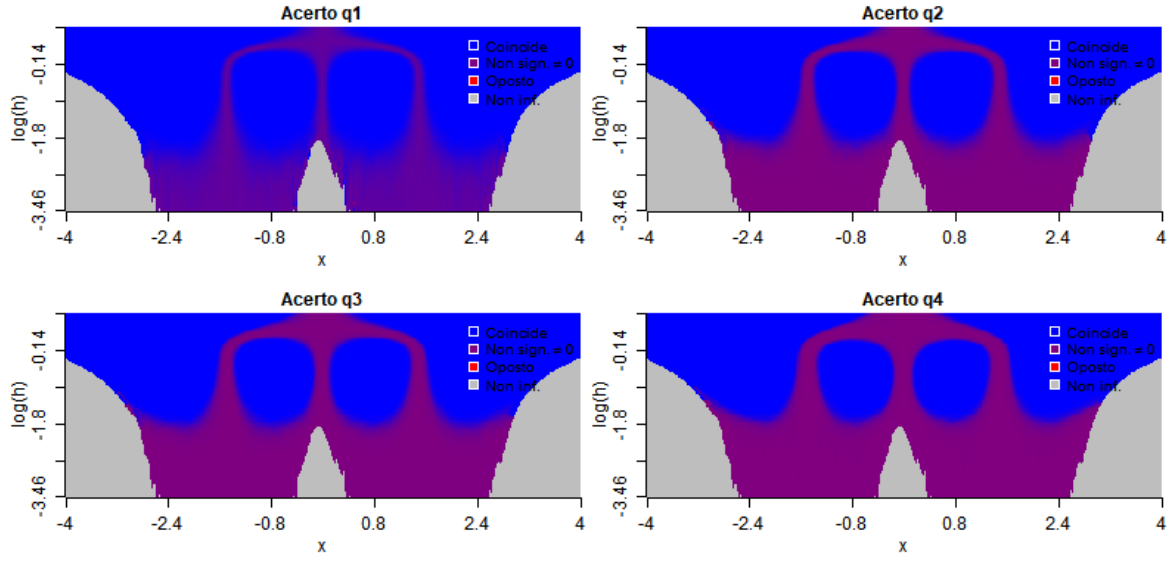

Figura C.15: Mapa de acerto para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 500$, xeradas a partir dunha variable aleatoria con función de densidade *Sep. Bimodal* (#7) de Marron e Wand (1992).

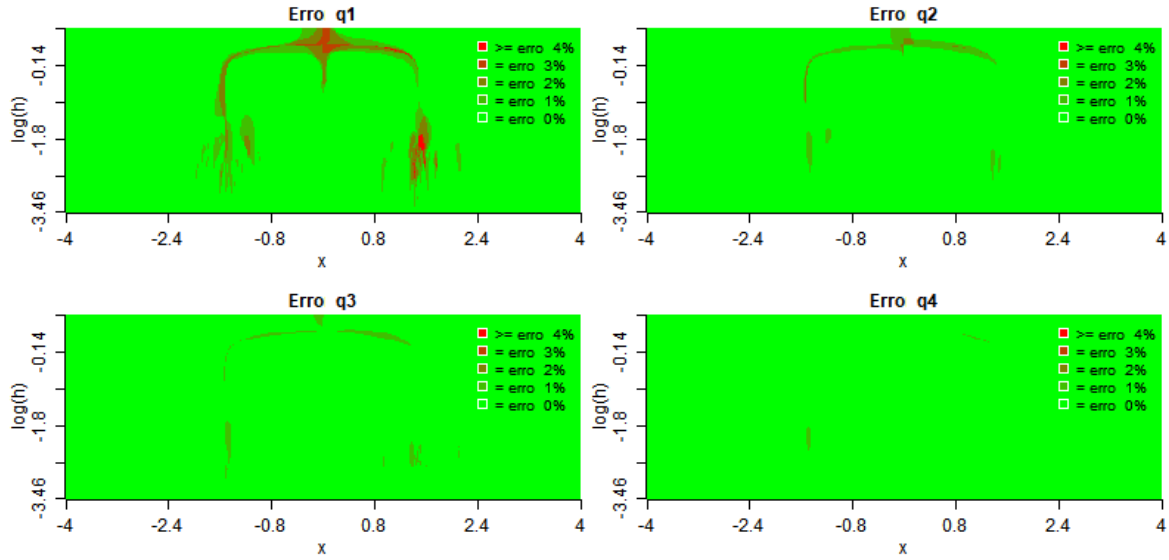

Figura C.16: Mapa de erro para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 50$, xeradas a partir dunha variable aleatoria con función de densidade *Sep. Bimodal* (#7) de Marron e Wand (1992).

