USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Estimación da densidade no plano

María Bugallo Porto

2019/2020

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Estimación da densidade no plano

María Bugallo Porto

Xullo 2020

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Estatística e Investigación Operativa
Título: Estimación da densidade no plano
Breve descrición do contido
Neste traballo abórdase o problema de estimar non parametricamente unha densidade bidimensional. No grao estúdase brevemente este problema en dimensión un. O obxectivo deste TFG é afondar no problema de estimación non paramétrica da densidade. A estimación da densidade bidimensional permite detectar rexións no plano con alta concentración de datos. Este aspecto é moi importante en diversas aplicacións, como acontece en epidemioloxía, xa que permite detectar as rexións onde se concentran os enfermos, e comparalas coas correspondentes da poboación non enferma.

Índice xeral

Resumo

A estimación da densidade é un problema que xorde co obxectivo de coñecer como é a concentración dunha poboación a partir dunha mostra da mesma. Neste TFG abordaremos a estimación non paramétrica da densidade, permitindo así que a función a estimar adopte case calquera forma posible, sen máis que esixir que sexa unha densidade.

Comezando co caso unidimensional para, posteriormente, centrarnos no caso bidimensional, presentaremos e analizaremos o estimador histograma e o estimador tipo núcleo. Con ese fin introduciremos diferentes criterios de erro e métodos de selección dos parámetros de suavizado de cada un dos estimadores, así como da función núcleo no segundo deles. Diferentes exemplos simulados con datos procedentes de densidades coñecidas axudan a mostrar os distintos resultados teóricos así como proporcionan representacións gráficas de gran utilidade para a compresión do traballo.

Para finalizar, ilustraranse as ideas expostas sobre dous conxuntos de datos reais. O primeiro deles analízase a medida que se desarrolla a teoría e está relacionado coas erupcións dun geyser nos Estados Unidos. O segundo corresponde coas posicións dos niños de avespa velutina en Galicia entre os anos 2016 e 2018, polo que é de gran interese biolóxico e social. Estas dúas aplicacións mostran a utilidade das técnicas descritas ao longo deste traballo en áreas tan dispares como as que se aplican.

Abstract

Density estimation is a problem that arises with the aim of knowing how a population is concentrated from a sample of it. In this TFG we will address non-parametric density estimation, thus allowing the function to be estimated to take almost any possible form, requiring it to be a density.

Starting with the one-dimensional case and then focusing on the two-dimensional case, we will present and analyze the histogram estimator and the kernel estimator. To this end we will introduce different error criteria and methods for selecting the smoothing parameters of each of the estimators, as well as the kernel function in the second of them. Different simulated examples with data from known densities help to show the different theoretical results as well as provide graphical representations of great utility for the comprehension of the work.

Finally, the ideas presented on two real data sets will be illustrated. The first of these is analyzed as the theory develops and is related to the eruptions of a geyser in the United States. The second corresponds to the positions of wasp nests in Galicia between 2016 and 2018, so it is of great biological and social interest. These two applications show the usefulness of the techniques described throughout this work in areas as dissimilar as those applied.

Introdución

Un dos obxectivos da estatística é a análise de información e de datos, que poden ser xa de por si numéricos ou que se poden transformar en números para o seu manexo matemático. Na actualidade, a estatística é unha parte esencial de case todas as áreas de coñecemento e abarca dende a recolección de datos (mostraxe), pasando polo tratamento dos mesmos para, por último, obter conclusións sobre eles. É esencial en áreas tan diversas como as matemáticas, a socioloxía, a economía, a administración, a bioloxía ou a informática.

Dado un conxunto de datos reais, unha cuestión importante é determinar o seu comportamento empregando unha función que asigne un reparto teórico da poboación de procedencia, en canto á posibilidade de que se den os distintos sucesos relativos a esta. Con este fin, unha técnica moi empregada é recorrer a representación gráfica dos datos para obter unha idea de como é a súa estrutura. Deste xeito, resulta razoable querer determinar unha función que permita asignar a certos sucesos definidos a probabilidade de que estes teñan lugar, é dicir, partindo dunha mostra determinar a posibilidade de que un suceso concreto ocorra ou non e, en caso afirmativo, cal é a frecuencia de veces que isto sucede.

Un dos obxectivos fundamentais da asignación de probabilidades, así como da representación gráfica, é interpretar os resultados para obter conclusións do comportamento dos datos co fin de facer predicións, estimacións,... É dicir, unha das labores fundamentais da estatística é estudar os posibles valores que pode tomar unha certa variable ou vector aleatorio e a probabilidade de ocorrencia dos mesmos. Intuitivamente, chegamos ao concepto de *distribución de probabilidade* dunha variable ou vector aleatorio. Calquera cálculo ou proceso que involucre á variable ou ao vector aleatorio, como pode ser a obtención da súa media, varianza, función característica,..., os contrastes de hipóteses, os modelos de regresión,... involucran á distribución probabilística do mesmo.

A partir dunha mostra e co fin de estudar a concentración da poboación a que pertence, o máis habitual é que se empreguen a *función de distribución* e a *función de densidade*,

ambos conceptos moi estreitamente relacionados. Por iso, un problema estatístico fundamental é a estimación da función de densidade dunha variable ou vector aleatorio a partir da información proporcionada por unha mostra do mesmo. En moitas ocasións, e como veremos ao longo deste traballo, a visualización dos datos así como a representación gráfica da súa función de densidade, ou no seu defecto, da súa estimación por algún método matemático, facilitan en gran medida a interpretación dos datos, así como permiten extraer conclusións do comportamento destes.

A densidade sempre é non negativa e a súa interpretación pódese intuír como segue: as zonas de alta densidade representan zonas de alta concentración de datos e as de baixa ou nula densidade, de concentración de datos baixa ou nula, respectivamente. Como ferramenta de visualización, o histograma unidimensional consiste en agrupar os datos en intervalos que discreticen a recta real e considerar unha escala que indique a cantidade de datos que caen en cada un deles. Na Figura 1 represéntanse tres histogramas distintos, cada un relativo a unha mostra tomada dunha poboación dada polos modelos M1, M4 e M5 do Apéndice B, respectivamente.

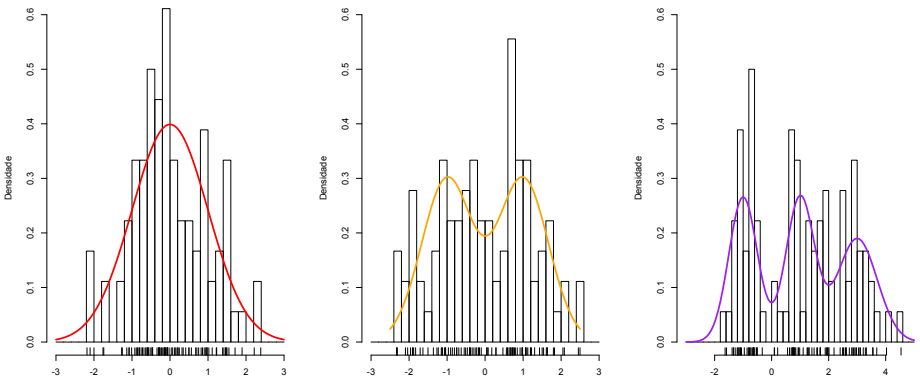

Figura 1: En cor negro, histogramas de tres mostras procedentes das mesturas de distribucións normais dadas polos modelos M1, M4 e M5 do Apéndice B; noutra cor, represéntase a curva da función real correspondente. As barras verticais no eixe de abscisas indican os valores mostrais e as súas posicións. O tamaño da mostra é $n = 90$.

Nas tres representacións da Figura 1, o comportamento do histograma é similar ao da curva teórica, correspondendo as barras verticais máis altas a rexións onde a densidade real presenta modas, e as máis baixas, a zonas de pouca densidade, como son os extremos da estimación e os vales. Parece razoable que ao aumentar o tamaño mostral aumente a

calidade da estimación, é dicir, que o comportamento do histograma sexa cada vez máis similar ao da curva teórica, en consecuencia de que un aumento do tamaño mostral implica obter máis información sobre a poboación orixinal e, polo tanto, poder estimala mellor. Como veremos ao longo do Capítulo 1 e posteriormente no Capítulo 2, isto ocorrerá con todos os estimadores da densidade expostos e ten a súa base en resultados teóricos.

Unha función de densidade caracteriza o comportamento probable dunha poboación en tanto que especifica a probabilidade relativa de que unha variable ou vector aleatorio continuo tome un valor na recta real próximo a x, ou no plano e próximo a (x, y), respectivamente. O estudo da densidade dun conxunto de datos dispoñibles non é máis que o estudo da concentración de ditos datos, é dicir, da concentración da mostra. Entón a representación gráfica dunha mostra de datos continuos proporciónanos os lugares onde se concentran os datos e así, permítenos estimar dun xeito gráfico, e posteriormente analítico, a densidade da poboación á que pertencen. Así, xorde o problema da estimación da densidade, entendido como un problema de visualización de datos que ten as súas raíces no histograma, que procede dos histogramas de Pearson, que explicaremos en máis profundidade en capítulos posteriores deste traballo, así como a súa historia.

En ocasións coñecemos información externa á mostra que condiciona a estimación da función de densidade, restrinxíndoa a certa clase de funcións. Nese caso, un posible enfoque consiste en supoñer que a función de densidade que queremos estimar pertence a algunha familia paramétrica de funcións coñecidas, como poden ser a normal, a expoñencial, a uniforme ou a Poisson, no caso de variables aleatorias. Nesta situación, o obxectivo é estimar os parámetros descoñecidos de tales distribucións a partir dos datos da mostra. Este método de estimación da función de densidade coñécese como nome de *estimación paramétrica da función de densidade*. A validez de tales suposicións sobre a distribución dos datos pode ser testada con posterioridade empregando tests de bondade de axuste.

Sen embargo, en moitos casos, o enfoque paramétrico non ten sentido porque, ou ben non coñecemos información externa á mostra, ou ben existen dúbidas sobre a validez desta información. No caso de que a suposición inicial de que a mostra siga un modelo paramétrico fixo sexa certa a estimación será moi boa pero, en caso contrario, as conclusións poderían ser totalmente erróneas.

Na Figura 2 móstranse tres estimacións da función de densidade da variable X_1 do conxunto de datos "$OldFaithful$" de R, que se presentarán en detalle en vindeiros capítulos

deste traballo, e que denota a duración das erupcións dun geyser. En cor negra, o estimador histograma (non paramétrico); en cor verde, o estimador tipo núcleo (non paramétrico); e en cor azul, o estimador de máxima verosimilitude supoñendo que os datos seguen unha distribución normal (paramétrico).

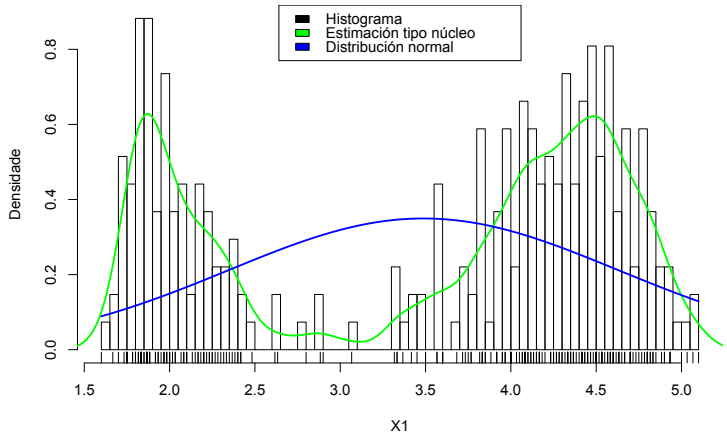

Figura 2: Tres estimacións distintas da función de densidade para a variable X_1, que recolle a duración das erupcións dun geyser, do conxunto de datos "$OldFaithful$" de R. En cor negra represéntase o estimador histograma; a gráfica verde corresponde coa estimación tipo núcleo e a azul co estimador de máxima verosimilitude supoñendo normalidade. As barras verticais no eixe de abscisas indican os valores mostrais e as súas posicións.

Fixándonos na concentración dos datos da Figura 2, tanto o histograma como o estimador tipo núcleo son fieis á súa distribución, pero non ocorre o mesmo coa curva paramétrica, baseada no modelo normal. Isto suxire que probablemente os datos non procedan dunha distribución normal, que non respecta a súa bimodalidade, senón dun modelo non paramétrico máis complexo, enfatizando así a importancia da estimación non paramétrica da densidade. Neste caso está bastante claro que a mostra non segue un modelo paramétrico normal, pois este asígnalle unha moda nunha zona de escaseza de datos e baixa densidade nas dúas zonas correspondentes coas dúas modas das funcións estimadas mediante os modelos non paramétricos.

Como acabamos de ver, unha posible alternativa á estimación paramétrica consiste en non impoñer ningún modelo paramétrico fixo para a función de densidade e así deixar que, ao estimar dita función, a estimación adopte calquera forma posible. Agora ben, debemos

impoñer que cumpra as condicións necesarias para ser unha función de densidade, é dicir, dada unha función de densidade f, esta debe ser non negativa e a integral no seu dominio (chamado *soporte* cando traballamos con variables e vectores aleatorios, por ser o conxunto de posibles valores que estes poden tomar) igual a un. A esixencia destas dúas condicións é moi intuitiva: a función de densidade indícanos como é a concentración dos datos, sendo esta unha cantidade non negativa por definición; ademais, ao considerar todo o o soporte, estamos a tomar todos os posible valores, polo que é de esperar que sumen un (a integral é unha xeralización da suma para o caso continuo). No caso unidimensional, que a integral no soporte sexa igual a un equivale a que a área total encerrada baixo a curva valga un.

A idea é estimar a función de densidade localmente, considerando os datos da mostra similares e, como veremos, considerando ou non a proximidade dos mesmos ao punto onde a queremos estimar. No caso de que a proximidade sexa un factor a ter en conta, as estimacións serán mais suaves, como ocorre co estimador tipo núcleo, que emprega todos os datos da mostra, cada un cun peso proporcional a súa proximidade ao valor no cal queremos estimar a función de densidade.

Este traballo céntrase no enfoque non paramétrico, menos restritivo pero tamén máis interesante, polo feito de que a carencia de restricións en canto á forma da función de densidade dá lugar á construción de todo tipo de estimadores para esta, sen ter que limitarnos a familias de funcións xa coñecidas. Dito enfoque coñécese co nome de *estimación non paramétrica da función de densidade*. Sen embargo, como xa veremos cando analicemos os distintos erros cometidos ao estimar a función de densidade con diferentes métodos non paramétricos, esta liberdade ten un custo maior, tanto computacional como conceptual, con respecto ao caso paramétrico.

O obxectivo é estudar diferentes estimadores da densidade no plano pero, en moitos aspectos esta estimación é unha xeneralización dos modelos unidimensionais polo que abordaremos a estimación da densidade en ambos casos. Na Figura 2 recóllese un exemplo da estimación unidimensional da densidade empregando, entre outros métodos, dous modelos non paramétricos distintos: o histograma e o estimador tipo núcleo. Os estimadores tipo núcleo, pola súa capacidade de converter conxuntos de datos en resumos útiles, interpretables e moi visuais, constitúen unha ferramenta fundamental do análise de datos; ademais resolven unha das limitacións do histograma: a súa irregularidade e falta de suavidade, por ser unha función definida a trozos e constante en cada un deles.

O caso bidimensional inspírase na estimación unidimensional de cada unha das dúas variables, tendo en conta a orientación dos datos e, polo tanto, a posible dependencia dunha variable coa outra. Se a orientación considerada é a estándar, é dicir, se consideramos matrices diagonais, aproximamos o noso modelo eliminando a relación local entre as variables e conservándoa globalmente. Este último enfoque carece de sentido en moitas ocasións (se os datos están claramente orientados cunha orientación non estándar) pero facilita en gran medida a estimación.

A continuación ilustraremos cun exemplo en que consiste a estimación bidimensional non paramétrica da densidade empregando os dous estimadores que estudaremos ao longo do Capítulo 2: o histograma e a estimación tipo núcleo. Na Figura 3 representamos dúas estimacións da densidade dos datos que analizaremos no Capítulo 3, e que corresponden á posición dos niños de avespa velutina no ano 2018 en Galicia, en coordenadas cartesianas. A esquerda o diagrama de calor resultante de considerar unha vista aérea do estimador histograma e a dereita as curvas de nivel procedentes da estimación tipo núcleo, xunto co mapa do contorno galego e o gráfico de calor asociado, onde a escala de cores é a mesma que a da representación da esquerda. Os eixos de ambas gráficas correspóndense cos eixos cartesianos, atopándose o mapa entre os 41.75°N e 44°N, e os 6.5°W e 9.5°W. A escala vertical da representación da esquerda indica os cores das barras do histograma en función da cantidade de datos que pertenzan a cada celda (base da barra), que é proporcional á altura da mesma. De branco a verde, pasando por laranxa e amarelo, os cores reflexan a concentración de niños nos diferentes puntos do territorio galego.

En primeiro lugar, é claro que a segunda representación proporciona resultados máis precisos que a primeira. Isto débese, como xa veremos, a que a estimación tipo núcleo é mellor que o histograma, en termos de suavidade, converxencia e interpretabilidade. Por iso mesmo, no Capítulo 2 centrarémonos na estimación tipo núcleo e unicamente introduciremos o histograma bidimensional como unha xeneralización do caso unidimensional, sen entrar en máis detalle acerca dos distintos criterios de erro, que si analizaremos no Capítulo 1 para o caso unidimensional.

Observando ambas estimacións, apreciamos que a concretación dos niños non é uniforme en todo o territorio galego e tampouco presentan unha estrutura unimodal, senón multimodal. Existen rexións do mapa onde a concentración de datos é elevada (véxase a zona de Coruña ou as Rías Baixas) e outras onde é case nula (véxase a parte oriental da provincia de Ourense). No Capítulo 3, adicado ao análise destes datos, explicaremos con

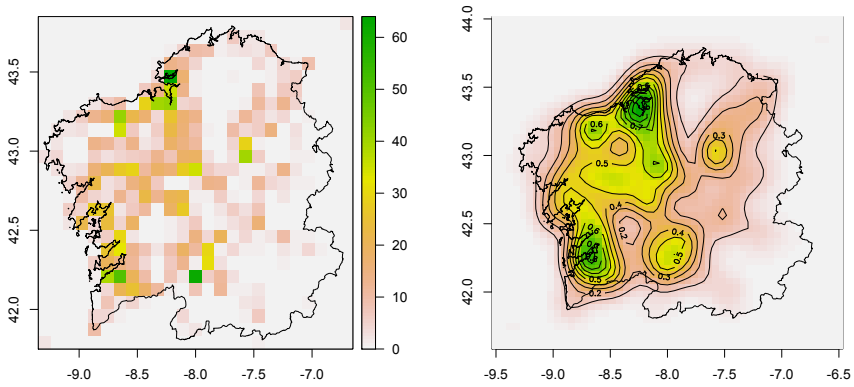

Figura 3: Dúas estimacións distintas da función de densidade dos niños de avespa velutina en Galicia no ano 2018. A esquerda, gráfica de calor asociada a un histograma bidimensional; a dereita, curvas de nivel da estimación tipo núcleo, xunto co mapa de calor e o contorno de Galicia en negro, igual que as curvas de nivel.

máis detalle a que se debe, en gran medida, esta distribución dos niños. Como adianto, dado que as velutinas son unha especie procedente de Asia e están acostumadas a un clima subtropical, a súa adaptación é mellor nas zonas costeiras, pois as temperaturas alí son máis suaves que no interior. Ademais a chegada das velutinas a Galicia no ano 2012 sitúase en dous focos distintos, que non é casualidade que correspondan coas dúas modas dos gráficos da Figura 3.

Unha cuestión importante é obter unha función densidade que, segundo o punto do territorio galego no que nos atopemos, estime a concentración de niños de avespa velutina. Ademais, adiantemos que na estimación tipo núcleo esta función dependerá dunha matriz de suavizado e doutra función de densidade, coñecida como función núcleo.
Para cada punto, a estimación tipo núcleo emprega os valores mostrais próximos, ponderándoos segunda a súa proximidade ao valor onde queremos estimar a densidade. Os parámetros de suavizado e a función núcleo determinan as ponderacións que corresponden a cada dato mostral, polo que unha cuestión importante é determinar cales son os valores dos parámetros óptimos e a función núcleo óptima para minimizar o erro da estimación.

Como veremos, existen varias medidas de discrepancia entre a función a estimar e a función estimada, polo que tamén hai varios criterios para a elección dos termos variables da estimación. Logo de tratar o caso unidimensional no Capítulo 1, no Capítulo 2 analiza-

remos todos estes detalles técnicos. Algo similar ocorre co histograma pero cunha expresión e interpretabilidade máis sinxela, de aí a que comecemos con este o estudo da estimación da función de densidade.

No Capítulo 1 imos estudar a estimación da densidade unidimensional por ser o caso bidimensional, e en xeral o multidimensional, unha xeneralización deste a máis dunha dimensión. Ademais as expresións dos estimadores e dos seus respectivos erros adoptan unha forma máis simple, facilitando a representación dos conceptos fundamentais. En orde cronolóxico e de dificultade conceptual, trataremos o estimador histograma, o estimador Naive (estimador intermedio por ser un caso particular do estimador tipo núcleo e unha xeneralización do histograma; tamén chamado histograma móbil, facendo referencia á relación que presenta co histograma, como xa veremos) e o estimador tipo núcleo. Ademais comentaremos diferentes métodos de elección do parámetro ventá na estimación tipo núcleo e como implementalos en R.

No Capítulo 2 estudaremos a estimación da densidade bidimensional abordando, como no capítulo anterior, o estimador histograma e o estimador tipo núcleo. Ademais, procederemos ao análise do conxunto de datos "$OldFaithful$", concentrándonos na súa distribución conxunta así como analizando a distribución de probabilidade marxinal de cada unha das dúas variables que contén, para ilustrar as diferentes ideas teóricas a un caso práctico. Por outra banda, analizaremos en detalle os distintos erros do estimador tipo núcleo, co fin de seleccionar unha matriz de parámetros ventá óptima e un núcleo axeitado. Por último comentaremos brevemente as dificultades de extender estes métodos a problemas con moitas variables.

No Capítulo 3 preséntaremos a aplicación dos diferentes estimadores da función de densidade e ideas expostas sobre outro conxunto de datos reais. Estes datos recollen información sobre a posición de niños de avespa velutina no territorio galego entre os anos 2016-2018, polo que son de gran interese biolóxico pero tamén social, debido ao impacto e preocupación que orixinan estas avespas na sociedade galega. Estudaremos a densidade de niños nas diferentes rexións galegas así como a evolución temporal da posición e cantidade destes ao longo dos tres anos de mostraxe. Para o manexo e representación gráfica dos dous conxuntos de datos expostos neste traballo empregaremos o entorno e linguaxe de programación estatística R Core Team (2020) (software libre).

Capítulo 1

Estimación da densidade unidimensional

Un dos grandes avances da estatística é a creación da coñecida como estatística non paramétrica, complementaria da estatística paramétrica, desarrollada fundamentalmente por grandes matemáticos como Pearson, Fisher e Student que, entre outras cousas, crearon procedementos de decisión estatística. A estimación da densidade é unha rama relativamente recente da inferencia estatística pero moi estudada actualmente polas súas múltiples aplicacións na gran maioría de campos do coñecemento. Neste traballo centrarémonos no enfoque non paramétrico da estimación da función de densidade, que ten os seus orixes nos traballos de Fix and Hodges (1951) que buscan liberar as ríxidas restricións do enfoque paramétrico sobre a distribución das variables implicadas. En certo modo o enfoque non paramétrico permite que os datos determinen a forma da función de densidade libremente, sen restricións.

Co fin de abordar a estimación da densidade no plano, neste primeiro capítulo presentaremos á estimación da densidade dunha variable aleatoria empregando distintos tipos de estimadores non paramétricos, por ser estes modelos unidimensionais máis sinxelos e con maior interpretabilidade. De menor a maior dificultade técnica e conceptual, comezaremos co estimador histograma e, pasando polo estimador Naive, introduciremos o estimador tipo núcleo, que é un dos máis empregados estimadores non paramétricos da densidade.

A representación gráfica de diferentes estimacións da densidade de mostras cuxa distribución real é coñecida será de gran axuda para visualizar o comportamento dos diferentes estimadores en situacións concretas, así como para comparan as distintas estimacións e

as discrepancias coa función orixinal. Estas representacións complementarán os diferentes criterios de erro que obteremos para os tres estimadores expostos neste capítulo. Por último, abordaremos a elección do parámetro ventá en R, referida á estimación da densidade tipo núcleo, para xustificar a súa elección nos diferentes exemplos, así como para estudar como esta ferramenta de programación estatística aborda a estimación non paramétrica da densidade.

1.1. Histograma unidimensional

En primeiro lugar, e co obxectivo de estimar a función de densidade a partir dun conxunto de datos, consideremos o estimador histograma. A motivación deste estimador é relativamente antiga dado que os seus orixes se remontan a mediados do século XVI, cando René Descartes propuxo un sistema que representaba un conxunto de números en dúas dimensións, unha no eixe vertical e outra no horizontal, pero non se empregou o termo 'histograma' ata finais do século XIX, cando o introduciu Pearson (1891) nun dos seus escritos matemáticos. A finalidade do histograma de Pearson era crear bloques temporais para representar gráficos sobre a duración de reinados, gobernos soberanos ou de diferentes mandatarios. O termo histograma é un composto do termos gregos '*histos*', que significa tela, e '*gramma*', que se pode traducir ao galego como gráfica ou texto. Deste xeito, ata o ano 1891, o histograma como tal non existía como o estimador da densidade que coñecemos actualmente e que se expoñerá neste traballo.

Antes de proceder á formulación matemática do histograma, recordemos como se definen a función de distribución e a de densidade dunha variable aleatoria X, e cal é a relación entre ambas:

$$F(x) = \mathbb{P}(X \leqslant x) = \int_{-\infty}^{x} f(u)du \text{ e } F'(x) = f(x), \text{ para case todo } x \in sop(X),$$

onde $sop(X)$ denota o soporte de X, é dicir, o conxunto de todos os posibles valores que pode tomar X; a expresión "para case todo" refírese a que se verifica para todos os puntos salvo un subconxunto de medida de Lebesgue nula. Estas dúas definicións son tamén válidas para o caso multidimensional, e en particular para o bidimensional recollido no Capítulo 2, sen máis que extender a definición como segue: A función de distribución e a de densidade conxunta dun vector aleatorio (X, Y) defínese, por extensión do caso unidimensional, como

$$F(x, y) = \mathbb{P}(X \leqslant x, Y \leqslant y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)dudv \text{ e } f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y},$$

para case todo $(x, y) \in sop(X, Y)$.

Deste xeito, consideremos unha variable aleatoria unidimensional X con función de distribución F e con función de densidade f, ambas descoñecidas e supoñamos que dispoñemos dunha mostra aleatoria simple (abreviadamente m.a.s) de X, que denotaremos por $X_1, ..., X_n$. Destaquemos que o tamaño mostral é n. Salvo que se indique o contrario, sexa $x \in \mathbb{R}$ un valor real arbitrario.

Discretizamos a recta real en intervalos disxuntos e definimos unha función (que exigiremos que integre 1 en $\mathbb{R}$ e que sexa non negativa, para que sexa unha función de densidade) constante en cada un dos intervalos. O valor constante que acadará en cada un deles será a proporción de datos mostrais que se atopen no mesmo dividida pola lonxitude do intervalo, e polo tanto, un valor decimal non negativo. Como consecuencia da facilidade na interpretación e expresión, o histograma é o máis sinxelo e coñecido estimador non paramétrico da densidade. En moitos casos, dada unha mostra unidimensional, a gráfica do histograma constitúe unha ferramenta moi útil para obter unha primeira idea de como é a concentración dos datos e cal é o seu rango estimado.

Consideremos unha orixe $t_0 \in \mathbb{R}$, tamén chamado punto de anclaxe do histograma, e un parámetro ventá $b > 0$ en función do tamaño mostral, é dicir, $b \equiv b(n) \equiv b_n$, que seguiremos denotando por b no sucesivo. O parámetro ventá mide o ancho dos intervalos de igual lonxitude cos que discretizamos $\mathbb{R}$ e a notación de b procede do inglés *bin*, que traducido ao galego sería envase, por ser o *binwidth*, é dicir, o ancho dos envases. Definimos unha densidade constante en intervalos da forma $(t_0 + bk, t_0 + b(k + 1)]$ con $k \in \mathbb{Z}$, e denotemos por $t_k = t_0 + kb$ polo que $t_{k+1} = t_k + b$.

