USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

MODELADO DE DATOS NO PLANO

Noa Hermida Costas

Xuño, 2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

MODELADO DE DATOS NO PLANO

Noa Hermida Costas

Xuño, 2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Estatística e Investigación Operativa
Título: Modelado de datos no plano
Breve descrición do contido
Os datos no plano representan localizacións nun mapa e a súa análise permite atopar posibles estruturas espaciais neles. Neste TFG trátase de facer unha introdución á estimación da densidade bidimensional. Os métodos serán ilustrados/aplicados coa súa correspondente análise de datos reais.
Recomendacións
Outras observacións

Índice

Resumo VIII

Introdución XI

1. Estimación da densidade multivariante 1

1.1. Fundamentos teóricos 1

1.2. Modelos simulados de referencia 4

1.3. Histograma 7

1.4. Estimador de tipo núcleo 11

1.4.1. Matriz de ancho de banda 14

1.4.2. Erro cadrático medio 17

1.4.3. Propiedades asintóticas 22

1.4.4. Anchos de banda óptimos 27

2. Selectores de ancho de banda 29

2.1. Selectores de escala normal 29

2.2. Validación cruzada 32

2.3. Plug-in 35

3. Análise de casos de leucemia 39

3.1. Estimación da densidade dos casos 40

3.2. Estimación da densidade dos controis 42

3.3. Conclusión 44

A. Notación e propiedades **47**

Bibliografía **49**

Resumo

Os datos espaciais representan localizacións xeográficas cuxa análise permite detectar posibles patróns e estruturas no espazo. Neste Traballo de Fin de Grao preséntase unha introdución á estimación non paramétrica da densidade bidimensional, centrada no uso do estimador tipo núcleo, co fin de obter representacións suaves da distribución espacial dos datos. Os diferentes métodos ilustraranse mediante a implementación de códigos en R, aplicados a datos reais de casos e controis de leucemia rexistrados no noroeste de Inglaterra. O obxectivo principal é identificar posibles agrupacións espaciais significativas que poidan contribuír á comprensión dos patróns epidemiolóxicos observados.

Abstract

Spatial data represent geographic locations whose analysis enables the detection of potential spatial patterns and structures. This Bachelor's Thesis presents an introduction to nonparametric bivariate density estimation, with a focus on the use of the kernel estimator in order to obtain smooth representations of the spatial distribution of the data. The different methods will be illustrated through the implementation of R code, applied to real data on leukemia cases and controls recorded in the northwest of England. The main objective is to identify significant spatial clusters that may contribute to the understanding of the observed epidemiological patterns.

Introdución

Os estudos epidemiolóxicos son investigacións científicas que analizan como as enfermidades afectan a distintas poboacións e que factores inflúen na súa aparición ou distribución, co obxectivo de determinar os riscos, patróns ou incluso posibles solucións que melloren a saúde poboacional ou preveñan problemas maiores. Existen dous tipos de estudos principais: os observacionais, que son aqueles que observan o que ocorre na poboación de maneira natural, e os experimentais, que realizan unha intervención inicial (por exemplo, unha vacina), e posteriormente observan os efectos que ten sobre a poboación. Neste traballo imos abordar a análise estatística dun estudo observacional de casos e controis, no cal se identifican pacientes que desenvolveron unha certa enfermidade e se compara a súa exposición ó factor de estudo coa dos controis, que son os referentes que non presentan a enfermidade pero que si comparten características similares en factores externos, como a idade, o sexo ou o nivel socioeconómico. O obxectivo é que a única diferenza significativa entre os casos e controis sexa a exposición ó factor de risco que se atopa en estudo, de forma que calquera cambio observado na incidencia da enfermidade se lle atribúa con seguridade a ese factor e non a outros sen determinar. Centrarémonos en observar as posibles diferenzas entre os patróns espaciais dos casos e controis.

Un dos estudos máis recoñecidos de epidemioloxía é o dos superviventes das bombas atómicas de Hiroshima e Nagasaki, realizado pola Radiation Effects Research Foundation (RERF) e publicado na revista *Radiation Research* (vol.137, nº 2). Nel, Preston et al. (1994) demostraron que existe unha relación directa entre a exposición á radiación e a incidencia de certas enfermidades como a leucemia. Outras investigacións mostraron un aumento significativo no risco de ter leucemia debido á exposición a produtos químicos durante a infancia, como pesticidas ou benceno (Environmental Health Perspectives, 2018).

A leucemia é unha enfermidade do sangue que se orixina na médula ósea. Caracterízase pola produción de células anormais que se acumulan e desprazan ás células sanguíneas sans, afectando na capacidade do organismo para combater infeccións, transportar oxíxeno ou coagular sangue. É o cáncer máis frecuente na poboación infantil, aínda que tamén se presenta en adultos (American Cancer Society, 2019).

As causas exactas da leucemia non están completamente definidas, pero considérase que interveñen unha combinación de factores xenéticos e ambientais. Ademais, existen diversos estudos epidemiolóxicos que suxiren a presenza doutro tipo de factores de risco, como a exposición a radiación ou a produtos químicos. Non obstante, non todos os resultados son concluíntes, polo que se requiren máis investigacións.

A falta de resultados decisivos acerca dos factores de risco da leucemia motivou a realización do estudo que presentamos neste Traballo de Fin de Grao (TFG), no que se recollen os casos de leucemia no noroeste de Inglaterra rexistrados entre 1982 e 1998. O obxectivo é ver se estes dependen da localización da súa residencia. Os datos foron extraídos da páxina web de Peter J. Diggle, na base de datos acerca do libro *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. Están compostos polas coordenadas (x, y) residenciais de tódolos casos de leucemia mieloide crónica (LMC) nun área delimitada de 100 km $\times$ 120 km, ademais de conter outras variables como a idade, o sexo e o índice de privación de Townsend, que mide o nivel socioeconómico da área de residencia.

A Figura 1 mostra a distribución espacial dos 233 casos de LMC no noroeste de Inglaterra e dos 988 controis realizados. Para a selección dos controis, utilizouse información do censo de 1991. Nel, foron tomados conteos da poboación nos 8131 distritos censais, que se utilizaron para obter unha mostra aleatoria estratificada de controis, con coordenadas asignadas según os respectivos centroides (lixeiramente desprazadas para evitar puntos coincidentes).

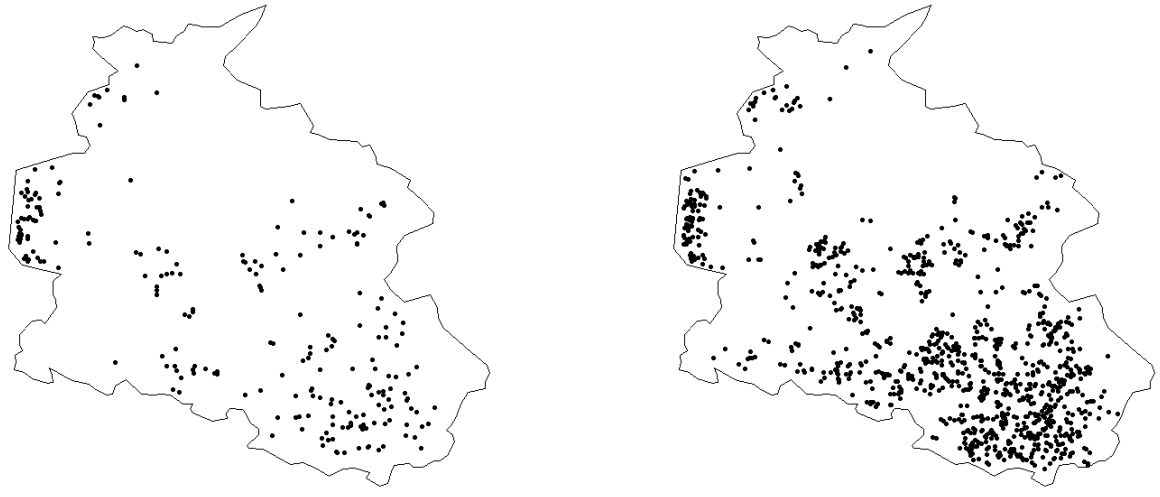

Figura 1: Ubicacións residenciais de 233 casos de leucemia no noroeste de Inglaterra entre 1982 e 1998 (á esquerda) e dos respectivos 988 controis (á dereita).

A distribución dos casos de leucemia podería ser aleatoria e non depender da súa localización. Nese caso sería esperable que a distribución espacial dos casos fose similar á da poboación reco-

llida na mostra de controis. Non obstante, a Figura 1 dos casos mostra unha alta concentración no sueste e na zona central da imaxe cara ó oeste, a pesar de existir casos dispersos por case toda a rexión. Isto podería indicar unha estrutura bimodal nos datos con dous grupos diferenciados. Unha moda é unha medida de tendencia central que representa o valor ou os valores máis frecuentes na distribución dos datos; en poboacións continuas, a moda considérase o punto ou os puntos onde a función ten un máximo. Aínda que esta estrutura bimodal suxire diferenzas entre o patrón espacial dos casos e dos controis, isto non é evidente mediante unha inspección visual. Porén, veremos na Sección 1.4 que os controis seguen unha distribución unimodal, enfocada na zona do sueste da imaxe, mentres que os casos teñen esa estrutura bimodal que mencionamos. Este feito anticipa a presenza dun posible patrón espacial na incidencia da enfermidade.

A alta incidencia de casos na zona do noroeste de Inglaterra suxire a posible influencia de factores ambientais, xeográficos ou socioeconómicos. Por un lado, a existencia de fábricas, industrias químicas ou centrais nucleares son factores importantes, xa que poden liberar substancias tóxicas ou emitir radiación. Ademais, os vertidos industriais ou agrícolas tamén contaminan a calidade da auga e doutros produtos. Por outro lado, as desigualdades socioeconómicas son outro factor a ter en conta, pois o mal acceso á atención médica ou a mala alimentación poderían ser moi influíntes.

Ó longo deste traballo presentamos as técnicas estatísticas que permiten analizar a distribución espacial dos datos. A estrutura deste TFG é a seguinte.

No Capítulo 1 tratamos a estimación da densidade, ilustrando as técnicas presentadas cos datos de LMC no noroeste de Inglaterra entre 1982 e 1998. Como descoñecemos a distribución orixinal dos datos, empregamos modelos non paramétricos, que van a permitir explorar patróns máis complexos e relacións espaciais sen a necesidade de facer suposicións previas sobre a distribución poblacional. En primeiro lugar, na Sección 1.3 introducimos o estimador de densidade non paramétrico máis simple, o histograma, que da unha representación intuitiva e visual da estrutura dos datos, pero non de todo precisa. Polo tanto, na Sección 1.4 presentamos o estimador de densidade de tipo núcleo e as súas propiedades estatísticas principais, como o erro cadrático medio integrado (MISE).

O Capítulo 2 aborda os distintos selectores de ancho de banda $\mathbf{H}$, unha serie de parámetros que condicionan o funcionamento do estimador de tipo núcleo visto no Capítulo 1. En primeiro lugar, na Sección 2.1 presentamos o selector de escala normal, que asume unha distribución normal para a estimación de $\mathbf{H}$. Esta suposición dá lugar, en xeral, a malas estimacións cando a estrutura dos datos non segue a dunha normal, polo que na Sección 2.2 explicamos o método de validación cruzada, un enfoque que selecciona unha matriz de ancho de banda que non depende da hipótese de normalidade. Por último, na Sección 2.3 tratamos o método plug-in, baseado na aproximación asintótica do MISE, que ten unha dependencia menor que o selector de escala nor-

mal da hipótese de normalidade, permitindo unha maior flexibilidade que este pero presentando un maior custo computacional.

O Capítulo 3 analiza con detalle os datos de leucemia no noroeste de Inglaterra, aplicando os conceptos teóricos desenvoltos ó longo do traballo. O obxectivo é avaliar a existencia dun patrón espacial na incidencia de leucemia, analizando e comparando as estimacións das densidades correspondentes ós casos e ós controis do estudo. O TFG finaliza cunha sección de conclusión onde se presentan os principais achados da análise de datos realizada.

Finalmente, o Anexo A recolle certas definicións e notacións empregadas ó longo do traballo, así como as propiedades principais dos operadores máis comúns tratados nas demostracións de resultados.

Capítulo 1

Estimación da densidade multivariante

Como dixemos na Introdución, a análise estatística é fundamental para a identificación de patróns na distribución dunha enfermidade. Existen dous enfoques principais: a estatística paramétrica e a non paramétrica. A diferenza é que a paramétrica asume que os datos seguen unha distribución específica e utiliza modelos predefinidos, como pode ser a normal, mentres que a non paramétrica non require de asumir unha forma concreta a priori. Polo tanto, asumimos modelos non paramétricos para o estudo dos casos de leucemia no noroeste de Inglaterra.

Ó longo deste capítulo estudamos dous estimadores non paramétricos moi utilizados na práctica. A estrutura do capítulo é como segue. Na Sección 1.1 presentamos o concepto de función de densidade, seguido na Sección 1.2 de modelos simulados que permitirán ilustrar os conceptos estudados. A continuación, na Sección 1.3 introducimos a estimación da densidade co histograma, un método máis sinxelo pero pouco preciso nalgúns casos. Finalmente, na Sección 1.4 describimos a estimación da densidade mediante o estimador tipo núcleo e analizamos as súas propiedades estatísticas principais. O desenvolvemento teórico inspírase principalmente no Capítulo 2 do libro *Multivariate Kernel Smoothing and Its Applications* (Chacon & Duong, 2018), complementado con achegas propias e adaptacións específicas ó contexto de estudo.

1.1. Fundamentos teóricos

Un concepto fundamental en estatística é a función de densidade de probabilidade. Consideremos unha variable aleatoria X cuxa función de densidade sexa f. Temos que f é unha función integrable e non negativa tal que

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Esta función describe como se distribúe a probabilidade ó longo dos posibles valores dunha variable aleatoria continua, de forma que para cada intervalo $[a, b]$, a probabilidade de que X tome un valor nel é

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

O concepto de función de densidade pódese estender ó caso multivariante, mantendo as mesmas propiedades de normalización e integrabilidade. Neste contexto, a función de densidade $f_{\mathbf{X}}$ describe a distribución de probabilidade dun vector aleatorio $\mathbf{X} = [X_1, ..., X_d]^\top$, de forma que a probabilidade de que $\mathbf{X}$ esté nun conxunto medible $A \subset \mathbb{R}^d$ vén dada como $P(\mathbf{X} \in A) = \int_A f_{\mathbf{X}}(\mathbf{x})d\mathbf{x}$. Esta integral debe entenderse como unha integral de Lebesgue respecto da medida de Lebesgue no espazo d-dimensional.

Na práctica é habitual ter unha mostra de datos observados $\mathbf{X}_1, ..., \mathbf{X}_n$ que proveñen dunha distribución ou función de densidade que se descoñece e se debe estimar (Silverman, 1986, Cap.1). Unha forma clásica e habitual de estimar unha densidade é mediante métodos paramétricos. Neste enfoque, supónse que os datos seguen unha distribución coñecida (p.e. a normal multivariante), e estímanse os seus parámetros a partir da mostra. Esta aproximación é moi útil cando se coñece de antemán que os datos se axustan ben ó modelo escollido, pero presenta moitos inconvenientes cando esta hipótese non se verifica.

Como dixemos, un modelo común é a normal multivariante, que vén dada pola expresión

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(\frac{-1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right), \tag{1.1}$$

onde $\boldsymbol{\mu} \in \mathbb{R}^d$ é o vector de medias e $\boldsymbol{\Sigma} \in \mathcal{M}_{d\times d}$ é a matriz de varianzas e covarianzas, simétrica e definida positiva, sendo $\mathcal{M}_{d\times d}$ o conxunto das matrices reais de dimensión $d \times d$. Se asumimos que temos unha mostra aleatoria simple $\mathbf{X}_1, ..., \mathbf{X}_n$ dunha poboación normal, con vector de medias $\boldsymbol{\mu}$ e matriz de varianzas e covarianzas $\boldsymbol{\Sigma}$, entón podemos estimar estes parámetros mediante o método de máxima verosimilitude, que consiste en atopar os valores de $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ que maximizan a función de verosimilitude (Shao, 2003). Esta función reflicte a probabilidade ou densidade de probabilidade que cada valor do parámetro outorga á realización mostral obtida, é dicir, mide como de compatible é cada un deles cos datos observados e identifica aqueles que mellor se axustan á mostra. Neste caso, as estimacións coinciden coa media e coa matriz de varianzas e covarianzas mostral:

- Media: $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i$
- Matriz de varianzas e covarianzas: $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{X}_i - \hat{\boldsymbol{\mu}})(\mathbf{X}_i - \hat{\boldsymbol{\mu}})^\top$

Substituíndo os estimadores na función de densidade da normal multivariante enunciada na

ecuación (1.1), chegamos a que a densidade estimada é a seguinte:

$$\hat{f}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\hat{\boldsymbol{\Sigma}}|^{1/2}} \exp\left(\frac{-1}{2}(\mathbf{x}-\hat{\boldsymbol{\mu}})^{\top}\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}})\right).$$

Consideremos agora a estimación da densidade dos datos de leucemia no noroeste de Inglaterra, supondo que seguen unha distribución normal multivariante. Vemos que só é necesario estimar dous parámetros co método de máxima verosimilitude, o vector de medias e a matriz de varianzas e covarianzas, para despois substituílos na función de densidade da normal. A Figura 1.1 mostra o axuste dos casos de leucemia. Na Figura 1 da Introdución vimos que a distribución dos casos é bimodal, é dicir, os datos presentan dous agrupamentos principais. Non obstante, a estimación paramétrica non é capaz de captar esta segunda moda, senón que asume unha estrutura elíptica e simétrica arredor da media, que ademais é a zona de máxima concentración estimada. Como podemos ver, esa zona non se corresponde coa rexión de maior concentración de casos. Esta rixidez estrutural dos modelos paramétricos motiva a busca doutros métodos máis flexibles, os métodos non paramétricos, que presentan maior capacidade de adaptación a estruturas que se descoñecen inicialmente. Por suposto, como veremos, esta flexibilidade tamén terá un prezo en termos do erro de estimación.

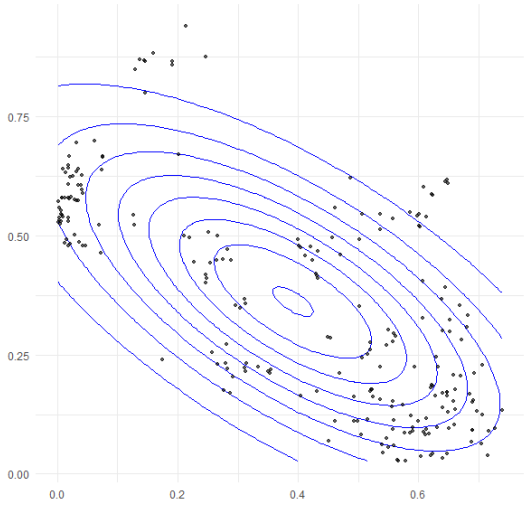

Figura 1.1: Estimación paramétrica baixo a hipótese de normalidade dos casos de leucemia no noroeste de Inglaterra, coa superposición da nube de puntos correspondente ós datos observados.

1.2. Modelos simulados de referencia

Para o estudo da estatística non paramétrica, é útil dispor de modelos controlados cuxa estrutura sexa coñecida, de forma que nos permitan verificar o bo comportamento dos estimadores e as súas limitacións. Neste traballo empregamos 3 distribucións simuladas en R, seleccionadas específicamente para comprobar os problemas xerados ó facer unha mala selección dos parámetros da estimación. Para a reprodución dos datos simulados imos empregar a semente *set.seed*(123).

Consideramos unha distribución normal multivariante en $\mathbb{R}^d$ da forma $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, onde $\mathbf{X}$ é un vector columna d-dimensional, $\boldsymbol{\mu} \in \mathbb{R}^d$ é o vector de medias e $\boldsymbol{\Sigma} \in \mathcal{M}_{d\times d}$ é a matriz de varianzas e covarianzas, simétrica e definida positiva.