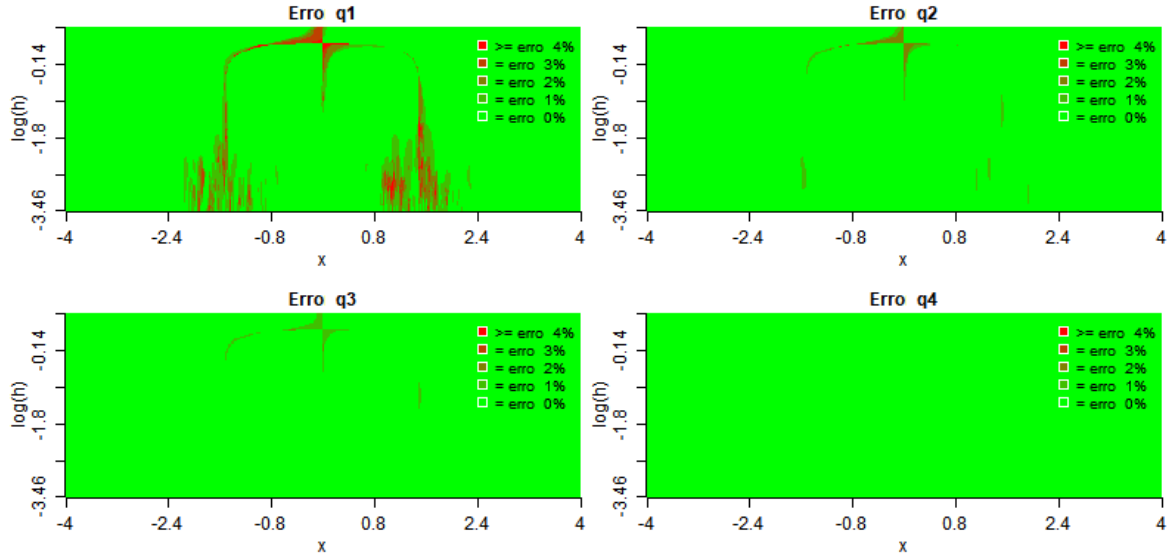

Figura C.17: Mapa de erro para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 200$, xeradas a partir dunha variable aleatoria con función de densidade *Sep. Bimodal* (#7) de Marron e Wand (1992).