Deste xeito para cada $k \in \mathbb{Z}$ o estimador da $\mathbb{P}(X \in (t_k, t_{k+1}])$ vén dado por:

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \in (t_k, t_{k+1}]),$$

por tratarse da conta dos datos que caen en dito intervalo, sendo $\mathbb{I}$ a función indicador, e dividindo polo tamaño mostral para obter un promedio ponderado. A función de densidade estimada para o histograma nese intervalo é

$$\hat{f}_{hist}(x) = \frac{1}{nb} \sum_{i=1}^{n} \mathbb{I}(X_i \in (t_k, t_{k+1}]), \ x \in (t_k, t_{k+1}], \tag{1.1}$$

por ser a estimación da $\mathbb{P}(X \in (t_k, t_{k+1}])$ entre a lonxitude do intervalo $(t_k, t_{k+1}]$, que é b.

Denotando $B_k = (t_k, t_{k+1}]$,

$$\hat{f}_{hist}(x) = \frac{1}{nb} \sum_{i=1}^{n} \mathbb{I}(X_i \in B_k),\ x \in B_k.$$

Se # denota a cantidade de elementos do conxunto que precede, é dicir, o seu cardinal, sexa $N_k = \#\{X_i : X_i \in B_k, i = 1, ..., n\}$, de modo que

$$\hat{f}_{hist}(x) = \frac{1}{nb} \sum_{k} N_k \mathbb{I}(x \in B_k) = \frac{N_k}{nb},$$

onde a última igualdade é válida se $x \in B_k$.

É fácil ver que, efectivamente, $\hat{f}_{hist} \geqslant 0$, sen máis que decatarse de que todos os termos da suma anterior son non negativos, e que $\int_{\mathbb{R}} \hat{f}_{hist}(x)dx = 1$, polo que a función estimada $\hat{f}_{hist}$ é unha función de densidade.

No Apéndice A recóllese un breve repaso da distribución normal, de gran utilidade no que segue de traballo. Moitos dos exemplos que presentaremos ao longo deste capítulo están inspirandos no artigo de Marron and Wand (1992, Apartado 3) e atópanse no Apéndice B, onde podemos atopar modelos procedentes de mesturas de densidades normais. Como veremos, a natureza destes modelos (e en concreto a multimodalidade) e a flexibilidade que proporcionan as mesturas de distribucións normais serán de gran axuda para ilustrar conceptos teóricos.

Recordemos que a moda dunha poboación é o valor máis probable, é dicir, aquel con maior frecuencia absoluta. En xeral, a moda non ten porque ser única, pois pode haber varios valores -ou incluso infinitos- que sexan os máis probables; tampouco ten por que existir, se todos os valores teñen a mesma probabilidade de ocorrer, dicindo nese caso que a distribución é amodal. Se a distribución ten unha moda dise que é unimodal, se ten dúas, bimodal e se ten múltiples, multimodal. Por exemplo, a $N(0, 1)$ é unimodal sendo $x = 0$ a súa única moda; a $U[0, 1]$ é amodal por ter todos os valores posibles a mesma frecuencia.

Na gráfica esquerda da Figura 1.1, represéntase un histograma de cor vermello empregando unha m.a.s de dimensión $n = 30$ da mestura de distribucións normais con densidade correspondente ao modelo M2 do Apéndice B. Eliximos un paso de $b = \frac{1}{2}$ e un punto de anclaxe $t_0 = -8$. En negro engadiuse a curva da densidade teórica. A pesares de que o tamaño mostral non é moi elevado, podemos apreciar que o histograma tende a achegarse á curva. Sen embargo, non estima con exactitude as tres modas que presenta a distribución

orixinal dos datos.

Ao aumentar o tamaño mostral, a estimación mellora notablemente e as modas estímanse con moita máis exactitude, como se observa co histograma da dereita, en cor azul, onde tamén engadimos a densidade teórica en negro, e o tamaño mostral é $n = 210$. Seguimos empregando un punto de anclaxe de $t_0 = -8$ e un paso $b = \frac{1}{2}$. Isto ilustra o feito de que o aumento do tamaño mostral mellora á estimación da densidade, neste caso empregando un histograma.

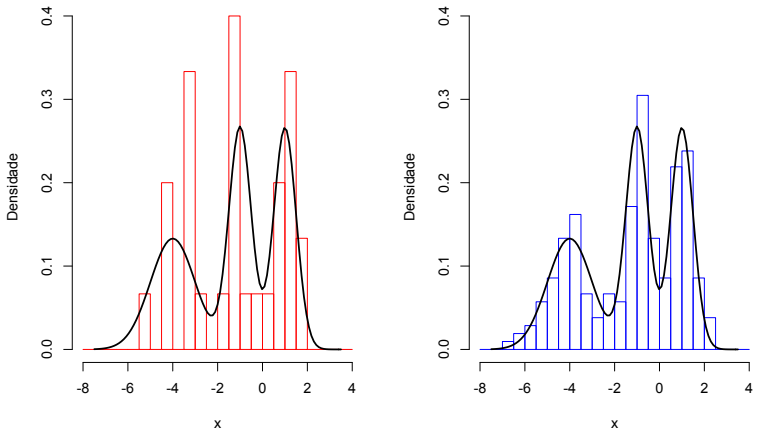

Figura 1.1: Histogramas de dúas m.a.s do modelo M2 recollido no Apéndice B, para $n = 30$ (esquerda) e para $n = 210$ (dereita), xunto coa súa función de densidade en cor negro. En ambos casos eliximos $t_0 = -8$ e $b = \frac{1}{2}$.

A orixe t_0 do histograma é o punto de partida dos intervalos empregados para discretizar $\mathbb{R}$. É dicir, é o extremo esquerdo do intervalo $B_0 = (t_0, t_1]$, tomando $k = 0$. Veremos que inflúe notablemente na forma do histograma: por estar a definir unha función constante a trozos en intervalos da forma $(t_0 + bk, t_0 + b(k + 1)]$, $k \in \mathbb{Z}$, que t_0 varíe pode dar lugar a que datos mostrais que antes caeran nun intervalo agora xa non o fagan, por tratarse de intervalos diferentes. Isto débese a que t_0 inflúe directamente na definición dos extremos dos intervalos B_k. Por exemplo, se tomamos $t_0 = 0$ e $b = 1$, $B_k = (k, k + 1]$; considerando $t_0 = \frac{1}{2}$ e o mesmo valor de b, $\tilde{B_k} = (k + \frac{1}{2}, k + \frac{1}{2}]$.

Na Figura 1.2 ilustramos como o punto de anclaxe t_0 é un parámetro que inflúe na forma do histograma. Consideramos a mesma m.a.s da mestura de distribucións normais

de tamaño $n = 100$ que vén dada polo modelo M3 do Apéndice B e fixemos a lonxitude dos intervalos $b = \frac{1}{4}$. De esquerda a dereita, consideramos valores de t_0 cada vez máis pequenos. Observemos como, dunha gráfica a outra, o reparto dos datos mostrais nos intervalos cambia notablemente, por cambiar os mesmos de posición (non de lonxitude). Este cambio apréciase sobre todo nas barras correspondentes ás modas dos histogramas, por proceder os datos dunha distribución multimodal con catro modas moi próximas. Enfaticemos que nos tres casos se considerou a mesma mostra e polo tanto o mesmo tamaño mostral, pero a forma dos histogramas varía como consecuencia dos cambios de t_0.

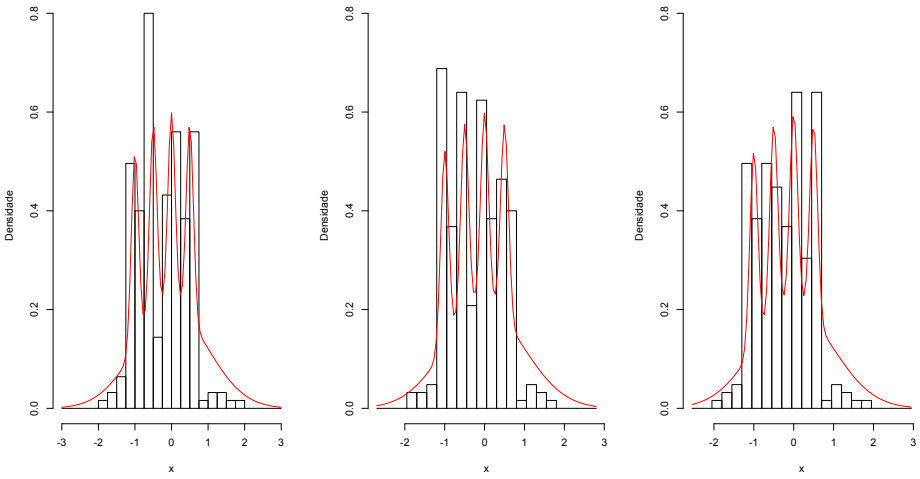

Figura 1.2: Histogramas dunha m.a.s de tamaño $n = 100$ da mestura de normais dada polo modelo M3 do Apéndice B cuxa densidade se representa en color vermello. Fixado $b = \frac{1}{4}$, de esquerda a dereita, os valores do punto de anclaxe t_0 son: $-2{,}55, -2{,}7$ e -3.

Ata o de agora vimos que o tamaño mostral e o punto de anclaxe inflúen na forma do histograma. Empregando a mostra do exemplo da Figura 1.2, ilustremos como a lonxitude dos intervalos tamén inflúe considerablemente na súa forma e así, na estimación da densidade. Por este motivo introduciremos diferentes criterios de erro que nos permitan determinar cal é o valor de b óptimo e que condicións debemos impoñer á función f para que iso teña sentido. Ademais, co fin de comparar o histograma con outros estimadores da densidade, calcularemos o orde deste estimador.

Na Figura 1.3 apréciase como o valor do parámetro ventá b é decisivo para determinar o estimador histograma. Para valores de b pequenos as barras son moi estreitas e altas en rexións cercanas ás modas e moi baixas nos extremos da estimación, dando lugar a unha

función con multitude de picos e vales. A medida que b aumenta, a estimación é cada vez máis suave ata que, na gráfica da dereita, as modas da estimación resultan inapreciables. Pasamos dun histograma moi dentado a un con escasas divisións, polo que determinar un valor de b axeitado resulta fundamental.

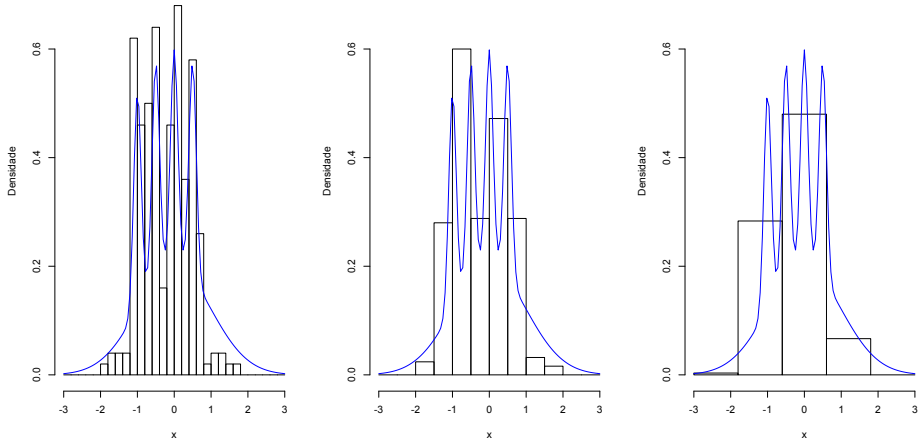

Figura 1.3: Histogramas dunha m.a.s de tamaño $n = 100$ do modelo M3 presentado no Apéndice B, cuxa densidade figura en color azul. Fixado $t_0 = -3$, de esquerda a dereita, os valores da lonxitude dos intervalos b son: 0,2, 0,5 e 1,2.

1.2. Criterios de erro

Nesta sección consideraremos algunhas definicións e resultados necesarios para analizar distintos criterios de erro do estimador histograma, así como para o resto dos estimadores que estudaremos posteriormente. O obxectivo será impoñer condicións á densidade teórica f para obter criterios que nos permitan escoller un parámetro ventá axeitado para a estimación, así como para comparar os diferentes estimadores. A súa formulación está adaptada á estimación da función de densidade e atópase, para máis xeneralidade e posterior uso, no marco multidimensional. Bastará tomar $d = 1$ para o caso unidimensional e $d = 2$ para o caso bidimensional.

Definición 1.1. Chámase nesgo dun estimador $\hat{f}$ dunha función de densidade f á diferenza entre a media (valor esperado) do estimador e f (valor verdadeiro do parámetro a estimar).

Denótase Nesgo$[\hat{f}(\cdot)]$ e a súa expresión en cada punto vén dada por

$$\text{Nesgo}[\hat{f}(x)] = \mathbb{E}[\hat{f}(x)] - f(x).$$

Dicimos que o estimador $\hat{f}$ é inesgado para estimar a función de densidade f se o seu nesgo é nulo, é dicir, se a súa media coincide coa función que desexamos estimar. Noutras palabras, $\hat{f}$ é un estimador inesgado de f se:

$$\text{para case todo } x \in \mathbb{R}^d,\ \mathbb{E}[\hat{f}(x)] = f(x),\ \text{ou equivalentemente, Nesgo}[\hat{f}(x)] = 0.$$

En caso contrario, dicimos que é un estimador nesgado e que a estimación presenta nesgo. Como o estimador se calcula en función da mostra, cuxa función de densidade é a función descoñecida f, o estimador tamén é función do tamaño mostral n. Así dicimos que o estimador $\hat{f} \equiv \hat{f}_n$ é asintóticamente inesgado para estimar f se:

$$\text{para case todo } x \in \mathbb{R}^d,\ \lim_{n\to\infty} \mathbb{E}[\hat{f}(x)] = f(x),\ \text{ou equivalentemente, } \lim_{n\to\infty} \text{Nesgo}[\hat{f}(x)] = 0.$$

Definición 1.2. Chámase varianza dun estimador $\hat{f}$ dunha función de densidade f á media do cadrado da desviación de dito estimador con respecto a súa media e defínese así:

$$\text{Var}[\hat{f}(x)] = \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2] = \mathbb{E}[\hat{f}^2(x)] - (\mathbb{E}[\hat{f}(x)])^2,$$

onde a última igualdade é consecuencia das propiedades da media.

Definición 1.3. Diremos que un estimador $\hat{f} \equiv \hat{f}_n$ dunha función de densidade f é consistente en media cadrática, e empregaremos abreviadamente consistente salvo en casos aos que leve a confusión, se verifica que

$$\text{para case todo } x \in \mathbb{R}^d,\ \lim_{n\to\infty} \text{Var}[\hat{f}(x)] + \text{Nesgo}^2[\hat{f}(x)] = 0,$$

onde Var$[\hat{f}(x)]$ + Nesgo$^2[\hat{f}(x)]$ é o que posteriormente definiremos como erro cadrático medio do estimador $\hat{f}$ ao estimar f, descomposto como a suma da varianza e o nesgo ao cadrado (ver Definición 1.4). Entón que un estimador sexa consistente en media cadrática significa que o seu erro cadrático medio converxa a cero cando $n \longrightarrow \infty$.

Definición 1.4. Definimos o erro cadrático medio dun estimador $\hat{f}$ ao estimar unha función de densidade f como a medida de erro puntual tal que a cada $x \in \mathbb{R}$ lle asigna o valor

$$MSE[\hat{f}(x)] = \mathbb{E}[(\hat{f}(x) - f(x))^2] = \text{Var}[\hat{f}(x)] + \text{Nesgo}^2[\hat{f}(x)],$$

onde a última igualdade se segue das propiedades da media.

Como veremos, o erro cadrático medio serve de guía para a elección do parámetro b. Sen embargo, este erro dependerá dunha forma complexa de b e da densidade poboacional f. Para poder ver de xeito máis claro como b inflúe no erro empregaremos aproximacións asintóticas (para mostras grandes) baseadas en desarrollos de Taylor. En concreto, imos calcular desarollos de Taylor da media e da varianza dos nosos estimadores, polo que o seguinte teorema é de gran utilidade no que segue de capítulo:

Teorema 1.5. ***(Teorema de Taylor).*** *Sexa $g \in C([a,d],\mathbb{R})$ $m+1$ veces diferenciable en (a,d), con $m \in \mathbb{N}$. Logo,*

$$g(x) = P_{m,x_0}(x) + R_{m,x_0}(x),$$

para todo $x, x_0 \in [a,d]$ onde

$$P_{m,x_0}(x) = \sum_{k=0}^{m} \frac{g^{k)}(x_0)}{k!}(x-x_0)^k$$

é o polinomio de Taylor de orde m centrado en x_0,

$$R_{m,x_0}(x) = \frac{g^{m+1)}(c_{x_0,x})}{(m+1)!}(x-x_o)^{m+1}$$

é a fórmula de Lagrange do resto de orde m, e $c_{x_o,x} \in \mathring{J}$, onde $J = L[x_0,x]$ denota o intervalo real pechado con extremos x_0 e x.

Demostración: Ver o libro Bartle and Sherbert 2002, Capítulo 6, Sección 6.4.

Observación 1.6. Se só esiximos que $g \in C([a,d],\mathbb{R})$ sexa $m \geqslant 1$ veces diferenciable no punto $x_0 \in (a,d)$, entón existe unha función $g_m \colon [a,d] \longrightarrow \mathbb{R}$ tal que

$$g(x) = P_{m,x_0}(x) + g_m(x)(x-x_0)^m \text{ e } \lim_{x \to x_0} g_m(x) = 0.$$

Notación 1.7 (Notación de Landau 1). Sexa $g \colon (-\epsilon,\epsilon) \longrightarrow \mathbb{R}$ unha función continua definida nun entorno de cero e sexa $k \geqslant 0$ un número enteiro. Dise que

$$g(h) = o(h^k) \Leftrightarrow \lim_{h \to 0} \frac{g(h)}{h^k} = 0.$$

Así reescribimos a Observación 1.6 como segue: $g(x) = P_{m,x_0}(x) + o((x-x_0)^m)$.

Agora formularemos outro resultado imprescindible para vindeiros cálculos, que está incluído no Teorema 1.5, sendo así menos xeral pero tamén de gran utilidade:

Teorema 1.8. ***(Teorema do Valor Medio de Lagrange).*** *Sexa $g \colon [a,d] \longrightarrow \mathbb{R}$ continua en $[a,d]$ e derivable en (a,d). Entón, existe $c \in (a,d)$ tal que $g(d) - g(a) = g'(c)(d-a)$.*

Demostración: Ver o libro Bartle and Sherbert 2002, Capítulo 6, Sección 6.2.

1.2.1. Media, varianza e criterios de erro do histograma

Agora imos proceder ao análise do erro que comete o estimador histograma ao estimar a función de densidade. Co obxectivo de calcular o valor do b óptimo para a estimación, analizar a influencia do orixe dos intervalos, e de obter distintos criterios de erro, comecemos calculando a súa media e varianza.

Fixemos un $x \in \mathbb{R}$ e supoñamos que F é continuamente derivable en $(x-\delta, x+\delta)$, para algún $\delta > 0$, e así f será continua aí. Dado que recubrimos a recta real con intervalos da forma B_k, con $k \in \mathbb{Z}$ e que $x \in \mathbb{R}$, é inmediato que para cada $n \in \mathbb{N}$, $\exists k_n \equiv k \in \mathbb{Z}$ tal que $x \in B_k$. Impoñamos a condición de que:

$$\lim_{n\to\infty} l(B_k) = \lim_{n\to\infty} l(t_k, t_{k+1}] = \lim_{n\to\infty} b_n = 0,$$

polo que para n suficientemente grande $B_k \subset (x-\delta, x+\delta)$, onde $\mu(A) = l(A)$ denota a medida de Lebesgue (coincidente, nunha dimensión, co concepto de lonxitude habitual) do conxunto Lebesgue-medible $A \subset \mathbb{R}$. Recordemos que a expresión $l(B_k)$ está ben definida por ser B_k un intervalo e polo tanto un conxunto Lebesgue-medible.

Deste xeito atopamos nas hipóteses do Teorema 1.8, polo que se cumpren as súas teses. Antes de aplicalo, destaquemos que para cada $k \in \mathbb{Z}$ se ten que

$$\int_{B_k} f(t)dt = \mathbb{P}(t_k, t_{k+1}] = F(t_{k+1}) - F(t_k),$$

empregando a definición de función de distribución e as propiedades da probabilidade, polo que $N_k \in B(n, F(t_{k+1}) - F(t_k))$. Así, N_k é unha variable aleatoria que depende da mostra, do parámetro ventá b e do punto de anclaxe t_0 do histograma. Como para cada $x \in B_k$, $\hat{f}_{hist}(x) = \frac{N_k}{nb}$, o estimador histograma é un múltiplo dunha binomial obtida como cociente de numeradores binomiais de n ensaios e probabilidades de éxito $\frac{F(t_{k+1})-F(t_k)}{nb}$, $k \in \mathbb{Z}$, divididos por nb, un número fixo (non aleatorio).
Este cociente non é unha binomial: unha binomial é unha variable aleatoria discreta que só toma valores enteiros entre 0 e n, e o histograma non ten esa propiedade. Recordemos que a media dunha binomial con n ensaios e probabilidade de éxito $p \in [0,1]$, $X \in B(n,p)$, é $\mathbb{E}[X] = np$ e a súa varianza é $\mathrm{Var}[X] = np(1-p)$.

O seguintes resultado danos condicións suficientes baixo as cales o estimador histograma e asintóticamente inesgado e consistente para estimar á densidade e, polo tanto, "bo" en termos de erro cadrático medio.

Proposición 1.9. *Sexa f unha función de densidade, $x \in \mathbb{R}$ e F a función da que f é derivada nun entorno de x. Supoñamos que* $\lim_{n\to\infty} b_n = 0$ *e que* $\lim_{n\to\infty} nb_n = \infty$. *Entón $\hat{f}_{hist}$ é un estimador asintóticamente inesgado e consistente para estimar f.*

Demostración. Comecemos calculando a media de $\hat{f}_{hist}$:

$$\mathbb{E}[\hat{f}_{hist}(x)] = \frac{F(t_{k+1}) - F(t_k)}{b} = \frac{F'(\xi_{x,n})(t_k + b - t_k)}{b} = f(\xi_{x,n}) = f(x) + o(1),$$

onde a última igualdade é consecuencia de que f é a derivada de F nun entorno de x que contén a $\xi \equiv \xi_{x,n}$, posto que $x, \xi \in B_k$, e que $\lim_{n\to\infty} \xi = x$.
Como $\lim_{n\to\infty} b_n = 0$, $\hat{f}_{hist}$ é un estimador asintóticamente inesgado de f.
Empregando cálculos realizados para a media e fixándonos en que $\lim_{n\to\infty} nb_n = \infty$ ademais de que $\lim_{n\to\infty} b_n = 0$, procedamos ao cálculo da varianza de $\hat{f}_{hist}$:

$$\begin{aligned}\mathrm{Var}[\hat{f}_{hist}(x)] &= \frac{(F(t_k + h) - F(t_k))(1 - (F(t_k + h) - F(t_k)))}{nb^2} = \\ &= \frac{1}{nb}(f(x) + o(1))(1 - bf(x) + o(b)) = \frac{1}{nb}(f(x) + o(1)) = \frac{f(x)}{nb} + o(\frac{1}{nb}).\end{aligned}$$

Baixo estas hipóteses obtemos que $\hat{f}_{hist}$ é un estimador consistente de f, é dicir, que $\lim_{n\to\infty} \mathrm{Var}[\hat{f}_{hist}(x)] = 0$, dado que tamén é asintóticamente inesgado.

□

Se esiximos que $F \in \mathcal{C}^2(x - \delta, x + \delta)$, en lugar de F derivable nun entorno de x, que supoñamos que é da forma $(x-\delta, x+\delta)$ para algún $\delta > 0$, grazas ao Teorema 1.5 e razoando de forma similar a como fixemos na demostración da Proposición 1.9, séguese que

$$\mathbb{E}[\hat{f}_{hist}(x)] = f(x) + \frac{3}{2}f'(x)b + o(b), \text{ e así que } \mathrm{Nesgo}[\hat{f}_{hist}(x)] = \frac{3}{2}f'(x)b + o(b).$$

Observemos que fixado un $b > 0$, se o valor absoluto de f' é grande, entón o nesgo do estimador será grande. Ademais pola regularidade esixida á función f, un valor absoluto da derivada grande interprétase como un gran crecemento (se $f' > 0$) ou decrecemento (se $f' < 0$) da función f. Isto é bastante razoable por ser o histograma unha función constante polo que terá dificultades ao estimar en zonas de gran crecemento ou decrecemento de f. Se nos atopamos nun punto crítico (son aqueles que verifican a ecuación $f'(x) = 0$), $\mathrm{Nesgo}[\hat{f}_{hist}(x)] = o(b)$ unicamente depende do termo de error $o(b)$.
A expresión do nesgo conduce a que o erro cadrático medio veña dado por:

$$MSE[\hat{f}_{hist}(x)] = \frac{1}{12}(f'(x))^2 b^2 + \frac{f(x)}{nb} + o(\frac{1}{nb} + b^2). \tag{1.2}$$

Supoñamos que $\int_{\mathbb{R}}(f'(x))^2dx < \infty$ e definamos o erro cadrático medio integral como unha medida de erro global, que se detallará con máis profundidade no Capítulo 2, e se obtén como se indica a continuación:

$$MISE[\hat{f}_{hist}(\cdot)] = \int_{\mathbb{R}} \text{Var}[\hat{f}_{hist}(x)]dx + \int_{\mathbb{R}} \text{Nesgo}^2[\hat{f}_{hist}(x)]dx =$$
$$= \frac{b^2}{12}\int_{\mathbb{R}}(f'(x))^2dx + \frac{1}{nb} + o(\frac{1}{nb} + b^2),$$

onde debemos impoñer algunha condición máis sobre a función f para garantir que a integral conmute co límite na igualdade

$$\int_{\mathbb{R}}[\frac{1}{12}(f'(x))^2b^2 + \frac{f(x)}{nb} + o(\frac{1}{nb} + b^2)]dx = \frac{b^2}{12}\int_{\mathbb{R}}(f'(x))^2dx + \frac{1}{nb} + o(\frac{1}{nb} + b^2) \quad (1.3)$$

e en concreto na igualdade $\int_{\mathbb{R}} o(\frac{1}{nb} + b^2)dx = o(\frac{1}{nb} + b^2)$ pois, en realidade $o(\frac{1}{nb} + b^2)$ é función de x por vir de desarrollos de Taylor da función f centrados en x. Con este fin, consideremos o seguinte teorema:

Teorema 1.10. ***Teorema da converxencia dominada de Lebesgue.*** *Sexa $\{g_n\}_{n\in\mathbb{N}}$ unha sucesión de funcións integrables con respecto da medida de Lebesgue μ, que converxe puntualmente a unha función medible g. Se existe unha función integrable p cumprindo que, $\forall n \in \mathbb{N}$, se dá a desigualdade $|g_n| \leqslant p$, entón a función g é integrable e $\int gd\mu = \lim_{n\to\infty}\int g_n d\mu$.*

Demostración: Ver o libro Sattinger 2004, Capítulo 3, Sección 3.3.

En virtude do Teorema 1.10, baixo as condicións nas que nos atopamos, se esiximos que f'' exista, sexa integrable e este limitada, temos garantizado que se cumpre a igualdade da expresión (1.3).

Finalmente, a expresión aproximada para o erro cadrático medio integral é o que se coñece como erro cadrático medio integral asintótico e ten a seguinte expresión:

$$AMISE[\hat{f}_{hist}(\cdot)] = \frac{b^2}{12}\int_{\mathbb{R}}(f'(x))^2dx + \frac{1}{nb}. \quad (1.4)$$

O valor b que minimiza a expresión anterior coñécese como b_{AMISE} e obtense derivando dita expresión, igualando a cero e despexando. A derivada vén dada por:

$$\frac{b}{6}\int_{\mathbb{R}}(f'(x))^2dx - \frac{1}{nb^2} \text{ polo que, } b_{AMISE} = (\frac{6}{\int_{\mathbb{R}}(f'(x))^2dx})^{\frac{1}{3}}n^{\frac{-1}{3}},$$

sempre que $\int_{\mathbb{R}}(f'(x))^2dx \neq 0$, é dicir, que f' non se anule nun conxunto de medida non nula. En caso contrario -como ocorre por exemplo coa $U[0,1]$- vimos que o nesgo depende

unicamente do termo de erro $o(b)$ e ademais $MISE[\hat{f}_{hist}(\cdot)] = \frac{1}{nb} + o(\frac{1}{nb} + b^2)$; neste caso non é posible definir b_{AMISE} pero si calcular un b axeitado para á estimación sen máis que decatarse de que a función a estimar é constante a trozos en conxuntos de medida non nula, polo que probando diferentes valores de b resulta fácil decatarse de cales son máis axeitados.