(M1) Modelo normal multivariante:

$$\mathcal{N}_2\left(\begin{bmatrix}5\\5\end{bmatrix}, \begin{bmatrix}9 & 0\\0 & 9\end{bmatrix}\right).$$

Consideramos un modelo de referencia baseado na distribución normal bivariante M1 e simulamos 100 observacións en R. A varianza de ambas compoñentes é 9, mentres que a covarianza entre elas é nula, o que indica que non teñen correlación lineal. No caso da normal, isto é equivalente á independencia entre as compoñentes. Como a dispersión en ambas direccións é a mesma e a correlación é nula, os datos distribúense arredor do centro $(5, 5)$ formando contornos aproximadamente circulares, cuxa dispersión está determinada pola desviación típica, 3. Na Figura 1.2 representamos a mostra simulada de puntos e a densidade teórica, que nos servirán para ter unha idea intuitiva da forma da distribución cando a estimemos mediante métodos non paramétricos.

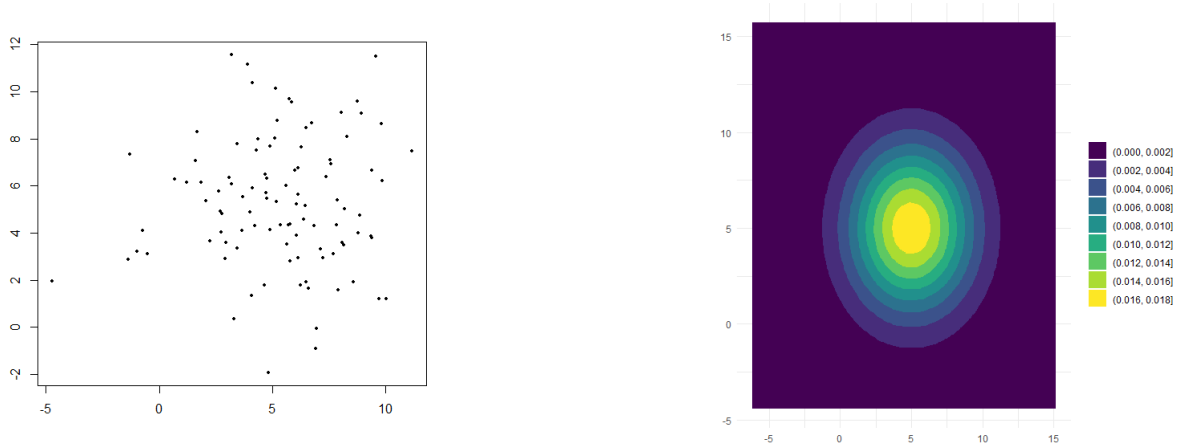

Figura 1.2: Mostra simulada de 100 datos do modelo M1 (esquerda) coa respectiva densidade teórica da distribución (dereita).

Este modelo será útil para ilustrar as limitacións do histograma multidimensional derivadas da elección do punto de anclaxe e do ancho de banda, que veremos na Sección 1.3, ademais de para comparar as estimacións de tipo núcleo con distintas matrices de ancho de banda, que abordaremos na Sección 1.4.

Observación 1.1. A densidade nos gráficos represéntase mediante un mapa de calor, empregando o paquete *ggplot2* de R e a función *geom_tile()*. Nos exemplos facemos uso da escala *viridis* de R, que asigna cores azuis oscuras a densidades baixas e cores amarelas a valores altos.

(M2) Mixtura bimodal de normais univariantes:

$$\frac{1}{2}\mathcal{N}(0,1) + \frac{1}{2}\mathcal{N}(3,1).$$

Sexa M2 outro modelo de referencia dado pola combinación equitativa de dúas distribucións normais unidimensionais con distinta media pero mesma varianza. Simulamos unha mostra de 1000 datos e representamos na Figura 1.3 o conxunto de datos á esquerda, e a densidade teórica á dereita. Observamos que a distribución é bimodal, pois a densidade presenta dous picos simétricos centrados en 0 e 3, cunha depresión entre eles. Isto débese a que cada unha das normais ten o seu máximo na media, e como están separadas o suficiente unha da outra, non se solapan.

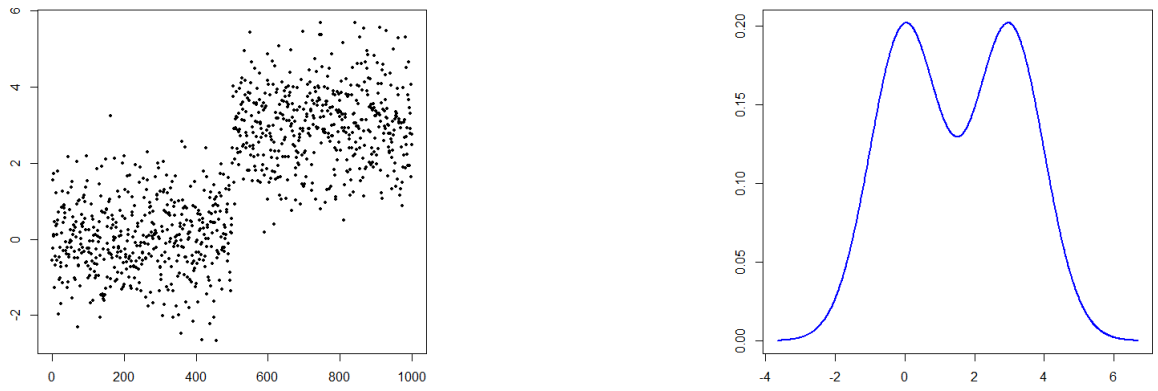

Figura 1.3: Mostra simulada de 1000 datos do modelo M2 (á esquerda) coa súa respectiva densidade teórica (á dereita). O eixe X representa o índice de cada punto e o eixe Y o valor da mostra simulada.

Este modelo permitirá ver a problemática da estimación do histograma en función do ancho de banda seleccionado.

(M3) Mixtura bimodal de normais bivariantes:

$$\frac{3}{8} \cdot \mathcal{N}_2 \left(\begin{bmatrix} 2 \\ 6 \end{bmatrix}, \begin{bmatrix} 1 & 0{,}5 \\ 0{,}5 & 1 \end{bmatrix} \right) + \frac{5}{8} \cdot \mathcal{N}_2 \left(\begin{bmatrix} 6 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & -0{,}3 \\ -0{,}3 & 1 \end{bmatrix} \right).$$

Simulamos unha mostra de 400 datos pertencentes ó modelo M3. Na Figura 1.4 representamos o conxunto de puntos e a correspondente densidade teórica. Observamos que a mostra é bimodal, posto que presenta dous clústers moi diferenciados. O agrupamento superior presenta unha tendencia lineal positiva, mentres que o inferior ten unha estrutura un pouco máis dispersa.

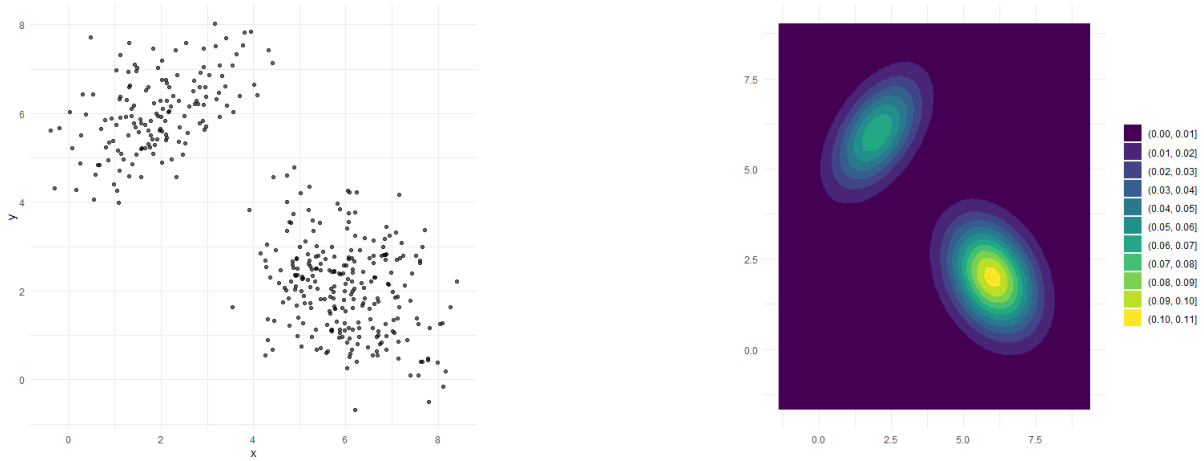

Figura 1.4: Mostra simulada de 400 datos do modelo M3 (á esquerda) coa súa respectiva densidade teórica (á dereita).

Este modelo de referencia empregarémolo para estudar a relación entre o nesgo e a varianza no estimador de tipo núcleo en función da matriz de ancho de banda. Ademais, a estrutura bimodal do exemplo dificultará a estimación e permitirá ilustrar os principais problemas asociados.

Ó longo deste traballo empregamos os tres modelos simulados para mostrar distintos aspectos e dificultades que xorden na estimación da función de densidade, como a sensibilidade do ancho de banda e do punto de anclaxe no histograma. Deste xeito será posible comparar o comportamento das estimacións en modelos controlados, o que facilitará a discusión e análise dos resultados. Pasamos agora a definir os estimadores non paramétricos da densidade.

1.3. Histograma

O primeiro método non paramétrico empregado para estimar densidades foi o histograma. Aínda que xa existían representacións gráficas similares no século XIX, Scott (2015) sinala que a súa formalización foi realizada por Pearson (1891), quen nomeu e definíu rigurosamente o histograma como unha ferramenta estatística (non só gráfica) que representaba frecuencias en intervalos (bins). Ademais, tamén mencionou que existían limitacións, como a elección do ancho de banda, do punto de anclaxe e a súa aparencia escalonada. Unha importante aportación de Pearson foi a normalización das alturas das barras para que a área total baixo o histograma fose 1. Esta simplicidade converte ó histograma no método non paramétrico máis sinxelo e intuitivo para a estimación da densidade, pero non no máis preciso.

No caso unidimensional, supónse unha mostra $X_1, ..., X_n$ de variables continuas con función de densidade común f. O histograma constrúese dividindo o eixe da variable X en intervalos disxuntos $I_1, ..., I_m$ de ancho b, de forma que cobren todo o rango das observacións, e contabilizando o número de datos N_j que caen en cada un, sendo por construción $\sum_{j=1}^{m} Nj = n$ o número total de observacións. Polo tanto, o estimador é unha función escalonada na que as alturas representan a proporción da mostra en cada intervalo dividido polo seu ancho (Wand & Jones, 1995, p.5). Entón, a estimación nun punto x pertencente ó intervalo I_j vén dada por

$$\hat{f}(x; b) = \frac{N_j}{n \cdot b}.$$

O ancho de banda b dos intervalos ten un papel fundamental. Para ilustrar como afecta na estimación da densidade co histograma, consideramos o modelo M2 e representamos os histogramas unidimensionais resultantes para dous anchos de banda distintos, $b = 2$ e $b = 0.2$, respectivamente. A Figura 1.5 mostra as estimacións correspondentes. Nela, observamos que un tamaño grande de b suaviza en exceso a distribución e perde precisión, pois neste caso o histograma non mostra a estrutura bimodal do exemplo. Por outro lado, un ancho moi pequeno xera un histograma moi fragmentado, con moitas subidas e baixadas esporádicas, e polo tanto con moito ruído, que non está presente no modelo poboacional. Ademais, a posición dos intervalos tamén é crucial para unha boa estimación, pois un mal punto de inicio pode afectar á estimación, como veremos con máis detalle no caso multidimensional.

Este método pode estenderse ó caso de dimensións superiores o dividir o espazo en celdas multidimensionais, en lugar de intervalos, xerando un estimador baseado na frecuencia de puntos en cada celda. Sexa $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ unha mostra aleatoria d-dimensional extraída dunha distribución de probabilidade multivariante con función de densidade f, onde cada $\mathbf{X}_i = [X_{i1}, X_{i2}, ..., X_{id}]^\top \ \forall i \in \{1, ..., n\}$ representa unha observación en $\mathbb{R}^d$.

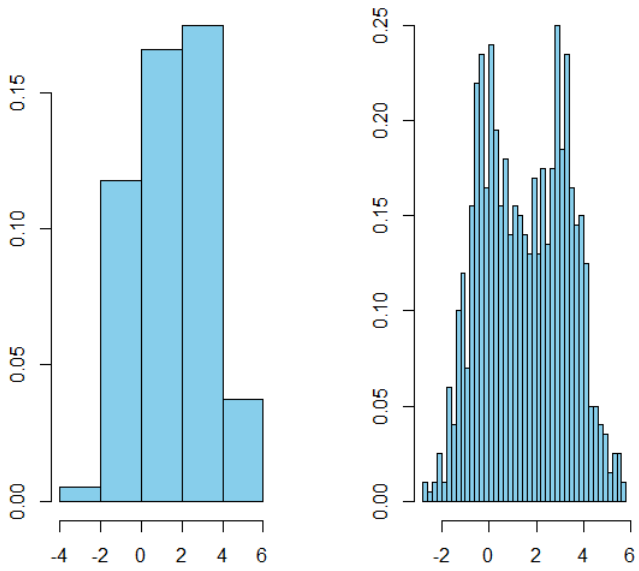

Figura 1.5: Histogramas dunha mostra de 1000 datos procedentes ó modelo M2. A figura da esquerda presenta un ancho de banda $b = 2$ e a da dereita $b = 0.2$.

O histograma multidimensional é un estimador non paramétrico de f que divide o espazo $\mathbb{R}^d$ en caixas $B_1, ..., B_m$ disxuntas que cobren o rango das observacións, onde cada unha representa unha rexión do espazo d-dimensional de volume constante e finito. Concretamente, as caixas son hiperrectángulos en $\mathbb{R}^d$, é dicir, unha xeneralización dos rectángulos que se forman mediante o produto de intervalos rectangulares para as dimensións superiores. Para cada rexión B_j do espazo, calcúlase, igual que no caso unidimensional, a proporción de datos que contén no seu interior. Ademais, a súa medida ó longo da i-ésima coordenada vén dada por b_i, que representa a amplitude do intervalo nesa dimensión. Por exemplo, para $d = 2$, podemos definir unha caixa xenérica como $B = (x, x + b_1] \times (y, y + b_2]$. En xeral, o vector de ancho de banda $\mathbf{b} = [b_1, ..., b_d]^\top$ define o tamaño das caixas en todas as dimensións de $\mathbb{R}^d$.

Polo tanto, a estimación nun punto $\mathbf{x} \in \mathbb{R}^d$ pertencente a B_j está dada por

$$\hat{f}_{hist}(\mathbf{x}; \mathbf{b}) = \frac{N_j}{n \cdot b_1 \cdots b_d},$$

onde $N_j = \sum_{i=1}^{n} \mathbf{1}\{\mathbf{X}_i \in B_j\}$ é o número de observacións na caixa B_j. Por construción, $\sum_{j=1}^{m} N_j = n$.

Como xa mencionamos no caso unidimensional, outro parámetro fundamental xunto co ancho de banda $\mathbf{b}$ é o punto de anclaxe, un vector d-dimensional que define o inicio da partición das

caixas B_j en cada unha das dimensións. A elección destes parámetros é crucial para garantir unha estimación precisa da distribución dos datos, xa que unha mala selección do ancho de banda, ó igual que acontecía en dimensión 1, pode introducir demasiado ruído ou ocultar patróns, mentres que unha mala selección do punto de anclaxe pode variar moito a estimación.

Para ilustrar os efectos destes parámetros, imos considerar o modelo simulado M1, procedente dunha distribución normal bidimensional con vector de medias $\boldsymbol{\mu} = [5, 5]^\top$ e matriz de varianzas e covarianzas $\boldsymbol{\Sigma} = 9\mathbf{I}_2$. Este modelo presenta unha densidade suave, continua, simétrica e unimodal, de forma que permite detectar fácilmente as deformacións causadas por unha mala elección dos valores de ancho de banda e punto de anclaxe. Na Figura 1.6 temos representados dous histogramas con distinto ancho de banda $\mathbf{b}$. Imos considerar o ancho de banda de escala normal para a i-ésima dimensión, sinalado en Chacon e Duong (2018, p.13), dado por

$$b_{\mathrm{NS},i} = 2 \cdot 3^{1/(d+2)} \cdot \pi^{d/(2d+4)} \cdot s_i \cdot n^{-1/(d+2)},$$

onde d é a dimensión do conxunto de datos, n o tamaño da mostra e s_i a desviación típica para a coordenada i. No noso exemplo, temos que $\mathbf{b}_{\mathrm{NS}} = [b_{\mathrm{NS},1}, b_{\mathrm{NS},2}]^\top = [3.21, 3.03]^\top$. O histograma da esquerda presenta un ancho de banda pequeno, $\mathbf{b}_{\mathrm{NS}}/2$, que divide o espazo en caixas considerablemente máis pequenas, proporcionando unha estimación máis detallada pero con moita variabilidade. Pola contra, a figura da dereita ten un ancho de banda maior, $2 \cdot \mathbf{b}_{\mathrm{NS}}$, o que dá lugar a unha estimación máis suave pero que oculta información, xa que por exemplo non capta a forma simétrica e suave da densidade normal.

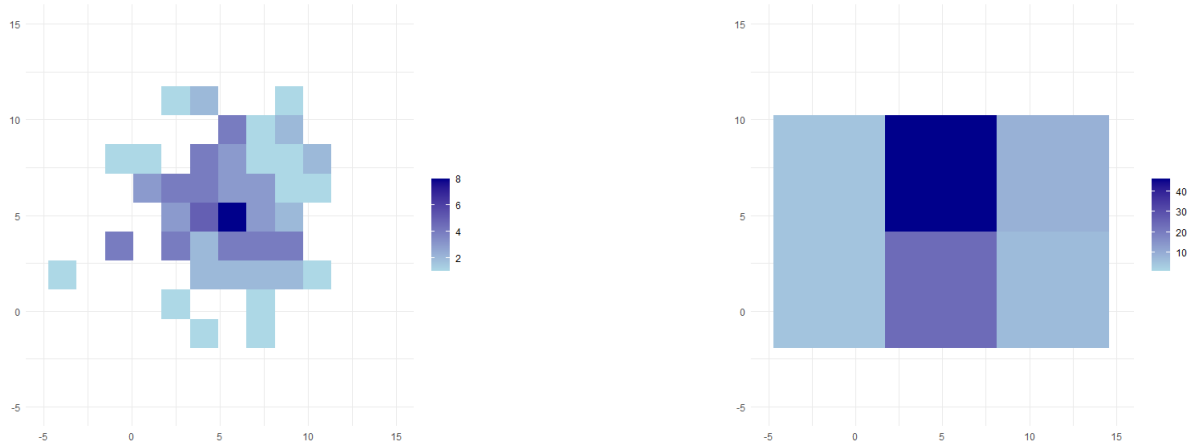

Figura 1.6: Comparación de histogramas con distintos anchos de banda do modelo M1, sendo $\mathbf{b} = [3.21, 3.03]^\top$. O histograma da esquerda utiliza un ancho de banda $\mathbf{b}/2$, mentres que o da dereita emprega $\mathbf{b} \cdot 2$.

Ademais do problema de suavizado da estimación ó aumentar ou diminuír $\mathbf{b}$, unha mala elección do mesmo pode dar lugar a outros inconvenientes, como unha localización errónea das

modas. Na Figura 1.7 realizamos a estimación do histograma dos casos de leucemia no noroeste de Inglaterra tomando como anchos de banda o de escala normal $\mathbf{b} = [0.22, 0.20]^\top$ e $\mathbf{b}/2$, respectivamente. No histograma da esquerda, a moda está localizada no suroeste do gráfico, mentras que o da dereita presenta a súa moda principal no noroeste. Esta variación no ancho de banda da lugar a dúas interpretacións moi diferentes da distribución que seguen os datos, indicando que polo menos unha delas é errónea.

Figura 1.7: Estimación do histograma dos casos de leucemia no noroeste de Inglaterra con distintos anchos de banda. O histograma da esquerda emprega $\mathbf{b} = [0.22, 0.20]^\top$ e o da dereita $\mathbf{b}/2$.

