Universidad Autónoma de Madrid | Escuela de Doctorado

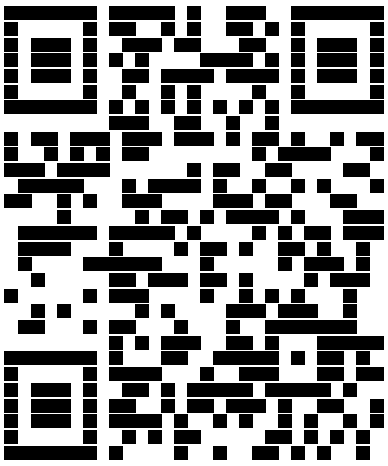
Estimación de características diferenciales y topológicas de las funciones de densidad de probabilidad

Javier Fernández Serrano

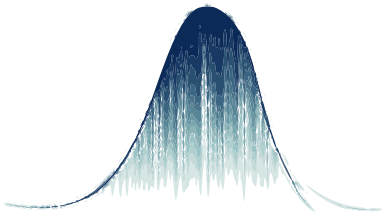
Programa de Doctorado en Matemáticas
Departamento de Matemáticas
Facultad de Ciencias

Madrid, 2025

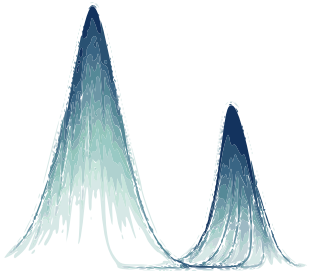
Javier Fernández Serrano es licenciado en Matemáticas e ingeniero en Informática por la Universidad Autónoma de Madrid (UAM). Posee másteres por la UAM en Matemáticas y Aplicaciones (especialidad en aplicaciones) y en Investigación e Innovación en TIC. Antes de acceder al Doctorado en Matemáticas, trabajó cinco años en una empresa aseguradora, ocupando distintos puestos en el ámbito de los datos y la tecnología. Al margen de la tesis, destaca entre sus intereses científicos la teoría de cópulas. En la actualidad, es miembro del proyecto de investigación *Statistical techniques in high-dimensional spaces*.



https://orcid.org/0000-0001-5270-9941

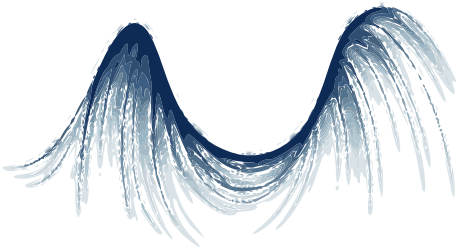


Universidad Autónoma de Madrid



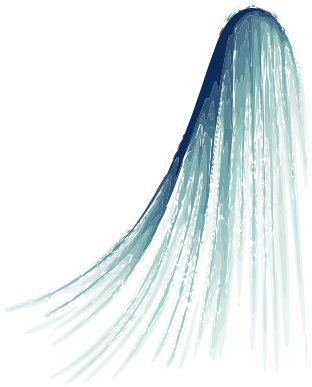
Estimación de características diferenciales y topológicas de las funciones de densidad de probabilidad

Memoria para optar al grado de doctor en el
Programa de Doctorado en Matemáticas
en la línea de investigación en
Estadística y Probabilidad



Presentada por
Javier Fernández Serrano

Dirigida por
José Enrique Chacón



Universidad Autónoma de Madrid
Facultad de Ciencias
Departamento de Matemáticas

– 2025 –

El autor declara que el contenido generado con inteligencia artificial en esta tesis se ha limitado *exclusivamente* a ilustraciones de carácter artístico.

Los derechos sobre dichas imágenes *pertenecen* al autor.

El autor *no* se responsabiliza de la naturaleza potencialmente ofensiva o imprecisa de dichas imágenes.

La fotografía de fondo en la anteportada, que está en el dominio *público*, es obra de Klaus-Dieter Keller.

El resto de las fotos empleadas como fondo de página son *obra* del autor.