Pódese comprobar que tal e como definimos b_{AMISE} minimiza $AMISE$, avaliando a segunda derivada e vendo que o resultado é < 0. Substituíndoo na expresión (1.4) obtemos:

$$\left(\frac{9}{16}\int_{\mathbb{R}}(f'(x))^2dx\right)^{\frac{1}{3}} n^{\frac{-2}{3}}. \tag{1.5}$$

Notación 1.11 (Notación de Landau 2). Sexa $g\colon (-\epsilon,\epsilon) \longrightarrow \mathbb{R}$ unha función continua definida nun entorno de cero e sexa $k \geqslant 0$ un número enteiro. Diremos que $g(h)$ é de orde $h^k \equiv h(n)^k$ cando $n \longrightarrow \infty$, e escribiremos

$$g(h) = \mathcal{O}(h^k), \text{ se e só se } \limsup_{n\to\infty} |\frac{g(h)}{h^k}| < \infty.$$

Deste xeito, no histograma o $AMISE$ óptimo é da orde $\mathcal{O}(n^{\frac{-2}{3}})$, moito peor que o que imos conseguir cos vindeiros estimadores da densidade, como explicaremos nas seguintes seccións. Ademais a constante é proporcional a $\int_{\mathbb{R}}(f'(x))^2dx$, polo que o estimador histograma funcionara moi ben estimando densidades constantes a trozos, como era de esperar.

Na Figura 1.4 consideramos unha variable aleatoria cuxa función de densidade non é continua en $\mathbb{R}$. Trátase dunha Exp(1). Recordemos que a función de densidade dunha exponencial con parámetro $\lambda > 0$ vén dada por:

$$f(x) := \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & x \leqslant 0, \end{cases}$$

e que presenta unha descontinuidade evitable na orixe para cada $\lambda > 0$.

En entornos da orixe, o estimador histograma pode funcionar mal posto que, en función de como elixamos o t_0, pode asignarlle densidade positiva a valores de $(-\infty, 0)$, asignación que carece de sentido neste caso. Algo similar ocorre en todos aqueles casos nos que a función de densidade presente descontinuidades evitables e polo tanto, na gráfica de ditas funcións poidamos apreciar bordes, que serán os lugares onde o estimador histograma poida chegar a ter un mal comportamento sen unha axeitada elección de t_0, aínda que b cumpra as condicións pedidas na Proposición 1.9. Ademais non son "uns poucos casos concretos" dado

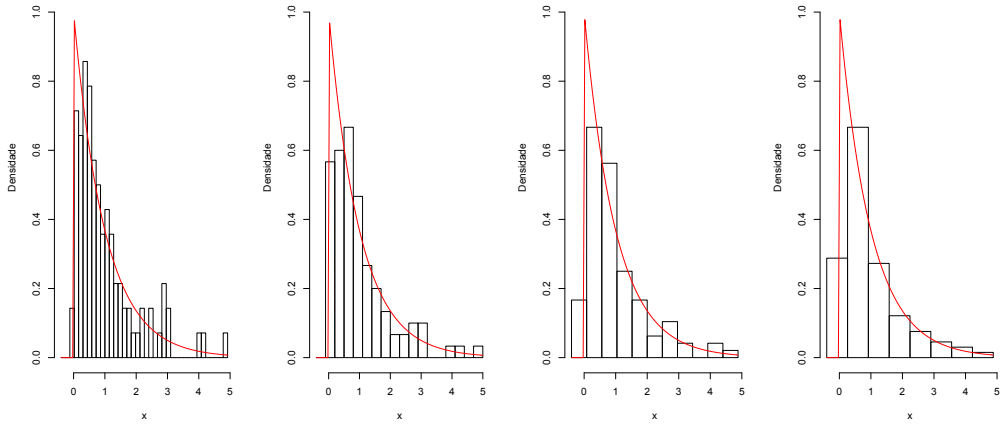

Figura 1.4: Histogramas da mesma m.a.s dunha Exp(1) de tamaño $n = 100$, para $t_0 = -0{,}4$ e, de esquerda a dereita, para $b = 0{,}14, b = 0{,}30, h = 0{,}48$ e $b = 0{,}66$, de cor negro. En vermello a función de densidade da Exp(1). En todos os casos asignamos densidade positiva a intervalos non disxuntos con $\mathbb{R}^-$.

que se presentan en familias de funcións paramétricas moi coñecidas, como a exponencial ou a uniforme.

Particularizando ao exemplo da Exp(1) e ao extremo $x = 0$, se escollemos t_0 tal que $\exists k \in \mathbb{Z}$ con $0 \in (t_k, t_{k+1}]^{\text{o}}$ e $\exists i \in \{1, ..., n\}$ con $X_i \in (t_k, t_{k+1}]$, ao intervalo $(t_k, 0)$ asígnaselle densidade positiva. Así, nestes puntos o estimador presentará nesgo positivo.

Ademais en $x = 0$ a densidade da exponencial vale 1, decrecendo estritamente na recta real positiva. A medida que a lonxitude dos intervalos diminúe e o tamaño mostral aumenta, a cuantía deste problema do histograma tende a diminuír: atopámonos fóra das hipóteses de suavidade que supuxemos ata o de agora e esiximos na Proposición 1.9; sen embargo, se restrinximos a función teórica a $[0, \infty)$ esta resulta continua polo que nos atopamos, en gran medida, " baixo as condicións da Proposición 1.9 " e así tamén podemos asegurar que $\hat{f}_{hist}$ é un estimador da densidade asintóticamente inesgado.

Para rematar co histograma unidimensional, destaquemos que a influencia do orixe dos intervalos é clave para que o estimador non presente nesgo nas descontinuidades evitables. Se temos ideas previas a mostraxe sobre a natureza dos datos e, por exemplo, non ten sentido asignar densidade positiva a valores negativos, podemos solucionar este problema escollendo $t_0 = 0$ dado que así non asignaremos densidade positiva a ningún intervalo que

interseque con $\mathbb{R}^-$ posto que ningún valor mostral caerá nel.

Por outra banda, o $MISE$ é practicamente insensible a variacións de posición do punto de anclaxe, agás se unha descontinuidade da función de densidade orixinal se atopa no interior dun intervalo, en lugar de coincidir cun extremo do mesmo (ver o libro Scott (1992, páxinas 11-12)). Ademais, o aspecto do histograma pode variar, para algunhas mostras, se consideramos parámetros ventá pequenos, por exemplo influíndo no número de modas.

1.3. Estimador Naive

O estimador Naive é un histograma móbil no sentido de que cada x será o centro do intervalo empregado para construír o estimador. Deste xeito resolvemos o problema de escoller o punto inicial t_0 do histograma convencional, pero non o de elección do ancho do intervalo nin ponderamos os pesos dos datos en función da súa proximidade a x, e segue a ser un estimador descontinuo -ao igual que o histograma tradicional-. Foi proposto por primeira vez por Rosenblatt (1956).

Definimos o estimador Naive como:

$$\hat{f}(x) = \frac{1}{2nh}\sum_{i=1}^{n}\mathbb{I}(X_i \in (x-h, x+h]) = \frac{1}{2nh}\sum_{i=1}^{n}\mathbb{I}(x-h < X_i \leqslant x+h), x \in \mathbb{R}. \quad (1.6)$$

O estimador Naive, $\hat{f}$, é unha función con $2n$ descontinuidades, dúas por cada observación mostral: teremos descontinuidades evitables nos puntos $X_i - h$ e $X_i + h$, $i \in \{1, ..., n\}$, onde h é o parámetro ventá e debémolo fixar en función de n: $h \equiv h(n) \equiv h_n$, que seguiremos denotando por h para simplificar notación. Entón empregando o estimador Naive, en lugar do histograma, non resolvemos o problema da descontinuidade da estimación nin da dependencia de h.

Consideremos agora a función de distribución empírica:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i \leqslant x),\ x \in \mathbb{R}, \quad (1.7)$$

con $\sum_{i=1}^{n}\mathbb{I}(X_i \leqslant x) \in B(n, F(x))$ por ser $F_n(x)$ un estimador de $F(x) = \mathbb{P}(X \leqslant x)$, e ademais, o Teorema de Glivenko-Canteilli asegúranos que

$$\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|F_n(x) - F(x)| = 0 \text{ (converxencia a 0 con probabilidade 1).}$$

A súa demostración pode atoparse no libro Billingsley 1995, páxina 269.

Deste xeito, $\mathbb{E}(F_n(x)) = F(x)$, $\mathrm{Var}(F_n(x)) = \frac{F(x)(1-F(x))}{n}$ e para cada x fixo, pola expresión (1.7), a función de distribución empírica é múltiplo dunha variable aleatoria que segue unha distribución binomial.
O estimador Naive esta motivado polo feito de que

$$f(x) = \lim_{h\to\ 0} \frac{1}{2h}\mathbb{P}(x-h < X \leqslant x+h) = \lim_{h\to\ 0} \frac{1}{2h}[F(x+h) - F(x-h)],\ x \in \mathbb{R}, \quad (1.8)$$

dado que podemos reescribir o noso estimador empregando a función de distribución empírica do seguinte xeito:

$$\hat{f}(x) = \frac{1}{2h}[F_n(x+h) - F_n(x-h)],\ x \in \mathbb{R}. \quad (1.9)$$

Destaquemos que para a construción do estimador Naive estamos a considerar intervalos de lonxitude $2h$, dado que son intervalos centrados en cada observación mostral de extensión h cara a dereita e cara a esquerda da mesma, e no histograma considerábamos intervalos de lonxitude b. De aí a que apareza o termo $\frac{1}{2}$ na expresión do estimador Naive e non na do histograma, para conseguir que este sexa tamén unha función de densidade (isto permítenos obter que integre 1 en $\mathbb{R}$).

Na Figura 1.5 observamos a converxencia da expresión (1.8) no caso de que f sexa a densidade correspondente ao modelo M4 do Apéndice B, debuxada en cor vermello. Na vindeira subsección imos repetir o análise de erro do histograma, pero para o estimador Naive, calculando a súa media e varianza. Teñamos en conta que a expresión exacta da media vén dada por $\frac{1}{2h}[F(x+h) - F(x-h)]$ polo que, neste caso, a expresión exacta do nesgo é coñecida e vén dada como suma de funcións de densidade e distribución normais (ϕ_{μ,σ^2} e $\Phi_{\mu,\sigma}$ respectivamente para unha $N(\mu, \sigma^2)$).

En consecuencia, o nesgo deste exemplo vén dado por

$$\frac{1}{4h}((\Phi_{-1,(\frac{2}{3})^2} + \Phi_{1,(\frac{2}{3})^2})(x+h) - (\Phi_{-1,(\frac{2}{3})^2} + \Phi_{1,(\frac{2}{3})^2})(x-h)) - \frac{1}{2}((\phi_{-1,(\frac{2}{3})^2} + \phi_{1,(\frac{2}{3})^2})(x)) =$$
$$= \frac{1}{4h}((\Phi_{-1-h,(\frac{2}{3})^2} + \Phi_{1-h,(\frac{2}{3})^2})(x) - (\Phi_{-1+h,(\frac{2}{3})^2} + \Phi_{1+h,(\frac{2}{3})^2})(x)) - \frac{1}{2}((\phi_{-1,(\frac{2}{3})^2} + \phi_{1,(\frac{2}{3})^2})(x)).$$

Nunha escala decrecente de grises, a medida que decrece h, representamos a función media $\frac{1}{2h}[F(x+h) - F(x-h)]$, para $h \in \{2, 1{,}5, 1{,}25, 1, 0{,}75, 0{,}5, 0{,}25, 0{,}15\}$ e F a función de distribución asociada. A medida que o valor de h diminúe (tende a cero) considerando valores sempre positivos por ser o parámetro ventá, as gráficas tende a achegarse máis á curva vermella. Sen embargo, observamos o fenómeno do nesgo tanto nas modas (a

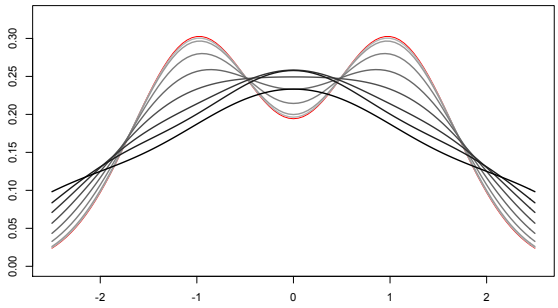

Figura 1.5: Ilustración de que $\frac{1}{2h}[F(x+h) - F(x-h)] \longrightarrow f(x)$ cando $h \longrightarrow 0$, no caso particular de que f siga a distribución do modelo M4 do Apéndice B. En vermello a densidade orixinal e en diferentes tonos de gris as funcións estimadas, correspondendo a curva máis escura a $h = 2$ e aclarando o cor ata acadar $h = 0{,}15$.

distribución é bimodal, cunha moda en $x = 1$ e outra en $x = -1$) onde as funcións estimadas se quedan por debaixo da verdadeira función, como nos vales, onde quedan por encima.

1.3.1. Criterios de erro do estimador Naive

A continuación imos estudar o análise do erro que comete o estimador Naive ao estimar a función de densidade, co fin de calcular unha expresión para o parámetro ventá óptimo no sentido de que minimice certos criterios de erro que xa consideramos no histograma e que posteriormente tamén consideraremos no estimador tipo núcleo. Isto tamén nos permitirá comparar estes tres estimadores e ver cal é máis eficiente para estimar á función de densidade dun conxunto de datos. Con este obxectivo, consideremos o seguinte resultado, cuxa proba involucra o cálculo da media e varianza do estimador Naive.

Proposición 1.12. *Sexa f unha función de densidade, F a función de distribución asociada e $x \in \mathbb{R}$. Supoñamos que F é continuamente derivable nun entorno de x e que $\lim_{n\to\infty} h_n = 0$ e $\lim_{n\to\infty} nh_n = \infty$. Entón $\hat{f}$ é un estimador asintóticamente inesgado e consistente para estimar f.*

Demostración. En primeiro lugar, como F é continuamente derivable nun entorno de x, f é continua en dito entorno. Ademais podemos considerar, sen perda de xeneralidade, que o entorno é da forma $(x-\delta, x+\delta)$, para algún $\delta > 0$. Deste xeito atopámonos nas condicións do Teorema 1.8, polo que se cumpren as súas teses.

Comecemos calculando a media de $\hat{f}$:

$$\begin{aligned}\mathbb{E}[\hat{f}(x)] &= \frac{1}{2h}\mathbb{E}[F_n(x+h) - F_n(x-h)] = \frac{1}{2h}[F(x+h) - F(x-h)] = \\ &= \frac{1}{2h}[F(x+h) - F(x) + F(x) - F(x-h)] = \frac{1}{2h}[F'(\xi_1)h + F'(\xi_2)h] = \\ &= \frac{1}{2}[F'(\xi_1) + F'(\xi_2)] = F'(x) + o(1) = f(x) + o(1),\end{aligned}$$

con $\xi_1 \in (x, x+h)$ e $\xi_2 \in (x-h, x)$; empregamos que $\lim_{n\to\infty} \xi_1 = \lim_{n\to\infty} \xi_2 = x$ e a continuidade de f nun entorno de x. Polo tanto, baixo estas hipóteses $\hat{f}$ é un estimador asintóticamente inesgado de f.

Empregando cálculos realizados para a media e fixándonos en que $\lim_{n\to\infty} nh_n = \infty$, ademais de que $\lim_{n\to\infty} h_n = 0$, procedamos ao cálculo da varianza de $\hat{f}$.

$$\begin{aligned}\mathrm{Var}[\hat{f}(x)] &= \frac{1}{4h^2}\mathrm{Var}[F_n(x+h) - F_n(x-h)] = \\ &= \frac{1}{4h^2 n}(F(x+h) - F(x-h))(1 - (F(x+h) - F(x-h))) = \\ &= \frac{1}{2hn}(f(x) + o(1))(1 - 2hf(x) + o(h)) = \frac{f(x)}{2nh} + o(\frac{1}{nh}),\end{aligned}$$

onde a igualdade

$$\mathrm{Var}[F_n(x+h) - F_n(x-h)] = \frac{1}{n}(F(x+h) - F(x-h))(1 - (F(x+h) - F(x-h)))$$

é válida dado que

$$F_n(x+h) - F_n(x-h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \in (x-h, x+h]) = \mathbb{P}_n(x-h, x+h] \in B(n, F(x+h) - F(x-h)),$$

onde o subíndice n de $\mathbb{P}_n$ indica que dita probabilidade se calcula supoñendo que a distribución dos datos vén dada pola función de distribución empírica. Baixo estas hipóteses, $\lim_{n\to\infty} \mathrm{Var}[\hat{f}(x)] = 0$ e así, $\hat{f}$ é un estimador consistente de f.

□

Como a media real do estimador Naive é $\frac{1}{2h}[F(x+h) - F(x-h)]$, na Figura 1.5 xa adiantábamos graficamente o feito de que é un estimador asintóticamente inesgado, no caso de que f seguise a distribución bimodal dado polo modelo M4 do Apéndice B, representando a media do estimador para distintos valores de h. Recordemos que se apreciaba que dita media tendía a achegarse á gráfica de f cando $h \longrightarrow 0$, quedándose por encima nas modas e por debaixo nos vales, sendo estas rexións zonas onde a curva das distintas medias non chegaba a tocar á da función a estimar.

Paralelamente ao que ocorría no caso do histograma, restrinxindo F a unha clase de funcións máis suave, poderemos garantir a existencia de desarrollos de Taylor de maior orde, e obter unha expresión do erro cadrático medio suficientemente desarrollada. Se esiximos que $F \in \mathcal{C}^3(x-\delta, x+\delta)$, en lugar de que $F \in \mathcal{C}^1(x-\delta, x+\delta)$, é dicir, de que sexa continuamente derivable, en virtude do Teorema 1.5, obtemos que:

$$\mathbb{E}[\hat{f}(x)] = f(x) + \frac{1}{6}f''(x)h^2 + o(h^2), \text{ e así, Nesgo}[\hat{f}(x)] = \frac{1}{6}f''(x)h^2 + o(h^2).$$

Fixando un $h > 0$, se o valor absoluto de f'' é grande, o nesgo do estimador será grande. Ademais, pola regularidade esixida á función f, un valor absoluto da derivada segunda grande está directamente relacionada cunha curvatura da función f moi marcada, cóncava se $f'' > 0$ e convexa se $f'' < 0$. Deste xeito, o estimador Naive ten limitacións ao estimar a función de densidade en rexións de vales e modas desta, quedándose por encima no primeiro dos casos e por debaixo no segundo. Na Figura 1.6 ilústrase este feito no caso de que f sexa a distribución trimodal correspondente ao modelo M2 do Apéndice B: nas tres modas e nos vales a estimación ten dificultades para axustarse á función a estimar, mellorando este problema ao aumentar o tamaño mostral.

Se nos atopamos nun punto de inflexión (son aqueles que verifican a ecuación $f''(x) = 0$), o Nesgo$[\hat{f}(x)] = o(h^2)$ unicamente depende do termo de error $o(h^2)$.
A expresión do nesgo conduce a que a do erro cadrático medio sexa a seguinte:

$$MSE[\hat{f}(x)] = \frac{1}{36}(f''(x))^2h^4 + \frac{f(x)}{2nh} + o(\frac{1}{nh} + h^4). \tag{1.10}$$

Pola definición de cadrático medio integral, impoñendo que $\int_{\mathbb{R}}(f''(x))^2dx < \infty$, obtemos que a súa expresión vén dada por

$$MISE[\hat{f}(\cdot)] = \frac{h^4}{36}\int_{\mathbb{R}}(f''(x))^2dx + \frac{1}{2nh} + o(\frac{1}{nh} + h^4), \tag{1.11}$$

onde a igualdade é válida se impoñemos algunha condición sobre a función f que garanta que a integral conmute co límite na vindeira igualdade:

$$\int_{\mathbb{R}}[\frac{1}{36}(f''(x))^2h^4 + \frac{f(x)}{2nh} + o(\frac{1}{nh} + h^4)]dx = \frac{h^4}{36}\int_{\mathbb{R}}(f''(x))^2dx + \frac{1}{2nh} + o(\frac{1}{nh} + h^4),$$

e en concreto na igualdade $\int_{\mathbb{R}} o(\frac{1}{nh} + h^4)dx = o(\frac{1}{nh} + h^4)$, por ser o integrando unha función de x. Isto é consecuencia de que o termo de erro procede de operar desarrollos de Taylor de segundo orde de $F(x+h)$ e $F(x-h)$. En concreto, polo Teorema 1.5, vén dado por:

$$\frac{h^3}{12}(f'''(c_{x,x+h}) - f'''(c_{x,x-h})),$$

con $c_{x,x+h} \in (x, x+h)$ e $c_{x,x-h} \in (x-h, x)$. En virtude do Teorema 1.10, é suficiente engadir as hipóteses de que f''' sexa integrable e limitada, para que a súa integral conmute co límite.

Finalmente, a expresión do erro cadrático medio integral asintótico é

$$AMISE[\hat{f}(\cdot)] = \frac{h^4}{36}\int_{\mathbb{R}}(f''(x))^2dx + \frac{1}{2nh}, \tag{1.12}$$

e así a expresión da súa derivada vén dada por

$$\frac{h^3}{9}\int_{\mathbb{R}}(f''(x))^2dx - \frac{1}{2nh^2},$$

polo que o mínimo do $AMISE$, é dicir, o h_{AMISE}, é $h_{AMISE} = (\frac{9}{2\int_{\mathbb{R}}(f''(x))^2dx})^{\frac{1}{5}}n^{\frac{-1}{5}}$, sempre e cando $\int_{\mathbb{R}}(f''(x))^2dx \neq 0$. Efectivamente, isto último pódese comprobar avaliando a derivada segunda de $AMISE$ en h_{AMISE} e vendo que o valor obtido é negativo. Substituíndoo na expresión (1.12) obtemos que

$$2^{\frac{-4}{5}}9^{\frac{-1}{5}}\frac{5}{4}(\int_{\mathbb{R}}(f''(x))^2dx)^{\frac{1}{5}}n^{\frac{-4}{5}}, \tag{1.13}$$

polo que o $AMISE$ óptimo é proporcional a $n^{\frac{-4}{5}}$ e a $\int_{\mathbb{R}}(f''(x))^2dx$.

Destaquemos que, en virtude das expresións (1.5) do histograma e (1.13) do estimador Naive, o primeiro deles é da orde $\mathcal{O}(n^{-\frac{2}{3}})$ e o segundo, $\mathcal{O}(n^{-\frac{4}{5}})$, polo que o estimador Naive é mellor en termos de orde de converxencia que o estimador histograma. Isto é moi intuitivo, sen máis que recordar que o estimador Naive é un histograma móbil e, polo tanto, unha mellora do histograma tradicional. Agora ben, tamén esiximos máis regularidade á función de densidade no caso do estimador Naive que no do histograma, como se reflexa nas Proposicións 1.9 e 1.12 respectivamente.

Na gráfica da esquerda da Figura 1.6 representamos en vermello unha estimación Naive empregando unha mostra aleatoria simple de tamaño $n = 30$ da mestura de distribucións normais dada polo modelo M2 do Apéndice B. No gráfico da dereita, o tamaño mostral aumenta ata valer $n = 210$, obtendo así unha estimación máis precisa e representada en azul. Analogamente ao que ocorría co histograma, vimos que o tamaño mostral ten unha influencia decisiva na estimación da densidade co estimador Naive, mellorándoa notablemente ao aumentar este. Ademais pode apreciarse que o resultado é mellor con este estimador que co histograma da Figura 1.1, para un mesmo valor de n. A función real represéntase en negro.

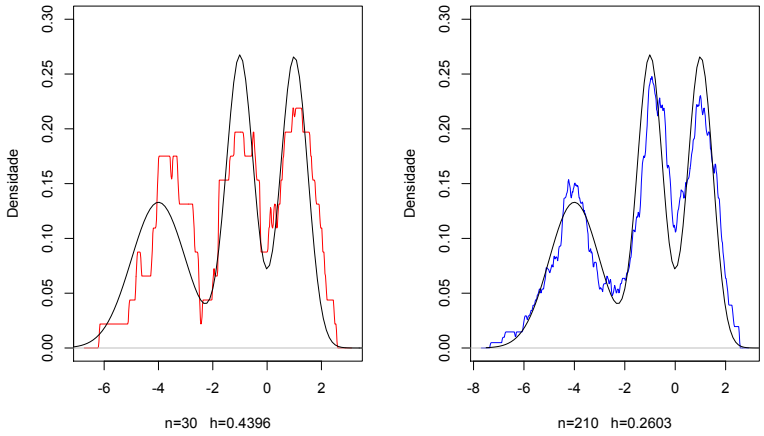

Figura 1.6: Estimacións Naive de dúas m.a.s da mestura de distribucións normais dada polo modelo M2 do Apéndice B, para $n = 30$ (esquerda) en vermello e para $n = 210$ (dereita) en azul, xunto coa súa función de densidade en negro. O h elixido en ambos casos obtívose empregando o criterio de validación cruzada inesgada.

A partir da expresión (1.12) deducimos a expresión do h_{AMISE} para o estimador Naive. Sen embargo, esta expresión da ventá óptima en termos do $AMISE$ ten o inconveniente de que depende de $\int_{\mathbb{R}}(f''(x))^2dx$ que é unha integral descoñecida. Por iso mesmo, ao final deste capítulo adicaremos un apartado para estudar como elixir o h óptimo a partir da mostra, centrándonos nas eleccións que fai R, na estimación tipo núcleo e, na anterior sección, veremos que o estimador Naive é un caso particular do estimador tipo núcleo, polo que quedará xustificada a elección do h neste exemplo e nos vindeiros. Adiantemos que o criterio de selección de h empregado neste exemplo é o criterio de validación cruzada inesgada, dependente da mostra e polo tanto, do tamaño mostral, variando dunha gráfica á outra.
Por outra banda, se $\int_{\mathbb{R}}(f''(x))^2dx = 0$ non é posible obter unha expresión do $AMISE$ óptima, polo que tamén veremos como proceder neste caso.

1.4. Estimador tipo núcleo unidimensional

Ata o de agora traballamos co histograma e co histograma móbil (ou estimador Naive) pero ambos presentan desvantaxes relacionadas coa súa descontinuidade e co feito de que, fixado un intervalo, non ponderan os pesos das observacións mostrais en función da súa proximidade a x, sendo x o punto onde queremos estimar a función de densidade. Co ob-

xectivo de resolver estes inconvenientes introduciuse o estimador tipo núcleo. Os primeiros artigos coñecidos que empregan a estimación tipo núcleo co obxectivo de estimar densidades son de Akaike (1954), Rosenblatt (1956) e Parzen (1962).

Como consecuencia das súas boas propiedades así como da súa flexibilidade de forma, na actualidade son múltiples as aportacións ao estudo dos estimadores tipo núcleo da densidade e o seu uso. Nesta sección imos introducir a expresión do estimador tipo núcleo dunha función de densidade, analizaremos os distintos erros cometidos na estimación e compararémolo cos dous estimadores xa presentados neste capítulo. Tamén abordaremos a selección do parámetro ventá h a partir da mostra, centrándonos nos principais métodos utilizados por R.

Sexa K unha función núcleo (densidade unimodal simétrica de media cero) e $h > 0$ o coñecido como parámetro ventá, cuxo papel é medir a veciñanza. Recordemos que unha variable aleatoria X ou, equivalentemente, a súa distribución, é simétrica entorno a un punto $\alpha \in \mathbb{R}$ se $\mathbb{P}[X \leqslant \alpha + x] = \mathbb{P}[X \geqslant \alpha - x],\ \forall x \in \mathbb{R}$. Esta definición xeralízase inmediatamente a dúas dimensións sen máis que considerar un vector aleatorio (X, Y) e un punto (α_x, α_y) do plano. Neste caso, a función núcleo é simétrica entorno á súa media, é dicir, entorno á orixe.

Como K é unha función de densidade, é claro que $\int_{\mathbb{R}} K(x)dx = 1$ e que $K \geqslant 0$, e como ademais é simétrica, $\int_{\mathbb{R}} xK(x)dx = 0$.
Consideraremos que os vecindarios son intervalos abertos, de lonxitude constante h. Igual que nos outros dous estimadores, o valor de h axeitado depende da rugosidade da función f descoñecida e do tamaño mostral n, e así $h \equiv h(n) \equiv h_n$. O obxectivo é escoller h de xeito óptimo para obter unha equilibrio entre a varianza e o nesgo do estimador tipo núcleo que presentaremos a continuación.

Abusando da notación, denotaremos ao estimador tipo núcleo do mesmo modo que ao Naive dado que, como xa veremos, este último será un caso particular.
Defínese o estimador tipo núcleo de función núcleo K como:

$$\hat{f}(x, h) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i),\ \forall x \in \mathbb{R}, \tag{1.14}$$

onde $K_h(x) = \frac{1}{h} K(\frac{x}{h}),\ \forall x \in \mathbb{R}$.