Por outra parte, a Figura 1.8 mostra a representación da estimación do modelo M1 mediante dous histogramas con ancho de banda $\mathbf{b}_{\text{NS}} = [3.21, 3.03]^\top$ desprazando o seu punto de anclaxe, o cal determina como se distribúen as caixas ó longo do espazo. Polo tanto, a súa variación pode influír na percepción das modas da distribución. No histograma da esquerda consideramos como punto de anclaxe o dado polo mínimo valor dos datos en cada coordenada, $[-4.72, -1.92]^\top$. Os datos aparecen concrentrados no centro do histograma, próximos á media da distribución, mentres ao redor a densidade diminúe a medida que se alonxa do centro. Esta estrutura aseméllase a densidade da normal teórica que vimos na Figura 1.2, o que suxire que podería ser un bo punto de anclaxe para estes datos. Na Figura 1.2 tamén vimos que só existe un dato simulado da mostra próximo a este punto de anclaxe, o que suxire que poderíamos considerar outro máis cercano ó conxunto de datos simulados. Polo tanto, no histograma da dereita tomamos o punto $[-1.2, 2.1]^\top$. Neste caso, a estimación presenta dúas zonas con alta densidade, suxerindo unha posible bimodalidade, que non pode ser posible xa que coñecemos de antemán que a distribución real é unha normal.

Este exemplo mostra a importancia de realizar unha boa elección do punto de anclaxe e do ancho de banda á hora de estimar a función de densidade co histograma, pois unha pequena variación pode dar lugar a resultados moi diferentes e incluso contraditorios. Estas son algunhas

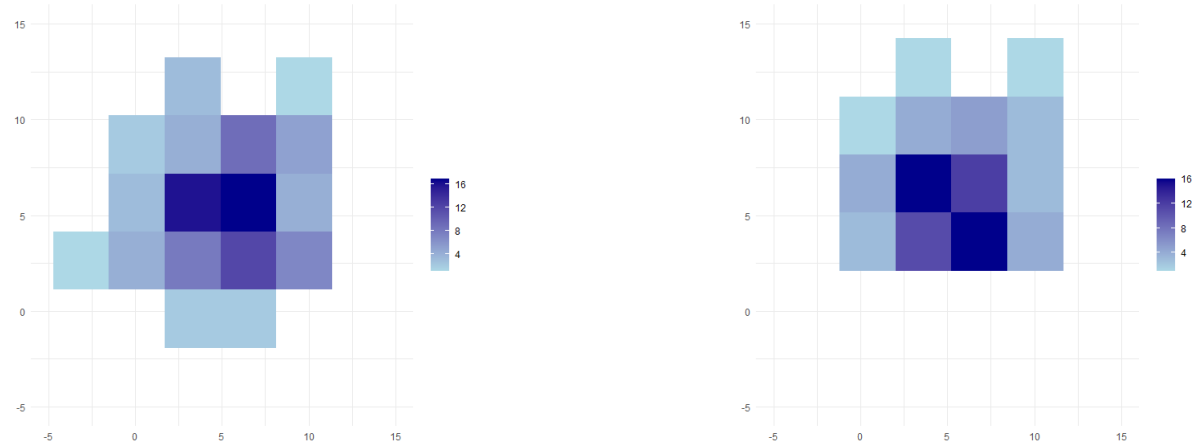

Figura 1.8: Estimación de dous histogramas con distintos puntos de anclaxe para o modelo M1. O histograma da esquerda utiliza $[-4.72, -1.92]^\top$, mentres que o da dereita emprega $[-1.2, 2.1]^\top$. Ambos gráficos teñen como ancho de banda $\mathbf{b}_{\text{NS}} = [3.21, 3.03]^\top$.

das limitacións do histograma, ademais da súa apariencia escalonada e discontinua.

Para mellorar a estimación da densidade, propuxéronse técnicas como os histogramas desprazados promediados (Average Shifted Histograms, ASH), que en lugar de utilizar un único conxunto de caixas d-dimensionais, realiza un promedio de varios histogramas con distinto punto de anclaxe (Scott, 2015, Cap.5). Este método elimina o problema da elección do punto de anclaxe, pero segue a ter limitacións, como a elección do ancho de banda e que a estimación resultante tampouco é suave. Ademais, engádese outro parámetro, m, que se corresponde co número de histogramas promediados e que controla o equilibrio entre o suavizado e o detalle. Aínda que a elección precisa de m non é crítica sempre que $m \geq 3$, según indica Scott (2015, p.133), é interesante notar que cando $m \to \infty$, o ASH converxe a un estimador de tipo núcleo, que xa non depende da elección deste parámetro m. Por conseguinte, na Sección 1.4 introducimos o estimador de tipo núcleo, que proporciona unha estimación da densidade suave e continua, dando unha visión máis precisa da distribución dos datos.

1.4. Estimador de tipo núcleo

O estimador de tipo núcleo é un dos métodos non paramétricos máis empregados para a estimación de densidades de variables aleatorias continuas a partir dunha mostra de datos. A diferenza do histograma d-dimensional, que divide os datos en caixas, o estimador de tipo núcleo estima a densidade en cada punto utilizando unha función de suavización K que pondera as observacións cercanas a cada un.

No caso unidimensional, consideramos K a función núcleo, que suporemos simétrica, non negativa e integrable, e que verifica $\int_{\mathbb{R}} K(x)dx = 1$. Tomamos un conxunto de observacións $X_1, X_2, ..., X_n$ procedentes dunha densidade continua f, e consideramos $h > 0$ o ancho de banda. Temos que a densidade nun punto x estímase como a suma ponderada da contribución de tódolos datos, onde o peso de cada un depende da súa distancia a x e da función núcleo. Entón,

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

onde K_h é a densidade da variable $h \cdot Y$ onde Y ten densidade K. Por exemplo, se K é o núcleo gaussiano estándar, entón K_h é unha normal de media cero e varianza h^2. Así, h actúa como un factor de escala que determina a dispersión do núcleo. No caso particular de que K sexa a densidade U[-1,1], vemos que $\hat{f}(x)$ é moi similar a un histograma, neste caso con intervalo centrado no punto x e con ancho de banda $b = 2h$. Convén indicar que, se K é unha función suave, entón $\hat{f}$ tamén o será.

Na Figura 1.9 representamos a estimación da densidade de tipo núcleo construída sobre cinco observacións aleatorias, tomando $h = 0.7$ e empregando a función núcleo dada pola distribución normal $\mathcal{N}(0,1)$, é dicir, $K(x) = (2\pi)^{-1/2}\exp\{-x^2/2\}$. As liñas discontinuas representan unha función kernel gaussiana escalada por $\frac{1}{5}$ e centrada en cada unha das observacións. Finalmente, a estimación total promedia o efecto das estimacións individuais e representa a densidade estimada.

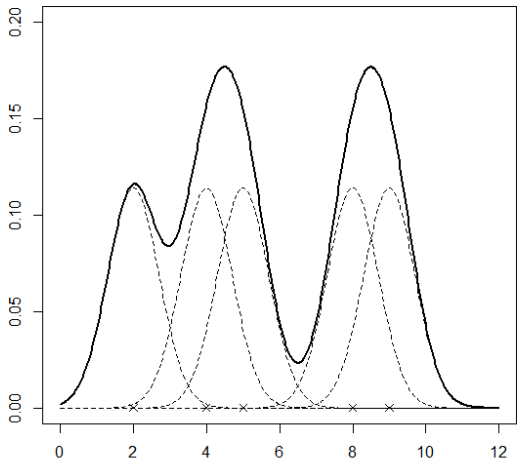

Figura 1.9: Estimación de tipo núcleo unidimensional baseada en 5 observacións. As liñas discontinuas mostran a función kernel gaussiana escalada e a liña negra a densidade estimada.

Análogamente ó que vimos na Sección 1.3 do histograma, a elección do ancho de banda h na estimación de tipo núcleo é moi importante. Un h máis pequeno mostrará con máis detalle a estimación, pero dará lugar a máis ruído, mentre que un h moi grande pode perder información relevante. Veremos con máis detalle, no Capítulo 2, a obtención a partir dos datos do óptimo no caso multidimensional, que requerirá da especificación de máis parámetros de ancho de banda. En calquera caso, agora xa non é necesario escoller un punto de anclaxe inicial, que era unha das grandes limitacións do histograma.

O estimador de densidade de tipo núcleo no caso multidimensional xeneraliza o caso anterior, tomando unha función núcleo multidimensional, que seguiremos denotando por K, e unha matriz $\mathbf{H}$ de ancho de banda en lugar dun escalar. Consideremos ó longo desta sección unha mostra aleatoria d-dimensional $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ con función de densidade f. Denotamos como $\mathbf{X}_i = [X_{i1}, X_{i2}, ..., X_{id}]^\top$ ás compoñentes de cada variable $\mathbf{X}_i$, tomamos como vector xenérico $\mathbf{x} \in \mathbb{R}^d$ dado por $\mathbf{x} = [x_1, x_2, ..., x_d]^\top$, e consideramos a matriz identidade $\mathbf{I}_d$ de dimensión $d \times d$.

Definimos o estimador de densidade de tipo núcleo nun punto $\mathbf{x}$ como

$$\hat{f}(\mathbf{x}; \mathbf{H}) = \frac{1}{n} |\mathbf{H}|^{-1/2} \sum_{i=1}^{n} K(\mathbf{H}^{-1/2}(\mathbf{x} - \mathbf{X}_i)), \tag{1.2}$$

onde o núcleo K é unha función de densidade d-dimensional e $\mathbf{H}$ é a matriz de ancho de banda, simétrica, definida positiva e de dimensión $d \times d$. A matriz $\mathbf{H}$ controla a orientación e a extensión do suavizado a través de $K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x})$, onde $|\mathbf{H}|$ é o determinante da matriz. Isto débese, por unha parte, a que o termo $|\mathbf{H}|^{-1/2}$ controla o tamaño global do suavizado, é dicir, canto máis pequeno sexa, máis localizada é a distribución. Por outro lado, os termos que non pertencen á diagonal de $\mathbf{H}$ son os que controlan a orientación, pois reflexan a correlación entre as variables. Temos entón que

$$\hat{f}(\mathbf{x}; \mathbf{H}) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i), \tag{1.3}$$

onde $K_{\mathbf{H}}$ é a densidade de $\mathbf{H}^{1/2} \cdot \mathbf{X}_K$ do vector $\mathbf{X}_K$ con densidade K. Ó longo deste traballo, consideraremos que K é suave (infinitamente diferenciable, de clase $\mathcal{C}^\infty$), unimodal e esféricamente simétrica, é dicir, o seu valor depende da distancia euclídea $||\cdot||$ de $\mathbf{x}$ á orixe no espazo $\mathbb{R}^d$ ($K(\mathbf{x}) = k(||\mathbf{x}||)$, $k : [0, \infty) \to \mathbb{R}$). A suavidade do kernel garante que a estimación $\hat{f}$ sexa regular.

Dado o estimador da ecuación (1.3), temos que $\mathbf{K}_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i)$ representa o peso de cada observación, que é maior cando os puntos están máis próximos a $\mathbf{x}$. Ademais, cada función núcleo $\mathbf{K}_{\mathbf{H}}(\cdot - \mathbf{X}_i)$ é unha densidade de probabilidade centrada en cada dato $\mathbf{X}_i$, e a súa suma proporciona unha estimación suave da densidade. Chacón e Duong (2018, p.14) sinalan que o estimador $\hat{f}$ pode interpretarse de dúas formas distintas: para un punto $\mathbf{x}$, como un promedio

ponderado local onde o peso de $\mathbf{X}_i$ diminúe cando aumenta a súa distancia a $\mathbf{x}$; e, por outro lado, como unha masa de probabilidade que se suaviza na vecindade local, xerando un modelo global a través dunha mixtura destas probabilidades locais suavizadas.

Existen diversos núcleos utilizados na práctica. O máis habitual é o kernel gaussiano, que usaremos na nosa análise de datos, que escalado e trasladado vén dado por

$$K_{\mathbf{H}}(\mathbf{x}-\mathbf{X}_i) = \frac{1}{(2\pi)^{d/2}}|\mathbf{H}|^{-1/2}\exp\{-\frac{1}{2}(\mathbf{x}-\mathbf{X}_i)^\top \mathbf{H}^{-1}(\mathbf{x}-\mathbf{X}_i)\},$$

que é unha densidade normal con vector de medias $\mathbf{X}_i$ e con matriz de varianzas e covarianzas $\mathbf{H}$. Este é un dos motivos polo que se parametriza $\mathbf{H}$ na escala dos datos ó cadrado no caso d-dimensional, é dicir, como matriz de varianzas, garantizando que o termo cadrático na exponencial sexa adimensional e o ancho de banda sexa dimensionalmente consistente coa dispersión dos datos. Deste xeito, $\mathbf{H}$ reflicte a dispersión e a correlación dos datos.

1.4.1. Matriz de ancho de banda

Para obter un bo estimador de tipo núcleo da función de densidade, a elección da matriz de ancho de banda é de gran importancia, pois controla a forma, a orientación e a suavidade do estimador no espazo multivariante. Existen diversas clases de matrices, pero neste traballo ímonos enfocar nas principais.

A opción máis simple é considerar matrices diagonais. Por un lado, están as pertencentes ó grupo $\mathcal{A} = \{h^2\mathbf{I}_d : h > 0\}$, sendo h un escalar, que poden chegar a ser moi restritivas se as variables teñen distintas escalas ou dependencia entre elas. O estimador redúcese a

$$\hat{f}(\mathbf{x};h) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right).$$

Por outro lado, están as pertencentes ó conxunto $D = \{\text{diag}(h_1^2, h_2^2, \dots, h_d^2) \ : \ h_1, \dots, h_d > 0\}$, onde h_j é o ancho de banda para a dimensión j, o que permite unha suavización diferente en cada dimensión. Esta forma é útil cando as escalas en cada dimensión son moi distintas. Non obstante, segue sen captar as posibles correlacións entre as variables.

O grupo máis xeral son as matrices sen restricións, simétricas e definidas positivas $\mathcal{F} = \{\mathbf{H} \in \mathcal{M}_{d\times d} : \mathbf{H} \succ 0, \mathbf{H} = \mathbf{H}^\top\}$, onde $\mathbf{H} \succ 0$ denota a matriz definida positiva. Esta clase permite captar as correlacións entre as diferentes dimensións dos datos, a diferenza das anteriores, permitindo adaptarse mellor á orientación. Non obstante, require dun custo computacional grande, xa que é necesario seleccionar máis parámetros. Por exemplo, para $d = 2$, o número de parámetros de $\mathcal{D}$ é 2 e o de $\mathcal{F}$ é 3. En xeral, non sempre é mellor tomar unha matriz sen restricións, senón que a súa elección depende da estrutura dos datos: se non están correlacionados, a matriz

diagonal pode ser suficiente e máis sinxela de estimar; se si o están, a matriz sen restricións é máis adecuada para captar a dependencia entre as variables e obter así unha mellor estimación nese caso.

Tendo en conta os distintos tipos de matrices de ancho de banda mencionados, imos aplicar dous enfoques distintos da estimación de tipo núcleo bidimensional nos casos de leucemia, de maneira que ilustren como afecta a elección desta matriz nos resultados obtidos. Na Figura 1.10 mostramos as estimacións dos casos de leucemia correspondentes coa matriz diagonal pertencente ó grupo $\mathcal{D}$ e coa matriz sen restricións, respectivamente. Para a súa elaboración, realizamos un código de R que describiremos a continuación brevemente. No caso da matriz diagonal, estimamos os valores diagonais mediante a función *bandwidth.nrd* do paquete *MASS* (Ripley et al., 2025). Por outra parte, obtemos a matriz sen restricións directamente coa función *Hpi* do paquete *ks*. Finalmente, calculamos a estimación en cada caso coa función *kde* do paquete *ks* (Duong, 2007). As estimacións resultantes das diferentes matrices de ancho de banda son:

$$H_{\text{diag}} = \begin{bmatrix} 0{,}1215 & 0{,}0000 \\ 0{,}0000 & 0{,}1010 \end{bmatrix} \quad \text{e} \quad H_{\text{full,1}} = \begin{bmatrix} 0{,}0033 & -0{,}0015 \\ -0{,}0015 & 0{,}0030 \end{bmatrix}. \tag{1.4}$$

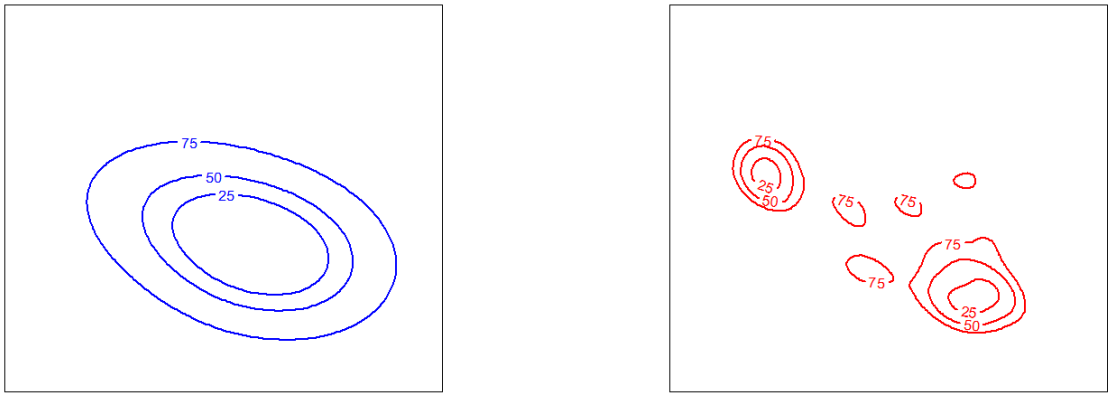

Figura 1.10: Estimación de tipo núcleo dos casos de leucemia no noroeste de Inglaterra cunha matriz diagonal, H_{diag}, e cunha sen restricións, $H_{\text{full,1}}$, respectivamente. Ver (1.4).

A elección da matriz na estimación de tipo núcleo inflúe na forma e precisión do resultado. Na Figura 1.10, observamos que cando se utiliza unha matriz diagonal, o resultado xera núcleos simétricos e aliñados cos eixes. Isto suaviza moito a densidade e non permite captar certas estruturas. No caso da matriz sen restricións, o gráfico capta con máis precisión a correlación dos datos, dando lugar a unha estimación máis flexible e probablemente máis próxima á distribución

Figura 1.11: Estimación de tipo núcleo coa matriz sen restricións $H_{\text{full},2}$ dos controis realizados no noroeste de Inglaterra. Ver (1.5).

observada. Empregando o mesmo método de estimación con matriz sen restricións para os controis tomados no noroeste de Inglaterra, podemos comparar a densidade estimada dos casos e dos controis. Neste caso, tense que

$$H_{\text{full},2} = \begin{bmatrix} 0{,}0014 & -0{,}0005 \\ -0{,}0005 & 0{,}0015 \end{bmatrix}. \tag{1.5}$$

Á vista das Figuras 1.10 e 1.11, vemos que existen diferenzas importantes nas densidades estimadas con matrices sen restricións nos casos e nos controis. O gráfico da estimación dos casos presenta dúas zonas con alta concentración, que dan lugar a dous focos principais de incidencia da enfermidade. Non obstante, na estimación da densidade dos controis a estrutura parece unimodal, ou polo menos a moda secundaria non é tan pronunciada. Isto suxire que a enfermidade non está distribuída do mesmo xeito que a poboación, senón que presenta clústers, indicando un posible patrón espacial.

En definitiva, a elección da matriz de ancho de banda inflúe na suavidade do estimador, polo que é importante facer unha boa elección. A obtención do seu valor óptimo implica un equilibro entre a precisión do estimador, representada polo nesgo (denotado por Nesgo$\{\cdot\}$), e a súa estabilidade, que vén dada pola varianza (indicada por Var$\{\cdot\}$). Estes dous elementos constitúen a base sobre a que se define o erro cadrático medio, unha medida que presentamos a continuación para avaliar a calidade do estimador.