Esta tesis y otros documentos relacionados se alojarán en

zenodo

https://doi.org/10.5281/zenodo.16207726

Escuela de Doctorado Multidisciplinar de la UAM
Calle Francisco Tomás y Valiente 2
Ciudad Universitaria de Cantoblanco
28049 MADRID
MADRID
SPAIN
https://www.uam.es

Depósito legal: M-19745-2025

Campus de Cantoblanco, frente a la Facultad de Ciencias de la UAM

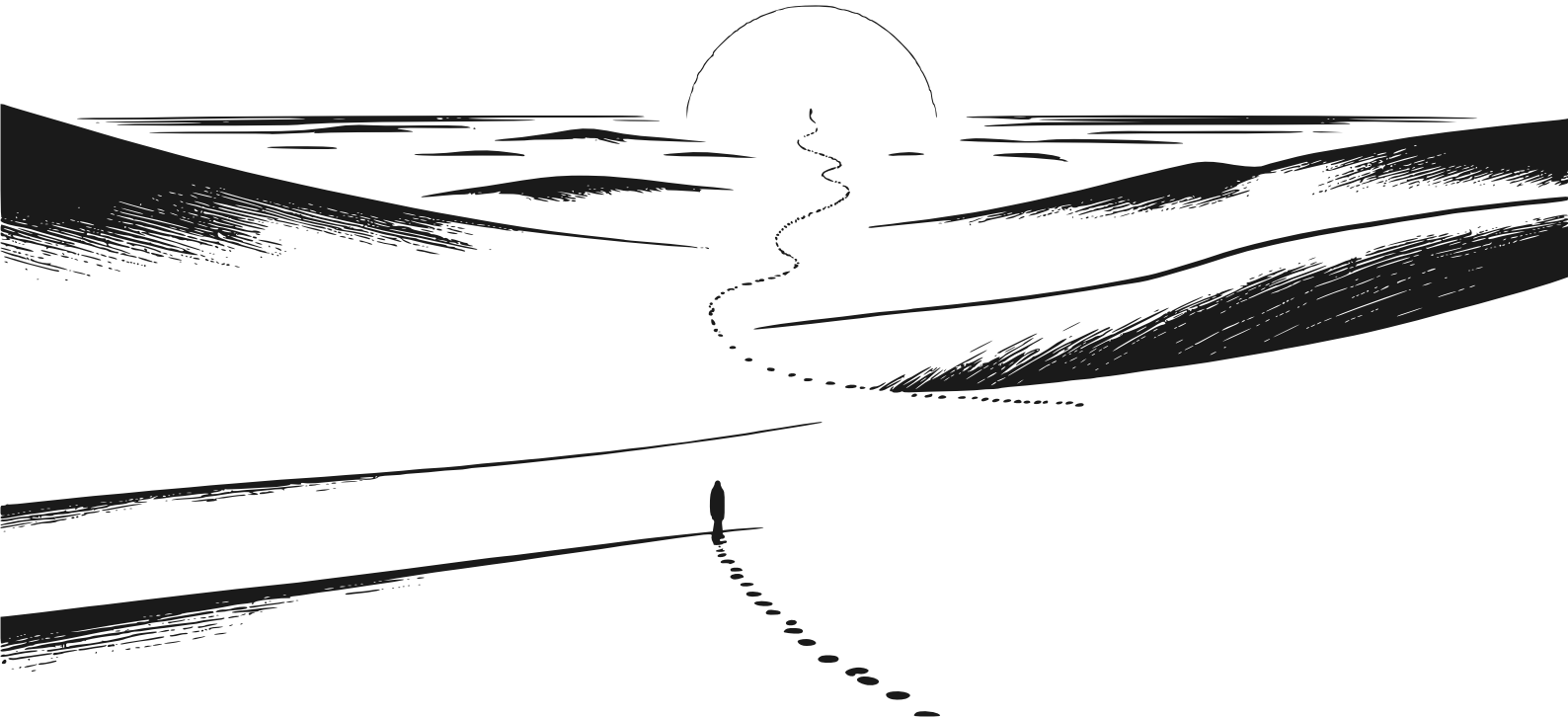

UAM
Universidad Autónoma
de Madrid

Caminante, son tus huellas
el camino, y nada más;
caminante, no hay camino,
se hace camino al andar.
Al andar se hace camino,
y al volver la vista atrás
se ve la senda que nunca
se ha de volver a pisar.
Caminante, no hay camino,
sino estelas en la mar.

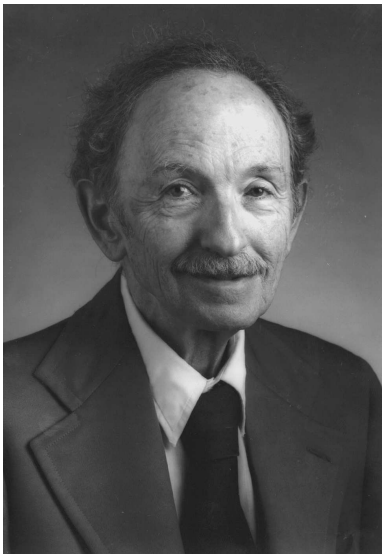
Antonio Machado
Campos de Castilla

With that disappearance, he
knew, even as Noÿs moved
slowly into his arms, came the
end, the final end of Eternity.
— And the beginning of Infinity.