Pode interpretarse como unha transformación da función de distribución empírica antes definida, para dar lugar a unha función continua onde a función núcleo redistribúe a

probabilidade $\frac{1}{n}$ no vecindario de cada valor mostral. Logo K_h é a densidade da variable aleatoria hZ, onde Z segue unha distribución con densidade K e o núcleo reescalado trasladado $K_h(\cdot - X_i)$ é a densidade da variable $X_i + hZ$, para cada $i \in \{1, ..., n\}$.

Fixémonos en que se consideramos como núcleo a función $K(x) = \frac{1}{2}\mathbb{I}(-1, 1]$, que otorga o mesmo peso a cada dato do intervalo $(-1, 1]$, obtemos o Estimador Naive. Cómpre destacar que, neste caso, K é unha función de densidade simétrica, pero non é unimodal. Logo o estimador tipo núcleo pode verse como unha xeneralización do estimador Naive, pois este último pode formularse no contexto do estimador tipo núcleo do seguinte xeito:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right), \text{ con } K(x) = \frac{1}{2}\mathbb{I}(-1, 1].$$

Para fixar ideas consideremos que o núcleo é unha normal estándar univariante: $K(x) = \phi(x)$. Entón o estimador tipo núcleo vén dado por

$$\hat{f}(x, h) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{(x - X_i)^2}{2h^2}\right) = \frac{1}{nh\sqrt{2\pi}}\sum_{i=1}^{n} \exp\left(-\frac{(x - X_i)^2}{2h^2}\right), \tag{1.15}$$

onde para cada $i = 1, ..., n$, $K_h(x - X_i)$ é a densidade dunha variable aleatoria con distribución normal de media X_i e varianza h^2. Graficamente estamos considerando n campás de Gauss, cada un delas centrada en cada un dos datos da mostra e con varianza h^2. Así, este estimador tipo núcleo é un promedio de n distribucións normais con varianza común pero medias distintas, localizadas en cada un dos datos observados.

Na Figura 1.7 representamos en cor azul dez campás de Gauss reescaladas correspondentes a centrar dez distribucións normais de varianza $h^2 = 0{,}124^2$ en dez observacións mostrais procedentes dunha distribución $U[0, 1]$. Marcamos cun punto negro cada unha das observacións mostrais, que serán o centro das correspondentes campás de Gauss (curvas das funcións de densidade $N(X_i, 0{,}124^2) : i \in \{1, .., 10\}$). Están reescaladas no sentido de que están divididas por 10, xa que é así como entran no promedio que dá lugar ao estimador tipo núcleo. En cor negro engádese a estimación tipo núcleo da función de densidade.

Na gráfica da esquerda da Figura 1.8 representamos en vermello unha estimación tipo núcleo con núcleo normal estándar, empregando unha m.a.s de tamaño $n = 30$ da densidade correspondente ao modelo M2 do Apéndice B. Pode apreciarse que a estimación é mellor que no caso do histograma da Figura 1.1 e da estimación Naive da Figura 1.6,

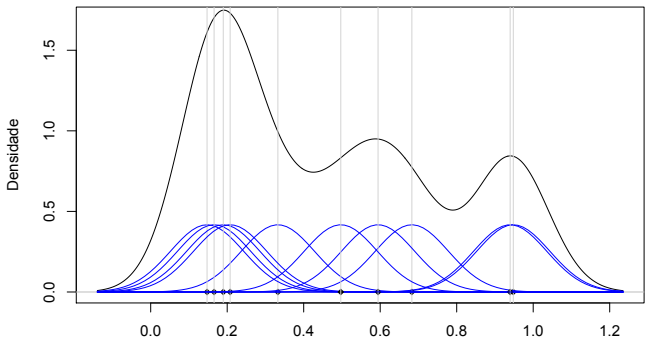

Figura 1.7: En cor azul, dez campás de Gauss reescaladas de varianza $h = 0{,}124^2$ centradas nos dez datos dunha m.a.s procedente dunha distribución $U[0,1]$ de tamaño $n = 10$. En cor negro a estimación tipo núcleo resultante.

para un mesmo valor de n. Veremos que, en termos de suavidade, o estimador tipo núcleo é mellor que os dous anteriores, e en termos de converxencia, equiparable ao estimador Naive (xeralizao) e mellor que o histograma.

Na gráfica da dereita, o tamaño mostral aumenta ata valer $n = 210$, obtendo así unha estimación máis precisa, representada en cor azul. En ambos casos debúxase a curva da densidade real en cor negro. No último apartado deste capítulo veremos os criterios que emprega R para axustar o h destas estimacións tipo núcleo. Neste caso, o criterio empregado é o de validación cruzada inesgada. Podemos apreciar que para $n = 210$ e $h = 0{,}2811$ a estimación tipo núcleo é bastante similar á densidade teórica, pero segue a presentar certo nesgo nos extremos e nas modas da estimación.

Observemos que unha desvantaxe propia do estimador tipo núcleo é que a dependencia do parámetro ventá h se suma á da función núcleo K. Ademais, o estimador tipo núcleo herda as propiedades de suavidade do núcleo.

1.4.1. Criterios de erro do estimador tipo núcleo

Analicemos os distintos erros que comete o estimador tipo núcleo ao estimar a función de densidade, co obxectivo de calcular o valor de h óptimo para a estimación, o núcleo máis eficiente e de obter expresións comparables ás dos estimadores anteriores, que neste

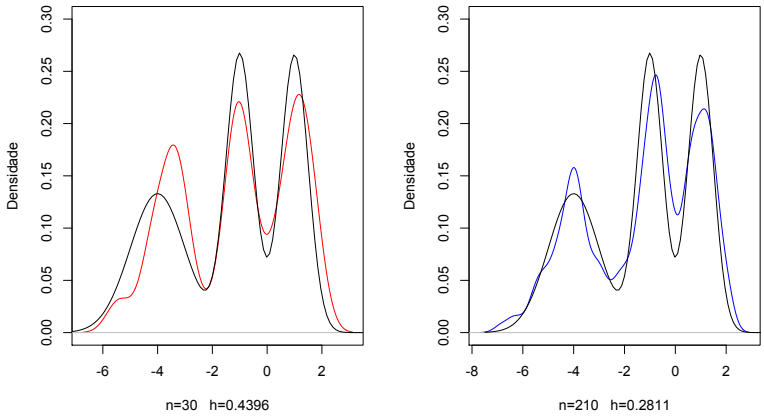

Figura 1.8: Estimacións tipo núcleo con núcleo normal estándar de dúas m.a.s do modelo M2 recollido no Apéndice B, para $n \in \{30, 210\}$ en cor vermello e azul, respectivamente. En cor negro represéntase a densidade teórica. O h elixido en ambos casos obtívose empregando o criterio de validación cruzada inesgada.

caso tamén virán dadas en termos do núcleo K. En primeiro lugar, imos medir o erro ao estimar a densidade f nun punto fixo x por $\hat{f}(x, h)$, polo que será unha medida de erro puntual. En apartados anteriores xa definíamos o erro cadrático medio como segue:

$$MSE[\hat{f}(x, h) = \text{Var}[\hat{f}(x, h)] + \text{Nesgo}^2[\hat{f}(x, h)]. \tag{1.16}$$

As convolucións xogan un papel clave na análise do erro do estimador tipo núcleo. Polo tanto, a seguinte definición será de gran utilidade e, como veremos, poderase xeneralizar ao caso bidimensional.

Definición 1.13. A convolución de dúas funcións de densidade f e g de dúas variables aleatorias independentes X e $\tilde{X}$ defínese como

$$(f * g)(x) = \int_{\mathbb{R}} f(x - \tilde{x})g(\tilde{x})d\tilde{x}, \tag{1.17}$$

onde a expresión ten sentido por ser f e g funcións de densidade e, polo tanto, integrables e con integral finita. A expresión (1.17) correspóndese coa densidade da suma de $X + \tilde{X}$. A noción de convolución de funcións pode extenderse á convolución de distribucións de probabilidade, e así a distribución de probabilidade da suma de dúas variables aleatorias independentes é a convolución de cada unha das súas distribucións de probabilidade conxuntas. Destaquemos que esta notación se pode xeneralizar ao caso multidimensional e, en particular, ao bidimensional, polo que será válida no Capítulo 2 simplemente reemprazando a integral en $\mathbb{R}$ pola correspondente integral en $\mathbb{R}^2$.

Dado que a densidade da suma de dúas variables aleatorias se pode expresar en termos de convolución das súas funcións de densidade e o estimador tipo núcleo é un promedio de n densidades -distribuídas segundo o núcleo K e o parámetro ventá h -, pódese considerar que o estimador da densidade tipo núcleo é a convolución da medida de probabilidade inducida polo núcleo reescalado K_h coa distribución empírica dos datos.

Co obxectivo de obter unha expresión do erro cadrático medio, reescribamos a media e a varianza do estimador tipo núcleo, en termos da función f e empregando a convolución de funcións:

$$\mathbb{E}[\hat{f}(x,h)] = \mathbb{E}[K_h(x) - X] = \int_{\mathbb{R}} K_h(x-\tilde{x}) f(\tilde{x}) d\tilde{x} = (K_H * f)(x) \text{ e}$$

$$\text{Var}[\hat{f}(x,h)] = \frac{1}{n}\text{Var}[K_h(x-X)] = \frac{1}{n}\int_{\mathbb{R}} K_h^2(x-\tilde{x}) f(\tilde{x}) d\tilde{x} - \frac{1}{n}(\int_{\mathbb{R}} K_h(x-\tilde{x}) f(\tilde{x}) d\tilde{x})^2 =$$

$$= \frac{1}{n}((K_h^2 * f)(x) - (K_h * f)^2(x)).$$

Obtemos que $\text{Nesgo}[\hat{f}(x,h)] = (K_H * f)(x) - f(x)$, polo que combinando esta expresión coa da varianza, reescribimos o erro cadrático medio dunha forma máis explícita:

$$MSE[\hat{f}(x,h)] = \frac{1}{n}((K_h^2 * f)(x) - (K_h * f)^2(x)) + ((K_h * f)(x) - f(x))^2. \tag{1.18}$$

Como a expresión (1.18) depende dun xeito complexo de h, pois involucra integrais de convolución, introduciremos algunha notación co obxectivo de calcular expansións de Taylor e reformulala. A finalidade é obter un resultado que nos garanta a existencia dunha expresión equiparable á do histograma e á do estimador Naive.

Notación 1.14. Sexa $g\colon \mathbb{R} \longrightarrow \mathbb{R}$ unha función integrable tal que $\int_{\mathbb{R}} x^2 g(x) dx < \infty$. Entón:

$$\mu_2(g) = \int_{\mathbb{R}} x^2 g(x) dx.$$

A expresión anterior pódese extender a máis dunha variable inmediatamente, polo que tamén será válida no caso bidimensional simplemente reemprazando o espazo de integración.

Notación 1.15. Sexa $g\colon \mathbb{R} \longrightarrow \mathbb{R}$ unha función cadrado integrable. Entón:

$$R(g) = \int_{\mathbb{R}} g^2(x) dx.$$

No caso multidimensional tamén empregaremos esta notación sen máis que considerar integrais múltiples en lugar de simples, xustificando así o seu uso no vindeiro capítulo.

Proposición 1.16. *Sexa f unha función de densidade limitada e $x \in \mathbb{R}$ tal que, nun entorno de x, f é dúas veces derivable, cadrado integrable, con f' e f'' continuas e f'''*

limitada. Supoñamos que o núcleo K é unha función de densidade simétrica entorno á orixe, cadrado integrable e con momento de segundo e terceiro orde finito. Ademais, esixamos que $\lim_{n\to\infty} h_n = 0$ *e que* $\lim_{n\to\infty} nh_n = \infty$*. Entón* $\hat{f}$ *é un estimador asintóticamente inesgado e consistente para estimar* f*.*

Demostración. Comecemos calculando a media de $\hat{f}$. Empregando o Teorema 1.5, obtemos que $f(x-hz) = f(x) - hzf'(x) + \frac{1}{2}h^2z^2f''(x) + o(h^2)$. A continuación realizaremos un cambio de variable $y = x - hz$ e operemos:

$$\mathbb{E}[\hat{f}(x,h)] = \mathbb{E}[K_h(x-X)] = \int_{\mathbb{R}} K_h(x-y)f(y)dy = \int_{\mathbb{R}} K_h(x-x+hz)f(x-hz)(hdz) =$$
$$= \int_{\mathbb{R}} hK_h(hz)f(x-hz)dz = \int_{\mathbb{R}} K(z)f(x-hz)dz = \int_{\mathbb{R}} K(z)[f(x) - hzf'(x) + \frac{1}{2}h^2z^2f''(x) + o(h^2)]dz =$$
$$= f(x) + \frac{1}{2}h^2f''(x)\int_{\mathbb{R}} z^2K(z)dz + o(h^2) = f(x) + \frac{1}{2}h^2f''(x)\mu_2(K) + o(h^2),$$

empregando que:

- $\int_{\mathbb{R}} K(z)dz = 1$, por ser K unha función de densidade.
- $\int_{\mathbb{R}} zK(z)dz = 0$, por ser K simétrica e así, o integrando unha función impar.
- $\int_{\mathbb{R}} z^2K(z)dz = \mu_2(K) < \infty,$ por ser o momento de segundo orde de K finito.

Falta comprobar que f cumpre algunha condición de regularidade e /ou limitación para asegurarnos de que a integral conmute co límite na igualdade

$$\int_{\mathbb{R}} K(z)[f(x) - hzf'(x) + \frac{1}{2}h^2z^2f''(x) + o(h^2)]dz = f(x) + \frac{1}{2}h^2f''(x)\int_{\mathbb{R}} z^2K(z)dz + o(h^2),$$

e en concreto, na igualdade $\int_{\mathbb{R}} o(h^2)dz = o(h^2)$, porque en realidade $o(h^2)$ é función de z por ser o termo de erro do desarrollo de Taylor de grao dous de $f(x-hz)$ entorno a x. En concreto, polo Teorema 1.5,

$$o(h^2) = -\frac{1}{3!}f'''(c_{x,x-hz})h^3z^3, \text{ con } c_{x,x-hz} \in L[x, x-hz]^{\text{o}}.$$

En virtude do Teorema 1.10, é suficiente aplicar as hipóteses de que f''' sexa limita e núcleo teña momento de terceiro orde finito, é dicir, que $\int_{\mathbb{R}} z^3K(z) < \infty$, para que a súa integral conmute co límite. Dado que $\lim_{n\to\infty} h_n = 0$, obtemos que $\hat{f}$ é asintóticamente inesgado:

$$\text{Nesgo}[\hat{f}(x,h)] = \frac{1}{2}h^2f''(x)\mu_2(K) + o(h^2).$$

Calculemos agora a varianza de $\hat{f}$ co obxectivo de obter que é un estimador consistente de f e finalizar a proba da proposición. Como consecuencia da continuidade de f temos

que $f(x-hz) = f(x) + o(1)$. Realizando o mesmo cambio de variable que no cálculo da media e operando, obtemos que

$$\begin{aligned}
&\mathrm{Var}[\hat{f}(x,h)] = \frac{1}{n}\mathrm{Var}[K_h(x-X)] = \frac{1}{n}\int_{\mathbb{R}} K_h^2(x-y)f(y)dy - (\int_{\mathbb{R}} K_h(x-y)f(y)dy)^2 = \\
&= \frac{1}{nh}\int_{\mathbb{R}} K^2(z)f(x-hz)dz - \frac{1}{n}(\mathbb{E}[\hat{f}(x,h)])^2 = \\
&= \frac{1}{nh}\int_{\mathbb{R}} K^2(z)f(x-hz)dz - \frac{1}{n}(\int_{\mathbb{R}} K(z)f(x-hz)dz)^2 = \\
&= \frac{1}{nh}\int_{\mathbb{R}} K^2(z)\{f(x)+o(1)\}dz - \frac{1}{n}(\int_{\mathbb{R}} K(z)\{f(x)+o(1)\}dz)^2 = \\
&= \frac{f(x)}{nh}\int_{\mathbb{R}} K^2(z)dz + o(\frac{1}{nh}) - \frac{1}{n}\{f(x)+o(1)\}^2 = \frac{f(x)}{nh}\int_{\mathbb{R}} K^2(z)dz + o(\frac{1}{nh}),
\end{aligned}$$

empregando que:

- $\int_{\mathbb{R}} K(z)dz = 1$, por ser K unha función de densidade.
- $\int_{\mathbb{R}} K^2(z)dz < \infty$, por ser K unha función cadrado integrable.

Ademais nas, igualdades $\int_{\mathbb{R}} K(z)o(1)dz = o(1)\int_{\mathbb{R}} K(z)dz$ e $\int_{\mathbb{R}} K^2(z)o(1)dz = o(1)\int_{\mathbb{R}} K^2(z)dz$ debemos ter en conta que o termo $o(1)$ procede do desarrollo de Taylor de $f(x-hz) = f(x)+o(1)$ polo que é función da variable z. Se empregamos que f é unha función limitada, o Teorema 1.10 garante que se cumpre a igualdade das expresións posto que:

$$\lim_{h\to 0} K(z)f(x-hz) = K(z)f(x),$$

e se f está limitada por unha constante $\eta \in \mathbb{R}^+, |K(z)f(x-hz)| < \eta|K(z)| = \eta K(z)$.

Tendo en conta a notación exposta antes desta proposición obtemos que

$$\mathrm{Var}[\hat{f}(x,h)] = \frac{1}{nh}f(x)R(K) + o(\frac{1}{nh}), \text{ e empregando que } \lim_{n\to\infty} h_n = 0 \text{ e } \lim_{n\to\infty} nh = \infty$$

resulta que $\hat{f}$ é un estimador consistente de f.

□

A expresión do erro cadrático medio do estimador tipo núcleo vén dada por:

$$MSE[\hat{f}(x,h)] = \frac{1}{4}h^4(f''(x))^2\mu_2^2(K) + \frac{1}{nh}f(x)R(K) + o(\frac{1}{nh} + h^4).$$

Observemos que cando o valor absoluto de f'' é grande, o nesgo é grande. Isto ocorre en zonas onde a curvatura de f é moi marcada. Como un dos termos do erro cadrático medio é o cadrado do nesgo, estas zonas serán zonas de erro cadrático medio grande e así,

zonas onde a estimación será pouco precisa. Entón, veciñanzas de picos e vales, serán rexións onde a estimación tipo núcleo será máis imprecisa, con respecto a zonas de curvatura menor ou incluso nula.

Na Figura 1.9 ilústrase esta conclusión teórica a un caso particular. Consideramos a mestura de dúas distribucións normais dada polo modelo M4 do Apéndice B. A súa función de densidade real presenta un mínimo na orixe e dous máximos, un en $x = -1$ e outro en $x = 1$. Entornos destes tres puntos son rexións onde f presenta moita curvatura; é claro que $f''(0) > 0$, por ser $x = 0$ o mínimo de f, e que $f''(1) < 0$ e $f''(-1) < 0$, por ser estes os puntos máximos de f. En vermello figura a función de densidade; en gris representamos diferentes estimacións tipo núcleo, con núcleo normal estándar, para tres valores de h e para 20 mostras de tamaño 100 distintas, e en negro a media das 20 estimacións anteriores, para os tres valores de h distintos. Como indica a lenda, de esquerda a dereita, $h = 0{,}65$, $h = 0{,}25$ e $h = 0{,}1$. O nesgo é a diferenza entre a media do estimador e a cantidade poboacional, polo que neste caso, unha estimación do nesgo en cada punto vén dada pola distancia da curva vermella (gráfica de f) e a curva negra (gráfica da estimación promedio).

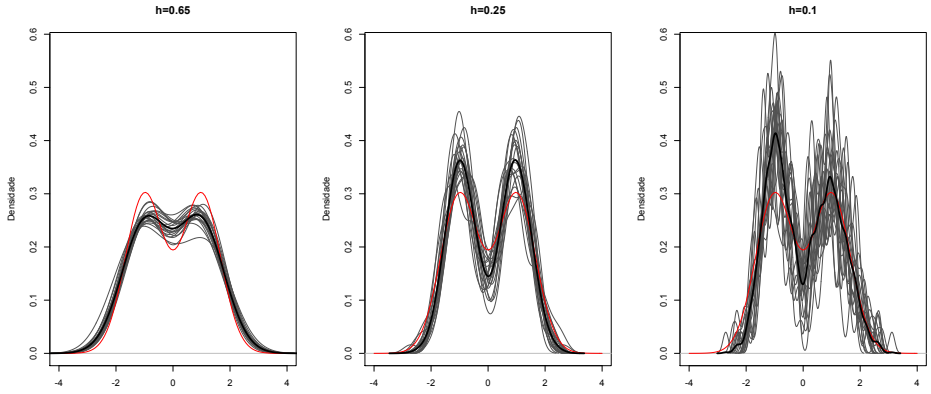

Figura 1.9: En vermello represéntase a gráfica da función de densidade a estimar; en gris, as estimacións tipo núcleo con núcleo normal estándar para 20 m.a.s desta densidade, e en negro a súa estimación promedio.

Cando a ventá é máis pequena o nesgo diminúe pero a varianza aumenta demasiado (gráfica da dereita). Considerando unha ventá maior, as estimación son moi similares, pero aloxadas en promedio da curva a estimar (gráfica da esquerda). Por isto último, o obxectivo é obter un valor de h axeitado que equilibre a cantidade de varianza e nesgo.

Nestas gráficas tamén podemos apreciar outro inconveniente da estimación tipo núcleo: dado que o parámetro ventá h é fixo ao longo de toda a recta real e a densidade de datos varía dunhas rexións a outras, hai tendencia a presentar distorsións -e así resultados peores- nas colas da estimación.

Tanto na Figura 1.9 como na Figura 1.10 o tamaño mostral é $n = 100$. Nesta última, apréciase de novo o problema que ten este tipo de estimación para aproximar correctamente os valores onde a curvatura da función orixinal é marcada. Consideramos que f segue unha distribución expoñencial de parámetro $\lambda = \frac{1}{2}$. Así engádese o problema de que sexa descontinua na orixe, punto de máxima curvatura, polo que non nos atopamos nas condicións da Proposición 1.16. Observamos como en entornos de $x = 0$ a estimación presenta moito nesgo e, en consecuencia, un erro cadrático medio elevado. Xa vimos anteriormente que este problema é común ao histograma e o estimador Naive. Tamén se aprecian as distorsións da estimación na cola dereita da gráfica, onde a densidade de datos diminúe polo que a calidade da estimación empeora.

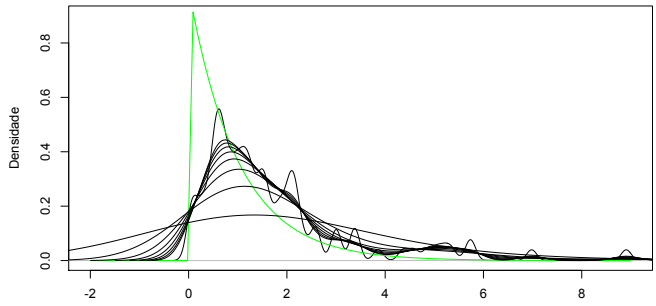

Figura 1.10: Nove estimacións tipo núcleo con núcleo normal estándar dunha mesma m.a.s dunha Exp($\frac{1}{2}$), para $h \in \{\frac{1}{10}\} \cup \{\frac{2}{i} : i = 1, ..., 8\}$, de cor negro. En verde a función de densidade teórica.

Ao longo de todo este apartado de criterios de erro do estimador tipo núcleo, a función de densidade real era continua. Sen embargo, a función de densidade da Exp($\frac{1}{2}$) presenta unha descontinuidade na orixe. Na gráfica apreciamos como o estimador tipo núcleo de f preto da descontinuidade non estima ben a función orixinal. Para valores de x positivos e próximos a cero pola dereita, o estimador debe estimar densidades relativamente altas (preto de 1), mentres que para x negativos e próximos a cero pola esquerda, o estimador desexa estimar unha densidade nula. Consecuentemente, o nesgo da estimación en entornos

da orixe é bastante elevado.

Volvendo ao análise en media, varianza e erro do estimador tipo núcleo, temos que:

$$\text{Nesgo}[\hat{f}(x)] = \frac{1}{2}h^2 f''(x)\mu_2(K) + o(h^2)$$

polo que, engadindo ás condicións expostas na Proposición 1.16 que f'' sexa cadrado integrable, o erro cadrático medio integral adopta a seguinte expresión:

$$MISE[\hat{f}(\cdot)] = \frac{h^4}{4}\mu_2^2(K)R(f'') + \frac{1}{nh}R(K) + o(\frac{1}{nh} + h^4), \tag{1.19}$$

empregando que f é unha función de densidade e que $\mu_2^2(K), R(K) \in \mathbb{R}$ son valores fixos, e non funcións de x.

Dado que a expresión do $MISE$ non depende de h dun xeito coñecido, pois o termo de erro $o(\frac{1}{nh} + h^4)$ é función de h e é un dos sumandos na expresión (1.19), definamos un erro asintótico cuxa dependencia de h sexa máis simple. Con este obxectivo, concretemos a que nos estamos a referir co termo " asintótico".

Definición 1.17. O análise asintótico é un método de descrición do comportamento no límite. Por iso mesmo, un erro asintótico dun estimador da densidade é un erro que comete dito estimador ao aproximar a función de densidade descoñecida, cando o tamaño mostral n tende a infinito.

Baixo as condicións impostas ata o de agora, obtemos que

$$AMISE[\hat{f}(\cdot)] = \frac{h^4}{4}\mu_2^2(K)R(f'') + \frac{1}{nh}R(K) \tag{1.20}$$

polo que a súa derivada vén dada por

$$h^3\mu_2^2(K)R(f'') - \frac{1}{nh^2}R(K)$$

e así o mínimo do $AMISE$, é dicir, o h_{AMISE}, é

$$h_{AMISE} = (\frac{R(K)}{\mu_2^2(K)R(f'')})^{\frac{1}{5}} n^{\frac{-1}{5}}, \tag{1.21}$$

sempre e cando $R(f'') \neq 0$, ao igual que ocorría co estimador Naive. Do mesmo xeito, se $R(f'') = 0$, non existe unha expresión do h_{AMISE} do estimador tipo núcleo polo que, para calcular un valor de h axeitado para a estimación debemos recorrer a criterios de minimización baseados no $MISE$, como veremos ao final do capítulo.

Substituíndo na fórmula (1.20) obtemos a expresión $\frac{5}{4}[\mu_2^2(K)R^4(K)R(f'')]^{\frac{1}{5}}n^{\frac{-4}{5}}$ polo que $\hat{f}$ é un estimador de f da orde de $\mathcal{O}(n^{\frac{-4}{5}})$. En termos de orde de converxencia, danse as seguintes desigualdades:

$$\mathcal{O}(n^{\frac{-2}{3}}) \leqslant \mathcal{O}(n^{\frac{-4}{5}}) = \mathcal{O}(n^{\frac{-4}{5}}) \leqslant \mathcal{O}(n^{-1}),$$

onde a primeira orde corresponde ao histograma, as dúas intermedias ao estimador Naive e ao estimador tipo núcleo, e a maior á estimación paramétrica da función de densidade supoñendo que os datos seguen un modelo normal, é dicir, que proceden dunha $N(\mu, \sigma^2)$ e debemos estimar μ e σ^2, por exemplo empregando o método de máxima verosimilitude. Deste modo, en termos de orde de converxencia, é mellor a estimación paramétrica pero, como xa dixemos, a falta de flexibilidade desta pode dar lugar a obter estimacións erróneas en moitos casos.

Unha vez resaltado o efecto do parámetro ventá h e vendo a expresión do $AMISE[\hat{f}(\cdot)]$ da fórmula (1.20), é o momento de estudar cal é o efecto da función núcleo K, dado que é o único factor que se pode optimizar da expresión (1.20), por ser $R(f'')$ un valor constante que depende da función descoñecida f e que, polo tanto, non podemos modificar. No que segue deste capítulo omitiremos gran parte dos cálculos. Para máis detalles ver o libro Wand and Jones (1995, Sección 2.8).

Ao igual que coa elección do parámetro ventá, o obxectivo é minimizar a expresión (1.20). Con este fin, impoñamos as seguintes condicións sobre K, que xa consideráramos na formulación da Proposición 1.16 :

$$K \geqslant 0, \int_{\mathbb{R}} K(x)dx = 1, \int_{\mathbb{R}} xK(x)dx = 0 \text{ e que } \int_{\mathbb{R}} x^2K(x)dx = \mu_2^2(K) \in (0, \infty).$$

Tamén é habitual impoñer que K sexa unimodal. Como xa vimos, substituíndo o h_{AMISE} na expresión do $AMISE$ obtemos o $AMISE$ óptimo. Considerando únicamente os termos que dependen de K desa expresión, definimos $C(K) := (\mu_2^2(K)R^4(K))^{\frac{1}{5}}$. A función $C(K)$ é invariante fronte e a transformacións do tipo $K_\delta(\cdot) = \frac{1}{\delta}K(\frac{\cdot}{\delta})$, con $\delta > 0$, e a función núcleo óptima é o núcleo Epanechnikov, definida así por ser o matemático Epanechnikov (1969) o primeiro en usala no contexto da regresión non paramétrica.