1.4.2. Erro cadrático medio

Unha vez definido o estimador de tipo núcleo e presentada a importancia da matriz de ancho de banda, xorde a dúbida sobre a calidade da estimación obtida. A forma máis común de medida do rendemento de $\hat{f}(\mathbf{x};\mathbf{H})$ é o erro cadrático medio (MSE), que avalía o desvío medio do estimador respecto da densidade real. Defínese como

$$\mathrm{MSE}\{\hat{f}(\mathbf{x};\mathbf{H})\} = \mathbb{E}\{[\hat{f}(\mathbf{x};\mathbf{H}) - f(\mathbf{x})]^2\} = \mathrm{Var}\{\hat{f}(\mathbf{x};\mathbf{H})\} + \mathrm{Nesgo}^2\{\hat{f}(\mathbf{x};\mathbf{H})\}, \tag{1.6}$$

onde $\mathrm{Nesgo}^2\{\hat{f}(\mathbf{x};\mathbf{H})\} = (\mathbb{E}[\hat{f}(\mathbf{x};\mathbf{H})] - f(\mathbf{x}))^2$. Este equilibrio entre o nesgo e a varianza é fundamental. Veremos que un ancho de banda pequeno reduce o nesgo pero aumenta a varianza, mentres que un ancho de banda grande reduce a varianza pero incrementa o nesgo. O obxectivo é seleccionar a matriz $\mathbf{H}$ que minimice o MSE. Para obter unha fórmula máis explícita, calcularemos o valor da esperanza e da varianza do estimador nun punto $\mathbf{x}$. Antes de proceder ó seu cálculo, é necesario introducir o concepto de convolución, que será clave para expresar estes valores de forma máis manexable.

Definimos a convolución entre dúas funcións integrables f e g como unha nova función $f * g$, dada por $(f * g)(\mathbf{x}) = \int_{\mathbb{R}^d} f(\mathbf{x}-\mathbf{y})g(\mathbf{y})d\mathbf{y}$. Considerando dous vectores $\mathbf{X}$ e $\mathbf{Y}$ con densidades f e g respectivamente, temos que $\mathbf{X}+\mathbf{Y}$ ten densidade $f * g$. Por exemplo, se $f \sim \mathcal{N}(0,1)$ e $K \sim \mathcal{N}(0,1)$, temos que o núcleo escalado $K_h \sim \mathcal{N}(0,h^2)$, polo que chegamos a que a convolución $(K_h * f) \sim \mathcal{N}(0, 1+h^2)$. Consideremos agora que os $\{\mathbf{X}_i\}_{i=1}^n$ son idénticamente distribuídos, entón, polas propiedades do operador esperanza,

$$\begin{aligned}\mathbb{E}[\hat{f}(\mathbf{x};\mathbf{H})] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x}-\mathbf{X}_i)\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x}-\mathbf{X}_i)\right] \\ &= \frac{n}{n}\mathbb{E}[K_{\mathbf{H}}(\mathbf{x}-\mathbf{X}_1)] = \int_{\mathbb{R}^d} K_{\mathbf{H}}(\mathbf{x}-\mathbf{y})f(\mathbf{y})\,d\mathbf{y} = (K_{\mathbf{H}} * f)(\mathbf{x}),\end{aligned} \tag{1.7}$$

polo que a esperanza é igual á convolución entre o núcleo e a función de densidade da mostra. Isto implica que o estimador non é insesgado en xeral, é dicir, o valor esperado da estimación non é f exactamente, senón unha versión suavizada. Por exemplo, no caso de $f \sim \mathcal{N}(0,1)$, vimos antes que $(K_h * f) \sim \mathcal{N}(0, 1+h^2)$. Se h é moi pequeno, esta convolución será moi similar a f. Para o espazo d-dimensional, tal como veremos na Sección 1.4.3, para que o nesgo converxa a cero será suficiente que $\mathbf{H}$ tenda a cero a medida que n se fai grande. Por outra parte, temos que

a varianza nun punto $\mathbf{x}$, supondo as observacións independentes, é da forma

$$\begin{aligned}\mathrm{Var}(\hat{f}(\mathbf{x};\mathbf{H})) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x}-\mathbf{X}_i)\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x}-\mathbf{X}_i)\right)\\ &= \frac{1}{n}\mathrm{Var}(K_{\mathbf{H}}(\mathbf{x}-\mathbf{X})) = \frac{1}{n}\left(\mathbb{E}[K_{\mathbf{H}}^2(\mathbf{x}-\mathbf{X})] - \mathbb{E}[K_{\mathbf{H}}(\mathbf{x}-\mathbf{X})]^2\right)\\ &= \frac{1}{n}\int_{\mathbb{R}^d} K_{\mathbf{H}}^2(\mathbf{x}-\mathbf{y})f(\mathbf{y})d\mathbf{y} - (K_{\mathbf{H}} * f)^2(\mathbf{x}) = \frac{1}{n}\left((K_{\mathbf{H}}^2 * f)(\mathbf{x}) - (K_{\mathbf{H}} * f)^2(\mathbf{x})\right).\end{aligned} \tag{1.8}$$

Polo tanto, tendo en conta as ecuacións (1.7) e (1.8), chegamos a que o erro cadrático medio vén dado por

$$\mathrm{MSE}\{\hat{f}(\mathbf{x};\mathbf{H})\} = \frac{1}{n}\{(K_{\mathbf{H}}^2 * f)(\mathbf{x}) - (K_{\mathbf{H}} * f)^2(\mathbf{x})\} + \{(K_{\mathbf{H}} * f)(\mathbf{x}) - f(\mathbf{x})\}^2. \tag{1.9}$$

O equilibrio entre o nesgo e a varianza é crucial para a obtención dun bo estimador de densidade. Comentamos anteriormente que o obxectivo é tomar a matriz de ancho de banda $\mathbf{H}$ que minimice o erro cadrático medio. Non obstante, na práctica isto pode resultar complexo de conseguir. Na ecuación (1.9) vemos que o MSE depende da función de densidade f, que na práctica se descoñece, polo que se implementan diferentes métodos de aproximación como a validación cruzada ou o pulg-in, que estudiaremos no Capítulo 2 en detalle.

Como ilustración dos resultados que acabamos de obter, consideramos o modelo M1 presentado na Sección 1.2 e a matriz de ancho de banda $\mathbf{H} = h^2\mathbf{I}_2$, con $h > 0$. Representaremos e analizaremos a dependencia de h co nesgo e ca varianza nun punto concreto $\mathbf{x} = [5,5]^\top$, de forma que nos permita visualizar e entender como funciona a estimación cando h varía. Sexa $\mathbf{x}_0 = [0,0]^\top$. Temos que o kernel gaussiano $K_{\mathbf{H}} \sim \mathcal{N}(\mathbf{x}_0, \mathbf{H})$, polo que para calcular o nesgo e a varianza seguindo as ecuacións (1.7) e (1.8), necesitamos o cálculo das convolucións correspondentes. Tense que $(K_{\mathbf{H}} * f) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \mathbf{H})$ e $\left(K_{\mathbf{H}}^2 * f\right) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \mathbf{H}/2)$, sendo $K_{\mathbf{H}}^2$ equivalente ó núcleo con matriz de varianzas e covarianzas $\mathbf{H}/2$. Realizamos un código en R para calcular o nesgo ó cadrado, a varianza e o MSE no punto $\mathbf{x} = [5,5]^\top$ para unha secuencia de valores de h. Na Figura 1.12, observamos que o nesgo ó cadrado aumenta con h, pois a suavización do estimador é maior e polo tanto afástase da densidade real. Non obstante, un h maior reduce a varianza do estimador, posto que está menos suxeito ás fluctuacións dos datos. O MSE é unha medida que combina o nesgo e a varianza, polo que o valor óptimo de h será aquel que minimice o erro cadrático medio, conseguindo un equilibrio entre ambas medidas. Neste caso, obsérvase fácilmente que o mínimo é alcanzado en $h_{\text{opt}} = 1.3$. Este equilibrio obtido para un punto concreto aporta unha idea visual do que ocorre no caso xenérico.

A continuación, seguindo co modelo M1, representamos o estimador de tipo núcleo para dous anchos de banda h distintos. Xeramos un código en R e estimamos os valores da diagonal coa

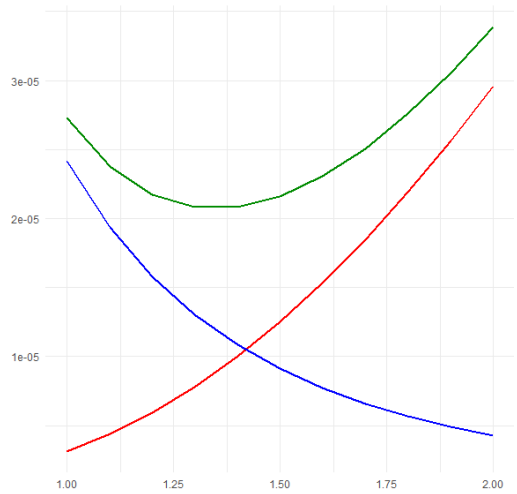

Figura 1.12: Relación entre o nesgo ó cadrado (vermello), a varianza (azul) e o erro cadrático medio (verde) en función do parámetro h.

función *bandwidth.nrd*, e tomamos como parámetro de referencia común a media entre ambos, $h_o = 4.63$. Seleccionamos unha matriz con ancho de banda máis pequeno, $h_1 = 2$, e outra con ancho de banda máis grande, $h_2 = 6$. Na Figura 1.13 observamos as estimacións dadas polas matrices diagonais con ancho de banda h_1 e h_2, respectivamente. O ancho de banda pequeno da lugar a unha densidade estimada pouco suave, mentres que o ancho de banda grande estima unha función moi suave, que pode chegar a perder información. Concluímos entón a importancia de elixir un bo ancho de banda, incluso nas matrices diagonais máis simples, xa que este afecta na suavidade do estimador.

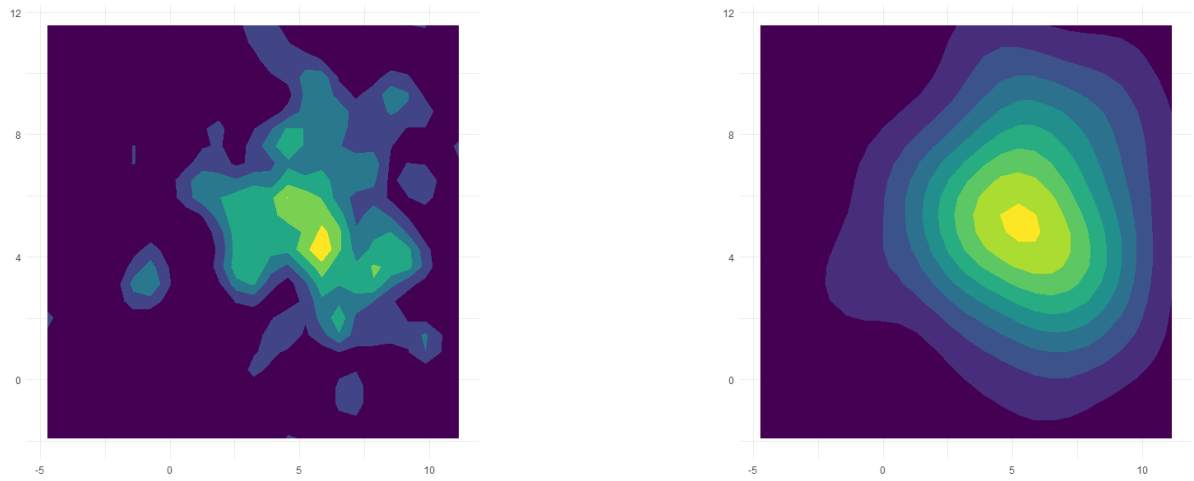

Figura 1.13: Estimación da función de densidade tipo núcleo de M1 con $\mathbf{H} = h^2\mathbf{I}_2$. Á esquerda, a estimación con $h_1 = 2$, e á dereita, con $h_2 = 6$.

Temos analizado como o nesgo e a varianza cambian nun punto concreto en función do parámetro h da matriz de ancho de banda diagonal. Ademais, tamén representamos a relación entre o tamaño de h e a precisión da estimación. Non obstante, no caso xeral, isto non nos permite estudar como varía o nesgo en todos os puntos $\mathbf{x}$. Con ese obxectivo, imos fixar un valor de ancho de banda e seleccionar unha grade de puntos, que nos permita avaliar o nesgo en cada un e de lugar a un mapa de calor. Desta forma, podemos observar cando aumenta ou diminúe o nesgo no caso completo.

Consideramos o modelo M3 de datos simulados pertencentes a unha mixtura de dúas normais bidimensionais, e realizamos un programa en R que calcula a densidade real e o promedio do estimador de tipo núcleo en cada punto da grade, empregando o kernel gaussiano con matriz diagonal $\mathbf{H} = h^2\mathbf{I}_2$. Desta forma, obtemos a diferenza entre os dous valores, que representa o nesgo en cada punto. Coa axuda do paquete *ggplot2* de R, realizamos os gráficos de calor correspondentes para $h_1 = 0.5$ e $h_2 = 2.5$ coa mesma escala diverxente, onde as cores cálidas indican nesgo positivo e as frías negativo.

O nesgo indica a media de canto se alonxa a densidade estimada da real. A medida que h aumenta, o estimador suavízase e tende a subestimar os picos e sobreestimar os vals, de forma que presenta nesgo negativo nas zonas de alta densidade e nesgo positivo nas baixas. Na Figura 1.14 observamos que o nesgo alcanza valores máis grandes no caso de $h_2 = 2.5$, e máis próximos a cero cando $h_1 = 0.5$. Este enfoque permite comprender como o ancho de banda da matriz diagonal h afecta na estimación da densidade

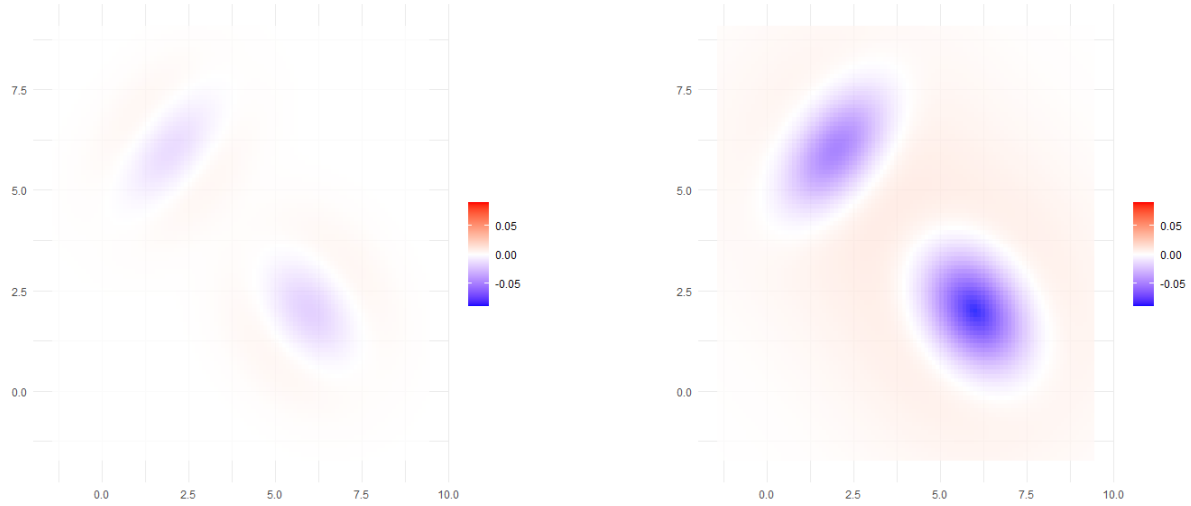

Figura 1.14: Mapa de calor do nesgo nunha greda de puntos para $h = 0.5$ (esquerda) e $h = 2.5$ (dereita).

Vexamos agora o que ocorre co gráfico de calor da varianza para un parámetro concreto

$h = 2.5$, e analicemos como varía respecto do nesgo. Na Figura 1.15 vemos que a rexión con maior varianza é o clúster inferior dereito, mentres que a varianza é próxima a cero nas zonas onde apenas hai datos. Se diminuísemos h, a varianza aumentaría, seguindo unha estrutura similar á do gráfico. Este patrón inverso reflicte o compromiso fundamental entre nesgo e varianza que ten lugar na estimación de tipo núcleo. Por ese motivo, empregamos o MSE para obter o ancho de banda que mellor axusta os datos.

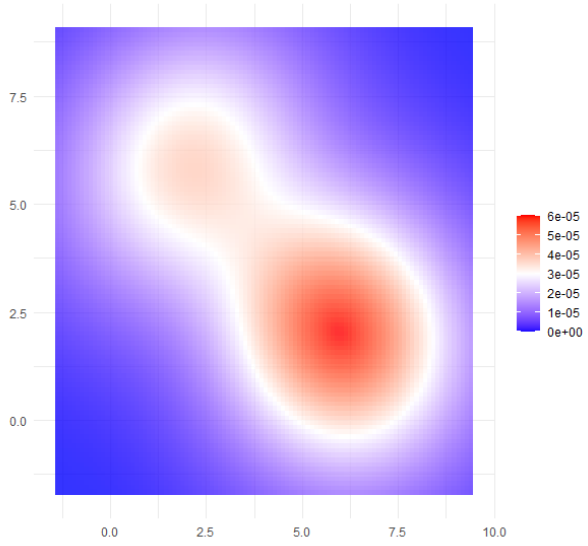

Figura 1.15: Mapa de calor da varianza nunha greda de puntos para $h = 2.5$.

O $\text{MSE}\{\hat{f}(\mathbf{x};\mathbf{H})\}$ é unha medida de discrepancia local, pois mide a precisión do estimador localmente en cada punto $\mathbf{x}$. Para estudar o comportamento global de $\hat{f}(\mathbf{x};\mathbf{H})$ en todo o dominio, integramos MSE respecto $\mathbf{x}$,

$$\text{MISE}\{\hat{f}(\cdot;\mathbf{H})\} = \int_{\mathbb{R}^d} \text{MSE}\{\hat{f}(\mathbf{x};\mathbf{H})\} d\mathbf{x} = \mathbb{E} \int_{\mathbb{R}^d} \{\hat{f}(\mathbf{x};\mathbf{H}) - f(\mathbf{x})\}^2 \, d\mathbf{x},$$

obtendo o erro cadrático medio integrado (MISE), que se corresponde coa distancia ó cadrado esperada L_2 entre $\hat{f}$ e f. A conmutación da integral e da esperanza na segunda igualdade é válida baixo as hipóteses do Teorema de Tonelli (Trinchet, p.134), posto que a función é integrable e non negativa.