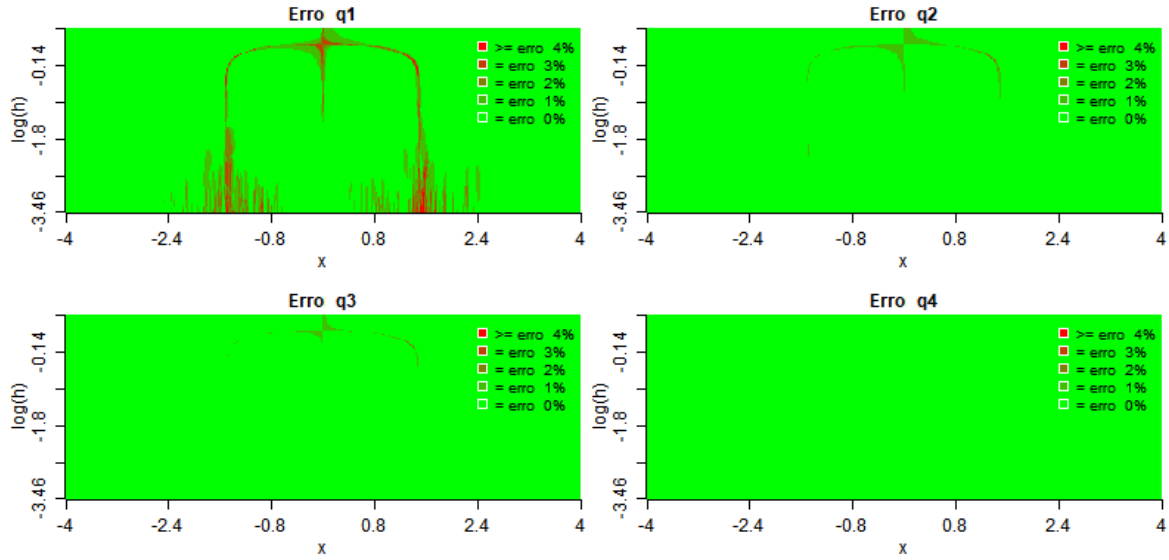

Figura C.18: Mapa de erro para os cuantís q_1, q_2, q_3 e q_4 para $N = 1000$ mostras, con tamaño de mostra $n = 500$, xeradas a partir dunha variable aleatoria con función de densidade *Sep. Bimodal* (#7) de Marron e Wand (1992).

Bibliografía

[1] Ameijeiras-Alonso J, Crujeiras RM, and Rodríguez-Casal A (2016). Mode testing, critical bandwidth and excess mass. *arXiv:1609.05188*, Submitted.

[2] Bowman AW (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71: 353–360.

[3] Chaudhuri P, Marron JS (1997). Scale-space view of curve estimation. *Mimeo Series* #2357, North Carolina Institute of Statistics.

[4] Chaudhuri P, Marron JS (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 447: 807–823.

[5] Davison AC, Hinkley DV (1997). *Bootstrap Methods and their Application.* Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge.

[6] Efron B (1979). SiZer for bootstrap methods: Another look at the jacknife. *Annals of Statistics*, 7: 1–26.

[7] Efron B (1993). *An Introduction to the Bootstrap.* Boca Raton, Florida.

[8] Fisher NI, Marron JS (1951). Mode testing via the excess mass estimate. *Biometrika*, 88, 419–517

[9] Fix E, Hodges JL (1951). Discriminatory analysis, nonparametric estimation: consistency properties. *Report No 4, Project no 21-49-004*, USAF School of Aviation Medicine, Randolph Field, Texas.

[10] Hall P (1988). Theoretical comparison of bootstrap confidence intervals. *Annals of Statistics*, 16: 927–953.

[11] Hall P (1991). Edgeworth expansions for nonparametric density estimators. *Statistics*, 2: 215–232.

[12] Hardle W, Marron JS (1991). Bootstrap simultaneous error bars for nonparametric regression. *The Annals of Statistics*, 19: 778–796.

[13] Jones MC, Lotwick HW (1983). On the errors involved in computing the empirical characteristic function. *Journal of Statistical Computation and Simulation*, 17: 133–149.

[14] Lindeberg T (1994). *Scale-Space Theory in Computer Vision.* Kluwer, Boston.

[15] Marron JS, Chung SS (2001). Presentation of smoothers: the family Approach. *Computational Statistics*, Volume 16, Issue 1, 195–207.

[16] Marron JS, Wand MP (1992). Exact mean integrated squared error. *The Annals of Statistcs*, 20:712–736.

[17] Minnotte MC, Scott D (1993). The mode tree: a tool for visualization of nonparametric density estimates. *Journal of Computational and Graphical Statistics*, 2:51–68.

[18] Parzen E (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33 (3):1065-1076.

[19] Rosenblatt M (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 832-837.

[20] Sheather SJ, Jones MC (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B*, 53: 683–690.

[21] Silverman BW (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B*, 43: 97–99.

[22] Silverman BW (1982). Kernel density estimation using the fast Fourier transform. *Applied Statistics*, 31: 93–99.

[23] Silverman BW (1986). *Density Estimation for Statistics and Data Analysis.* Champman and Hall, London.

[24] Taylor CC (1989). Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrika*, 76: 705-712.

[25] Witkin AP (1983). Scale-space filtering. *8th International Joint Conference of Artificial Intelligence, Karlsruhe, West Germany*, pp. 1019–1022.