Na esquerda da Figura 1.11 represéntanse algunhas das funcións núcleo máis comúns reescaladas, é dicir, empregando un sistema de referencia común, incluíndo entre elas a

normal estándar empregada anteriormente e a Epanechnikov. Na lenda da gráfica aparecen os seus nomes xunto coa cor empregada para a súa representación. O parámetro ventá elixido é $h = 1$ para todos os núcleos.
A expresión analítica dos núcleos representados é:

- Núcleo Epanechnikov: $K^E(x) = \frac{3(1-x^2)}{4},\ x \in (-1,1)$.
- Núcleo normal ou gaussiano: $K(x) = \frac{1}{\sqrt{2\pi}} \exp -\frac{x^2}{2},\ x \in \mathbb{R}$.
- Núcleo triangular: $K(x) = 1 - |x|,\ x \in (-1,1)$.
- Núcleo uniforme ou rectangular: $K(x) = \frac{1}{2},\ x \in (-1,1]$.
- Núcleo biweight: $K(x) = \frac{15(1-x^2)^2}{16},\ x \in (-1,1)$.
- Núcleo triweight: $K(x) = \frac{35(1-x^2)^3}{32},\ x \in (-1,1)$.

Na dereita da Figura 1.11 represéntase, para unha mostra aleatoria simple dunha $N(0,1)$ de tamaño mostral $n = 100$, a estimación tipo núcleo resultante, tomando en todos os casos un parámetro ventá común $h = 0{,}15$ e os núcleos anteriormente formulados. A lenda indica o cor da curva estimada en función do núcleo empregado.

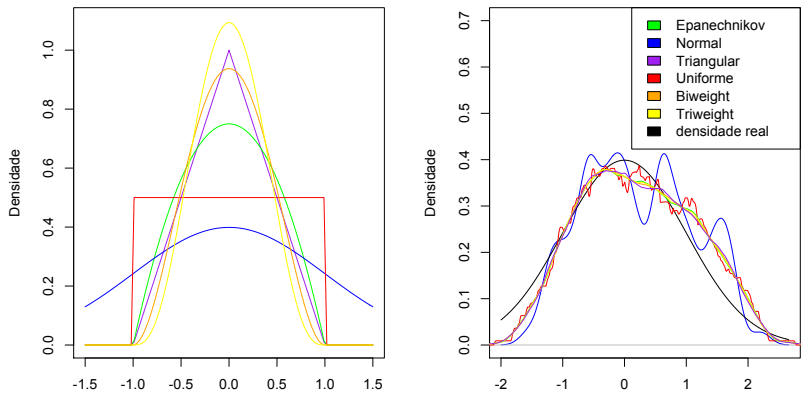

Figura 1.11: A esquerda, as curvas dalgunhas das funcións núcleo univariantes máis comúns, reescaladas para $h = 1$ e a dereita, as gráficas da estimación tipo núcleo empregando seis núcleos diferentes. Emprégase unha m.a.s de tamaño $n = 100$ dunha $N(0,1)$ e, en todos os casos, un parámetro ventá $h = 0{,}15$.

Denotemos por K^E a función núcleo Epanechnikov, cuxa expresión analítica vén dada por $K^E(x) = \frac{3(1-x^2)}{4}\mathbb{I}(x \in (-1,1))$. Co obxectivo de comparar o $AMISE$ de dúas estimacións tipo núcleo dunha mesma densidade f e para un mesmo parámetro ventá h, pero para dúas funcións núcleo K e K^E, introdúcese o concepto de eficiencia.

Defínese a eficiencia de K relativa a K^E como $\text{ef}(K) = (\frac{C(K)}{C(K^E)})^{\frac{5}{4}} = \frac{n_K}{n_K^E}$. É dicir, a eficiencia representa o cociente entre os tamaños mostrais necesarios para o obter o mesmo $AMISE$ empregando o núcleo K -necesítase n_K,- e o núcleo K^E -necesítase n_K^E-.

Na Táboa 1.1 móstranse os valores da eficiencia para os núcleos definidos na anterior lista. É claro que o núcleo Epanechnikov é o máis eficiente (a súa eficiencia vale 1). Sen embargo existen núcleos "case" óptimos en canto a eficiencia, como son o Triweight, o Biweight ou incluso o normal. Ademais, dado que o núcleo Epanechnikov proporciona unha función estimada con derivada primeira descontinua, en moitas ocasións é preferible empregar como función núcleo algunha con máis propiedades de regularidade, como ocorre coa densidade normal dado que é unha función de clase $\mathcal{C}^\infty$. Por iso mesmo, en moitas ocasións se emprega como función núcleo a normal e non a Epanechnikov, a pesar de ser esta última máis eficiente.

Núcleo	**Eficiencia**
Epanechnikov	1.000
Normal ou gaussiano	0.951
Triangular	0.986
Uniforme ou rectangular	0.930
Biweight	0.994
Triweight	0.987

Cadro 1.1: Eficiencia dos núcleos máis comúns comparados co núcleo óptimo. Recordemos que a eficiencia dun núcleo K vén dada por $\text{ef}(K) = (\frac{C(K)}{C(K^E)})^{\frac{5}{4}}$.

1.4.2. Elección do parámetro ventá en R

Ata o de agora determinamos teoricamente cal é o parámetro ventá óptimo e a función núcleo óptima, ambos en termos do $AMISE$. Sen embargo, tamén empregamos exemplos gráficos ilustrativos do estimador tipo núcleo, na maioría dos casos con núcleo gaussiano, salvo cando empregamos o núcleo uniforme para obter o estimador Naive, pero non

falamos de como escoller h xa que as ventás óptimas son, en xeral, descoñecidas. É un tema bastante amplo que abordaremos en maior profundidade no Capítulo 2, adicado á estimación bidimensional, pero aquí comentaremos brevemente como selecciona a ventá o software libre R Core Team (2020).

En primeiro lugar, R é un entorno e linguaxe de programación con un enfoque á análise estatística. Trátase dunha das linguaxes de programación máis empregadas en investigación científica, dispoñible para a gran maioría de sistemas operativos. Unha das súas maiores vantaxes é a posibilidade de cargar diferentes bibliotecas ou paquetes con funcionalidades de cálculo e representación gráfica.

Empregaremos a función *density* da librería *stats*, librería con ferramentas estatísticas dispoñible de base en R, sen a necesidade de cargar ningún paquete. Por defecto, R escolle a opción *bw='nrd0'* onde *bw* é o parámetro ventá empregado- abreviatura do termo *bandwidch* en inglés- e *'nrd0'* implementa a Regra do Polgar de Silverman, exposta por primeira vez no libro Silvermann (1986, páxina 48, ecuación (3.31)), en inglés *Rule of Thumb*, para elixir dito parámetro ventá dun estimador tipo núcleo con núcleo gaussiano, polo que é unha regra baseada en distribución paramétricas. Sen embargo, nos imos empregar unha versión máis común do método de Silverman, formulada por Scott (1992), e que corresponde con elixir a opción *bw='nrd'*.

A expresión (1.21) do h_{AMISE} reflexa que o parámetro de suavizado óptimo depende do valor descoñecido $R(f'')$, que supoñeremos que é non nulo, polo que imos escoller un "valor piloto" inicial para h para estimar $R(f'')$ e logo usar a estimación formulada na expresión (1.21) para calcular o h_{AMISE}. Silverman propuxo asumir que f segue unha distribución paramétrica para calcular o valor inicial de h. A opción de R elixida emprega que f segue unha distribución $N(\mu, \sigma^2)$ e así,

$$R(f'') = \int_{\mathbb{R}} \phi_{\mu,\sigma^2}(x)^2 dx = \frac{1}{\sigma^2 2\pi} \int_{\mathbb{R}} \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right) dx = \frac{3}{8\sqrt{\pi}\sigma^5},$$

onde a última igualdade non é inmediata xa que o integrando non ten primitivas elementais. Salvo cambio de variable, no libro Nieto and Albés 2017, Sección 3.4 pódese ver como se realiza o cálculo da integral $\int_{\mathbb{R}} \exp\left(-x^2\right) dx$ e cal é o seu resultado.
Empregando un núcleo normal estándar, e pola expresión (1.21):

$$h_{pilot} = (4\pi)^{-\frac{1}{10}} [(\frac{3}{8\sqrt{\pi}})]^{-\frac{1}{5}} \sigma n^{-\frac{1}{5}} \approx 1{,}06\sigma n^{-\frac{1}{5}}, \tag{1.22}$$

que logo se emprega para obter o h_{AMISE}, de novo coa expresión (1.21).

Cómpre destacar que este método necesita calcular h_{pilot} e así estimar σ, por exemplo, coa desviación típica mostral $S_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{x})^2}$ onde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$ é a media mostral, e logo h_{AMISE}. Na práctica, unha simplificación consiste en tomar h_{pilot} como o parámetro ventá empregado na estimación tipo núcleo, é dicir, considerar que $h_{AMISE} = h_{pilot}$, saltándose o paso de calcular h_{AMISE}. Este novo método recibe o nome de Regra do Polgar con referencia normal, dado que o parámetro ventá se escolleu supoñendo que f segue unha distribución normal. Esta simplificación resulta axeitada se a poboación se asemella na súa distribución á da normal, pero se traballamos con poboacións multimodais prodúcese unha sobresuavización da estimación.

A gran diferenza entre a Regra do Polgar de Silvermann e a de Scott é a estimación da desviación típica. Defínese o rango intercuartílico estandarizado como

$$\hat{\sigma_{IQR}} = \frac{\text{Rango intercuartílico mostral}}{\Phi^{-1}(\frac{3}{4}) - \Phi^{-1}(\frac{1}{4})}, \text{ onde } \Phi \text{ denota a función de distribución da } N(0,1).$$

Silverman estima σ dunha forma máis robusta que Scott, co obxectivo de que este estimador non se vexa tan afectado pola presenza de datos atípicos ou polo exceso de peso das colas na estimación, tomando como $\hat{\sigma} = \text{mín}(S_n, \hat{\sigma_{IQR}}/1{,}349)$, logo de comprobar que o resultado desta aproximación proporciona parámetros ventás que funcionan relativamente ben con densidades unimodais e moderadamente bimodais, e considerando o mínimo para reducir o risco de sobresuavizar.

Existen outros métodos alternativos aos de Silvermann e Scott para a elección do parámetro ventá, como pode ser os métodos de validación cruzada nesgada ou inesgada, que expoñeremos en detalle na sección da estimación tipo núcleo bivariante do Capítulo 2, e que tamén están implementados en R sen máis que elixir a opción da función *density* *bw='bcv'* ou *bw='ucv'*, respectivamente. En dito capítulo formularanse para o caso bidimensional, e son facilmente extrapolables a unha dimensión, cunhas expresións máis sinxelas neste último caso. Adiantemos que os dous métodos se basean en calcular un valor de h que minimice un estimador de $MISE$ e de $AMISE$ respectivamente, obtendo resultados inesgados no primeiro dos casos -validación cruzada inesgada- e nesgados no segundo -validación cruzada nesgada-.

Co obxectivo de ilustrar as diferenzas orixinadas pola elección dos distintos parámetros ventá, na Figura 1.12 móstranse catro estimacións tipo núcleo distintas, empregando

como núcleo a distribución normal estándar. Na elección do parámetro ventá empregamos os catros métodos anteriormente citados: Regra do Polgar de Silverman (obtemos un $h = 0{,}3348$), Regra do Polgar de Scott (obtemos un $h = 0{,}3942$), validación cruzada nesgada (obtemos un $h = 0{,}1019$) e validación cruzada inesgada (obtemos $h = 0{,}1577$). Cómpre destacar que, en todos os casos, a estimación da densidade obtida é a da variable X_1 do conxunto de datos "$OldFaithful$" de R, que analizaremos no Capítulo 2 en detalle e que o único que necesitamos explicitar del, para este exemplo, é que o tamaño mostral é $n = 272$.

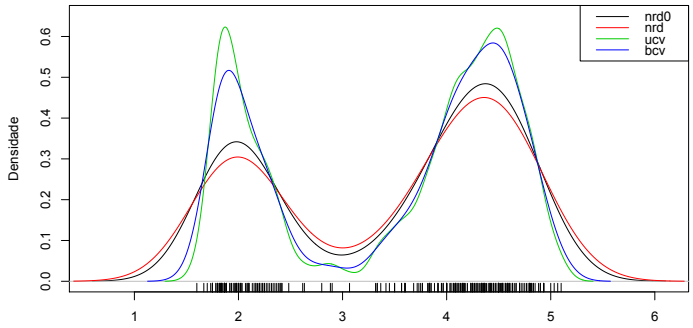

Figura 1.12: Catro estimacións tipo núcleo con núcleo gaussiano da función de densidade da variable X_1 do conxunto de datos "$OldFaithful$" de R. O valor de h en cada estimación obtívose do seguinte xeito: en negro coa Regra do Polgar de Silvermann, en vermello coa Regra do Polgar de Scott, en verde coa validación cruzada inesgada e en azul coa validación cruzada nesgada. As liñas verticais no eixe de abscisas correspóndense coa posición dos datos mostrais.

Neste caso, os métodos de validación cruzada proporcionan un h considerablemente menor que as Regras do Polgar (aproximadamente o triple destas últimas), sendo a validación cruzada nesgada a que obtén un h menor e a Regra do Polgar de Scott a que obtén un h maior. Apreciemos tamén que os valores de h máis pequenos dán lugar a estimacións con máis picos, en consecuencia de que se adaptan máis a mostra, e os valores de h maiores dán lugar a estimacións máis suaves pero, como vimos teoricamente, máis nesgadas.

Sen embargo, estes datos proveñen dunha función de densidade "fácil" de analizar, pois presentan dúas modas moi marcadas (no vindeiro capítulo veremos que ata no histograma se aprecian claramente) e o tamaño mostral é aceptable. Consideremos unha mostra

que proveña dunha distribución multimodal onde as modas non estean tan marcadas e coñezamos a función de densidade da que proveñen, para así comparar as estimacións coa realidade. Na Figura 1.13 represéntase catro estimacións tipo núcleo con núcleo normal dunha m.a.s de tamaño $n = 150$ procedente da mestura de distribucións normais dada polo modelo M5 do Apéndice B.

En cor celeste debúxase a función de densidade real e, como indica a lenda, nos outros catro colores as catro estimacións tipo núcleo empregando o parámetro ventá calculado segundo os correspondentes métodos. A densidade real presenta tres modas claramente distinguidas, pero isto pérdese nas estimacións tipo núcleo, dado que as catro estiman unha densidade lixeiramente bimodal, omitindo a moda correspondente a $x = 3$ e estimando escasamente as outras dúas. O valor do parámetro ventá obtido polos diferentes métodos é o seguinte: $h = 0{,}5714$ coa Regra do Polgar de Silverman, $h = 0{,}6730$ coa Regra do Polgar de Scott, $h = 0{,}4628$ coa validación cruzada nesgada e $h = 0{,}7235$ coa validación cruzada inesgada. Nos catro casos as estimacións son moi semellantes e nesgadas con respecto á curva teórica.

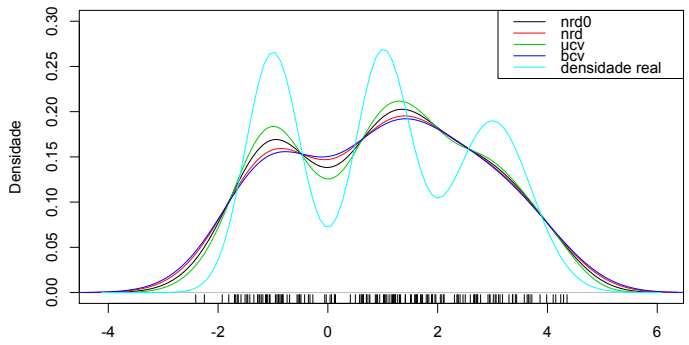

Figura 1.13: Catro estimacións tipo núcleo con núcleo gaussiano da función de densidade do modelo M5 do Apéndice B. O valor de h en cada estimación obtívose do seguinte xeito: en negro coa Regra do Polgar de Silvermann, en vermello coa Regra do Polgar de Scott, en verde coa validación cruzada inesgada e en azul coa validación cruzada nesgada. As liñas verticais no eixe de abscisas correspóndense coa posición dos datos mostrais e a liña celeste coa densidade real.

Unha vez máis, ilustramos como a estimación tipo núcleo ten dificultades en estimar correctamente as modas e os vales da función orixinal dado que son zonas de moita cur-

vatura. Unha axuda a este problema é aumentar o tamaño mostral, de ser posible. Unha densidade con pouca curvatura dará lugar a un valor de $R(f'')$ pequeno e, en consecuencia, un parámetro ventá grande; do mesmo modo, valores de $R(f'')$ grandes corresponderán a parámetros ventá pequenos.

No caso de que $R(f'') = 0$, excluído ata o de agora e que ocorre por exemplo na distribución uniforme, a expresión (1.20) do $AMISE$ redúcese a: $\frac{R(K)}{nh}$, que non depende da función f. Remontándonos a expresión (1.19) do $MISE$, obtemos a expresión

$$\frac{R(K)}{nh} + o(\frac{1}{nh} + h^4) \text{ onde o termo de erro é función de } f.$$

Poderíamos usar algún criterio de minimización do $MISE$, como a validación cruzada inesgada que formularemos no Capítulo 2. Cómpre destacar que unha clase de densidades moi sinxela e verificando que $R(f'') = 0$ son as funcións definidas a trozos (non negativas e integrando 1 en $\mathbb{R}$) como unión de rectas e funcións constantes. Sen embargo, ao igual que ocorre coa distribución uniforme, estas estimacións presentar problemas na unión dos trozos.

Ao longo deste capítulo abordamos a estimación non paramétrica da densidade univariante empregando o estimador histograma e o estimador tipo núcleo, que inclúe ao estimador Naive. Tanto os estimadores estudados como o seu comportamento serán de gran axuda para abordar a estimación da densidade no plano, dado que esta última é a extensión a $d = 2$ dimensións da estimación unidimensional. Reciprocamente, veremos que dado un conxunto de datos bidimensionais, en moitos casos, é interesante estudar o comportamento marxinal de cada unha das variables, volvendo así ao caso unidimensional. Sen embargo, no plano a flexibilidade dos estimadores aumenta aínda máis que na recta pois, como xa veremos, podemos considerar a orientación que mellor se axuste aos nosos datos.

Capítulo 2

Estimación da densidade bidimensional

No Capítulo 1 abordábamos a estimación non paramétrica da densidade no caso unidimensional co obxectivo de, posteriormente, extender os conceptos e ideas ao caso bidimensional. Neste capítulo é o momento de estudar a estimación non paramétrica da densidade en dúas dimensións. Igual que no caso unidimensional, comezaremos co estimador histograma para logo traballar co estimador tipo núcleo e ilustraremos estes estimadores empregando os datos de "$OldFaithful$". Ademais tamén analizaremos en profundidade os distintos tipos de erro que se poden cometer na estimación tipo núcleo, así como algún criterio de selección dos parámetros de suavizado. Por último, presentaremos brevemente algunha das dificultades que aparecen cando abordamos a estimación da densidade multivariante no contexto xeral para un número de variables elevado.

Ata o de agora vimos que a estimación da densidade na recta real nos permite analizar como é o comportamento estimado dun mostra en canto á concentración de datos da mesma. Deste xeito, zonas de alta concentración de datos, eran zonas de alta densidade e zonas de baixa concentración, zonas de baixa densidade. Intuitivamente isto xeneralízase ao plano, onde os datos poden presentar unha determinada orientación, ligada á dependencia dunha variable coa outra. Adiantando acontecementos, ambas variables do conxunto de datos "$OldFaithful$" miden períodos de tempo, unha delas entre erupcións e a outra da duración das mesmas, polo que é razoable esperar que exista unha cerca relación entre ambas, que se traduce nunha correlación positiva.

Así, a estimación da densidade no plano permítenos estimar as zonas de alta concen-

tración de datos dunha mostra tendo en conta a relación entre as variables. Esta análise conxunta proporciona resultados moi distintos a realizar a estimación de cada unha das dúas variables por separado e logo extraer conclusións, porque este último razoamento esquece a dependencia das mesmas. Tendo en conta o exemplo que imos analizar no Capítulo 3 sobre a distribución dos niños de avespa velutina no mapa de Galicia, a estimación da densidade no plano permite analizar a distribución espacial de, por exemplo, unha especie, como ocorre neste caso coas velutinas. Un tema moi actual, pero que non imos abordar neste traballo, e cuxo estudo é similar ao que expoñeremos no Capítulo 3, é a evolución do coronavirus no mapa mundial, de modo que a estimación da densidade de casos axudaría a paliar as consecuencias deste problema de saúde pública e a respectar a liberdade de movementos, dando unha medida de risco á poboación e ás autoridades sanitarias dunhas zonas fronte a outras, permitindo así actuar de xeito consecuente en cada unha delas.

Nas vindeiras seccións deste capítulo consideraremos un vector aleatorio bidimensional (X, Y) con densidade descoñecida f e supoñeremos que observamos unha mostra aleatoria simple del, que denotaremos por $(X_1, Y_1), ..., (X_n, Y_n) \in \mathbb{R}^2$. O obxectivo será estimar a densidade f co estimador histograma e co estimador tipo núcleo, así como comparar estas estimacións por medio dos diferentes erros presentados xa no Capítulo 1 para o caso unidimensional. Salvo que se indique o contrario, sexa $(x, y) \in \mathbb{R}^2$ arbitrario.

2.1. Histograma bidimensional

Comecemos esta sección mostrando un exemplo de cal é a utilidade do histograma bidimensional empregando un conxunto de datos reais para, posteriormente, proceder a formalizar a súa expresión matemática. Consideremos os datos "$OldFaithful$" que conteñen información sobre a duración das erupcións, que denotaremos por X_1, e o tempo de espera entre unha erupción e outra, que denotaremos por X_2, do geyser "$OldFaithful$", no Parque Nacional de Yellowstone, en Wyoming, nos Estados Unidos. A mostra tomouse cunha medición continuada entre o 1 de agosto ata o 15 de agosto do ano 1985. Os datos extraéronse de R, estando a fonte de recollida orixinal no libro Azzalini. and Bowman (1990, páxinas 357–365).

Trátase dunha mostra aleatoria bidimensional de tamaño $n = 272$. Ambas variables se miden en minutos. O obxectivo é estimar a función de densidade conxunta da que proceden estes datos. Na Figura 2.1 represéntase o diagrama de dispersión da mostra. Claramente os datos están orientados positivamente e existe unha relación de dependencia entre ambos. A

súa distribución é moi variable, claramente bimodal e, como se recolle no libro Shaughnessy and Pfannkuch (2002, páxinas 252-259), se pretendemos coñecer os tempos de espera en función da duración da anterior erupción, a predición carecería de fiabilidade. Poderíamos esperar entre 49 e 58 minutos ou 89 e 102 minutos. Sen embargo, se temos en conta que hai dous grupos, un deles formado polos datos procedentes de tempos de espera pequenos e duracións curtas e o outro polos datos procedentes de tempos de espera grandes e duracións longas, a predición é moi fiel á realidade.

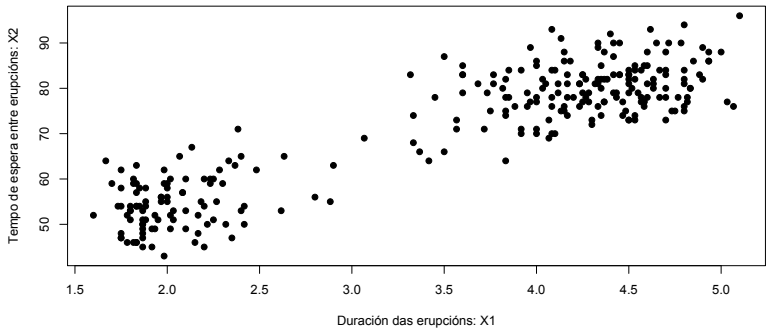

Figura 2.1: Diagrama de dispersión das dúas variables do conxunto de datos "$OldFaihful$". A variable X_1 recolle a duración das erupcións do geyser e a variable X_2 o tempo de espera entre unha erupción e outra. A medida de ambas son os minutos.

Podemos interpretar a bimodalidade dos datos como segue: o máis frecuente son tempos de espera pequenos e duración das erupcións curta (X_1 e X_2 toman valores pequenos, nas súas respectivas escalas), e tempos de espera grandes e duración das erupcións longa (X_1 e X_2 toman valores grandes, nas súas respectivas escalas). Como consecuencia, os datos concéntranse nos extremos superior-dereito e inferior-esquerdo do diagrama de dispersión da Figura 2.1. Deste xeito, é esperable que a duración do geyser sexa curta se agardamos pouco tempo para vela dende a anterior erupción, e longa en caso contrario.

Como veremos, o histograma bidimensional é unha xeneralización a dúas variables do caso unidimensional formulado no Capítulo 1. Para ter unha idea de como é a distribución marxinal de cada unha das variables podemos realizar unha análise individual de cada unha delas, representando para distintos valores de h, catro histogramas, dous para cada variable, como se mostra na Figura 2.2. Deste xeito, podemos obter unha idea previa de como se comportan X_1 e X_2 por individual. A escala das variables é moi diferente dado que X_1 toma valores entorno ao intervalo $[0, 6]$ e X_2 entorno a $[40, 100]$. Sen embargo,

ambas variables presentan un comportamento bimodal.

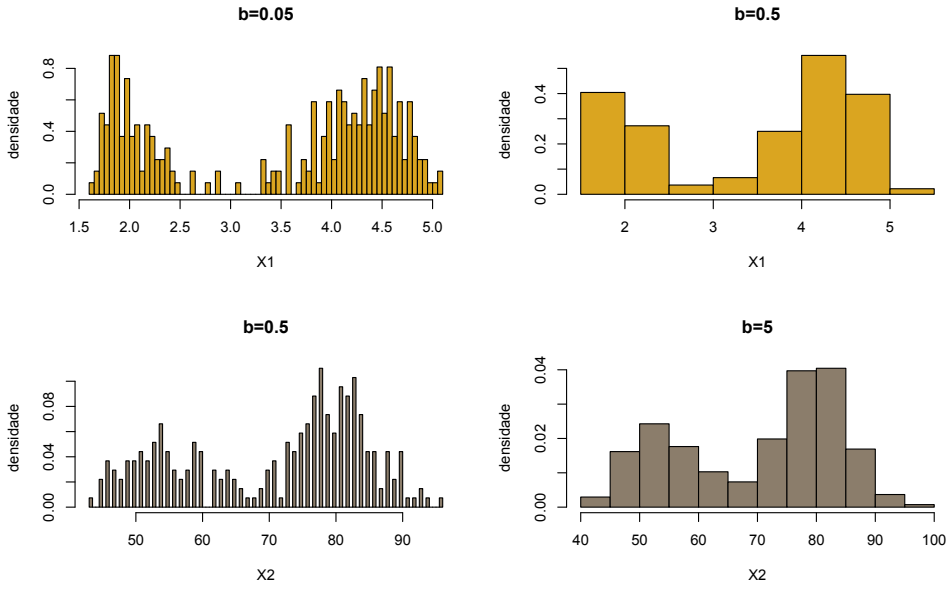

Figura 2.2: Na parte superior móstranse dous histogramas unidimensionais de X_1, e na inferior de X_2, para catro valores distintos de h. Ambas variables proceden do conxunto de datos "$OldFaithful$"; X_1 mide a duración das erupcións do geyser e X_2 o tempo de espera entre erupcións, as dúas en minutos.

Chegados a este punto, unha cuestión de interese sería preguntarnos como obter unha posible representación da distribución conxunta dos datos. Unha solución a esta cuestión é o histograma bidimensional, que nos proporciona unha representación en $\mathbb{R}^3$ de como se comportan, en termos de densidade e concentración de datos, as dúas variables en conxunto.

Supoñamos que discretizamos o plano en rectángulos disxuntos $B_0, B_1, B_2... \subset \mathbb{R}^2$ cada un cun ancho b_x na compoñente x e b_y na compoñente y. Podemos considerar calquera orientación do plano para estes rectángulos. Por exemplo, poden ser paralelos aos eixes coordenados e así simplificar a súa construción pero, en xeral, esta perda de xeneralidade pode dar lugar a unha peor estimación. A orientación dos rectángulos é un problema que xorde no caso multidimensional, a diferencia do unidimensional.