O MISE é unha cantidade teórica que non depende directamente dos datos observados. Para considerar unha versión que si sexa estocástica, definimos o erro cadrático integrado (ISE) como

$$\text{ISE}\{\hat{f}(\mathbf{x};\mathbf{H})\} = \int_{\mathbb{R}^d} \{\hat{f}(\mathbf{x};\mathbf{H}) - f(\mathbf{x})\}^2 d\mathbf{x}, \tag{1.10}$$

que mide o erro real para unha mostra de datos concreta. Polo tanto, para avaliar o comportamento medio do estimador, resulta máis sinxelo considerar o MISE, xa que avalía o valor esperado do

erro sobre todas as mostras posibles. Desenvolvemos a súa expresión integrando (1.9) e supondo que $\int_{\mathbb{R}^d} f^2(\mathbf{x})d\mathbf{x} < \infty$, de forma que chegamos á seguinte expresión:

$$\begin{aligned}\mathrm{MISE}\{\hat{f}(\cdot;\mathbf{H})\} = \frac{1}{n}\int_{\mathbb{R}^d}(K_{\mathbf{H}}^2 * f)(\mathbf{x})d\mathbf{x} + \left(1-\frac{1}{n}\right)\int_{\mathbb{R}^d}(K_{\mathbf{H}} * f)^2(\mathbf{x})d\mathbf{x} \\ - 2\int_{\mathbb{R}^d}(K_{\mathbf{H}} * f)(\mathbf{x})f(\mathbf{x})d\mathbf{x} + \int_{\mathbb{R}^d} f^2(\mathbf{x})d\mathbf{x}.\end{aligned}$$

Empregando a definición de convolución e a de $K_{\mathbf{H}}$, podemos escribir o primeiro termo como

$$\begin{aligned}\frac{1}{n}\int_{\mathbb{R}^d}(K_{\mathbf{H}}^2 * f)(\mathbf{x})d\mathbf{x} &= \frac{1}{n}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} K_{\mathbf{H}}^2(\mathbf{x}-\mathbf{y})f(\mathbf{y})d\mathbf{y}d\mathbf{x} \underset{\text{(a)}}{=} \frac{1}{n}\int_{\mathbb{R}^d} f(\mathbf{y})\int_{\mathbb{R}^d} K_{\mathbf{H}}^2(\mathbf{x}-\mathbf{y})d\mathbf{x}d\mathbf{y} \\ &\underset{\text{(b)}}{=} \frac{1}{n}|\mathbf{H}|^{-1/2}R(K)\int_{\mathbb{R}^d} f(\mathbf{y})d\mathbf{y} \underset{\text{(c)}}{=} \frac{1}{n}|\mathbf{H}|^{-1/2}R(K).\end{aligned}$$

onde $R(a) = \int_{\mathbb{R}^d} a(\mathbf{x})^2 d\mathbf{x}$, sendo $a : \mathbb{R}^d \to \mathbb{R}$ unha función cadrado integrable. En (a) empregamos o Teorema de Tonelli para intercambiar a orde de integración e en (c) utilizamos que f é unha densidade de probabilidade, e polo tanto integrable. Finalmente, en (b) aplicamos o cambio de variable $\mathbf{u} = \mathbf{x} - \mathbf{y}$ e a definición de $K_{\mathbf{H}}(\mathbf{x})$, de forma que

$$\begin{aligned}\int_{\mathbb{R}^d} K_{\mathbf{H}}^2(\mathbf{x}-\mathbf{y})d\mathbf{x} &= \int_{\mathbb{R}^d} K_{\mathbf{H}}^2(\mathbf{u})d\mathbf{u} = |\mathbf{H}|^{-1}\int_{\mathbb{R}^d} K^2(\mathbf{H}^{-1/2}\mathbf{u})d\mathbf{u} \\ &= |\mathbf{H}|^{-1/2}\int_{\mathbb{R}^d} K^2(\mathbf{z})d\mathbf{z} = |\mathbf{H}|^{-1/2}R(K),\end{aligned}$$

aplicando na terceira igualdade o cambio de variable $\mathbf{z} = \mathbf{H}^{-1/2}\mathbf{u}$, con $d\mathbf{u} = |\mathbf{H}|^{1/2}d\mathbf{z}$. Concluímos que o MISE é da forma

$$\begin{aligned}\mathrm{MISE}\{\hat{f}(\cdot;\mathbf{H})\} &= n^{-1}|\mathbf{H}|^{-1/2}R(K) - n^{-1}R(K_{\mathbf{H}} * f) \\ &+ R(K_{\mathbf{H}} * f) - 2\int_{\mathbb{R}^d}(K_{\mathbf{H}} * f)(\mathbf{x})f(\mathbf{x})d\mathbf{x} + R(f).\end{aligned}$$

Aínda que o MISE proporciona unha medida teórica moi útil para estudar o erro dun estimador, na práctica pode ser difícil de calcular, posto que depende da densidade real f dos datos, que adoita ser descoñecida. Ademais, a integración no espazo pode ser complexa cando o número de dimensións é grande, e a dependencia do MISE de $\mathbf{H}$ non é sinxela de analizar. Por este motivo, estudaremos o comportamento asintótico do MISE cando o tamaño n da mostra se fai grande. Desta forma, simplificamos a análise e obtemos información interesante sobre o comportamento do estimador en función de $\mathbf{H}$.

1.4.3. Propiedades asintóticas

Unha aproximación asintótica do MISE é o erro cadrático medio integrado asintótico (AMISE), que simplifica a súa expresión e permite derivar regras prácticas para a obtención da matriz

H. Para obter a súa fórmula, é necesario definir o operador diferencial de orde r e enunciar unha formulación particular do teorema de Taylor en expansión multivariante.

Consideramos o operador diferencial de orde r como o produto de Kronecker de $\mathbf{D}$ r veces consigo mesmo, $\mathbf{D}^{\otimes r}$, onde $\mathbf{D} = [\frac{\partial}{\partial x_1}, ..., \frac{\partial}{\partial x_d}]^\top$. Para $r = 2$, tense que $\mathbf{D} \otimes \mathbf{D}$ actúa como a vectorización da matriz hessiana H de dimensións $d \times d$, de forma que $\text{vec}(H) = \mathbf{D} \otimes \mathbf{D}$. O operador de vectorización (denotado como vec) transforma unha matriz nun vector columna, apilando as columnas unha debaixo da outra, de forma que se $\mathbf{A} = (a_{ij}) \in \mathcal{M}_{m\times n}$, temos que $\text{vec}(\mathbf{A}) \in \mathbb{R}^{mn}$ é igual a $\text{vec}(\mathbf{A}) = [a_{11}, ..., a_{m1}; ...; a_{1n}, ..., a_{mn}]^\top$. Unha descrición máis formal destes operadores, así como unha análise das súas principais propiedades, pode verse no Anexo A. Empregando as derivadas vectorizadas, podemos escribir a expansión de Taylor multivariante para unha función f continua e diferenciable r veces.

Teorema 1.2 (Expansión de Taylor Multivariante con Derivadas Vectorizadas)**.** *Sexa $f : \mathbb{R}^d \to \mathbb{R}$ unha función real con derivadas continuas ata orde r (de clase $\mathcal{C}^r$, nun entorno de $\mathbf{x} \in \mathbb{R}^d$). Entón, para calquera perturbación $\mathbf{a} \in \mathbb{R}^d$ con $||\mathbf{a}||$ o suficientemente pequeno, tense que*

$$f(\mathbf{x}+\mathbf{a}) = \sum_{j=0}^{r} \frac{1}{j!} \mathbf{D}^{\otimes j} f(\mathbf{x})^\top \mathbf{a}^{\otimes j} + Re(\mathbf{a}), \tag{1.11}$$

onde $\mathbf{a}^{\otimes j}$ é o produto de Kronecker do vector $\mathbf{a}$ consigo mesmo j veces, e $Re(\mathbf{a})$ é o resto de orde menor que $\|\mathbf{a}\|^r$ cando $\mathbf{a} \to 0$. Podemos expresalo como $o(\|\mathbf{a}\|^r)$, sendo $\|\cdot\|$ a norma euclidiana, ver Anexo A.

Demostración. Esta demostración non é obxecto de interese no desenvolvemento deste traballo, pero a súa estrutura pode derivarse do enunciado clásico do Teorema de Taylor Multivariante, adaptado á notación vectorizada e empregando produtos de Kronecker. Para unha demostración máis detallada do enfoque clásico, ver Burgos (2008, p.109). □

Este teorema permítenos aproximar valores dunha función próximos a un punto $\mathbf{x} \in \mathbb{R}^d$ mediante un desenvolvemento en serie dado en termos de derivadas de orde superior avaliadas nese punto, sempre que a función sexa o suficientemente suave como para que as derivadas existan e sexan continuas. Por exemplo, se $f : \mathbb{R}^2 \to \mathbb{R}$ é de clase $\mathcal{C}^2$ nun entorno do punto $\mathbf{x} = (x_1, x_2)$, e $\mathbf{a} \in \mathbb{R}^2$ é un desprazamento, escribimos a expansión de Taylor de segunda orde como

$$f(\mathbf{x}+\mathbf{a}) = f(x_1, x_2) + \nabla f(x_1, x_2)^\top \mathbf{a} + \frac{1}{2}\mathbf{a}^\top H_f(x_1, x_2)\mathbf{a} + o(||\mathbf{a}||^2),$$

onde $\nabla f(x_1, x_2)$ é o gradiente e $H_f(x_1, x_2)$ a matriz hessiana.

A continuación, imos presentar un teorema que desenvolve as fórmulas do nesgo e da varianza en función da expansión de Taylor, para posteriormente deducir a partir delas unha aproximación asintótica do MISE.

Teorema 1.3. *Sexa $\hat{f}(\mathbf{x};\mathbf{H})$ o estimador de densidade de tipo núcleo multivariado. Supoñamos que:*

(A1) *f é cadrado integrable ($f \in L_2$) e dúas veces diferenciable, onde as segundas derivadas son continuas, acotadas e cadrado integrables;*

(A2) *K é unha función cadrado integrable, esféricamente simétrica ($K(\mathbf{z}) = k(||\mathbf{z}||)$, para unha función escalar k), e ten un momento de segunda orde finito, o que implica que*

$$\int_{\mathbb{R}^d} \mathbf{z}K(\mathbf{z})d\mathbf{z} = \mathbf{0} \quad e \quad \int_{\mathbb{R}^d} \mathbf{z}^{\otimes 2}K(\mathbf{z})d\mathbf{z} = m_2(K)\mathrm{vec}(\mathbf{I}_\mathrm{d}),$$

onde $m_2(K) = \int_{\mathbb{R}^d} z_i^2 K(\mathbf{z})d\mathbf{z}$ para todo $i \in \{1,...,d\}$, é dicir, $m_2(K)$ non depende da coordenada i;

(A3) *As matrices de ancho de banda forman unha secuencia da forma $\mathbf{H} = \mathbf{H}_n$ de matrices definidas positivas e simétricas tal que*

$$\mathrm{vec}(\mathbf{H}) \to 0 \quad e \quad \frac{1}{n}|\mathbf{H}|^{-1/2} \to 0 \quad cando \quad n \to \infty.$$

Entón, cúmprense as seguintes aproximacións asintóticas:

1. $\mathrm{Nesgo}\{\hat{f}(\mathbf{x};\mathbf{H})\} = \mathbb{E}[\hat{f}(\mathbf{x};\mathbf{H})] - f(\mathbf{x}) = \frac{1}{2}m_2(K)\mathbf{D}^{\otimes 2}f(\mathbf{x})^\top \mathrm{vec}(\mathbf{H}) + o(\|\mathrm{vec}(\mathbf{H})\|)$;
2. $\mathrm{Var}\{\hat{f}(\mathbf{x};\mathbf{H})\} = n^{-1}|\mathbf{H}|^{-1/2}f(\mathbf{x})R(K) + o(n^{-1}|\mathbf{H}|^{-1/2})$.

Combinando ambas expresión e integrándoas en $\mathbb{R}^d$, dan lugar á aproximación asintótica do MISE:

$$\mathrm{AMISE}\{\hat{f}(\mathbf{x};\mathbf{H})\} = \frac{1}{n}|\mathbf{H}|^{-1/2}R(K) + \frac{1}{4}m_2(K)^2\{\mathrm{vec}^\top \mathbf{R}(\mathbf{D}^{\otimes 2}f)\}(\mathrm{vec}\mathbf{H})^{\otimes 2}, \tag{1.12}$$

sendo $\mathbf{R}(\mathbf{a}) = \int_{\mathbb{R}^d} \mathbf{a}(\mathbf{x})\mathbf{a}(\mathbf{x})^\top d\mathbf{x} \in \mathcal{M}_{p\times p}$, para $\mathbf{a} : \mathbb{R}^d \to \mathbb{R}^p$.

Demostración. A continuación, presentamos un esquema da demostración, baseada en Chacón e Duong (2018) e dividida en tres partes: (1) a expansión do nesgo, (2) o cálculo da varianza, e (3) a combinación dos resultados para obter o AMISE. Ó longo da demostración imos empregar propiedades do produto de Kronecker e do operador de vectorización, presentadas no Anexo A.

1. O valor esperado do estimador vén dado pola ecuación (1.7). Mediante o cambio de variable $\mathbf{z} = \mathbf{H}^{-1/2}(\mathbf{x} - \mathbf{y})$, podemos reescribir a integral como

$$\mathbb{E}[\hat{f}(\mathbf{x};\mathbf{H})] = \int_{\mathbb{R}^d} K(\mathbf{z})f(\mathbf{x} - \mathbf{H}^{1/2}\mathbf{z})d\mathbf{z},$$

onde o xacobiano vén dado como $d\mathbf{y} = |\mathbf{H}|^{1/2}d\mathbf{z}$. Esta transformación permite eliminar a dependencia de $\mathbf{H}$ no núcleo, que agora aparece na f. A continuación, dadas as hipóteses **(A1)** de f, e **(A3)** para que $\mathbf{H}^{1/2}\mathbf{z}$ sexa pequeno, podemos desenvolver a expansión de Taylor de $f(\mathbf{x}-\mathbf{H}^{1/2}\mathbf{z})$ en torno a $\mathbf{x}$ empregando o Teorema 1.2, de forma que

$$f(\mathbf{x}-\mathbf{H}^{1/2}\mathbf{z}) = f(\mathbf{x}) - \mathbf{D}f(\mathbf{x})^\top \mathbf{H}^{1/2}\mathbf{z} + \frac{1}{2}\mathbf{D}^{\otimes 2}f(\mathbf{x})^\top (\mathbf{H}^{1/2}\mathbf{z})^{\otimes 2} + o(\|\mathrm{vec}(\mathbf{H})\|).$$

Considerando as condicións do núcleo K de **(A2)**, chegamos a que

$$\begin{aligned}
\mathbb{E}\{\hat{f}(\mathbf{x};\mathbf{H})\} &= f(\mathbf{x})\int_{\mathbb{R}^d} K(\mathbf{z})\,d\mathbf{z} - \mathbf{D}f(\mathbf{x})^\top \mathbf{H}^{1/2}\int_{\mathbb{R}^d}\mathbf{z}K(\mathbf{z})\,d\mathbf{z} \\
&\quad + \frac{1}{2}\mathbf{D}^{\otimes 2}f(\mathbf{x})^\top (\mathbf{H}^{1/2})^{\otimes 2}\int_{\mathbb{R}^d}\mathbf{z}^{\otimes 2}K(\mathbf{z})\,d\mathbf{z} + o(\|\mathrm{vec}(\mathbf{H})\|) \\
&= f(\mathbf{x}) + \frac{1}{2}\mathbf{D}^{\otimes 2}f(\mathbf{x})^\top (\mathbf{H}^{1/2})^{\otimes 2}\,\mathrm{vec}(\mathbf{I}_d)m_2(K) + o(\|\mathrm{vec}(\mathbf{H})\|) \\
&= f(\mathbf{x}) + \frac{1}{2}\mathbf{D}^{\otimes 2}f(\mathbf{x})^\top m_2(K)\mathrm{vec}(\mathbf{H}) + o(\|\mathrm{vec}(\mathbf{H})\|),
\end{aligned}$$

onde empregamos as fórmulas $\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^\top \otimes \mathbf{A})\mathrm{vec}(\mathbf{B})$ e $(\mathbf{H}^{1/2}\mathbf{z})^{\otimes 2} = (\mathbf{H}^{1/2})^{\otimes 2}\mathbf{z}^{\otimes 2}$ do Anexo A.

2. Consideramos a varianza dada pola ecuación (1.8). Resolvemos o primeiro termo, xa que o segundo é inmediato ó ter calculada a expresión da esperanda do estimador. Realizamos novamente o cambio de variable $\mathbf{z} = \mathbf{H}^{-1/2}(\mathbf{x}-\mathbf{y})$, e chegamos a que

$$n^{-1}(K_{\mathbf{H}}^2 * f)(\mathbf{x}) = n^{-1}\int_{\mathbb{R}^d} K_{\mathbf{H}}^2(\mathbf{x}-\mathbf{y})f(\mathbf{y})d\mathbf{y} = n^{-1}|\mathbf{H}|^{-1/2}\int_{\mathbb{R}^d} K^2(\mathbf{z})f(\mathbf{x}-\mathbf{H}^{1/2}\mathbf{z})d\mathbf{z}.$$

Aplicando a expansión de Taylor de orde 1 de $f(\mathbf{x}-\mathbf{H}^{1/2}\mathbf{z})$ arredor de $\mathbf{x}$, e considerando as hipóteses **(A1)** e **(A2)**, obtemos que

$$n^{-1}(K_{\mathbf{H}}^2 * f)(\mathbf{x}) = n^{-1}|\mathbf{H}|^{-1/2}f(\mathbf{x})R(K) + o(n^{-1}|\mathbf{H}|^{-1/2}).$$

Para o segundo termo, cúmprese que

$$n^{-1}\left(\mathbb{E}\{\hat{f}(\mathbf{x};\mathbf{H})\}\right)^2 = n^{-1}f(\mathbf{x})^2 + o(n^{-1}|\mathbf{H}|^{-1/2}).$$

Este termo é asintóticamente desprezable fronte ó primeiro, que é o dominante. Así, debido á hipótese **(A3)**, na expresión asintótica da varianza só se conserva o termo de maior orde, é dicir,

$$\mathrm{Var}\{\hat{f}(\mathbf{x};\mathbf{H})\} = n^{-1}|\mathbf{H}|^{-1/2}f(\mathbf{x})R(K) + o(n^{-1}|\mathbf{H}|^{-1/2}).$$

3. Dadas as expresións asintóticas puntuais do nesgo e da varianza demostradas nos apartados 1 e 2, podemos deducir unha aproximación asintótica global do erro cadrático medio integrado (MISE). A proba consiste en integrar respecto $\mathbf{x}$ e combinar os resultados, de forma que chegamos a que

$$\mathrm{AMISE}\{\hat{f}(\mathbf{x};\mathbf{H})\} = \frac{1}{n}|\mathbf{H}|^{-1/2}R(K) + \frac{1}{4}m_2(K)^2\{\mathrm{vec}^\top \mathbf{R}(\mathbf{D}^{\otimes 2}f)\}(\mathrm{vec}\mathbf{H})^{\otimes 2}.$$

□

Este teorema establece as expresións asintóticas do nesgo, da varianza e do AMISE, que permiten entender o comportamento do estimador cando o tamaño da mostra n é grande. Para ilustrar o comportamento do estimador en situacións menos complexas, consideramos o caso particular $d = 2$ no que a matriz de ancho de banda é diagonal, de forma que $\mathbf{H} = h^2\mathbf{I}_2$. Este enfoque permite calcular con máis detalle as fórmulas do nesgo e da varianza, mostrando o compromiso entre ambos. Ademais, podemos derivar a expresión do AMISE, o que permitirá obter o h óptimo.

Sexa $\mathbf{H} = h^2\mathbf{I}_2$, con $h > 0$ o parámetro de suavizado, de xeito que $h \to 0$. As expresións do nesgo e da varianza asintóticas que vimos no Teorema 1.3 simplifícanse de maneira significativa. Por unha parte, temos que

$$\mathbf{D}^{\otimes 2} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} \\ \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix} \qquad \text{e} \qquad \operatorname{vec}(\mathbf{H}) = \begin{bmatrix} h^2 \\ 0 \\ 0 \\ h^2 \end{bmatrix},$$

polo que chegamos á expresión

$$\text{Nesgo}\{\hat{f}(\mathbf{x}; \mathbf{H})\} = \frac{1}{2} h^2 m_2(K) \triangle f(\mathbf{x}) + o(h^2),$$

sendo $\triangle f(\mathbf{x}) = \frac{\partial^2 f}{\partial x_1^2} + \frac{\partial^2 f}{\partial x_2^2}$ o laplaciano de f. Polo tanto, esta expresión diminúe cando o fai h. Se consideramos $\hat{\mathbf{x}}$ un punto crítico de f, temos que $\frac{\partial f}{\partial x_1}(\hat{\mathbf{x}}) = \frac{\partial f}{\partial x_2}(\hat{\mathbf{x}}) = 0$, e se a matriz hessiana de f é definida negativa, é dicir, $\nabla^2 f(\hat{\mathbf{x}}) \prec 0$, entón temos que $\hat{\mathbf{x}}$ é un máximo. Neste caso, $\triangle f(\hat{\mathbf{x}}) \leq 0$ e $\text{Nesgo}\{\hat{f}(\hat{\mathbf{x}}; \mathbf{H})\} \leq 0$, o que indica que o estimador subestima o verdadeiro valor nos puntos máximos, como xa vimos no exemplo da Figura 1.14.