Isaac Asimov
The End of Eternity




UAM
Universidad Autónoma
de Madrid



En reconocimiento a I. J. Good[1], cuya prolífica obra ha alumbrado y seguirá alumbrando a generaciones de estadísticos, matemáticos y científicos.

I think that once the theory of probability is taken for granted, the principle of maximizing the expected utility per unit time is the only fundamental principle of rational behaviour. It teaches us, for example, that the older we become the more important it is to use what we already know rather than to learn more.

I. J. Good[2]

There may be occasions when it is best to behave irrationally, but whether there are should be decided rationally.

I. J. Good[2]

Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the *last* invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.

I. J. Good[3]

Until an ultraintelligent machine is built perhaps the best intellectual feats will be performed by men and machines in very close, sometimes called "symbiotio," relationship, although the term "biomechanical" would be more appropriate.

I. J. Good[3]

[1] Foto en Banks (1996). Reimpresa con el permiso del Institute of Mathematical Statistics.
[2] Cita en Good (1952).
[3] Cita en Good (1966).


UAM
Universidad Autónoma
de Madrid

UAM | FACULTAD DE CIENCIAS

Declaración de originalidad

Esta tesis doctoral es un trabajo **original**, y libre de plagio, del autor.

Esta tesis doctoral sigue los estándares de **calidad** de la UAM en materia de investigación.

Partes de esta tesis doctoral han aparecido previamente en distintas **publicaciones**:

- Chacón, J. E. y Fernández Serrano, J. (2023). Bump hunting through density curvature features. *TEST* **32**, 1251-75. DOI: 10.1007/s11749-023-00872-z. arXiv: 2208.00174.
- Chacón, J. E. y Fernández Serrano, J. (2024). Bayesian taut splines for estimating the number of modes. *Computational Statistics & Data Analysis* **196**, 107961. DOI: 10.1016/j.csda.2024.107961. arXiv: 2307.05825.
- Chacón, J. E. y Fernández Serrano, J. (2025a). Mode-based estimation of the center of symmetry. *Annals of the Institute of Statistical Mathematics* **77**, 685-717. DOI: 10.1007/s10463-025-00942-z. arXiv: 2406.08241.
- Chacón, J. E. y Fernández Serrano, J. (2025b). Rejoinder to the discussion of "Mode-based estimation of the center of symmetry". *Annals of the Institute of Statistical Mathematics* **77**, 727-30. DOI: 10.1007/s10463-025-00945-w. arXiv: 2508.18909.

Madrid y Lugo, 2021–2025

Javier Fernández Serrano

DEPARTAMENTO DE MATEMÁTICAS
UNIVERSIDAD AUTÓNOMA DE MADRID


UAM
Universidad Autónoma
de Madrid

Contenidos

CONTENIDOS


UAM
Universidad Autónoma
de Madrid


UAM
Universidad Autónoma
de Madrid

Prefacio

Mucho antes de saber que quería ser doctor, comienza a fraguarse la historia de esta peculiar disertación. En contra de la corriente actual, que demanda al joven investigador no dejar nada al azar y sin optimizar, un servidor no lo supo hasta después de varios años trabajar, para luego acabar dando un giro radical. Echando la vista atrás, aunque los indicios abundasen, no siempre estuvo claro que ese fuese a ser mi camino. Un camino improbable, al igual que natural… ¡Una empresa sin más atisbo de locura que la eventualidad de haberse perdido tan espléndida aventura!

Como para tantos otros de mi generación, tras finalizar la carrera, encontrar un empleo era mi aspiración. No por una marcada vocación o una necesidad imperiosa, mas porque el trabajo en aquella época, por escaso, era piedra preciosa. En circunstancias tan adversas, aun teniendo una licenciatura y una ingeniería, la búsqueda fue infructuosa, viéndome abocado a seguir con estudios de maestría. La investigación se cruzó por primera vez de lleno en mi camino durante esa dura y tortuosa etapa, al término de la cual, de manera tan ansiada como inesperada, llegó la gran oportunidad laboral.