A posibilidade de orientar os rectángulos pode modelarse cun terceiro parámetro que, por exemplo, correspondería co ángulo que o lado do rectángulos asociado á lonxitude b_x forma co eixe de abscisas (eixe OX). Unha posible forma de incluír a orientación como pa-

rámetro sería denotar por θ este ángulo e definir unha matriz de xiros (matriz de rotación) do seguinte xeito:

$R(\theta) = \begin{pmatrix} \cos(\theta) & -\operatorname{sen}(\theta) \\ \operatorname{sen}(\theta) & \cos(\theta) \end{pmatrix} \in M_{2x2}$, que representa a rotación de ángulo θ graos do plano no sentido antihorario. Posteriormente deberíamos incluír esta transformación na expresión do estimador histograma sen máis que considerar a partición paralela ao eixe OX, aplicar o xiro a cada un dos rectángulos $B_0, B_1, B_2...$, e así obter a partición xirada θ graos no sentido antihorario.

Este novo parámetro, xunto co punto de anclaxe e a lonxitude dos lados dos rectángulos, aumenta o número de posibles formas que pode tomar o histograma, así como a súa complexidade. Xorden as seguintes cuestións: dado un conxunto de datos bidimensionais, cal é o tamaño dos rectángulos máis axeitado? e a súa orientación? Para simplificar a exposición deste traballo omitiremos no que segue o problema da orientación dos rectángulos e centrarémonos na elección do seu tamaño.

Consideraremos rectángulos paralelos aos eixos coordenados polo que para cada $j \in \mathbb{Z}$, $B_j = (t_{jx}, t_{(j+1)x}] \times (t_{jy}, t_{(j+1)y}]$ e a orixe (t_{0x}, t_{0y}) do histograma, que corresponde co punto de partida dos rectángulos empregados para discretizar $\mathbb{R}^2$, é o punto de $\mathbb{R}^2$ cuxas coordenadas corresponden cos extremos esquerdos dos intervalos do seguinte produto: $B_0 = (t_{0x}, t_{1x}] \times (t_{0y}, t_{1y}]$. Analogamente a como influía na forma do histograma unidimensional, o punto de anclaxe tamén inflúe na forma do histograma bidimensional.

Recordemos que $\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{(X_i, Y_i) \in B_j\}$ é a proporción de datos mostrais que caen no rectángulo B_j, para cada $j \in \mathbb{Z}$ e que $b_x b_y$ é a súa área. Deste xeito, e por analoxía ao caso unidimensional, a función estimada mediante o histograma en cada punto vén dada por:

$$\hat{f}_{hist}((x,y),(b_x,b_y)) = \frac{\sum_{i=1}^{n} \mathbb{I}_{\{(X_i,Y_i)\in B_j\}}}{n b_x b_y}, \ \forall (x,y) \in B_j \text{ e para cada } j \in \mathbb{Z}. \tag{2.1}$$

É unha función a trozos, constante en cada rectángulo B_j. Os parámetros a elixir son a lonxitude de cada lado dos rectángulos, que se determinará en función do tamaño mostral, e o punto de anclaxe. Este último aparece indirectamente na expresión (2.1) dado que as súas coordenadas determinan, xunto con b_x e b_y, o rectángulo B_0 e así, todos os demais.

Unha vez definido o estimador histograma $\hat{f}_{hist}$ é o momento de analizar conxuntamente a distribución dos datos de "$OldFaithful$". Na Figura 2.3 represéntanse dous histogramas bidimensionais distintos para este conxunto de datos, na esquerda elixindo una lonxitude

dos rectángulos menor e na dereita maior, correspondente a agrupar os datos en rectángulos da mesma lonxitude, de tal modo que o espazo mostral quedase dividido en 20 e 5 subintervalos no eixe de abscisas e outros 20 e 5 no de ordeadas, creando unha malla 20×20 e 5×5, respectivamente.
Considéranse en todos os casos rectángulos paralelos aos eixos coordenados a pesar de que os datos presentaban orientación positiva (ver Figura 2.1) porque non incluímos como parámetro na definición de $\hat{f}_{hist}$ a orientación. Esta restrición na forma do histograma non supón un inconveniente para recuperar a orientación da mostra, sobre todo para valores máis pequenos de b_x e b_y, por tratarse a orientación dun fenómeno máis global que local.

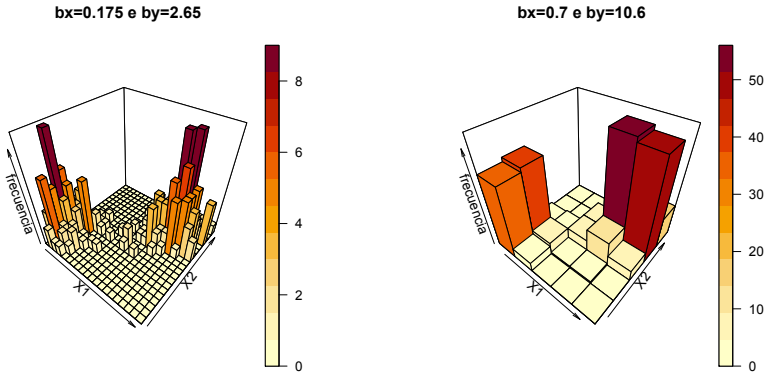

Figura 2.3: Dous histogramas bidimensionais das dúas variables do conxunto de datos "$OldFaithful$", para distintos b_x e b_y. Medidas en minutos, a variable X_1 recolle a duración das erupcións do geyser e a variable X_2 o tempo de espera entre unha erupción e outra.

O histograma da esquerda da Figura 2.3 é bastante dentado, en contraposición ao da dereita, que ten escasas divisións. En ambos o punto de anclaxe é común, con coordenadas $(t_{0x}, t_{0y}) = (1{,}6, 43)$, e corresponde coa esquina inferior esquerda da base de cada unha das gráficas. Nos dous histogramas observamos que hai dúas rexións do plano onde a concentración de datos é alta e no resto é baixa ou incluso nula.

A nosa mostra segue unha distribución claramente bimodal, como se aprecia nos dous mapas de calor da Figura 2.4, cada un deles relativo a cada un dos dous histogramas da Figura 2.3. Recordemos que os mapas de calor son unha ferramenta gráfica que serve para visualizar o comportamento dun conxunto de datos, empregando unha escala de cores para representar variacións, neste caso, da densidade dos datos. Noutras palabras, os mapas de

calor corresponden coa vista aérea dos histogramas, onde as modificacións de cor e a lenda nos permiten omitir a representación das alturas das barras e así reducir nunha dimensión a gráfica e facilitar a visualización.

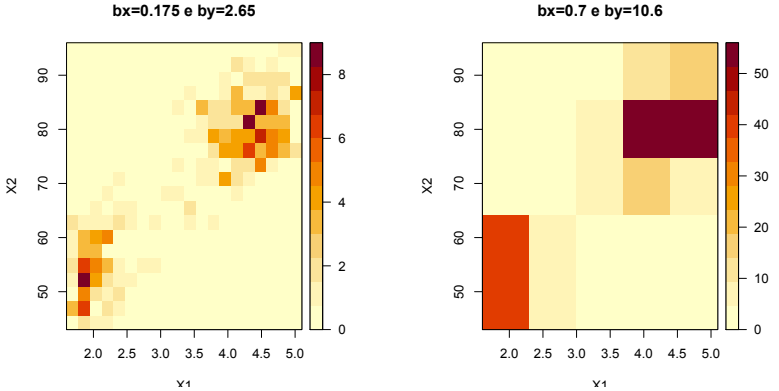

Figura 2.4: Dous mapas de calor das variables X_1 e X_2 do conxunto de datos "$OldFaithful$" de R, relativos a cada un dos histogramas da Figura 2.3. Conservouse a escala de cada unha das variables, ambas en minutos, sendo X_1 a duración das erupcións do geyser e X_2 o tempo de espera entre erupcións.

Como xa vimos no Capítulo 1, o estimador tipo núcleo proporciona resultados máis exactos acerca da distribución dos nosos datos no sentido de que obtemos erros de estimación menores. Ademais é un estimador continuo que dá lugar a unha estimación máis suave que empregando histogramas. Nese capítulo calculamos distintos criterios de erro do estimador histograma unidimensional; sen embargo, como vimos que o histograma é menos eficiente que o estimador tipo núcleo, no caso bidimensional imos omitir estes detalles e proceder directamente a traballar co estimador tipo núcleo, ao que si lle calcularemos en pormenor os distintos criterios de erro.

2.2. Estimación da densidade tipo núcleo bidimensional

Por analoxía ao caso unidimensional, o estimador tipo núcleo é unha mellora do histograma tradicional en canto a que proporciona estimacións máis precisas e suaves e elimina a dependencia do punto de anclaxe. Estas últimas son algunhas das razóns que fan que, logo do histograma tradicional e a súa sinxeleza, o estimador da densidade máis coñecido sexa o estimador da densidade tipo núcleo, pola súa eficacia na estimación.

Na Figura 2.4 representábamos dous diagramas de calor da duracións das erupcións dun geyser e do tempo de espera entre elas, que viñan a ser a visualización aérea dos dous histogramas bidimensionais correspondentes. Como xa observábamos, a estimación era moi pouco suave, e a pesar de que non chegou a representarse, en moitos casos un cambio no punto de anclaxe daría lugar a histogramas diferentes. Como mellora a estas estimacións calculamos os contornos de nivel da estimación tipo núcleo considerando o núcleo gaussiano. Esta nova estimación da densidade do conxunto de datos de "$OldFaithful$" corresponde coa Figura 2.5. Ademais, para comparar ambas gráficas, consideramos como matrices de suavizado $H = \text{diag}(0{,}175, 2{,}65)$ e $H = \text{diag}(0{,}7, 10{,}6)$, que serían equivalentes ás lonxitudes dos lados dos rectángulos e á orientación paralela ao eixo OX no caso do histograma. Os detalles técnicos quedan para máis adiante, onde se dará unha xustificación destas eleccións de H. De novo, observemos que hai dúas zonas do plano de alta concentración de datos e que estes están orientados positivamente.

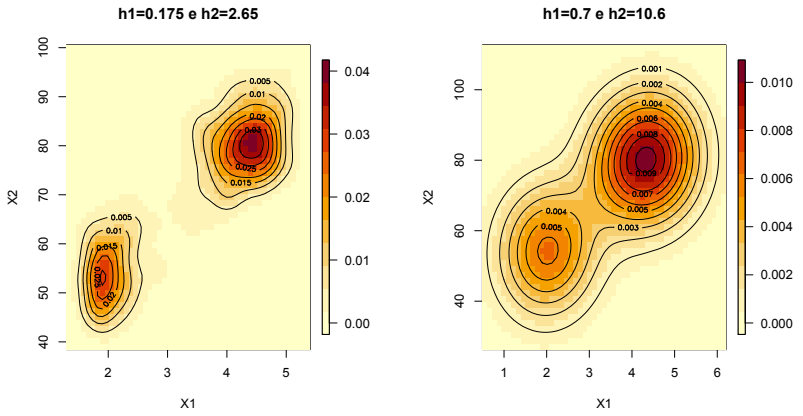

Figura 2.5: Curvas de nivel da estimación tipo núcleo con núcleo gaussiano de X_1 e X_2 do conxunto de datos "$OldFaithful$" de R xunto cos mapas de calor correspondentes. Ambas variables se miden en minutos, sendo X_1 a duración das erupcións do geyser e X_2 o tempo de espera entre erupcións.

Claramente a estimación tipo núcleo proporciona resultados máis suaves e estimacións máis precisas, como se observa ao comparar as gráficas das Figuras 2.4 e 2.5. A continuación imos definir a función de densidade estimada por este método. Para cada $(x, y) \in \mathbb{R}^2$,

$$\hat{f}((x,y), H) = \frac{1}{n} \sum_{i=1}^{n} K_H(x - X_i, y - Y_i), \tag{2.2}$$

onde K é a función núcleo (densidade unimodal simétrica de media cero) e $H \equiv H(n) \equiv$

$H_n \in M_{2x2}$ é a matriz de parámetros de suavizado, necesariamente simétrica, definida positiva e non singular, porque a parametrizamos como unha matriz de covarianzas, que controla a orientación e o nivel de suavizado do núcleo, e depende do tamaño mostral.

A matriz H é o homólogo ao parámetro h no caso unidimensional, pasando agora a ter que estimar ata 3 parámetros diferentes en relación a ela. Sen embargo, o papel da matriz de parámetros ventá no caso bidimensional non é exactamente o mesmo que no unidimensional: agora h_1 e h_2 son os termos asociados á varianza e no caso unidimensional, h correspondía á desviación típica. Esta diferenza débese unicamente á notación empregada: se no Capítulo 1 usásemos h^2 en lugar de h, esta correspondería tamén á varianza. Ademais agora aparece un termo relacionado coa covarianza, h_{12}, como consecuencia da interación das dúas variables.

Por analoxía ao caso unidimensional,

$$K_H(x,y) = |H|^{-1/2} K(H^{-1/2} \begin{pmatrix} x \\ y \end{pmatrix}), \tag{2.3}$$

onde $|H|$ é o determinante de H e $H^{-1/2}$ é a raíz cadrada da inversa de H, que existe por ser H non singular. Para calcular $H^{-1/2}$ podemos empregar a forma canónica de Jordan da matriz H, é dicir, descompoñela en $H = PJP^{-1}$, onde P é unha matriz invertible formada polos autovectores de H en columnas e J é unha matriz diagonal cuxos elementos diagonais son os autovalores de H (como H é simétrica e con coeficientes reais. é diagonalizable en $\mathbb{R}$ polo Teorema Espectral: ver demostración no libro Moscoso 2004, Sección 2.4, Teorema 2.22). Deste xeito, $H^{-1/2} = PJ^{-1/2}P^{-1}$, onde $J^{-1/2}$ non é máis que calcular a raíz cadrada de cada elemento diagonal de J.
No caso particular de que H fose diagonal, obteríamos que $H = J$ e así, $H^{-1/2} = J^{-1/2}$.

Ademais K_H é a densidade dun vector aleatorio HZ, onde Z segue unha distribución bivariante con densidade K, e o núcleo reescalado trasladado non é máis que centrar o núcleo reescalado K_H en cada observación mostral $(X_i, Y_i), i \in \{1, ..., n\}$. Deste xeito trasladamos a cada observación mostral o máximo que K presenta na orixe e así, o estimador tipo núcleo pódese interpretar como a densidade resultante de cambiar os valores da mostra por vectores aleatorios independentes e distribuídos de acordo ao núcleo reescalado K_H, centrado en cada dato mostral. Novamente o estimador tipo núcleo é un promedio de densidades centradas nos datos mostrais, convenientemente reescaladas.

Para fixar ideas supoñamos que o núcleo segue unha distribución normal estándar bi-

variante, é dicir, que $K(x,y) = \phi_{\overrightarrow{0},I_2}(x,y)$, cuxa expresión formal corresponde á expresión (A.4) do Apéndice A, e consideramos a matriz de suavizado

$$H = (h_1, h_{12}; h_{12}, h_2) = \begin{pmatrix} h_1 & h_{12} \\ h_{12} & h_2 \end{pmatrix} \in M_{2x2}.$$

Deste xeito, o núcleo reescalado vén dado por

$$K_H(x,y) = \phi_{\overrightarrow{0},H}(x,y) = \frac{1}{2\pi}|H|^{-1/2} \exp\left(\frac{-1}{2}(x,y)H^{-1}(x,y)'\right), \tag{2.4}$$

e o núcleo reescalado trasladado por

$$K_H(x - X_i, y - Y_i) = \frac{1}{2\pi}|H|^{-1/2} \exp\left(\frac{-1}{2}(x - X_i, y - Y_i)H^{-1}(x - X_i, y - Y_i)'\right), \tag{2.5}$$

polo que este último é unha distribución normal bivariante centrada no dato mostral (X_i, Y_i) e con matriz de varianzas e covarianzas H. Como vemos, a interpretación é análoga ao caso unidimensional pois, onde representábamos campás de Gauss unidimensionais, agora representámolas bidimensionais.

O estimador tipo núcleo de f en (x,y) adopta a seguinte forma:

$$\hat{f}((x,y),H) = \frac{1}{2n\pi}|H|^{-1/2} \sum_{i=1}^{n} \exp\left(\frac{-1}{2}(x - X_i, y - Y_i)H^{-1}(x - X_i, y - Y_i)'\right). \tag{2.6}$$

Nótese que o termo vinculado coa orientación da estimación é o termo h_{12}. Entón, tomar $h_{12} = 0$ na estimación tipo núcleo é o equivalente a considerar rectángulos paralelos aos eixes coordenados no histograma, xustificando así a elección das matrices de suavizado na Figura 2.5. Por analoxía ao que fixemos no histograma da Sección 2.1, supoñamos que a matriz H está restrinxida a ser diagonal e así:
$H = \text{diag}(h_1, h_2) = \begin{pmatrix} h_1 & 0 \\ 0 & h_2 \end{pmatrix} \in M_{2x2}, h_1 > 0, h_2 > 0$, polo que o núcleo reescalado será

$$K_H(x,y) = \frac{1}{\sqrt{h_1}\sqrt{h_2}}\phi(\frac{x}{\sqrt{h_1}})\phi(\frac{y}{\sqrt{h_2}}),$$

e o núcleo reescalado trasladado

$$K_H(x - X_i, y - Y_i) = \frac{1}{\sqrt{h_1}\sqrt{h_2}}\phi(\frac{x - X_i}{\sqrt{h_1}})\phi(\frac{y - Y_i}{\sqrt{h_2}}).$$

Deste xeito o estimador tipo núcleo de f en (x,y) é da forma

$$\hat{f}((x,y),H) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - X_i, y - Y_i) = \frac{1}{n\sqrt{h_1}\sqrt{h_2}}\sum_{i=1}^{n}\phi(\frac{x - X_i}{\sqrt{h_1}})\phi(\frac{y - Y_i}{\sqrt{h_2}}). \tag{2.7}$$

Na Figura 2.6 móstranse os contornos de nivel de dez núcleos individuais empregando como núcleo a distribución normal estándar e como matriz de suavizado $H = (1, \frac{1}{2}; \frac{1}{2}, 2)$. Os datos proceden dunha distribución uniforme no cadrado unidade, centrando os contornos de nivel de cada unha das campás de Gauss nos dez valores mostrais. O estimador resultante é o promedio destas dez campás de Gauss, sendo os seus contornos de nivel elipses con orientación positiva, como consecuencia de que $h_{12} = \frac{1}{2} > 0$.

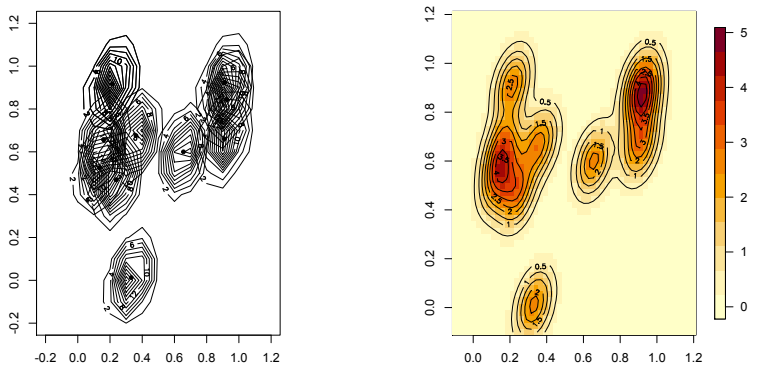

Figura 2.6: Construción da estimación tipo núcleo bivariante: a esquerda, contornos de nivel de dez núcleos individuais correspondentes de centrar campás de Gauss en dez observacións mostrais procedentes dunha distribución uniforme no cadrado unidade; a dereita contornos de nivel xunto co mapa de calor da estimación resultante. Destaquemos que o núcleo empregado segue unha distribución normal estándar e a matriz de suavizado é $H = (1, \frac{1}{2}; \frac{1}{2}, 2)$.

2.3. Criterios de erro da estimación da densidade tipo núcleo

Agora consideraremos diferentes erros co obxectivo de obter criterios para escoller a mellor matriz de parámetros ventá H. No Capítulo 1 comentamos cal era o núcleo óptimo no caso unidimensional, sen embargo, no caso bidimensional imos omitilo por non ser máis que unha xeneralización inmediata: o núcleo óptimo para $d = 2$ é o núcleo Epanechnikov bidimensional[1]. Paralelamente ao caso unidimensional, a análise da actuación do estimador

[1] A función núcleo Epanechnikov bivariante defínese como segue:

$$K^E(x, y) = 2\pi(1 - (x^2 + y^2))\mathbb{I}[x^2 + y^2 < 1]$$

tipo núcleo requirirá explicar con detalle criterios de erro puntuais, que midan o erro que o estimador comete cando estima a densidade nun punto, así como criterios de erro globais, que cuantifiquen o erro que comete ao estimala no plano.

2.3.1. Erro cadrático medio

Como xa adiantamos no Capítulo 1, o erro cadrático medio emprégase para medir o erro ao estimar a densidade f nun punto fixo (x, y) por $\hat{f}((x, y), H)$. Proporciona unha medida puntual: para cada $(x, y) \in \mathbb{R}^2$, medimos a discrepancia entre o número real descoñecido $f(x, y)$ e a súa estimación puntual $\hat{f}((x, y), H)$. Recordemos que se calcula como segue:

$$MSE[\hat{f}((x, y), H)] = \mathrm{Var}[\hat{f}((x, y), H)] + \mathrm{Nesgo}^2[\hat{f}((x, y), H)]. \tag{2.8}$$

Paralelamente ao caso unidimensional, podemos considerar que o estimador da densidade tipo núcleo bidimensional é a convolución da medida de probabilidade inducida polo núcleo reescalado K_H coa distribución empírica dos datos. Co obxectivo de obter unha expresión do erro cadrático medio, calculemos a media e a varianza do noso estimador, en termos da función f, e sen impoñer restricións á matriz H.
Comezando coa expresión da media, obtemos que

$$\mathbb{E}[\hat{f}((x, y), H)] = \mathbb{E}[K_H((x, y) - (X, Y))] = \int_{\mathbb{R}^2} K_H((x, y) - (\tilde{x}, \tilde{y}))f(\tilde{x}, \tilde{y})d\tilde{x}d\tilde{y} = (K_H * f)(x, y).$$

Do mesmo modo, a expresión da varianza redúcese a

$$\mathrm{Var}[\hat{f}((x, y), H)] = \frac{1}{n}\mathrm{Var}[K_H((x, y) - (X, Y))] = \frac{1}{n}\int_{\mathbb{R}^2} K_H^2((x, y) - (\tilde{x}, \tilde{y}))f(\tilde{x}, \tilde{y})d\tilde{x}d\tilde{y} -$$
$$- \frac{1}{n}(\int_{\mathbb{R}^2} K_H((x, y) - (\tilde{x}, \tilde{y}))f(\tilde{x}, \tilde{y})d\tilde{x}d\tilde{y})^2 = \frac{1}{n}((K_H^2 * f)(x, y) - (K_H * f)^2(x, y)).$$

Nesta mesma terminoloxía, temos que $\mathrm{Nesgo}[\hat{f}((x, y), H)] = (K_H * f)(x, y) - f(x, y)$, polo que substituíndo na expresión (2.8) do erro cadrático medio

$$MSE[\hat{f}((x, y), H)] = \frac{1}{n}((K_H^2 * f)(x, y) - (K_H * f)^2(x, y)) + ((K_H * f)(x, y) - f(x, y))^2. \tag{2.9}$$

Dado que f é descoñecida, a dependencia da matriz H é difícil de visualizar nestas fórmulas, igual que ocorría no caso unidimensional. Ademais, na sección da estimación tipo núcleo do Capítulo 1 vimos como este estimador ten dificultades en aproximar con exactitude rexións onde a curvatura da función á estimar é moi marcada, como mostramos no caso da mestura de dúas distribucións normais na Figura 1.9, onde a función orixinal presenta dúas modas moi marcadas.

Procedamos agora a analizar nun exemplo como é o erro cadrático medio do estimador tipo núcleo bivariante para aproximar a mestura de distribucións normais dada por

$$\frac{1}{2}N_2\left(\begin{pmatrix}1\\-1\end{pmatrix},\begin{pmatrix}(\frac{2}{3})^2 & \frac{14}{45}\\ \frac{14}{45} & (\frac{2}{3})^2\end{pmatrix}\right)+\frac{1}{2}N_2\left(\begin{pmatrix}-1\\1\end{pmatrix},\begin{pmatrix}(\frac{2}{3})^2 & 0\\ 0 & (\frac{1}{3})^2\end{pmatrix}\right) \tag{2.10}$$

na orixe e posteriormente ver o diagrama de calor que mostre como varía o erro cadrático medio no plano, fixado un tamaño mostral e un parámetro ventá. En primeiro lugar, fixemos $(x, y) = (0, 0)$ e supoñamos que $\hat{f}$ vén dado pola expresión (2.7). As curvas de nivel da función de densidade desta mestura represéntanse na gráfica da esquerda da Figura 2.7, para ter unha idea de como é a súa distribución e apreciar a súa bimodalidade. Co obxectivo de conseguir unha representación gráfica en $\mathbb{R}^2$ e así, unha mellor interpretabilidade, supoñamos que $h_1 = h_2 = h > 0$, é dicir, que $H = \text{diag}(h, h)$.

Empregaremos o linguaxe Maxima (2019) para efectuar algunhas contas. Maxima distribúese baixo a licencia GNU-GPL e é de libre acceso. Obtemos que:

$$(K_H^2 * f)(0,0) = \frac{1}{(4\pi)^3}\left[\frac{15\sqrt{3}\exp(\frac{30}{289}\frac{55125h^4-48705h^2-19652}{675h^4+1200h^2+272})}{\sqrt{675h^4+1200h^2+272}}+\frac{18\exp(\frac{-18}{9h^2+8})}{9h^2+8}\right] \text{ e que}$$

$$(K_H * f)(0,0) = \frac{h}{2\pi}\left[\frac{15\sqrt{3}\exp(\frac{15}{289}\frac{110250h^4-48705h^2-9826}{675h^4+600h^2+68})}{\sqrt{675h^4+600h^2+68}}+\frac{18\exp(\frac{-9}{9h^2+8})}{9h^2+8}\right]$$

e así, pola expresión (2.9) xa temos todos os termos necesarios para calcular $MSE[\hat{f}((0,0),h)]$.

Na gráfica central da Figura 2.7 móstranse as curvas do erro cadrático medio na orixe para $h \in (0, 1]$ e tamaños mostrais $n \in \{10, 20, 50, 100, 200\}$. Observamos que o erro diminúe lentamente ao aumentar n e que a ventá óptima é cada vez máis pequena, apreciando a dependencia de n que presenta o h óptimo, que posteriormente xustificaremos. Para valores de h menores que o mínimo a función é estritamente decrecente e logo crece estritamente pero con menos pendente ao longo da recta real positiva. Con tamaños mostrais máis pequenos e segundo indica a lenda, o mínimo acádase despois que con tamaños mostrais maiores. Observamos que en $h = 0$ todas as curvas presentan unha asíntota vertical.

Co fin de visualizar como varía o erro cadrático medio no plano e apreciar os cambios entre vales e modas, na gráfica da dereita da Figura 2.7 represéntase un diagrama de calor que mostra como varía o erro cadrático medio no cadrado $[-3, 3] \times [-3, 3]$. Fixamos un tamaño mostral $n = 200$ e, guiándonos pola gráfica central, un parámetro ventá $h = 0{,}05$, próximo ao h óptimo na orixe restrinxindo H á clase de matrices diagonais múltiplo da identidade. Nas dúas modas da densidade teórica, o erro cadrático medio acada os valores

máis elevados: por similitutde ao caso unidimensional, a estimación tipo núcleo ten dificultades na estimación en lugares onde a curvatura da función orixinal é moi marcada.

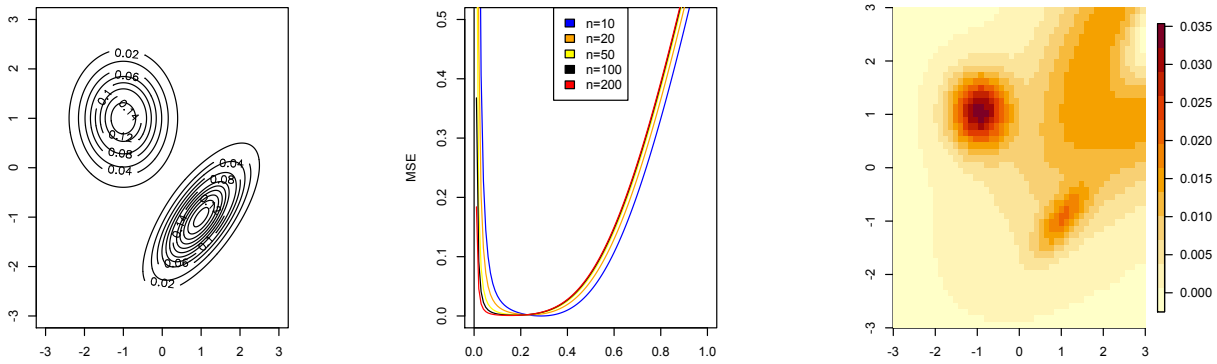

Figura 2.7: A esquerda, curvas de nivel da densidade correspondente á mestura de normais bivariantes dada por (2.10); no centro, erro cadrático medio na orixe para a estimación tipo núcleo dada pola expresión (2.7) con $H = \text{diag}(h, h)$, $h \in (0, 1]$ e $n \in \{10, 20, 50, 100, 200\}$; na dereita, gráfica de calor para o erro cadrático medio no plano correspondente á estimación tipo núcleo fixado $n = 200$ e $h = 0{,}05$.