Por outra parte, a varianza é da forma

$$\text{Var}\{\hat{f}(\mathbf{x}; \mathbf{H})\} = \frac{1}{nh^2} f(\mathbf{x}) R(K) + o\left(\frac{1}{nh^2}\right),$$

que aumenta cando diminúe h. Estas expresións mostran o compromiso entre o nesgo e a varianza, e a necesidade de acadar un equilibrio entre ambos, que será o h que minimice o AMISE. Temos que

$$\text{AMISE}\{\hat{f}(\mathbf{x}; \mathbf{H})\} = \frac{1}{nh^2} R(K) + \frac{1}{4} h^4 m_2(K)^2 R(\triangle f).$$

Entón, derivando con respecto a h e igualando a cero, chegamos a que

$$h_{opt} = \left(\frac{2R(K)}{n m_2(K)^2 R(\triangle f)} \right)^{1/6},$$

que representa o compromiso óptimo entre o nesgo e a varianza no caso bidimensional con matriz de ancho de banda $\mathbf{H} = h^2\mathbf{I}_2$.

Esta análise pódese estender ó caso de matrices de ancho de banda máis xerais, como $\mathbf{H} = \text{diag}(h_1^2, h_2^2)$. Neste caso, a matriz capta variacións direccionais en cada dimensión, obtendo polo tanto expresións máis complexas. Temos que

$$\text{Nesgo}\{\hat{f}(\mathbf{x};\mathbf{H})\} = \frac{1}{2}m_2(K)\left(h_1^2\frac{\partial^2 f}{\partial x_1^2} + h_2^2\frac{\partial^2 f}{\partial x_2^2}\right) + o\left(\sqrt{h_1^4 + h_2^4}\right)$$

e

$$\text{Var}\{\hat{f}(\mathbf{x};\mathbf{H})\} = \frac{1}{nh_1h_2}f(\mathbf{x})R(K) + o\left(\frac{1}{nh_1h_2}\right).$$

Novamente, se consideramos o máximo $\hat{\mathbf{x}}$, chegamos a que $\text{Nesgo}\{\hat{f}(\mathbf{x};\mathbf{H})\} \leq 0$, e polo tanto concluímos que o estimador subestima os valores reais.

Finalmente, chegamos á fórmula do AMISE como

$$\text{AMISE}\{\hat{f}(\mathbf{x};\mathbf{H})\} = \frac{1}{nh_1h_2}R(K) + \frac{1}{4}m_2(K)^2\left(\psi_{11,11}h_1^4 + 2h_1^2h_2^2\psi_{11,22} + \psi_{22,22}h_2^4\right), \tag{1.13}$$

onde $\psi_{ij,kl} = \int_{\mathbb{R}^d}\left(\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}\cdot\frac{\partial^2 f(\mathbf{x})}{\partial x_k \partial x_l}\right)d\mathbf{x}$, para $i,j,k,l \in \{1,2\}$. Neste caso, para obter os parámetros de ancho de banda óptimos, temos que derivar esta expresión con respecto a h_1 e h_2, e igualalas a cero, obtendo unha expresión que explicaremos na Subsección 1.4.4.

Estes exemplos concretos ilustran como as expresións asintóticas permiten obter expresións aproximadas da matriz de ancho de banda óptima. No caso non asintótico, isto non é posible, xa que a expresión do erro depende de $\mathbf{H}$ dunha forma complexa. A continuación, imos presentar o caso xeral d-dimensional para a obtención do óptimo.

1.4.4. Anchos de banda óptimos

O ancho de banda óptimo defínese como o que minimiza o MISE sobre a clase $\mathcal{F}$ de matrices sen restricións, é dicir, $\mathbf{H}_{\text{MISE}} = \text{argmin}_{\mathbf{H}\in\mathcal{F}}\text{MISE}\{\hat{f}(\cdot;\mathbf{H})\}$. Como dixemos na Subsección 1.4.2, a obtención deste parámetro é complexa debido á dependencia do MISE en función de $\mathbf{H}$, polo que se adoita tomar un ancho de banda óptimo asintótico aproximado, dado por $\mathbf{H}_{\text{AMISE}} = \text{argmin}_{\mathbf{H}\in\mathcal{F}}\text{AMISE}\{\hat{f}(\cdot;\mathbf{H})\}$. Temos que $\mathbf{H}_{\text{MISE}}$ e $\mathbf{H}_{\text{AMISE}}$ son asintóticamente equivalentes cando $n \to \infty$. Non obstante, ambos dependen de f, polo que non é posible calculalos a partir dos datos. Estes valores coñécense como anchos de banda oráculo (*oracle bandwidths*).

Se nos restrinximos ó caso das matrices diagonais, $\mathcal{A}$ ou $\mathcal{D}$, existen fórmulas explícitas para obter o ancho de banda óptimo. Por unha parte, se $\mathbf{H} = h^2\mathbf{I}_d$, a ecuación (1.12) simplifícase a

$$\text{AMISE}\{\hat{f}(\mathbf{x};h)\} = \frac{1}{nh^d}R(K) + \frac{1}{4}m_2(K)^2R(\Delta f)h^4dx, \tag{1.14}$$

que derivando e igualando a cero obtén o h óptimo como

$$h_{AMISE} = \left(\frac{dR(K)}{nm_2(K)^2 R(\Delta f)} \right)^{1/d+4}.$$

Podemos comprobar que no caso $d = 2$, feito na Subsección 1.4.3, a fórmula coincide. Por outra parte, se $\mathbf{H} = \text{diag}(h_1^2, ..., h_d^2)$, Wand e Jones (1994) demostraron que no caso multivariado só existe unha fórmula explícita para dimensión 2, que se obtén derivando a fórmula (1.13) con respecto a h_1 e h_2 e igualando a cero, de forma que resolvemos un sistema de dúas ecuación e dúas incógnitas. En Chacón e Duong (2018, p.35), vemos que a expresión resultante é

$$\mathbf{h}_{\text{AMISE}} = \left(\{\psi_{22,22}/\psi_{11,11}\}^{1/8}, \{\psi_{11,11}/\psi_{22,22}\}^{1/8} \right) \times \left(\frac{R(K)}{nm_2(K)^2(\psi_{11,11}^{1/2}\psi_{22,22}^{1/2} + \psi_{11,22})} \right)^{1/6}.$$

No caso das matrices sen restricións $\mathbf{H} \in \mathcal{F}$ non existe unha fórmula explícita. Non obstante, Chacon e Duong (2018, Sección 2.9.2) fixeron unha descrición da súa estrutura considerando $\mathbf{H} = \lambda\mathbf{A}$, sendo $\lambda = |\mathbf{H}|^{1/d}$ o parámetro que controla o tamaño e $\mathbf{A} = |\mathbf{H}|^{-1/d}\mathbf{H}$ unha matriz simétrica, definida positiva e con determinante 1, que define a orientación. Desta forma, tense que $|\mathbf{H}| = \lambda^d|\mathbf{A}| = \lambda^d$ e $(\text{vec}(\mathbf{H}))^{\otimes 2} = \lambda^2 (\text{vec}(\mathbf{A}))^{\otimes 2}$, que substituíndo na ecuación (1.12) da lugar á expresión

$$\text{AMISE}\{\hat{f}(\mathbf{x}; \lambda\mathbf{A})\} = \frac{1}{n\lambda^{d/2}} R(K) + \frac{1}{4} m_2(K)^2 \lambda^2 \mathcal{Q}(\mathbf{A}), \tag{1.15}$$

sendo $\mathcal{Q}(\mathbf{A}) = (\text{vec}^\top(\mathbf{A}))\mathbf{R}(\mathbf{D}^{\otimes 2} f)(\text{vec}(\mathbf{A}))$. Entón, o valor de $\lambda = \lambda_0(\mathbf{A})$ que minimiza a ecuación é

$$\lambda_0(\mathbf{A}) = \left(\frac{dR(K)}{nm_2(K)^2 \mathcal{Q}(\mathbf{A})} \right)^{2/(d+4)},$$

que aparece na expresión do AMISE na parte formada polo nesgo e pola varianza. Esta elección emprégase para equilibrar ambas contribucións, xa que cada un deses termos depende de λ de maneira contraposta. Substituíndo este valor na ecuación (1.15), obtemos que

$$\min_{\lambda_0 > 0} \text{AMISE}\{\hat{f}(\mathbf{x}; \lambda\mathbf{A})\} = \frac{d}{d+4} \left\{ d^{-1} m_2(K)^2 R(K)^{4/d} \mathcal{Q}(\mathbf{A}) \right\}^{d/(d+4)} n^{-4/(d+4)},$$

polo que a elección óptima de $\mathbf{A}$ é $\mathbf{A}_0 = \text{argmin}_{\mathbf{A} \in \mathcal{F}, |\mathbf{A}|=1} \mathcal{Q}(\mathbf{A})$, que non depende do tamaño da mostra n. Isto quere dicir que a orientación da matriz non depende do número de observacións, senón da forma de f. Ademais, vemos que o termo do AMISE que depende de $\mathbf{A}$ é o nesgo, polo que a orientación da matriz escóllese para minimizar o nesgo cadrático. Polo tanto, podemos escribir o ancho de banda óptimo como $\mathbf{H}_{\text{AMISE}} = \lambda_0(\mathbf{A})\mathbf{A}_0 = \mathbf{C}_0 n^{-2/(d+4)}$, sendo $\mathbf{C}_0$ unha matriz simétrica e definida positiva que non depende de n, concluíndo que o ancho de banda óptimo é de orde $n^{-2/(d+4)}$.

Capítulo 2

Selectores de ancho de banda

No Capítulo 1 estudamos dous estimadores non paramétricos da función de densidade: o histograma e o estimador de tipo núcleo. Vimos que no estimador tipo núcleo a calidade depende da selección da matriz de ancho de banda $\mathbf{H}$, que controla o suavizado en cada dimensión e as posibles correlacións entre variables. Un tamaño de $\mathbf{H}$ pequeno da lugar a moito ruído, mentres que un grande pode ocultar a estrutura real dos datos. Definimos o ancho de banda óptimo como aquel que minimiza o MISE e consideramos a súa aproximación asintótica dada por

$$\mathbf{H}_{\text{AMISE}} = \text{argmin}_{\mathbf{H}\in\mathcal{F}}\text{AMISE}\{\hat{f}(\cdot;\mathbf{H})\}.$$

Non obstante, ambos valores dependen de f, que na práctica se descoñece, polo que non se poden calcular a partir dos datos da mostra.

Os selectores de ancho de banda xorden de distintos estimadores do MISE e do AMISE, en particular, de diferentes aproximacións do nesgo ó cadrado integrado. Ó longo deste capítulo presentamos o selector de escala normal (Rule-of-Thumb), o método de validación cruzada insesgada e o método plug-in para matrices de ancho de banda sen restricións. Os fundamentos teóricos, tal e como se presentan no Capítulo 3 do libro *Multivariate Kernel Smoothing and Its Applications* (Chacon & Duong, 2018), constitúen a base principal deste desenvolvemento.

2.1. Selectores de escala normal

Os selectores de ancho de banda de escala normal son os máis sinxelos, polo que adoitan ser os primeiros en empregarse para un estimador de tipo núcleo. Coñécense como o método da regra do pulgar. A súa obtención consiste en substituír a densidade f por unha normal $\phi_{\mathbf{\Sigma}}(\cdot - \boldsymbol{\mu})$, onde $\boldsymbol{\mu}$ é a media e $\mathbf{\Sigma}$ a matriz de varianzas e covarianzas, e estimar as cantidades descoñecidas $\mathbf{H}_{\text{AMISE}}$ ou $\mathbf{H}_{\text{MISE}}$ cando se usa o estimador tipo núcleo con kernel gaussiano. Desta forma, obtemos que

$\mathbb{E}\{\hat{f}(\mathbf{x};\mathbf{H})\} = (K_{\mathbf{H}} * f)(\mathbf{x})$ segue unha distribución normal de media $\boldsymbol{\mu}$ e matriz de varianzas e covarianzas $\boldsymbol{\Sigma} + \mathbf{H}$. Polo tanto, pódense escribir as fórmulas do MISE e do AMISE de $\hat{f}$ como

$$\begin{aligned}\mathrm{MISE}_{\mathrm{NS}}\{\hat{f}(\cdot;\mathbf{H})\} = {} & n^{-1}|\mathbf{H}|^{-1/2}(4\pi)^{-d/2} + (2\pi)^{-d/2}\big\{(1-n^{-1})|2\mathbf{H}+2\boldsymbol{\Sigma}|^{-1/2} \\ & - 2|\mathbf{H}+2\boldsymbol{\Sigma}|^{-1/2} + |2\boldsymbol{\Sigma}|^{-1/2}\big\}\end{aligned}$$

$$\begin{aligned}\mathrm{AMISE}_{\mathrm{NS}}\{\hat{f}(\cdot;\mathbf{H})\} = {} & n^{-1}|\mathbf{H}|^{-1/2}(4\pi)^{-d/2} + \tfrac{1}{16}(4\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2} \\ & \times \big\{2\mathrm{tr}(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}\boldsymbol{\Sigma}^{-1}) + \mathrm{tr}^2(\mathbf{H}\boldsymbol{\Sigma}^{-1})\big\}.\end{aligned}$$

Para obter o mínimo do MISE non existe unha fórmula explícita para ningunha dimensión. Non obstante, a través de Chacon e Duong (2018, p.44), coñecemos que Wand (1992) demostrou que dada $\mathbf{H} \in \mathcal{F}$, o mínimo do $\mathrm{AMISE}_{\mathrm{NS}}$ é alcanzado en

$$\mathbf{H}_{\mathrm{NS}} = \left(\frac{4}{n(d+2)}\right)^{2/(d+4)} \boldsymbol{\Sigma}.$$

Substituíndo o valor da matriz de varianzas e covarianzas poboacional pola mostral estimada $\hat{\boldsymbol{\Sigma}}$, obtemos o ancho de banda de escala normal baseado na mostra de datos, dado como

$$\hat{\mathbf{H}}_{\mathrm{NS}} = \left(\frac{4}{n(d+2)}\right)^{2/(d+4)} \hat{\boldsymbol{\Sigma}}. \tag{2.1}$$

Aínda que a ecuación (2.1) é relativamente simple, cando se emprega este método para despois aplicar o resultado en problemas máis complexos, é útil considerar un ancho de banda restrinxido da forma $\mathbf{H} = h^2\mathbf{I}_d$ ou $\mathbf{H} = \mathrm{diag}(h_1^2, ..., h_d^2)$. Por exemplo, na Sección 2.3 veremos que para aplicar o método plug-in e calcular $\hat{\mathbf{H}}_{\mathrm{PI}}$, primeiro é necesario estimar unha matriz $\hat{\mathbf{G}}_{\mathrm{NS},6}$ cun selector normal, que non dará lugar a unha gran perda de eficiencia no resultado final. Se consideramos $\mathbf{H} = h^2\mathbf{I}_d$, tense que o ancho de banda que minimiza o $\mathrm{AMISE}_{\mathrm{NS}}\{\hat{f}(\cdot;\mathbf{H})\}$ é

$$h_{\mathrm{NS}} = \left(\frac{4d|\boldsymbol{\Sigma}|^{1/2}}{n[2\mathrm{tr}(\boldsymbol{\Sigma}^{-2}) + \mathrm{tr}^2(\boldsymbol{\Sigma}^{-1})]}\right)^{1/(d+4)}, \tag{2.2}$$

mentres que se $\mathbf{H} = \mathrm{diag}(h_1^2, ..., h_d^2)$, entón o ancho en cada coordenada i vén dado como

$$h_{\mathrm{NS},i} = \left(\frac{4d|\boldsymbol{\Delta}|^{1/2}}{n[2\mathrm{tr}(\boldsymbol{\Delta}^{-2}) + \mathrm{tr}^2(\boldsymbol{\Delta}^{-1})]}\right)^{1/(d+4)} \boldsymbol{\sigma}_i, \tag{2.3}$$

onde $\boldsymbol{\sigma}_i$ denota a desviación típica na i-ésima coordenada e $\boldsymbol{\Delta} = (\mathrm{diag}\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}$ (Chacón & Duong, 2018, Sección 5.8). De forma análoga ó caso non restrinxido, os anchos de banda baseados na escala normal obtéñense substituíndo a matriz de varianzas e covarianzas pola súa estimación mostral.

Observación 2.1. Aínda que as matrices óptimas obtidas a partir das ecuacións (2.2) e (2.3) están restrinxidas, a súa estimación vén dada a partir de toda a información, é dicir, os seus valores dependen da varianza e da covarianza das variables.

Aplicamos a estimación de tipo núcleo co selector de ancho de banda normal no modelo M3 da mixtura de dúas normais bidimensionais. En R, calculamos o selector $\hat{\mathbf{H}}_{\text{NS}}$ a partir da ecuación (2.1), que da lugar á matriz

$$\hat{\mathbf{H}}_{\text{NS}} = \begin{bmatrix} 0{,}6583 & -0{,}5155 \\ -0{,}5155 & 0{,}6323 \end{bmatrix}. \tag{2.4}$$

A continuación, para estimar a densidade de tipo núcleo, cargamos a libraría *ks* e aplicamos a función *kde*. O resultado da estimación móstrase na Figura 2.1. A estimación obtida parece ser

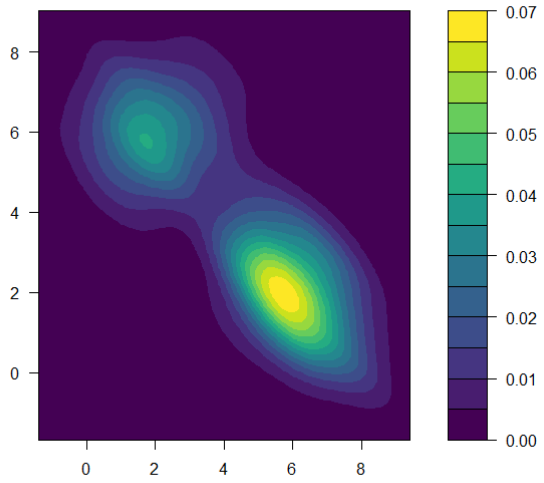

Figura 2.1: Estimación de tipo núcleo do modelo M3 co selector de ancho de banda normal $\hat{\mathbf{H}}_{\text{NS}}$ da ecuación (2.4).

unimodal, pois a moda superior esquerda non está moi marcada. Ademais, non diferenza os dous agrupamentos, senón que estes aparecen conectados. Polo tanto, o método non capta a estrutura correcta da densidade. Esta é unha das súas principais limitacións.

O selector de ancho de banda normal é moi empregado debido a súa simplicidade computacional. Non obstante, depende da suposición dunha distribución paramétrica: a normal. Cando os datos se desvían moito desta suposición, como por exemplo no caso de distribucións bimodais ou asimétricas, o ancho obtido pode non ser óptimo. Esta sensibilidade á forma da distribución limita a aplicación deste selector en moitos casos prácticos. Por ese motivo, estudaremos outros métodos que reduzan esta dependencia.

2.2. Validación cruzada

O método de validación cruzada busca optimizar empíricamente a elección do ancho de banda $\mathbf{H}$ do estimador de tipo núcleo. Existen dous enfoques principais: nesgado e insesgado. Nesta sección, trataremos o caso insesgado, posto que o outro enfoque non é de gran interés debido o seu gran custo computacional (Chacón & Duong, 2018, pp.47-49). Un estimador $\hat{\theta}$ é insesgado cando a súa media coincide co valor real, é dicir, $\mathbb{E}[\hat{\theta}] = \theta$. No caso contrario, dise que o estimador é nesgado.

O método de validación cruzada (UCV), tamén coñecido como validación cruzada por mínimos cadrados (LSCV), utiliza a técnica *leave-one-out* para seleccionar o ancho de banda óptimo. Este enfoque consiste en adestrar o modelo n veces, excluíndo cada vez unha das n observacións, e avaliando o erro sobre a observación eliminada. O obxectivo é minimizar o erro cadrático integrado (ISE).