Como este criterio de erro é puntual e o obxectivo da estimación da densidade é obter como actúa globalmente a distribución probabilística dos nosos datos, necesitamos definir criterios de erros globais que nos proporcionen información sobre como se comporta a estimación a nivel global, e non só localmente. Con este fin definiremos o erro cadrático integral e o erro cadrático medio integral.

2.3.2. Criterios de erro globais exactos

Supoñamos que f e a súa estimación son cadrado-integrables, é dicir, que $f, \hat{f}(\cdot, H) \in \mathscr{L}^2(\mathbb{R}^2)$. O erro cadrático integral é un criterio de erro global, dado que non depende do punto onde se avalía o estimador, definido do seguinte xeito:

$$ISE[\hat{f}(\cdot, H)] = \int_{\mathbb{R}^2} (\hat{f}((x, y), H) - f(x, y))^2 dxdy. \tag{2.11}$$

Trátase dunha variable aleatoria e polo tanto, dunha medida estocástica para cuantificar o erro que comentemos na estimación de f. Por iso mesmo definimos o erro cadrático medio integral. Este último proporciónanos unha medida de erro global cometida ao aproximar no plano a función f por $\hat{f}(\cdot, H)$. Trátase da media da distancia ao cadrado en $\mathscr{L}^2(\mathbb{R}^2)$

entre as funcións $\hat{f}(\cdot, H)$ e f, é dicir, da esperanza matemática do ISE.
A súa expresión analítica vén dada por:

$$MISE[\hat{f}(\cdot, H)] = \mathbb{E}[\int_{\mathbb{R}^2} (\hat{f}((x,y);H) - f(x,y))^2 dxdy] = \tag{2.12}$$

$$= \int_{\mathbb{R}^2} \mathrm{Var}[\hat{f}((x,y),H)]dxdy + \int_{\mathbb{R}^2} \mathrm{Nesgo}^2[\hat{f}((x,y),H)]dxdy, \tag{2.13}$$

onde a última igualdade é válida supoñendo que o orde da integración e a esperanza poidan intercambiarse. Para garantir iso, consideremos o seguinte teorema:

Teorema 2.1. ***(Teorema de Fubini-Tonelli).*** *Sexa $g\colon \mathbb{R}^{m+k} \longrightarrow [0,\infty]$ unha función medible non negativa. Logo a función da variable $y \in \mathbb{R}^k$, $g_x\colon y \longrightarrow g(y) = g(x,y)$ é medible para case todo $x \in \mathbb{R}^m$; a función l definida case para todo $x \in \mathbb{R}^m$ por $l(x) = \int_{\mathbb{R}^k} g(x,y)dy$ é medible. Ademais $\int_{\mathbb{R}^m} l(x)dx = \int_{\mathbb{R}^{m+k}} g(x,y)dxdy$ (é dicir, a integral de g coincide coas súas integrais iteradas e así temos asegurado que podemos intercambiar a orde da integración).*

Demostración: Ver o libro Sattinger 2004, Capítulo 4, Sección 4.1.

Tendo en conta que, ao existir densidade, a esperanza é unha integral con respecto á densidade conxunta e ademais é unha función medible, polo Teorema 2.1:

$$\mathbb{E}[\int_{\mathbb{R}^2} (\hat{f}((x,y);H) - f(x,y))^2 dxdy] = \int_{\mathbb{R}^2} \mathbb{E}(\hat{f}((x,y);H) - f(x,y))^2 dxdy = \int_{\mathbb{R}^2} MSE[\hat{f}((x,y),H)]dxdy.$$

Deste xeito, empregando a expresión (2.9) do MSE, reescribimos o $MISE$ como segue:

$$MISE[\hat{f}(\cdot,H)] = \int_{\mathbb{R}^2} \frac{1}{n}((K_H^2 * f)(x,y) - (K_H * f)^2(x,y)) + ((K_H * f)(x,y) - f(x,y))^2 dxdy =$$
$$= \frac{1}{n}\int_{\mathbb{R}^2} (K_H^2 * f)(x,y)dxdy - \frac{1}{n}\int_{\mathbb{R}^2} (K_H * f)^2(x,y)dxdy + \int_{\mathbb{R}^2} (K_H * f)^2(x,y)dxdy -$$
$$- 2\int_{\mathbb{R}^2} (K_H * f)(x,y)f(x,y)dxdy + \int_{\mathbb{R}^2} f^2(x,y)dxdy =$$
$$= \frac{1}{n}\int_{\mathbb{R}^2} (K_H^2 * f)(x,y)dxdy + (1 - \frac{1}{n})R((K_H * f) - 2\int_{\mathbb{R}^2} (K_H * f)(x,y)f(x,y)dxdy + R(f).$$

Ao igual que ocorría co erro cadrático medio, non podemos obter expresións simples do $MISE$ que nos permitan analizalo con facilidade. Isto débese a que depende de H dunha forma complexa, non lineal; ademais, consideramos integrais cuxo integrando involucra á función f, que é descoñecida. En consecuencia, definiremos erros asintóticos, cuxa expresión é moito máis sinxela e a dependencia da matriz de suavizado é máis clara.

2.3.3. Erro cadrático medio integral asintótico

Como xa dixemos, e paralelamente ao que ocorría no Capítulo 1, a expresión do $MISE$ presenta unha dependencia da matriz de suavizado moi complicada. Isto motiva a definir erros asintóticos como será, neste caso, o erro cadrático medio integral asintótico, como unha aproximación asintótica do erro cadrático medio integral.

Recordemos que os elementos da matriz $H \equiv H_n \equiv H(n)$ dependen do tamaño mostral n e definamos o $AMISE$ como unha aproximación asintótica do $MISE$ que satisfai que $MISE\{\hat{f}(\cdot;H)\} \backsim AMISE\{\hat{f}(\cdot;H)\}$ cando $n \longrightarrow \infty$. Veremos que na súa expresión está claramente dividida a procedencia do erro en nesgo e varianza, polo que será máis doado apreciar o efecto da matriz de parámetros ventá H no estimador tipo núcleo, é dicir, a súa influencia no suavizado da estimación.

Como obxectivo de obter unha expresión compacta do $AMISE$, consideremos a seguinte definición, e introduzamos notación que nos permita abreviar termos.

Definición 2.2. Dado un vector aleatorio (X, Y) e $k \in \mathbb{N}$, defínense os momentos non centrados de orde k como:

$$m_{k_X,k_Y} = \mathbb{E}[X^{k_X}Y^{k_Y}], \forall k_X, k_y \in \mathbb{N} \cup \{0\}, k_X + k_Y = k,$$

sempre que as expresións X^{k_X} e Y^{k_Y} estean ben definidas e existan ditas esperanzas. Nótese que tamén se poden definir os momentos centrados (en medias) e particularizar a definición para o caso dunha variable aleatoria.

Notación 2.3. Sexa K unha función núcleo de dúas variables simétrico-esférica, cadrado integrable e con momento non centrado de segundo orde finito (para que as seguintes integrais sexan finitas e coincidan); isto significa que $\int_{\mathbb{R}^2} xK(x,y)dxdy = \int_{\mathbb{R}^2} yK(x,y)dxdy = 0$, $\int_{\mathbb{R}^2} K^2(x,y)dxdy < \infty$ e que

$$\int_{\mathbb{R}^2} x^2K(x,y)dxdy = \int_{\mathbb{R}^2} y^2K(x,y)dxdy, \text{ que denotaremos por } m_2(K).$$

No caso de ser K a densidade asociada á distribución normal estándar bivariante,

$$m_2(K) = \int_{\mathbb{R}^2} y^2\phi(x)\phi(y)dxdy = \int_{\mathbb{R}} y^2\phi(y)(\int_{\mathbb{R}} \phi(x)dx)dy = \int_{\mathbb{R}} y^2\phi(y)dy = 1.$$

A continuación consideramos un teorema que nos proporciona condicións baixo as cales temos asegurada unha expresión asintótica do $MISE$, e como é a súa formulación no caso de que o núcleo sexa unha normal estándar e H diagonal. No libro de Chacón and Duong 2018, Capítulo 2, Sección 2.6 e 2.9, explícase con rigurosidade a proba deste teorema para o caso multivariable; aquí daremos unha idea esquemática da demostración para $d = 2$.

Teorema 2.4. *Sexa f unha función de densidade cadrado-integrable e dúas veces diferenciable, con todas as derivadas parciais de segundo orde limitadas, continuas e cadrado-integrables. Supoñamos que o núcleo K é unha densidade cadrado-integrable, simétrico-esférica e con momento de segundo orde finito; por analoxía ao caso unidimensional, supoñamos que a matriz de suavizado $H = H_n$ é diagonal e verifica que*

$$\frac{1}{n\sqrt{h_1}\sqrt{h_2}} = n^{-1}|H|^{\frac{-1}{2}} \longrightarrow 0 \text{ e } |H| = h_1 h_2 \longrightarrow 0, \text{ cando } n \longrightarrow \infty.$$

Baixo estas podemos obter unha expresión asintótica do $MISE$ do estimador $\hat{f}$.

Se ademais supoñemos que o núcleo K segue unha distribución normal estándar

$$AMISE\{\hat{f}(\cdot, H)\} = \frac{1}{n\sqrt{h_1}\sqrt{h_2}} R(K) + \frac{1}{4} m_2^2(K) \int_{\mathbb{R}^2} tr^2\{HHf(x,y)\} dxdy = \tag{2.14}$$

$$= \frac{1}{4\pi n\sqrt{h_1}\sqrt{h_2}} + \frac{1}{4}(h_1^2 + h_2^2)^2 \int_{\mathbb{R}^2} f^2(x,y) dxdy = \frac{1}{4\pi n\sqrt{h_1}\sqrt{h_2}} + \frac{1}{4}(h_1^2 + h_2^2) R(f), \tag{2.15}$$

onde para cada $A \in \mathcal{M}_{2\times 2}$, $tr(A)$ denota a traza de A, polo que tr denota o operador traza.

Idea da demostración. A proba do teorema reside en calcular expresións asintóticas do nesgo e da varianza de $\hat{f}$ empregando expansións de Taylor en dúas variables, que involucran derivadas de segundo orde mixtas. En consecuencia, baixo as condicións do Teorema pero sen esixir necesariamente que K siga unha distribución normal estándar, obtemos unha expresión asintótica do $MISE$ de $\hat{f}$, que dá lugar a expresión (2.14).
Se ademais engadimos que o núcleo K se distribúe baixo a normal estándar bivariante, a seguinte igualdade formulada no Teorema é inmediata recordando que:

- $m_2(K) = 1$.
- $R(K) = \frac{1}{4\pi}$.
- $\mathrm{tr}\{HHf(x,y)\} = f(x,y)\{\begin{pmatrix} h_1 & 0 \\ 0 & h_2 \end{pmatrix}\begin{pmatrix} h_1 & 0 \\ 0 & h_2 \end{pmatrix}\} = f(x,y)(h_1^2 + h_2^2)$.
- $\int_{\mathbb{R}^2} \mathrm{tr}^2\{HHf(x,y)\} dxdy = (h_1^2 + h_2^2)^2 \int_{\mathbb{R}^2} f^2(x,y) dxdy = (h_1^2 + h_2^2)^2 R(f)$.

Con isto finalizamos, a grandes rasgos, a demostración do Teorema. □

Concluímos que a expresión do $AMISE$, no caso de que o núcleo sexa a densidade dunha normal estándar e H sexa diagonal, ten unha dependencia moi sinxela da matriz de parámetros de suavizado, resultando unha expresión moi fácil de derivar con respecto ás variables h_1 e h_2. O primeiro sumando é procedente da integral da varianza e o segundo

da do nesgo ao cadrado. Debemos obter un equilibrio nesgo-varianza coa finalidade de minimizar o $AMISE$: o obxectivo é empregar unha cantidade de suavizado óptima, pois esta equilibraría os termos de nesgo e de varianza. Nesta liña, temos que:

- Un ancho de banda pequeno (h_1 e h_2 toman valores pequenos) conduce a que o estimador empregue poucas observacións na ponderación, obtendo pouco nesgo pero moita variabilidade: a estimación en cada punto está moi condicionada polos datos máis próximos. Deste xeito o estimador dependerá moito deles, influenciado pola varianza mostral, polo que considerar unha mostra distinta podería dar lugar a estimacións moi diferentes.
- Un ancho de banda grande (h_1 e h_2 toman valores grandes) produce un aumento do nesgo dado que empregaremos moitas máis observacións na ponderación no punto de interese, reducindo así a variabilidade da estimación. Deste xeito, a función estimada será máis robusta en canto a que, cambiando a mostra, a nova estimación será similar, pero nesgada.

Recordemos que no caso unidimensional a consistencia asintótica estaba garantizada se h -que correspondía coa desviación típica das campás de Gauss que formaban parte do promedio que daba lugar ao estimador tipo núcleo con núcleo normal- converxe a cero e nh a infinito. Isto pode reescribirse no caso bidimensional coas dúas condicións seguintes:

$$\frac{1}{n\sqrt{h_1}\sqrt{h_2}} = n^{-1}|H|^{\frac{-1}{2}} \longrightarrow 0 \text{ e } |H| = h_1 h_2 \longrightarrow 0, \text{ cando } n \longrightarrow \infty \tag{2.16}$$

onde agora os termos h_1 e h_2 corresponden á varianza. É dicir, se se cumpre as dúas condicións de (2.16) aseguramos que $AMISE[\hat{f}(\cdot, H)] \longrightarrow 0$ cando $n \longrightarrow \infty$.

Calculemos as derivadas parcias da expresión (2.14) do $AMISE$ co obxectivo de minimizalo con respecto de h_1 e h_2:

$$\frac{\partial}{\partial h_1} AMISE[\hat{f}(\cdot, H)] = -\frac{1}{8\pi n h_1^{\frac{3}{2}} h_2^{\frac{1}{2}}} + \frac{h_1 R(f)}{2},$$
$$\frac{\partial}{\partial h_2} AMISE[\hat{f}(\cdot, H)] = -\frac{1}{8\pi n h_1^{\frac{1}{2}} h_2^{\frac{3}{2}}} + \frac{h_2 R(f)}{2}.$$

Igualando a cero as expresións anteriores obtemos un sistema de dúas ecuacións non lineais e dúas incógnitas que, por exemplo, resolvendo polo método de substitución, e empregando que $h_1 > 0$ e $h_2 > 0$, ten unha solución dada por:

$$h_1 = \left(\frac{1}{4\pi R(f)}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}, \; h_2 = \left(\frac{R(f)^2}{4\pi}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}. \tag{2.17}$$

Calculando a matriz hessiana e substituíndo os valores de h_1 e h_2 de (2.17) obtemos que minimizan a expresión (2.14) (a matriz resultante é definida positiva). Sen embargo, ambos valores están dados en función do termo descoñecido $R(f)$, paralelamente ao que ocorría coa expresión (1.21) do h_{AMISE} do Capítulo 1, que dependía de $R(f'')$.

2.4. Selección da matriz de parámetros ventá

No Capítulo 1 comentouse a Regra do Polgar co obxectivo de explicar como escollemos en función da mostra o parámetro ventá na estimación tipo núcleo unidimensional, empregando R. Pola súa construción, a Regra do Polgar pódese extender á estimación da densidade para máis dunha variable sen máis que considerar agora dous h_{AMISE}, dados polos h_1 e h_2 da expresión (2.17), supoñendo que a matriz de parámetros de suavizado H é diagonal. Así estimaríamos $R(f)$ supoñendo que f segue unha distribución normal bivariante e o resto do razoamento é análogo ao caso unidimensional xa exposto.
Recordemos que esta estimación paramétrica no caso unidimensional consistía en estimar μ empregando a media mostral e σ^2 empregando a varianza mostral (método de Scott), ou o mínimo entre esta e o rango intercuartílico estandarizado/1,349 (método de Silvermann), entre outros. Para $d = 2$ dimensións aumenta o número de parámetros que debemos estimar, que agora son os elementos do vector de medias e da matriz de varianzas e covarianzas, pero o razoamento é o mesmo.

Existen outros métodos para estimar a matriz de parámetros ventá óptima baseados en minimizar algún criterio de erro xa definido, estimando $R(f)$ de forma non paramétrica. Entre eles destaca a xeneralización a dúas dimensións da Regra de Sheather and Jones (1991), baseada nunha estimación non paramétrica do termo descoñecido $R(f'')$ da expresión do $AMISE$ da estimación tipo núcleo unidimensional, e cuxo razoamento se podería xeralizar neste caso á estimación de $R(f)$ do $AMISE$ bidimensional.
Alternativamente, presentaremos dous métodos que seguen un camiño distinto, xa que se centran en minimizar un estimador do $MISE$ (validación cruzada inesgada) e do $AMISE$ (validación cruzada nesgada).

2.4.1. Validación cruzada de mínimos cadrados ou inesgada

Este método de selección da matriz de suavizado foi exposto por Rudemo (1982) e por Bowman (1984). O seu nome está motivado pola seguinte expansión do $MISE$, que recordemos que considera diferencias ao cadrado para evitar que se contrarresten termos

positivos e negativos, e cuxa relación xustifica o nome de mínimos cadrados:

$$MISE[\hat{f}(\cdot,H)] = \mathbb{E}[\int_{\mathbb{R}^2}(\hat{f}((x,y);H) - f(x,y))^2 dxdy] =$$
$$= \mathbb{E}[\int_{\mathbb{R}^2}\hat{f}((x,y);H)^2 dxdy] - 2\mathbb{E}[\int_{\mathbb{R}^2}\hat{f}((x,y);H)f(x,y)dxdy] + R(f).$$

Como $R(f)$ non depende de H, minimizar o $MISE$ respecto de H é equivalente a minimizar respecto de H a seguinte función:

$$\Phi(H) = \mathbb{E}[\int_{\mathbb{R}^2}\hat{f}((x,y);H)^2 dxdy - 2\int_{\mathbb{R}^2}\hat{f}((x,y);H)f(x,y)dxdy].$$

Un estimador inesgado da cantidade anterior vén dado por

$$LSCV(H) = \int_{\mathbb{R}^2}\hat{f}((x,y),H)^2 dxdy - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}((X_i,Y_i),H), \tag{2.18}$$

aínda que amiúdo tamén se denota por UCV, do inglés *unbiased cross validation*, onde

$$\hat{f}_{-i}((x,y),H) = \frac{1}{n-1}\sum_{j=1,j\neq i}^{n} K_H(x - X_j, y - Y_j)$$

é a estimación tipo núcleo de f empregando a matriz de suavizado H, o núcleo K e a mostra de tamaño $n-1$ $(X_1,Y_1),...(X_{i-1},Y_{i-1}),(X_{i+1},Y_{i+1}),...,(X_n,Y_n), i \in \{1,...,n\}$. Como o estimador $LSCV(H)$ é un estimador inesgado de $\Phi(H)$, é razoable pasar a minimizar a expresión (2.18) con respecto de H. Ata o de agora neste apartado non impuxemos ningunha restrición sobre a matriz de suavizado, sen embargo, en moitas ocasións e para simplificar os cálculos na minimización de $LSCV(H)$, suponse que H está restrixida a ser diagonal. A matriz H que minimice $LSCV(H)$ denótase por H_{LSCV}.

Cómpre destacar que se na práctica atopamos varios mínimos desta función, debemos restrinxir o proceso de búsqueda a un rango fixo axeitado, onde existirá un único mínimo. Se o núcleo K segue unha distribución normal estándar, no libro Chacón and Duong 2018, Capítulo 3, Sección 3.4, xustifícase a seguinte simplificación da expresión (2.18), operando o primeiro sumando e empregando propiedades da distribución normal para reducir o segundo a unha expresión máis coñecida:

$$LSCV(H) = \frac{1}{4\pi n}|H|^{\frac{-1}{2}} + \frac{1}{n(n-1)}\sum_{i,j=1,j\neq i}^{n}(\phi_{2H} - 2\phi_H)(X_i - X_j, Y_i - Y_j),$$

onde ϕ_Σ representa a función de densidade dunha $N_2(\overrightarrow{0},\Sigma)$. Sen embargo, pódese ver como o comportamento na práctica deste método de selección de parámetros ventá é moi variable e en moitos casos desacertado, polo que é conveniente formular outro método de selección que mellore as desvantaxes deste. Con ese obxectivo introdúcese a validación cruzada nesgada.

2.4.2. Validación cruzada nesgada

Agora imos considerar un método de selección de matriz de parámetros ventá baseado na expresión do $AMISE$, que conducirá a que sexa nesgado, e non na do $MISE$, como ocorría co método anterior, que era inesgado. Foi formulado por Scott and Terrel (1987), que consideraron que a función a minimizar era a obtida substituíndo a función f da expresión (2.14) do $AMISE$ pola súa estimación tipo núcleo, $\hat{f}(\cdot, H)$. Dado que a expresión do $AMISE$ unicamente a calculamos baixo o suposto de que H sexa diagonal, supoñámolo agora tamén. Así, coñecemos todos os termos da función que queremos minimizar, dado que vén dada por:

$$BCV(H) = \frac{1}{4\pi n\sqrt{h_1}\sqrt{h_2}} + \frac{1}{4}(h_1^2 + h_2^2)R(\hat{f}), \tag{2.19}$$

onde BCV procede do inglés *biased cross validation*. A matriz H que minimice $BCV(H)$ denótase por H_{BCV}. Paralelamente ao que ocorría co método inesgado, se o núcleo K segue unha distribución normal estándar, o termo $R(\hat{f})$ da expresión (2.19) pódese simplificar, reducindo dita expresión a:

$$BCV(H) = \frac{1}{4\pi n\sqrt{h_1}\sqrt{h_2}} + \frac{h_1^2 + h_2^2}{4n^2 h_1 h_2} R\left(\sum_{i=1}^{n} \phi(\frac{x - X_i}{\sqrt{h_1}})\phi(\frac{y - Y_i}{\sqrt{h_2}})\right)$$

Consideramos a validación cruzada nesgada coa esperanza de que o aumento do nesgo conducira a unha diminución da varianza, respecto á validación cruzada inesgada. Sen embargo, estudos posteriores demostraron que isto non era así polo que é preferible o criterio de validación cruzada inesgada.

Non existe ningún selector de matriz de parámetros ventá que sexa o máis competitivo e acade os mellores resultados en todas as mostras. En gran medida, o comportamento de cada método depende do conxunto de datos considerado e das súas propiedades (modalidade, simetría, dispersión,...).

2.5. Xeneralización ao caso multivariante: A maldición da dimensión

A maldición da dimensión ou fenómeno de Hughes é un importante problema que xorde ao xeneralizar a estimación da densidade ao caso multidimensional. Os primeiros en estudar este fenómeno foron Hughes (1968) e Bellman (1961). O primeiro deles tratou este problema mentres estudaba a relación de acerto esperado dun clasificador coa complexidade do mesmo e co número de mostras empregadas para entrealo. Ilustremos este fenómeno

co seguinte exemplo:
Chega con empregar $n = 100$ puntos para mostrear o intervalo unidade $[0, 1]$ de tal xeito que os puntos non disten máis de 10^{-2} entre eles. Para obter unha mostraxe equivalente no cadrado unidade $[0, 1] \times [0, 1]$ fan falta 10^4 puntos. Isto fai que a estimación tipo núcleo bivariante teña unha complexidade computacional maior que no caso univariante dado que precisamos máis datos para obter a mesma concentración mostral. O mesmo ocorre co histograma, que tamén padece este problema ligado á dimensión. Ademais, a natureza de moitas mostraxes pode imposibilitar a obtención de mostras de tamaño demasiado elevado.

Sen embargo, a gravidade do problema é baixa na estimación bidimensional e aumenta considerablemente para o caso d-dimensional, con $d \geqslant 3$. Neste último, o tamaño mostral debería ser moi elevado para garantir estimacións precisas pero o coste computacional pode chegar a incrementarse demasiado ou non ser posible obter mostras tan grandes. Ademais, para $n \geqslant 3$ tamén hai un problema de visualización dado que só somos capaces de ver en tres dimensións e gráfica atoparíase en catro ou máis. Entre outras cousas, poderíanse usar gráficos de calor, pero tamén sería difícil de visualizar.

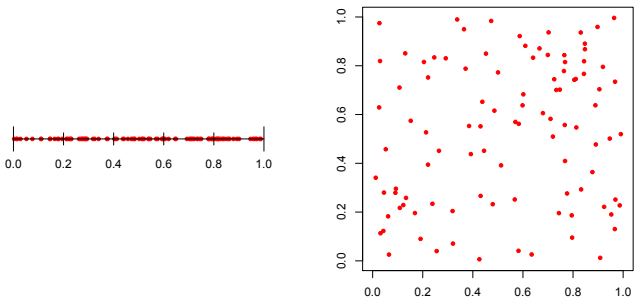

Figura 2.8: Representación gráfica do problema da dimensión para $d = 2$ dimensións: a esquerda, $n = 100$ valores aleatorios do intervalo $[0, 1]$ e a dereita, $n = 100$ valores aleatorios do cadrado $[0, 1] \times [0, 1]$.

Na Figura 2.8 ilustramos este exemplo. Así observamos como, para un mesmo tamaño mostral, ao pasar ao caso bidimensional, os datos se atopan máis espaciados que en dimensión un. Polo tanto para un mesmo tamaño das veciñanzas, a densidade de datos será moito menor e como consecuencia podemos atoparnos nunha situación de escaseza de datos nas veciñanzas e así, obter estimacións pouco precisas. Como a mostra debe aumentar considerablemente ao aumentar a dimensión, o coste computacional é moito maior.

Existen diversas solucións ao fenómeno de Hughes ligado á estimación da densidade que podemos citar a modo de exemplo, pero que non son de interese neste traballo dado que ao sumo manexaremos $d = 2$ dimensións, como pode ser a estimación paramétrica ou a semiparamétrica, que impoñen máis restricións sobre a función a estimar pero tamén que teñen un coste computacional menor e esixen un tamaño mostral máis pequeno. Este último xeito de estimación da función de densidade combina un termo paramétrico cun non paramétrico, flexibilizando así a estimación paramétrica pero sen lograr a flexibilidade da estimación non paramétrica. Agora ben, neste traballo non abordamos o enfoque paramétrico nin o semiparamétrico, dado que unicamente consideramos o caso unidimensional e o bidimensional, polo que é suficiente a estimación non paramétrica.

Capítulo 3

Análise de datos

Ao longo do anterior capítulo empregamos os datos de "$OldFaithful$" para ilustrar os dous tipos de estimadores da densidade bidimensional considerados: o histograma e o estimador tipo núcleo bidimensional; así como o histograma e o estimador tipo núcleo unidimensional en cada unha das dúas variables. Agora imos analizar un novo conxunto de datos, xa empregado brevemente na Introdución deste traballo, proceder ao seu análise e, entre outras cousas, á estimación da súa función de densidade. A natureza dos datos conducirá a que só estudemos a súa distribución conxunta, centrándonos na súa estimación tipo núcleo e na elección dunha matriz de suavizado. Dado que a mostra se tomou en anos distintos, tamén analizaremos a evolución temporal da poboación á que pertence.

Os datos que imos a empregar recollen información sobre a posición de niños de avespa velutina en Galicia entre os anos 2016 e 2018, ambos inclusive. Atópanse no sistema de coordenadas universal transversal de Mercator (en inglés *Universal Transverse Mercator*, UTM) que permiten debuxar as posicións no mapa galego. Este sistema de coordenadas esta baseado na proxección cartográfica transversa de Mercator, que se constrúe como a proxección de Mercator normal (proxección cilíndrica conforme, é dicir, que conserva os ángulos, onde os meridianos son paralelos e equidistantes e as liñas de latitude son paralelas pero vanse afastando unhas das outras cara os polos), pero en lugar de facela tanxente ao Ecuador, realízase secante a un meridiano.

A diferenza do sistema de coordenadas xeográficas, expresadas en lonxitude e latitude, ambas en graos, as magnitudes no sistema de coordenadas UTM exprésanse en metros ao nivel do mar. Sen embargo, para facilitar a interpretación dos resultados, e dado que o sistema de coordenadas xeográficas é máis sinxelo e empregado cotidianamente -non precisa coñecementos matemáticos para a súa compresión- transformaremos os datos mediante un

cambio de variable, renomeándoos para consideralos en termos de latitude e lonxitude.