Partimos da ecuación (1.10) do ISE e expandímola de forma que

$$\mathrm{ISE}\{\hat{f}(\mathbf{x};\mathbf{H})\} = \int_{\mathbb{R}^d} \hat{f}(\mathbf{x};\mathbf{H})^2 d\mathbf{x} - 2\int_{\mathbb{R}^d} \hat{f}(\mathbf{x};\mathbf{H}) f(\mathbf{x}) d\mathbf{x} + R(f),$$

onde o primeiro termo é coñecido, o último non afecta na estimación xa que non depende de $\mathbf{H}$ e para o segundo descoñecemos a densidade f. Podemos expresar a integral do segundo termo como a esperanza condicional $\mathbb{E}\{\hat{f}(\mathbf{X};\mathbf{H})|\mathbf{X}_1,...,\mathbf{X}_n\}$ para unha variable aleatoria $\mathbf{X} \sim f$ independente da mostra. Consideremos o estimador *leave-one-out* como o estimador de tipo núcleo eliminando o i-ésimo elemento da mostra, de forma que $\hat{f}_{-i}(\mathbf{x};\mathbf{H}) = (n-1)^{-1}\sum_{j=1,j\neq i}^{n} K_{\mathbf{H}}(\mathbf{x}-\mathbf{X}_j)$. Entón, podemos construír un estimador insesgado da esperanza condicional $\mathbb{E}\{\hat{f}(\mathbf{X};\mathbf{H})|\mathbf{X}_1,...,\mathbf{X}_n\}$ como $n^{-1}\sum_{i=1}^{n}\hat{f}_{-i}(\mathbf{X}_i;\mathbf{H})$. Polo tanto, o criterio de validación cruzada insesgada (UCV) vén dado por

$$\mathrm{UCV}(\mathbf{H}) = \int_{\mathbb{R}^d} \hat{f}(\mathbf{x};\mathbf{H})^2 d\mathbf{x} - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(\mathbf{X}_i;\mathbf{H}),$$

onde $\hat{\mathbf{H}}_{\mathrm{UCV}} = \mathrm{argmin}_{\mathbf{H}\in\mathcal{F}}\mathrm{UCV}(\mathbf{H})$ da lugar ó estimador de ancho de banda por validación cruzada.

Para facilitar a estimación de $\mathbf{H}$ na práctica, expresamos o criterio en función do núcleo $K_{\mathbf{H}}$. Aplicando a definición do estimador tipo núcleo no primeiro termo, temos que

$$\int_{\mathbb{R}^d} \hat{f}(\mathbf{x};\mathbf{H})^2 d\mathbf{x} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{\mathbb{R}^d} K_{\mathbf{H}}(\mathbf{x}-\mathbf{X}_i)K_{\mathbf{H}}(\mathbf{x}-\mathbf{X}_j)d\mathbf{x}.$$

Considerando o cambio de variable $\mathbf{u} = \mathbf{x} - \mathbf{X}_j$ e a definición de convolución, obtemos que

$$\begin{aligned}\int_{\mathbb{R}^d} \hat{f}(\mathbf{x};\mathbf{H})^2 d\mathbf{x} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(K_{\mathbf{H}} * K_{\mathbf{H}})(\mathbf{X}_i - \mathbf{X}_j) \\ &= \frac{n}{n^2}(K_{\mathbf{H}} * K_{\mathbf{H}})(0) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}(K_{\mathbf{H}} * K_{\mathbf{H}})(\mathbf{X}_i - \mathbf{X}_j) \\ &= \frac{1}{n}|\mathbf{H}|^{-1/2}R(K) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}(K_{\mathbf{H}} * K_{\mathbf{H}})(\mathbf{X}_i - \mathbf{X}_j),\end{aligned}$$

onde na segunda igualdade separamos os termos $i = j$ e $i \neq j$ e na terceira aplicamos a definición de convolución e do núcleo en $i = j$, da forma

$$\frac{1}{n^2}\sum_{i=1}^{n}(K_{\mathbf{H}} * K_{\mathbf{H}})(0) = \frac{1}{n^2}\sum_{i=1}^{n}\int_{\mathbb{R}^d} K_{\mathbf{H}}^2(\mathbf{u})d\mathbf{u} = \frac{n}{n^2}\sum_{i=1}^{n}|\mathbf{H}|^{-1/2}R(K) = \frac{1}{n}|\mathbf{H}|^{-1/2}R(K).$$

Por outra parte, empregando a definición do estimador *leave-one-out*, o segundo termo do criterio UCV pódese escribir como

$$\frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(\mathbf{X}_i;\mathbf{H}) = \frac{2}{n(n-1)}\sum_{\substack{i,j=1\\ j\neq i}}^{n}K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{X}_j).$$

Finalmente, reescribimos o criterio de validación cruzada insesgado como

$$\mathrm{UCV}(\mathbf{H}) = \frac{1}{n}|\mathbf{H}|^{-1/2}R(K) + \frac{1}{n(n-1)}\sum_{\substack{i,j=1\\ j\neq i}}^{n}\{(1-n^{-1})(K_{\mathbf{H}} * K_{\mathbf{H}}) - 2K_{\mathbf{H}}\}(\mathbf{X}_i - \mathbf{X}_j), \quad (2.5)$$

e obtemos a súa esperanza como $\mathbb{E}\{\mathrm{UCV}(\mathbf{H})\} = \mathrm{MISE}\{\hat{f}(\mathbf{x};\mathbf{H})\} - R(f)$. Polo tanto, ignorando a constante $R(f)$ que non depende do ancho de banda, temos que o UCV é un estimador insesgado do MISE.

Observación 2.2. A identificación $1 - n^{-1} \approx 1$ na ecuación (2.5) é común debido a súa simplicidade. Aínda que introduce un pouco de nesgo, Chacón e Duong (2018) afirman que o ancho de banda resultante é asintóticamente equivalente.

Existen dous problemas comúns neste método. Un deles é a existencia de varios mínimos locais na curva UCV no caso unidimensional, que pode dar lugar a resultados diferentes en función do mínimo local seleccionado. Por outra parte, está a tendencia ó subsuavizado, sobre todo en mostras pequenas. Para mitigar estes problemas, é habitual definir un rango fixo, afastado de valores pequenos de ancho de banda, e iniciar o proceso cun ancho de banda relativamente grande.

Estimamos a matriz de ancho de banda do modelo M3 co método de validación cruzada, co obxectivo de representar a estimación de densidade tipo núcleo. En R, mediante o paquete *ks*, podemos aplicar a función *Hscv* ós datos para obter o selector de validación cruzada insesgada,

$$\hat{\mathbf{H}}_{\text{UCV}} = \begin{bmatrix} 0{,}2606 & -0{,}1001 \\ -0{,}1001 & 0{,}2578 \end{bmatrix}. \tag{2.6}$$

Na Figura 2.2 mostramos a representación da densidade estimada. Podemos ver que a separación das modas é máis clara que no caso do selector normal da Figura 2.1. Isto pode deberse a que o selector $\hat{\mathbf{H}}_{\text{UCV}}$ é notablemente máis pequeno que $\hat{\mathbf{H}}_{\text{NS}}$. Ademais, a estrutura da estimación é bastante detallada, aproximándose á densidade real do modelo. Non obstante, como o tamaño da mostra non é moi grande, a estimación presenta bastante ruído, dando lugar a unha menor suavización.

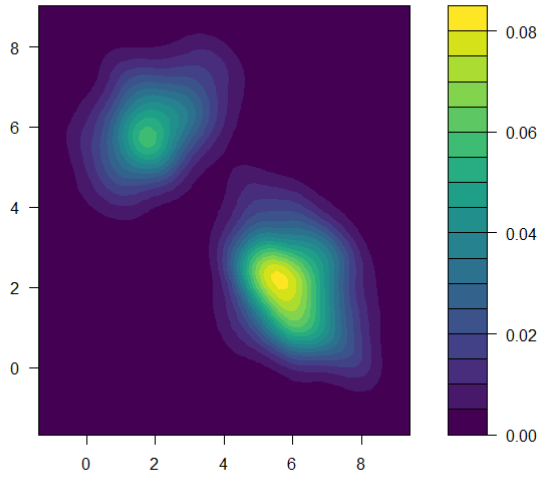

Figura 2.2: Estimación de tipo núcleo do modelo M3 co selector obtido mediante o método de validación cruzada insesgada $\hat{\mathbf{H}}_{\text{UCV}}$ da ecuación (2.6).

Unha alternativa á validación cruzada é o método plug-in, que busca estimar o ancho de banda mediante a substitución dos termos descoñecidos do erro cadrático medio asintótico (AMISE) por estimadores consistentes. A diferenza é que a validación cruzada estima directamente o ISE, que depende da mostra e varía en función dos datos dispoñibles, mentres que o método plug-in está enfocado na aproximación asintótica do MISE, que realiza un promedio e resulta máis sinxelo de estimar, xa que non depende tanto da variabilidade da mostra.

2.3. Plug-in

Os selectores de ancho de banda plug-in para datos multivariantes con matrices restrinxidas foron introducidos por Wand e Jones (1994), mentres que o enfoque actual para as matrices de ancho de banda sen restricións foi perfeccionado por Chacón e Duong (2010).

O método plug-in para a selección da matriz de ancho de banda $\mathbf{H}$ baséase na forma asintótica do erro cadrático medio integrado, o AMISE. Na ecuación (1.12) vimos que o único termo descoñecido é $\mathbf{R}(D^{\otimes 2}f) \in \mathcal{M}_{d^2\times d^2}$, polo que as distintas formas de estimar esa matriz dan lugar a diferentes versións do método plug-in. Neste traballo tratamos a estimación con matrices de ancho de banda sen restricións.

Definimos $\boldsymbol{\psi}_r = \int_{\mathbb{R}^d} D^{\otimes r} f(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \in \mathbb{R}^{d^r}$. Usando integración por partes baixo a condición de que f teña $2s$ derivadas continuas integrables, e aplicando as propiedades do Anexo A, demostramos que $\operatorname{vec}[\mathbf{R}(D^{\otimes s}f)] = (-1)^s \boldsymbol{\psi}_{2s}$. Polo tanto, o problema de estimar $\mathbf{R}(D^{\otimes s}f)$ é equivalente a estimar o vector $\boldsymbol{\psi}_{2s}$. Como $\boldsymbol{\psi}_{2s} = \mathbb{E}\{D^{\otimes 2s} f(\mathbf{X})\}$, para $\mathbf{X} \sim f$ un vector aleatorio, consideramos o estimador

$$\hat{\boldsymbol{\psi}}_{2s} \equiv \hat{\boldsymbol{\psi}}_{2s}(\mathbf{G}) = \frac{1}{n}\sum_{i=1}^{n} D^{\otimes 2s}\hat{f}(\mathbf{X}_i;\mathbf{G}) = \frac{1}{n^2}\sum_{i,j=1}^{n} D^{\otimes 2s} L_{\mathbf{G}}(\mathbf{X}_i - \mathbf{X}_j),$$

onde $\hat{f}$ é o estimador de tipo núcleo baseado en L, o núcleo piloto, e $\mathbf{G}$, a matriz de ancho de banda piloto. Considerando $s = 2$ baixo as hipóteses de **(A1)** do Teorema 1.3, obtemos $\hat{\boldsymbol{\psi}}_4$, polo que chegamos á estimación plug-in do AMISE substituíndoo en (1.12)

$$\operatorname{PI}(\mathbf{H};\mathbf{G}) = \frac{1}{n}|\mathbf{H}|^{-1/2}R(K) + \frac{1}{4}m_2(K)^2\hat{\boldsymbol{\psi}}_4(\mathbf{G})^{\top}(\operatorname{vec}(\mathbf{H}))^{\otimes 2}, \tag{2.7}$$

onde o selector de ancho de banda estimado é $\hat{\mathbf{H}}_{\mathrm{PI}} = \operatorname{argmin}_{\mathbf{H}\in\mathcal{F}}\operatorname{PI}(\mathbf{H};\mathbf{G})$.

Na ecuación (2.7) vemos que para obter $\mathbf{H}$ é necesario estimar a matriz $\mathbf{G}$. Non obstante, a elección de $\mathbf{G}$ ten un efecto secundario sobre $\hat{f}$ a través de $\hat{\boldsymbol{\psi}}_4$. Podemos ver que, dado o erro cadrático medio de $\hat{\mathbf{H}}_{\mathrm{PI}}$, $\operatorname{MSE}(\hat{\mathbf{H}}_{\mathrm{PI}}) = \mathbb{E}\{||\operatorname{vec}(\hat{\mathbf{H}}_{\mathrm{PI}} - \mathbf{H}_{\mathrm{AMISE}})||^2\}$, Duong e Hazelton (2005a, Lema 1) demostraron que este era igual a

$$\operatorname{MSE}(\hat{\mathbf{H}}_{\mathrm{PI}}) = cte \cdot \mathbb{E}\{||\hat{\boldsymbol{\psi}}_4(\mathbf{G}) - \boldsymbol{\psi}_4||^2\}\{1 + o(1)\} = cte \cdot \operatorname{MSE}(\hat{\boldsymbol{\psi}}_4(\mathbf{G}))\{1 + o(1)\},$$

o que implica que o rendemento de $\hat{\mathbf{H}}_{\mathrm{PI}}$ está directamente relacionado co erro cadrático medio de $\hat{\boldsymbol{\psi}}_4$. O criterio para a selección óptima de $\mathbf{G}$ consiste en minimizar o erro cadrático medio asintótico (AMSE), que Chacón e Duong (2010, Teorema 1) demostraron que pode escribirse como

$$\operatorname{AMSE}\{\hat{\boldsymbol{\psi}}_4(\mathbf{G})\} = ||\frac{1}{n}|\mathbf{G}|^{-1/2}(\mathbf{G}^{-1/2})^{\otimes 4}D^{\otimes 4}L(\mathbf{0}) + \frac{1}{2}m_2(L)(\operatorname{vec}^{\top}\mathbf{G}\otimes\mathbf{I}_{d^4})\boldsymbol{\psi}_6||^2, \tag{2.8}$$

onde se descoñece o valor de $\boldsymbol{\psi}_6$. Substituímos esta función polo estimador de tipo núcleo $\hat{\boldsymbol{\psi}}_6$, empregando como matriz de ancho de banda piloto a que minimiza o AMSE$\{\hat{\boldsymbol{\psi}}_6(\hat{\mathbf{G}})\}$ no caso normal, $\hat{\mathbf{G}}_{\mathrm{NS},6} = \{2/(d+6)^{2/(d+8)} 2\hat{\boldsymbol{\Sigma}} n^{-2/(d+8)}\}$ (Chacón & Duong 2010, Ecuación 8). O selector de ancho de banda normal $\hat{\mathbf{G}}_{\mathrm{NS},6}$ de $\hat{\boldsymbol{\psi}}_6$ ten unha perda de eficiencia menor que se considerásemos inicialmente o estimador $\hat{\mathbf{H}}_{\mathrm{NS}}$ de $\hat{f}$ enunciado na Sección 2.1, polo que é práctico para simplificar o proceso. Este método coñécese como selector de dúas etapas, xa que precisa estimar $\hat{\boldsymbol{\psi}}_4$ e $\hat{\boldsymbol{\psi}}_6$.

A continuación, estimamos a matriz de ancho de banda do modelo M3 co método plug-in de dúas etapas mediante R. A función *Hpi* do paquete *kde* calcula o selector directamente, permitindo obter a estimación de tipo núcleo fácilmente aplicando a función *kde* ós nosos datos co selector obtido,

$$\hat{\mathbf{H}}_{\mathrm{PI}} = \begin{bmatrix} 0{,}2487 & -0{,}1204 \\ -0{,}1204 & 0{,}2427 \end{bmatrix}. \tag{2.9}$$

Na Figura 2.3 mostramos a representación da densidade estimada. O resultado é moi similar á estimación obtida mediante o método de validación cruzada, aproximándose á densidade real. De feito, neste caso a matriz de ancho de banda obtida mediante o método plug-in é moi similar a do método de validación cruzada, de ahí a semellanza das estimacións.

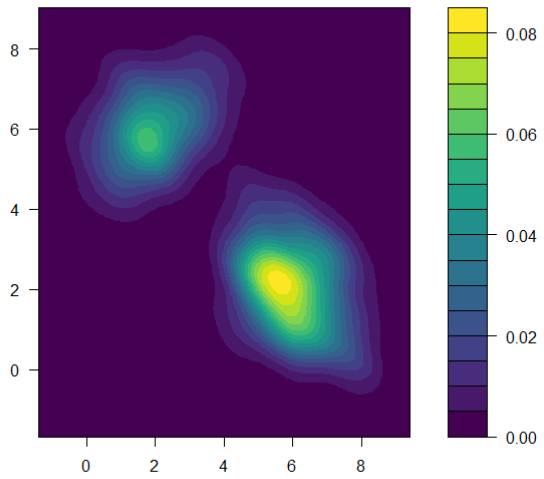

Figura 2.3: Estimación de tipo núcleo do modelo M3 co selector obtido co método plug-in $\hat{\mathbf{H}}_{\mathrm{PI}}$ da ecuación (2.9).

O parecido entre as estimacións presentadas mediante o método plug-in e o de validación cruzada insesgada, máis a súa similitude á densidade real, ilustra a capacidade de ambos métodos para captar a estrutura real dos datos. Non obstante, isto non sempre ocorre, polo que convén

ter en conta as vantaxes e limitacións de cada un á hora de escoller o selector que queremos empregar.

En resumo, o selector de ancho de banda normal destaca pola súa sinxeleza computacional e a súa rápida aplicación, pero a súa gran dependencia da hipótese de normalidade dos datos limita a súa aplicación en casos prácticos con estruturas máis complexas, como acontece no modelo sinalado M3. Por outra parte, o método de validación cruzada insesgado pode ser máis efectivo á hora de captar as estruturas complexas nos datos, pero pode presentar unha alta variabilidade e unha tendencia ó subsuavizado. Finalmente, o método plug-in adoita ser máis práctico e estable que o de validación cruzada, pero presenta unha maior complexidade computacional e pode ter dificultades.

Estos resultados resaltan a importancia de escoller un bo método de selección de ancho de banda, tendo en conta o compromiso entre precisión e complexidade, así como as características do conxunto de datos a estudar.

Capítulo 3

Análise de casos de leucemia

Logo de expoñer na Introdución o obxectivo deste estudo e de presentar os conceptos teóricos nos Capítulos 1 e 2, nesta última parte procedemos á aplicación práctica da teoría exposta. Cómpre sinalar que parte dos datos que se van analizar xa foron utilizados previamente para ilustrar algunhas das técnicas presentadas, pero reproducímolos aquí para ofrecer unha visión completa da análise realizada.

Os estudos de patróns espaciais en epidemioloxía buscan identificar factores de risco de distintas enfermidades. Neste caso, a distribución espacial dos casos de leucemia no noroeste de Inglaterra entre 1982 e 1998, fronte á distribución dos seus respectivos controis, suxire a posible presenza dun patrón espacial influínte na incidencia de leucemia. A metodoloxía empregada neste estudo baséase na estimación de densidade de tipo núcleo, presentada no Capítulo 1, que suaviza as distribucións espaciais sen necesidade de asumir unha estrutura previa. Esta técnica require dunha selección adecuada do ancho de banda, que é o que controla o grao de suavización.

A estrutura do capítulo presenta tres fases diferenciadas. En primeiro lugar, na Sección 3.1 realizamos a estimación de densidade de tipo núcleo dos casos de leucemia, empregando os distintos selectores de ancho de banda explicados no Capítulo 2. Unha vez identificado o selector que mellor se adapta á estrutura dos casos, realizamos o mesmo proceso para a estimación da densidade dos controis na Sección 3.2. Finalmente, na Sección 3.3 analizamos se existen diferenzas significativas entre as densidades dos casos e dos controis, co obxectivo de identificar a posible presenza de patróns espaciais.

3.1. Estimación da densidade dos casos

A identificación de posibles patróns espaciais na incidencia de leucemia no noroeste de Inglaterra require da aplicación de métodos de estimación non paramétrica. Concretamente, aplicamos a estimación de tipo núcleo con tres selectores de ancho de banda distintos, co obxectivo de seleccionar aquel que mellor axuste os datos.

En primeiro lugar, calculamos en R a ecuación (2.1) aplicada ós casos de leucemia, obtendo así o selector de ancho de banda normal

$$\hat{\mathbf{H}}_{\mathrm{NS}} = \begin{bmatrix} 0.0097 & -0.0061 \\ -0.0061 & 0.0080 \end{bmatrix}. \tag{3.1}$$

Unha vez estimada a matriz de ancho de banda, aplicamos a función *kde* do paquete *ks* para realizar a estimación de tipo núcleo. A representación gráfica correspondente lévase a cabo mediante a función *filled.contour*, que permite visualizar a estrutura espacial dos casos de leucemia nun mapa de calor.

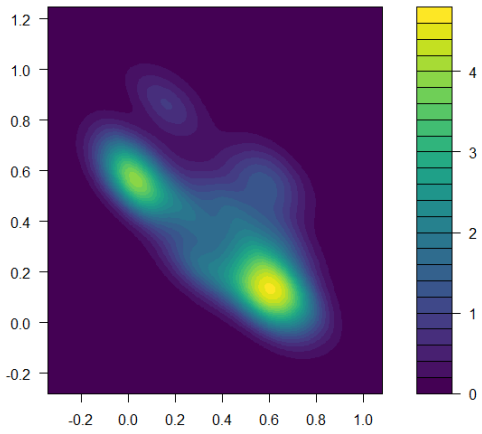

Figura 3.1: Estimación de tipo núcleo dos casos de leucemia co selector de ancho de banda normal $\hat{\mathbf{H}}_{\mathrm{NS}}$ da ecuación (3.1).

A Figura 3.1 mostra a estimación resultante. Podemos ver que o gráfico de calor capta dúas zonas de alta densidade, que se corresponden coas áreas de alta concentración de casos na Figura 1. Non obstante, o seleccionar o ancho de banda supondo unha distribución normal, a estimación tende a suavizarse en exceso, o que leva a unha perda de detalles locais. Por exemplo, podemos apreciar que o espazo situado entre as dúas agrupacións do gráfico aparece conectado, sen captar as posibles estruturas intermedias. De feito, veremos nas estimacións posteriores que esta rexión

presenta máis detalles que aquí non se poden apreciar. Este efecto de suavización excesiva, ademais de non mostrar certas estruturas, tamén pode ocultar modas secundarias. Detectar zonas de alta concentración de casos parece especialmente relevante nesta aplicación.

Por outra parte, vimos no Capítulo 2 que a función *Hscv* de R calcula directamente a matriz estimada polo método de validación cruzada, de forma que

$$\hat{\mathbf{H}}_{\mathrm{UCV}} = \begin{bmatrix} 0.0035 & -0.0015 \\ -0.0015 & 0.0033 \end{bmatrix}. \tag{3.2}$$

Substituíndo o seu valor na función *kde*, obtemos a estimación da densidade representada na Figura 3.2. Este método permite identificar estruturas máis detalladas que co selector normal non apreciábamos. Por exemplo, a zona que conecta ambas modas agora presenta pequenas áreas diferenciadas, que antes viamos como un camiño uniforme debido á suavización excesiva. Isto indica que estamos ante un método máis preciso, que capta con máis detalle as pequenas variacións locais.

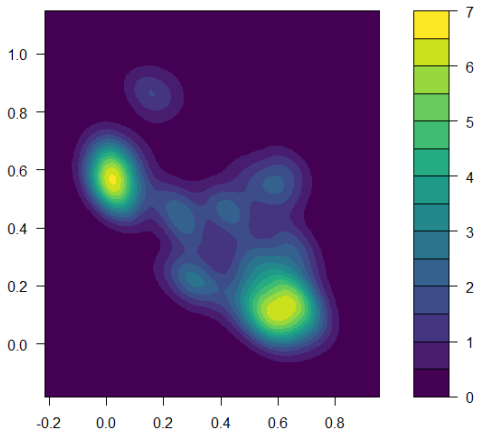

Figura 3.2: Estimación de tipo núcleo dos casos de leucemia co selector obtido mediante o método de validación cruzada $\hat{\mathbf{H}}_{\mathrm{UCV}}$ da ecuacion (3.2).

Finalmente, o último método que tratamos neste traballo para a obtención do selector de ancho de banda é o plug-in. A función *Hpi* de R calcula directamente a estimación da matriz,

$$\hat{\mathbf{H}}_{\mathrm{PI}} = \begin{bmatrix} 0.0033 & -0.0015 \\ -0.0015 & 0.0030 \end{bmatrix}, \tag{3.3}$$

que substituíndo na función *kde* da lugar á densidade estimada e representada na Figura 3.3. O resultado desta estimación é moi similar ó de validación cruzada, o que era de esperar debido

á semellanza entre as matrices de ancho de banda estimadas. Polo tanto, a estrutura capta novamente as pequenas variacións locais no espazo intermedio entre as dúas modas.

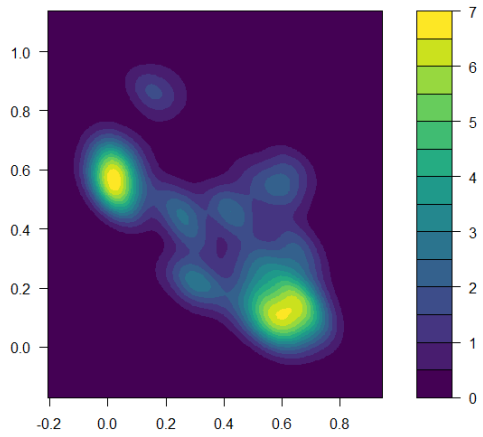

Figura 3.3: Estimación de tipo núcleo dos casos de leucemia co selector obtido mediante o método plug-in $\hat{\mathbf{H}}_{\mathrm{PI}}$ da ecuación (3.3).

Dado que o método de validación cruzada e o de plug-in dan lugar a estimacións moi semellantes e máis detalladas ca do caso normal, consideramos a estimación dada polo método de validación cruzada como a máis acertada.

3.2. Estimación da densidade dos controis

A continuación, realizamos un proceso análogo ó da Sección 3.1 cos datos correspondentes ós controis do estudo, co obxectivo de obter unha estimación precisa da súa densidade. En primeiro lugar, calculamos en R o selector de ancho de banda normal, que da lugar á matriz

$$\hat{\mathbf{H}}_{\mathrm{NS},2} = \begin{bmatrix} 0.0039 & -0.0024 \\ -0.0024 & 0.0039 \end{bmatrix}. \tag{3.4}$$

Empregando esta matriz na estimación de tipo núcleo dos controis, obtemos a densidade representada na Figura 3.4. Neste caso, a diferenza do que acontece coa densidade dos casos, a distribución é unimodal, cuxa zona de alta densidade está localizada no sueste da área de estudo. Dado que a estrutura dos datos parece ser unimodal e non é moi complexa, o selector normal proporciona unha estimación suave e bastante precisa, que ademais é sinxela de calcular.

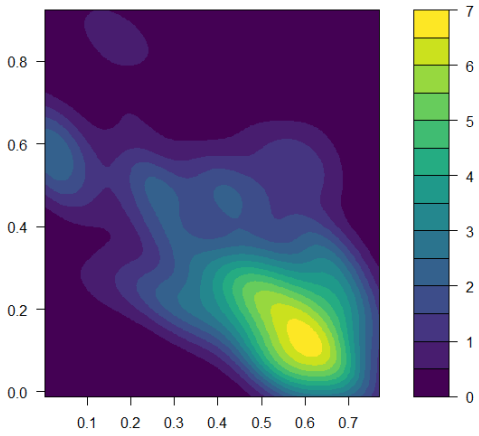

Figura 3.4: Estimación de tipo núcleo dos controis de leucemia cun selector de ancho de banda normal $\hat{\mathbf{H}}_{\mathrm{NS},2}$ da ecuación (3.4).

Por outra parte, o método de validación cruzada da lugar á seguinte matriz de ancho de banda,

$$\hat{\mathbf{H}}_{\mathrm{UCV},2} = \begin{bmatrix} 0.0035 & -0.0026 \\ -0.0026 & 0.0035 \end{bmatrix}. \tag{3.5}$$

A estimación da densidade correspondente, mostrada na Figura 3.5, é moi similar á obtida co selector normal.

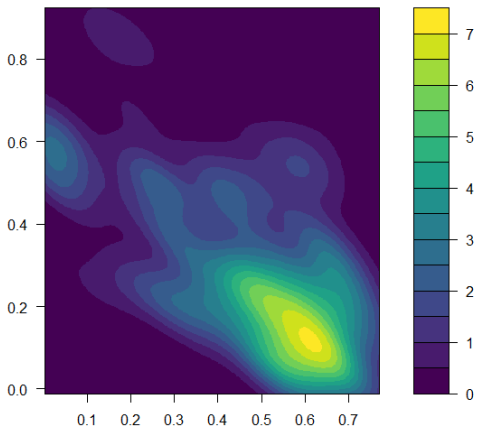

Figura 3.5: Estimación de tipo núcleo dos controis de leucemia co selector obtido mediante o método de validación cruzada $\hat{\mathbf{H}}_{\mathrm{UCV},2}$ da ecuación (3.5).

Finalmente, aplicando o método plug-in obtense a matriz

$$\hat{\mathbf{H}}_{\mathrm{PI},2} = \begin{bmatrix} 0.0014 & -0.0005 \\ -0.0005 & 0.0015 \end{bmatrix}. \tag{3.6}$$

Neste caso, os valores de ancho de banda son máis pequenos que os obtidos mediante outros selectores, polo que a estimación será máis localizada. Na Figura 3.6 vemos unha estrutura máis precisa, con áreas pequenas moi detalladas. Polo tanto, consideramos esta estimación como a máis adecuada para representar a densidade dos controis.

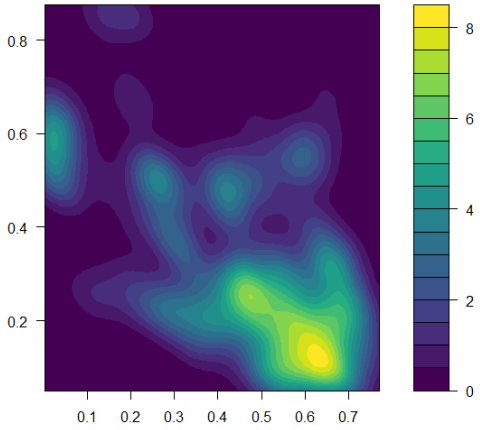

Figura 3.6: Estimación de tipo núcleo dos controis co selector obtido mediante o método plug-in $\hat{\mathbf{H}}_{\mathrm{PI},2}$ da ecuación (3.6).

3.3. Conclusión

Os resultados deste capítulo revelan diferenzas notables entre as distribucións espaciais dos casos e dos controis de leucemia. Mentres que os controis mostran unha estrutura unimodal, os casos amosan unha distribución bimodal, con dúas zonas de alta concentración diferenciadas. Isto suxire a posible presenza de focos de risco nas zonas onde non cabería esperar unha alta concentración de casos, sobre todo as situadas no noroeste da rexión estudada, que poden estar asociados a exposicións localizadas, como fontes de radiación ou complexos industriais.

A identificación destas áreas de risco son relevantes para futuras investigacións, pero queremos destacar que non constitúen unha proba definitiva da súa influencia na incidencia da enfermidade. Aínda que os controis foron tomados de forma rigurosa, non se pode descartar a intervención doutros factores externos non contemplados no estudo. En definitiva, os achados que permiten

acadar estos datos proporcionan unha base importante para comprender a distribución dos casos de leucemia e a posible influencia de patróns non aleatorios que merecen ser estudados con maior detalle, realizando estudos que integren datos ambientais, socioeconómicos ou outros de interese. Por suposto, aínda que se poidan tomar medidas preventivas que reduzan as posibilidades de presentar a enfermidade, isto non vai evitar por completo a incidencia da mesma, xa que existen moitos factores externos influíntes.

Anexo A

Notación e propiedades

A continuación recóllense as características fundamentais de operadores e matrices que teñen relevancia no desenvolvemento deste traballo. Estas propiedades son esenciais para a comprensión de resultados presentados ó longo dos Capítulos 1 e 2.

A.1 Produto de Kronecker

Dadas dúas matrices $\mathbf{A} = (a_{ij}) \in \mathcal{M}_{m\times n}$ e $\mathbf{B} \in \mathcal{M}_{p\times q}$, definimos o seu produto de Kronecker como a matriz $\mathbf{A}\otimes\mathbf{B} \in \mathcal{M}_{(mp)\times(nq)}$ da forma

$$\mathbf{A}\otimes\mathbf{B} = \left[\begin{array}{c|c|c} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \hline \vdots & \ddots & \vdots \\ \hline a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{array}\right],$$

onde o bloque (i,j) vén dado como $(a_{ij})\mathbf{B} \in \mathcal{M}_{p\times q}$. Polo tanto, o produto de Kronecker da matriz $\mathbf{A}$ consigo mesma r veces defínese como $\mathbf{A}^{\otimes r} = \mathbf{A}\otimes\cdots\otimes\mathbf{A} \in \mathcal{M}_{m^r\times n^r}$.

Dadas catro matrices conformables $\mathbf{A},\mathbf{B},\mathbf{C}$ e $\mathbf{D}$, temos que as propiedades máis importantes e empregadas ó longo do traballo son:

- $\mathbf{A}\otimes\mathbf{B}\otimes\mathbf{C} = (\mathbf{A}\otimes\mathbf{B})\otimes\mathbf{C} = \mathbf{A}\otimes(\mathbf{B}\otimes\mathbf{C})$
- $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = (\mathbf{AC})\otimes(\mathbf{BD})$
- $(\mathbf{A}\otimes\mathbf{B})^\top = \mathbf{A}^\top\otimes\mathbf{B}^\top$
- $\mathrm{tr}(\mathbf{A}\otimes\mathbf{B}) = \mathrm{tr}(\mathbf{A})\mathrm{tr}(\mathbf{B})$.

En xeral, o produto de Kronecker non é conmutativo. Ademais, se $\mathbf{A} \in \mathcal{M}_{n\times n}$ e $\mathbf{B} \in \mathcal{M}_{q\times q}$, entón $|\mathbf{A}\otimes\mathbf{B}| = |\mathbf{A}|^q|\mathbf{B}|^n$.

O produto de Kronecker tamén ten sentido para vectores. Se consideramos $\mathbf{a} = [a_1, ..., a_d] \in \mathbb{R}^d$ e $\mathbf{b} \in \mathbb{R}^p$, tense que $\mathbf{a} \otimes \mathbf{b} = [a_1\mathbf{b}; ...; a_d\mathbf{b}] \in \mathbb{R}^{dp}$ e $\mathbf{a} \otimes \mathbf{b}^\top = \mathbf{a}\mathbf{b}^\top = \mathbf{b}^\top \otimes \mathbf{a} \in \mathcal{M}_{d\times p}$.

A.2 Operador de vectorización

Sexa $\mathbf{A} = (a_{ij}) \in \mathcal{M}_{m\times n}$. O operador de vectorización transforma unha matriz nun vector columna, apilando as columnas unha debaixo da outra, de forma que $\text{vec}(\mathbf{A}) \in \mathbb{R}^{mn}$ é igual a

$$\text{vec}(\mathbf{A}) = [a_{11}, ..., a_{m1}; ...; a_{1n}, ..., a_{mn}]^\top.$$

Dadas tres matrices conformables $\mathbf{A}, \mathbf{B}$ e $\mathbf{C}$ e dous vectores $\mathbf{a}, \mathbf{b}$, as propiedades do operador vec máis empregadas son:

- $\text{vec}(\mathbf{a}) = \text{vec}(\mathbf{a}^\top) = \mathbf{a}$
- $\text{vec}(\mathbf{ABC}) = (\mathbf{C}^\top \otimes \mathbf{A})\text{vec}(\mathbf{B})$
- $\text{vec}(\mathbf{a}\mathbf{b}^\top) = \text{vec}(\mathbf{a} \otimes \mathbf{b}^\top) = \text{vec}(\mathbf{b}^\top \otimes \mathbf{a}) = \mathbf{b} \otimes \mathbf{a}$
- $\text{tr}(\mathbf{A}^\top\mathbf{B}) = (\text{vec}(\mathbf{A})^\top(\text{vec}(\mathbf{B}))$

A.3 Definicións e notacións variadas

A continuación, imos definir conceptos adicionais que se empregan ó londo do traballo.

- **Trasposta dun vector:** Sexa $\mathbf{x} \in \mathbb{R}^{1\times d}$ un vector fila. A súa trasposta, denotada por $\mathbf{x}^\top$, é o vector columna $\mathbf{x}^\top \in \mathbb{R}^d$.
- **Definición *o* pequena:** Sexan a_n e b_n dúas sucesións. Temos que $a_n = o(b_n)$ se

$$\lim_{n\to\infty} |\frac{a_n}{b_n}| = 0.$$

- **Operador diferencial d-dimensional:** Definimos o operador diferencial $\mathbf{D}$ como

$$\mathbf{D} = \left[\frac{\partial}{\partial x_1}, ..., \frac{\partial}{\partial x_d}\right]^\top.$$

- **Norma euclidiana dun vector:** Sexa $\mathbf{a} \in \mathbb{R}^d$, entón a norma euclidiana é $||\mathbf{a}|| = (\mathbf{a}\mathbf{a}^\top)^{1/2}$.
- **Integral dunha matriz:** Sexa $\mathbf{A}(\mathbf{x}) \in \mathcal{M}_{m\times n}$ unha matriz con entradas que dependen do vector $\mathbf{x} \in \mathbb{R}^d$. Entón, a integral da matriz nun conxunto $\Omega \subset \mathbb{R}^d$ é da forma

$$\left[\int_\Omega \mathbf{A}(\mathbf{x})d\mathbf{x}\right]_{ij} = \int_\Omega A_{ij}(\mathbf{x})d\mathbf{x},$$

para $i = 1, ..., m$ e $j = 1, ..., n$.

Bibliografía

[1] American Cancer Society (2019). *What Is Childhood Leukemia?*. https://www.cancer.org/cancer/types/leukemia-in-children/about/what-is-childhood-leukemia.html

[2] Burgos, J. de (2008). *Cálculo infinitesimal de varias variables (2.ª ed.)*. McGraw-Hill.

[3] Chacón, J. E. & Duong, T. (2010). Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test, 19*(3), 375-398.

[4] Chacón, J. E. & Duong, T. (2018). *Multivariate Kernel Smoothing and its applications*. CRC Press, Taylor & Francis Group.

[5] Duong, T. & Hazelton, M. L. (2005a). Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *Journal of Multivariate Analysis, 93*(2), 417-433.

[6] Duong, T. (2007). ks: Kernel Density Estimation and Kernel Discriminant Analysis for Multivariate Data in R. *Journal of Statistical Software, 21*(7), 1-16.

[7] Environmental Health Perspectives (2018). *Pesticides exposure and the risk of childhood acute leukemia*. https://ehp.niehs.nih.gov/doi/10.1289/isee.2011.00684

[8] Pearson, K. (1891). Contributions to the Mathematical Theory of Evolution. *Journal of the Royal Statistical Society, 56*(4), 675-679.

[9] Preston, D. L., Kusumi, S., Tomonaga, M., Izumi, S., Ron, E., Kuramoto, A., Kamada, N., Dohy, H., Matsui, T., Nonaka, H., Thompson, D. E., Soda, M., & Mabuchi, K. (1994). Cancer Incidence in Atomic Bomb Survivors. Part III: Leukemia, Lymphoma and Multiple Myeloma, 1950-1987. *Radiation Research, 137*(2, Suppl), S68-S97.

[10] Radiation Effects Research Foundation (RERF). (n.d.). *Leukemia Risks among Atomic-bomb Survivors*. https://www.rerf.or.jp/en/programs/roadmap_e/health_effects-en/late-en/leukemia/

[11] Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., & Firth, D. (2025). *Support Functions and Datasets for Venables and Ripley's MASS.* (Version 7.3.65). CRAN.

[12] Scott, D. W. (2015). *Averaged Shifted Histograms.* John Wiley & Sons.

[13] Shao, J. (2003). *Mathematical Statistics.* Springer.

[14] Silverman, B. W (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall.

[15] Trinchet Soria, R. M. (n.d.). *Cálculo vectorial e integración de Lebesgue. Medida e integral de Lebesgue.* Universidade de Santiago de Compostela.

[16] Wand, M. P. (1992). Error analysis for general multivariate kernel estimators, *Journal of Computational and Graphical Statistics 2*, 1-15.

[17] Wand, M. P. & Jones, M. C. (1994). Multivariate plug-in bandwidth selection. *Computational Statistics, 9*, 97-116.

[18] Wand, M. P. & Jones, M. C. (1995). *Kernel Smoothing.* Chapman & Hall.