Recordemos que a latitude (cartográfica) expresa a distancia angular entre o Ecuador e un punto determinado da Terra, medida ao longo do meridiano no que se atopa dito punto. Segundo o hemisferio no que se sitúe o punto, pode ser latitude norte ou sur. Exprésase en medidas angulares que varían dende os 0° do Ecuador ata os 90°N do polo Norte ou os 90°S do polo Sur. Por outra parte, a lonxitude (cartográfica) expresa a distancia angular entre o meridiano de Greenwich (é o considerado meridiano base, asignándoselle 0°) e un punto dado da superficie terrestre, medida ao longo do paralelo no que se atopa dito punto. Existen varias maneiras de medila pero aquí empregaremos unha angular que varía entre 0° e 180°, indicando se é cara o Oeste (denotaremos W, do inglés *west*) ou cara o Este (denotaremos E, do inglés *east*). Sen embargo, en R os graos varían entre 0° e 180° positivos, indicando que é cara o Este, ou negativos, cara o Oeste. Esta é a razón pola que na lenda dalgunhas das vindeiras gráficas o eixe de abscisas toma unidades negativas.

O conxunto dos nosos datos consta de tres variables:

- UTM.X: valor en coordenadas da variable X. Logo do cambio de coordenadas, é unha variable continua que se mide en graos. Correspóndese coa lonxitude cartográfica e toma valores entre 6.8778°W e 9.2654°W.
- UTM.Y: valor en coordenadas da variable Y. Logo do cambio de coordenadas, é unha variable continua que se mide en graos. Correspóndese coa latitude cartográfica e toma valores entre 41.8164°N e 43.7373°N.
- Ano: como ben di o nome, ano no que se tomou o dato. É unha variable discreta que só toma tres valores, que son os anos da realización da mostraxe: 2016, 2017 e 2018.

En primeiro lugar, imos considerar unicamente os datos do ano 2018, obtendo así unha submostra de tamaño $n = 2635$ da mostra orixinal. Isto é equivalente a considerar únicamente os datos que verifiquen que a variable discreta Ano sexa igual a 2018. Deste xeito, a submostra pódese ver como unha mostra bidimensional.

Na Figura 3.1 represéntase o diagrama de dispersión dos niños de avespa velutina en Galicia no ano 2018 xunto co contorno do mapa galego. Hai dúas rexións onde a concentración de datos é maior: as Rías Baixas e a zona de Coruña. Isto débese a que houbo dous puntos de expansión inicial das velutinas en Galicia, en inverno de 2012, un en O Rosal (sur da provincia de Pontevedra), onde se cree que procedían do norte de Portugal -lugar

onde xa había exemplares- e outro en Burela (Lugo), onde se cree que chegaron por medio dunha descarga no porto. A gran capacidade de multiplicación fai que a presencia destes insectos se dispárase en só uns anos, dende a súa chegada no 2012 ao diagrama que aquí presentamos, seis anos despois.

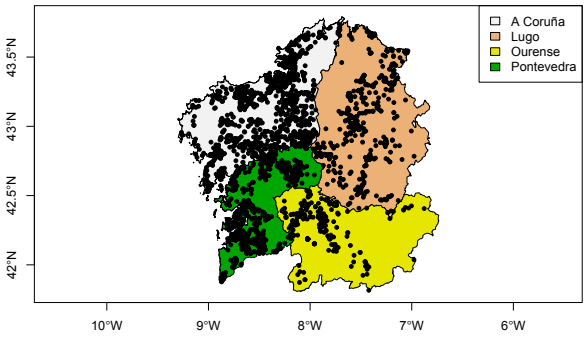

Figura 3.1: Diagrama de dispersión dos niños de velutina en Galicia no ano 2018.

Na zona costeira, aparentemente, a existencia de niños é máis elevada con respecto ao interior. En certo modo, créese que a avespa velutina sente predilección pola costa dado que é unha especie orixinaria do sudeste asiático, acostumada a temperaturas subtropicais, polo que en temperaturas suaves e a baixa latitude exténdese con máis facilidade. Isto explica a súa presencia en case toda a costa, de Ribadeo ata A Guarda, e polo contrario, a escaseza relativa de niños no interior da provincia de Lugo e na de Ourense. A súa expansión ao interior do territorio galego prodúcese principalmente seguindo cursos de auga, que son zonas nas que esta especie se adapta mellor.

Nun primeiro lugar, poderíamos pensar en realizar unha análise individual de cada unha das variables para estudar o seu comportamento por separado, e así estimar a súa función de densidade, empregando tanto o histograma como o estimador tipo núcleo unidimensional. Estas estimacións proporcionaríannos información acerca de se a latitude ou a lonxitude, por separado, inflúen na concentración de niños de velutina. Sen embargo, o noso obxectivo é estimar a densidade conxunta para determinar que zonas do mapa galego presentar concentracións de niños máis elevadas, co obxectivo de comprender mellor o seu patrón espacial. Para iso é preciso coñecer a latitude e a lonxitude simultaneamente polo que procedamos á análise conxunta dos datos.

Na Figura 3.2 móstranse dous histogramas distintos; empregando rectángulos paralelos ao eixos coordenados, o mesmo punto de anclaxe $(t_{0x}, t_{0y}) = (-9{,}2654, 41{,}8164)$ e, de esquerda a dereita, en cada unha das gráficas unha lonxitude dos rectángulos menor. Ningún dos histogramas permite afirmar nada sobre a modalidade dos datos: no histograma da esquerda existen varias barras con densidades elevadas situadas na zona central e nos extremos a densidade estimada é baixa; a escaseza de divisións dá lugar a que o histograma da dereita teña escasa interpretabilidade.

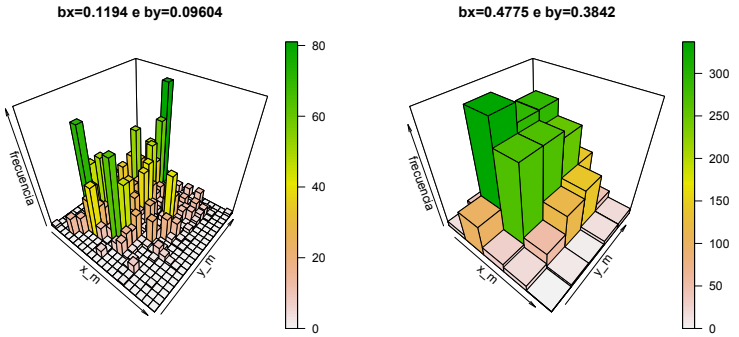

Figura 3.2: Dous histogramas bidimensionais dos datos dos niños de velutina en Galicia no ano 2018, para distintos b_x e b_y, sendo no gráfico da esquerda catro veces máis pequenos que no gráfico da dereita.

Neste caso, o procedemento de elección da lonxitude dos lados dos rectángulos é o seguinte: no primeiro dos histogramas agrupamos os datos en rectángulos de tal modo que o espazo mostral quedase dividido en 20 subintervalos no eixe de abscisas e outros 20 no de ordenadas, creando así unha malla 20×20; no segundo dos histogramas, as divisións reducíronse a 5 en ambos eixes. Este procedemento, sen base en ningún método matemático formulado en capítulos anteriores pero si na dependencia da lonxitude dos lados na forma dos histogramas, reflexa como varían os mesmos a medida que as celdas son de maior ou menor tamaño. Poderíase utilizar un método baseado na análise do $MISE$ do histograma como fixemos co estimador tipo núcleo pero dado que este último proporciona estimacións mellores, no contexto bidimensional só abarcamos o seu estudo dos criterios de erro, omitindo o do histograma.

Consideremos agora a vista aérea dos histogramas da Figura 3.2, é dicir, os mapas de calor da Figura 3.3, onde as variacións de cor e a lenda nos permiten omitir a represen-

tación das alturas das barras, para pasar a considerar gráficas no plano. Ademais, en cor negro engadiuse o contorno de Galicia. No diagrama da esquerda, observamos que as zonas de maior concentración de datos son a da Coruña e as Rías Baixas. A escaseza de divisións do diagrama da dereita, igual que ocorría co seu histograma asociado, impiden obter conclusións precisas. Neste último unicamente podemos afirmar que nas provincias de Coruña e Pontevedra parece haber concentracións máis elevadas que nas de Lugo e Ourense.

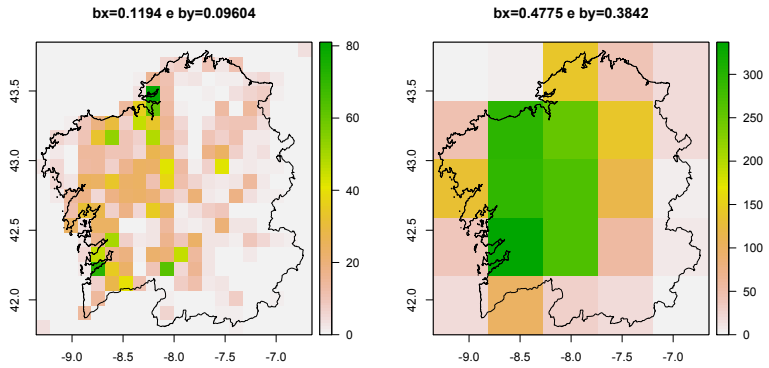

Figura 3.3: Dous mapas de calor dos datos dos niños de velutina en Galicia no ano 2018, relativos a cada un dos histogramas da Figura 3.2. Engadiuse o contorno de mapa galego en cor negro.

Unha vez analizada a distribución conxunta dos datos por medio do estimador histograma, é o momento de traballar co estimador tipo núcleo, dado que este último é mellor que o histograma en termos de suavidade, proporciona estimacións máis precisas e resolve o problema de escoller o punto de anclaxe. Ademais, por analoxía ao caso unidimensional, a orde de converxencia do estimador tipo núcleo é maior que a do histograma. Agora ben, cal é a mellor matriz de parámetros de suavizado para estimar esta densidade bidimensional?

Se nos centramos unicamente nas matrices diagonais, omitindo a orientación dos datos e a dependencia da lonxitude coa latitude, podemos considerar métodos de elección dos parámetros de suavizado h_1 e h_2 en cada unha das dúas variables. A esquerda da Figura 3.4 representamos as curvas de nivel xunto co diagrama de calor e o contorno do mapa galego da estimación tipo núcleo da densidade, e a dereita, o gráfico de perspectiva correspondente. Na parte superior consideramos a matriz de suavizado obtida co criterio de validación cruzada inesgada exposto na Sección 2.4 do Capítulo 2, e na inferior, a Regra do Polgar de Scott bidimensional. En ambos casos o núcleo empregado corresponde coa normal estándar.

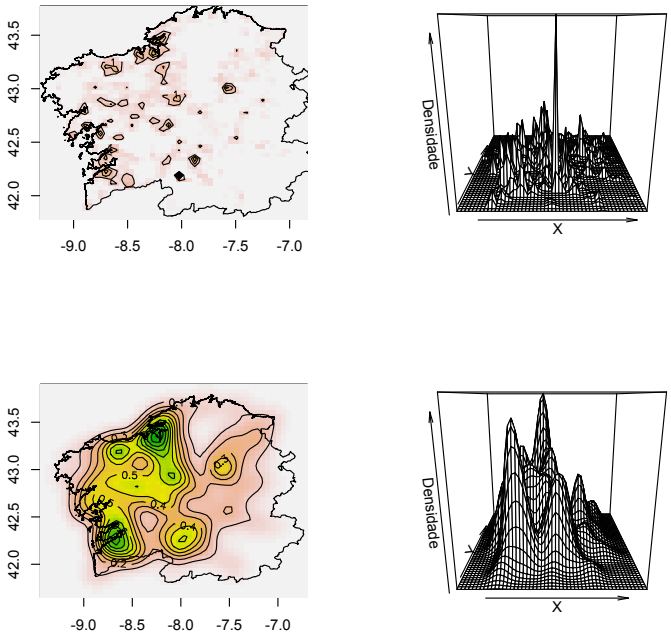

Figura 3.4: Na parte superior: a esquerda, curvas de nivel e diagrama de calor da estimación tipo núcleo das velutinas en Galicia no 2018 empregando como núcleo a normal estándar e como criterio de elección de H a validación cruzada inesgada, restrinxíndoa a ser diagonal; a dereita, gráfica de perspectiva asociada. Na parte inferior o cambio radica unicamente no criterio de elección da matriz de suavizado, que agora é a Regra do Polgar de Scott.

Os malos resultados que a validación cruzada inesgada proporcionan condúcennos a considerar outros métodos de elección dos valores h_1 e h_2 (elementos diagonais de H) como é a Regra do Polgar de Scott. A matriz de parámetros de suavizado, que recordemos que restrinximos a ser diagonal, en cada caso é:

$$H = \text{diag}(0{,}01154, 0{,}01070), \text{ e } H = \text{diag}(0{,}10263, 0{,}09530)$$

A estimación empregando a matriz obtida coa validación cruzada inesgada condúcenos a unha gráfica en perspectiva moi dentada, dado que os valores de h_1 e h_2 son moi pequenos. Como vimos no Capítulo 2, a interpretación dos resultados resulta case imposible como consecuencia da falta de suavidade. A Regra do Polgar proporciona estimacións máis axeitadas para estes datos, considerando case dez veces maiores os elementos diagonais da matriz de suavizado. Neste caso isto resolve os problemas da primeira estimación.
Observamos como no noroeste e suroeste do territorio galego a concentración de niños é

maior. Isto xa se apreciaba nos histogramas bidimensionais e a explicación reside en como chegou e se propagou a velutina en Galicia, como comentabamos ao comezo do capítulo.

Dado que dispoñemos de observacións procedentes da mostraxe realizada durante tres anos consecutivos, podemos preguntarnos cal foi a evolución temporal da distribución da densidade dos niños de velutinas no territorio galego. A anterior cuestión é de relevancia xa que, dende o punto de vista ecolóxico e social, existe un gran interese na evolución da expansión desta especie. Ademais, pola natureza da recollida de datos, é claro que hai unha marcada dependencia temporal dos mesmos.

Na Figura 3.5 represéntase a evolución estimada nos anos 2016, 2017 e 2018 da densidade dos niños de avespa velutina no mapa de Galicia. Para realizar esta representación gráfica, en primeiro lugar, dividíronse os datos mostrais en tres grupos segundo o ano no que foron tomadas cada unha das observacións. Posteriormente, calculouse o estimador da densidade tipo núcleo bivariante de cada un dos grupos, é dicir, empregando en cada caso os datos mostrais dos respectivos anos considerados. Como función núcleo tomouse a normal estándar bivariante e como criterio para escoller a matriz de parámetros de suavizado, a Regra do Polgar de Scott restrinxindo H a ser diagonal, por ser a validación cruzada inesgada un criterio inapropiado para o ano 2018 e, pola similitude dos datos, tamén para o resto dos anos. As matrices resultantes foron, respectivamente:

$$H = \text{diag}(0{,}09765, 0{,}11685),\ H = \text{diag}(0{,}08523, 0{,}09481),\ \text{e}\ H = \text{diag}(0{,}10263, 0{,}09530).$$

Unha vez calculados os tres estimadores tipo núcleo, procedeuse a súa representación gráfica. Dado que a mostra é bidimensional e os gráficos en perspectiva se atoparían en $\mathbb{R}^3$, para unha mellor visualización empregáronse os contornos de nivel, que dán lugar a gráficas en $\mathbb{R}^2$. De esquerda a dereita, e como ben indica a lenda, observamos os contornos de nivel desta estimación nos anos 2016, 2017 e 2018, respectivamente. Engadíronse os mapas de cor, onde a escala é a mesma que os da Figura 3.3, e o contorno de Galicia.

Observamos como, co avance do tempo, as velutinas se foron extendendo e ocupando case a totalidade do mapa galego, sendo dita expansión dende a costa -sobre todo a zona costeira noroeste e a suroeste- cara o interior. No ano 2016, a presenza de niños de velutina no interior, e en concreto na provincia de Ourense, é practicamente nula, seguindo a selo no ano 2017 e xa non no ano 2018 -a pesar de non chegar a ser tan elevada como nas zonas costeiras-. Na Figura 3.6 visualizamos as curvas de nivel das zonas con concentracións por encima do 50 %, correspondentes a lugares con alta presenza de niños de avespa velutina.

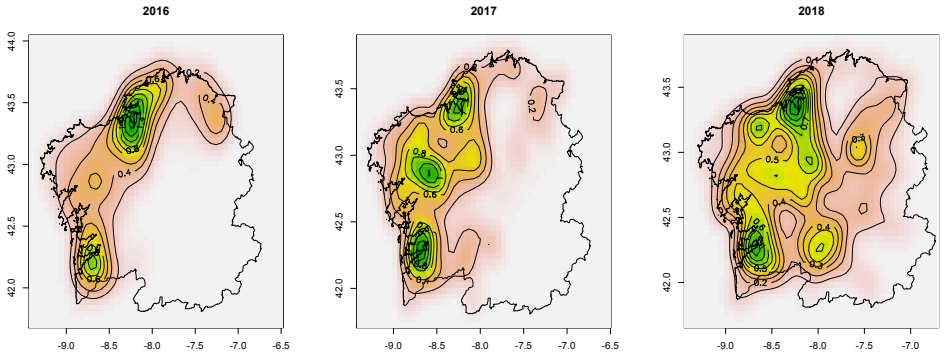

Figura 3.5: Evolución temporal da distribución da densidade dos niños de avespa velutina entre 2016 e 2018, ambos incluídos, empregando curvas de nivel e mapas de calor. A escala dos cores coincide coa dos mapas de calor da Figura 3.3.

No ano 2017 existen dúas zonas destacadas onde a concentración de velutinas é moito maior: a costa noroeste e as Rías Baixas. A explicación é que foi por eses lugares por onde entrou a velutina a Galicia no 2012. Por último, e como xa observamos na Figura 3.4, no ano 2018 as velutinas ocupan a maior parte de Galicia, relaxando a concentración de niños das dúas zonas destacadas no ano anterior. Do ano 2017 ao 2018 chama a atención o cambio de tendencia na comarca de Santiago pois, pasa a ser unha zona de alta densidade de niños a diminuír en gran medida esta concentración. Ademais, neste último ano, tamén é alta a densidade de niños no noroeste da provincia de Ourense.

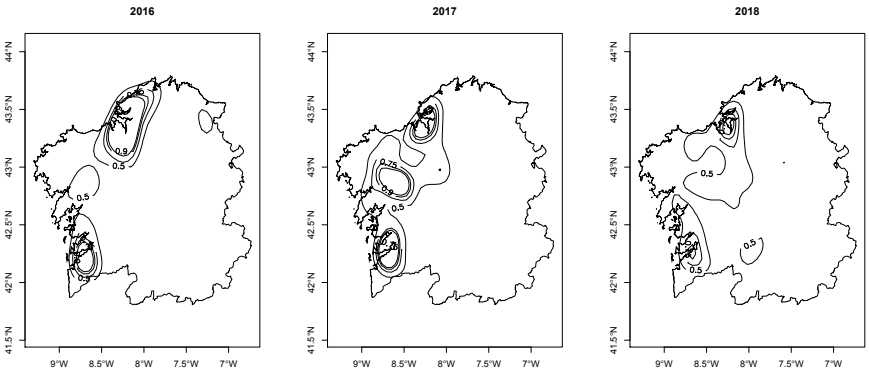

Figura 3.6: Curvas de nivel dos lugares con concentración de niños de velutina superior ao 50 % nos anos 2016, 2017 e 2018, xunto con contorno do mapa galego. Isto permite distinguir as zonas altamente densas.

Apéndice A

A distribución normal

A distribución normal é unha das distribucións máis coñecidas da estatística, publicada por Moivre (1733) e cuxo outro nome, distribución gaussiana ou campá de Gauss, se debe a Gauss (1823) debido a súa ampla contribución no estudo de propiedades desta.

Na maior parte dos exemplos simulados no Capítulo 1 empregamos mostras procedentes de distribucións normais ou de mesturas das mesmas. Esta familia constitúe un elemento fundamental para ilustrar calquera procedemento ou suceso no contexto da estimación da densidade. Ademais, vimos que a función de densidade normal estándar é unha boa elección de núcleo na estimación tipo núcleo, polas súas propiedades de regularidade que se transmiten á función estimada e porque a súa eficiencia é relativamente boa con respecto ao núcleo Epanechnikov. Por iso mesmo, e porque no Capítulo 2 tamén facemos uso da distribución normal, tanto univariante como bivariante, na estimación tipo núcleo, realizaremos un breve repaso desta. A función de densidade dunha variable aleatoria con distribución normal univariante $X \in N(\mu, \sigma^2)$ vén dada por

$$\phi_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \ x \in \mathbb{R}. \tag{A.1}$$

Deste xeito, no caso de que a normal sexa estándar, é dicir, $X \in N(0,1)$, tense que

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \ x \in \mathbb{R}. \tag{A.2}$$

Cómpre destacar que $X \in N(\mu, \sigma^2) \Leftrightarrow Z = \frac{X-\mu}{\sigma} \in N(0,1)$, onde Z é a estandarización univariante da variable aleatoria X. Ademais a función de densidade normal é simétrica entorno a súa media e dita variable toma valores en toda a recta real.

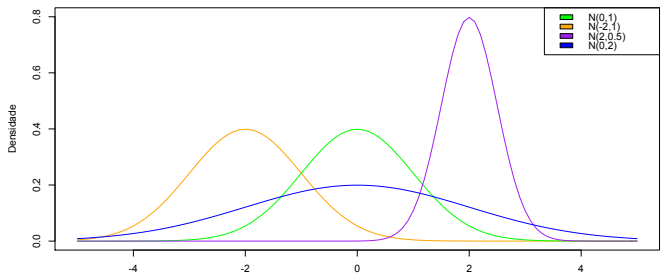

Figura A.1: Funcións de densidade de variables normais univariantes con distintas medias e varianzas. A curva verde correspóndese coa función de densidade normal estándar.

A función de densidade conxunta dun vector aleatorio con distribución normal bivariante $(X,Y) \in N_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{pmatrix}\right) = N_2(\overrightarrow{\mu}, \Sigma)$, con Σ matriz simétrica e semidefinida positiva vén dada por

$$\phi_{\overrightarrow{\mu},\Sigma}(x,y) = \frac{1}{2\pi}|\Sigma|^{\frac{-1}{2}} \exp\left(-\frac{(x-\mu_1, y-\mu_2)\Sigma^{-1}(x-\mu_1, y-\mu_2)'}{2}\right),\ (x,y) \in \mathbb{R}^2. \quad \text{(A.3)}$$

En concreto, se $(X,Y) \in N_2(\overrightarrow{0}, I_2 = \text{diag}(1,1))$ é unha normal estándar bivariante

$$\phi_{\overrightarrow{0},I_2}(x,y) = \frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right) = \phi(x)\phi(y),\ (x,y) \in \mathbb{R}^2. \quad \text{(A.4)}$$

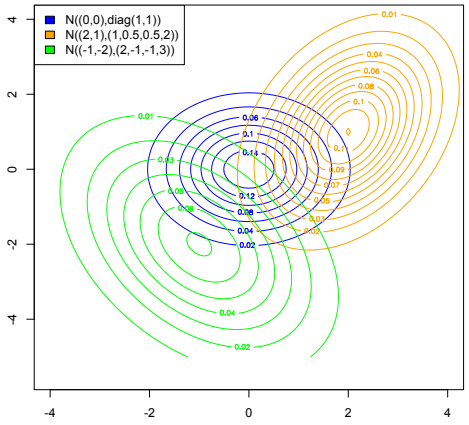

Figura A.2: Contornos de nivel de funcións de densidade conxunta de vectores normais para distintos vectores de medias e matrices de varianzas e covarianzas.

Apéndice B

Algunhas mesturas de densidades normais

As mesturas de densidades normais permítennos traballar cun gran abanico de densidades, creando modelos unimodais, bimodais, e multimodais. A súa vez, estes poden ser simétricos ou non e todos se obteñen realizando combinacións finitas de distribucións normais, onde os pesos suman 1, para que a función resultante sexa unha función de densidade. Como vemos ao longo do traballo, estes modelos resultan moi útiles para ilustrar o comportamento dos diferentes estimadores expostos. Na táboa B.1 recóllense as expresións das mesturas de densidades normais empregadas ao longo deste traballo, en concreto, no Capítulo 1. Gran parte dos modelos proceden do artigo de Marron and Wand (1992), conservando os seus nomes orixinais, e o resto creáronse co fin de ilustrar diversos fenómenos dos estimadores. Recordemos que $N(\mu_j, \sigma_j^2)$ é a normal de media μ_j e varianza σ_j^2.
As densidades de Marron e Wand empregadas corresponden aos modelos M1, M3 e M4.

Densidade	**Expresión:** $\sum_{j=1}^{k} \omega_j N(\mu_j, \sigma_j^2)$
M1: Gaussian	$N(0,1)$
M2: Asymmetric Trimodal	$\frac{1}{3}N(1,\frac{1}{2}) + \frac{1}{3}N(-1,\frac{1}{2}) + \frac{1}{3}N(-4,1)$
M3: Claw	$\frac{1}{2}N(0,1) + \frac{1}{10}\sum_{l=0}^{4} N(\frac{l}{2}-1, \frac{1}{100})$
M4: Bimodal	$\frac{1}{2}N(-1,(\frac{2}{3})^2) + \frac{1}{2}N(1,(\frac{2}{3})^2)$
M5: Another Asymmetric Trimodal	$\frac{1}{3}N(-1,\frac{1}{2}) + \frac{1}{3}N(1,\frac{1}{2}) + \frac{1}{3}N(3,\frac{7}{10})$

Cadro B.1: Algunhas densidades recollidas no artigo de Marron and Wand 1992 empregando mesturas de densidades normais, xunto con outras propias. Os nomes das densidades tomáronse do artigo orixinal, agás as de nova creación.

Bibliografía

Akaike, H. (1954). *An approximation to the density function.* Annals of the Institute of Statistical Mathematics.

Azzalini., A. y Bowman, A. (1990). *A look at some data on the Old Faithful geyser.* John wiley and Sons, Glasgow.

Bartle, R. G. y Sherbert, D. R. (2002). *Introducción al Análisis Matemático de una Variable.* Editorial Limusa, 2ª edition.

Bellman, R. (1961). *Adaptive control processes, A guided tour.* Information Theory, IEEE Transations on.

Billingsley, P. (1995). *Probability and Measure.* John Wiley and Sons, Canada, 3ª edition.

Bowman, A. (1984). *An Alternative Method of Cross-Validation for the Smoothing of Density Estimates.* Biometrika.

Chacón, J. E. y Duong, T. (2018). *Multivariate Kernel Smoothing and Its Applications.* Chapman and Hall.

Epanechnikov, V. (1969). *Non-Parametric Estimation of a Multivariate Probability Density.Teor. Veroyatnost. i Primenen.* Moscov.

Fix, E. y Hodges, J. (1951). *Discriminatory analysis, nonparametric es- timation: consistency properties.* USAF School of Aviation Medicine, Randolph Field, Texas.

Gauss, C. F. (1823). *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae.* Nabu Press (republicado o 1 de xaneiro do 2010).

Hughes, C. (1968). *On the mean accuracy of statistical pattern recognizers.* Information Theory, IEEE Transations on.

Marron, J. y Wand, M. (1992). *Exact mean integrated squared error.* The Annals of Statistics.

Maxima (2019). *Software Maxima.* Massachusetts Institute of Technology, United States.

Moivre, A. D. (1733). *La doctrina de posibilidades: un método de cálculo de las probabilidades de los sucesos en el juego.*

Moscoso, J. J. (2004). *Álgebra Lineal II [con aplicaciones en Estadśtica].* Universidad Nacional de Colombia, Bogotá.

Nieto, J. J. y Albés, I. M. (2017). *Variable Compleja.* Nino Centro de Impresión Digital S.L.

Parzen, E. (1962). *On estimation of a probability density function and mode.* Annals of Mathematical Statistics.

Pearson, K. (1891). *Maps and chartograms.* England.

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing.

Rosenblatt, M. (1956). *Remarks on some nonparametric estimatees of a density function.The Annals of Mathematical Statistics.*

Rudemo, M. (1982). *Empirical choice of histograms and kernel density estimators.* Scandinavian Journal of Statistics.Theory and Applications.

Sattinger, D. (2004). *Measure Theory and Integration.* Department od Mathematics, Yale University.

Scott, D. (1992). *Multivariate Density Estimation:Theory, Practice, and Visualization.* John Wiley and Sons, New Jersey, 2ª edition.

Scott, D. y Terrel, G. (1987). *Biased and Unbiased Cross-Validation in Density Estimation.* Journal of the American Statistical Association.

Shaughnessy, J. y Pfannkuch, M. (2002). *How faithful is Old Faithful? Statistical thinking: A story of variation and prediction.* Mathematics Teacher.

Sheather, S. y Jones, M. (1991). *A reliable data-based bandwidth selection method for kernel density estimation.* Journal of the Royal Statistical Society Series B 53.

Silvermann, B. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Wand, M. y Jones, M. (1995). *Kernel Smoothing.* Chapman and Hall, London, 1ª edition.