Merced a mi afán por agradar y honda preparación, mi debut en la empresa como analista fulgurante resultó. Sin embargo, en aquel torbellino de vivencias, tan pronto me sentí en lo más alto como sumido en mil miserias. De haber caído de pie, fui en parte culpable yo mismo, pero, inocente, por pretender ser yo mismo, sufrí un buen traspié. Anhelaba cotas más elevadas y, para alcanzarlas, debía levar anclas. Entretanto, el cielo se me abrió en la empresa con otro reto desafiante, en un departamento de investigación, de atmósfera elitista y estresante. Hubo quien veía en mi decisión un desatino, mas afronté con decisión mi destino.

Porque el riesgo es inherente a la investigación, y, en el seno de una empresa, fracasar no es una opción. No sabía entonces todavía que hasta el mayor naufragio, bien mirado, puede esconder en sí un éxito en ciernes, muy al contrario. Decidí dar un paso atrás en mi glorificada carrera profesional —que no era tal— y, mientras me fogueaba en un puesto de gestión, tuve la ocasión de reencontrarme, a espaldas de la empresa, con la investigación. Esta vez sí, para encauzar mi pasión. De aquella extraoficial indagación nació una prepublicación que habría de llevarme a conocer a quien es hoy mi director.

Con todo, la pandemia supuso el punto de inflexión. Superada la tempestad, volví a salir de casa y a ver la luz del sol. Pero entonces arreció otro temporal —este, artificial—, y volando se llevó mi paraguas; calado hasta los huesos, tuve que proseguir la marcha. Me despedí del que había sido mi hogar unos años para guarecerme en mi casa anterior, no demasiado lejana, de la que me fui casi sin decir adiós. El hijo pródigo regresaba, radiante de ilusión; entre la lucidez y la enajenación… ¡Iluminado por una razón y una convicción que, a todas luces, deslumbrarían al mismísimo Sol!

Sin otra "beca" que la que mi propio ahorro emuló, pero con un director que, pese a la abismal distancia sobre el mapa, con los brazos abiertos me acogió, mi singladura doctoral arrancó. Partiendo de un título para la tesis y del estimador núcleo de la densidad, trazamos rumbo a la meta, sin las escalas concretar; empero, izamos las velas y, con más voluntad que certezas, hicímonos a la mar. Sin prisa pero sin pausa, párrafo a párrafo, publicación a publicación, el trabajo avanzó, a la vez que el tiempo, de tan gozosa la experiencia, rápidamente voló. Así, hállome a las puertas de un excelso honor. Lo que pase después… solo lo sabe Dios.


UAM
Universidad Autónoma
de Madrid

Agradecimientos

Esta tesis no hubiese sido posible sin la colaboración, más o menos directa, de multitud de personas e instituciones: el ascenso hacia el grado de doctor es el camino de toda una vida. Para todas ellas va mi más profundo agradecimiento, que hago extensivo a las familias de aquellos que ya no están con nosotros.

Además de darme la vida y su amor, y procurar siempre mi bienestar, le debo a mi madre Isabel el que ha sido mi mayor pilar, la educación, tanto en el plano moral como cultural. De ella aprendí que nada que merezca la pena se consigue sin esfuerzo y que no debía hacer a los demás lo que no quisiera que me hiciesen a mí. También ella me inculcó el amor por el conocimiento, como un bien en sí mismo, y el valor de obrar de manera coherente. Me enseñó a amar las palabras y a respetar —que no temer— las matemáticas, cuyo andamiaje debía apuntalarse meticulosamente, curso a curso, para progresar.

La vertiente curricular de mi educación, paralelamente, ha recaído en numerosos maestros y profesores, responsables de una amplia variedad de materias. Incluso los más alejados de la temática de esta tesis, todos han contribuido a ella. Muy especialmente me acuerdo de aquellos, de cualquier ciclo educativo, de matemáticas, una rama del saber que acompaña al alumno desde sus primeros pasos hasta donde este se ose aventurar.

En el Colegio Luz Casanova de Embajadores, esenciales durante la Educación Primaria fueron Vicky, con las primeras operaciones, y don José, quien infundió realmente en mí el gusto por las matemáticas. En la Educación Secundaria Obligatoria, en el Colegio Salesianos de Atocha, Jesús nos abrió al lenguaje del álgebra, pero también nos descubrió todo un mundo físico y social, lleno de historia, en el que las matemáticas jugaban un papel preeminente. Finalmente, en Bachillerato, en el Colegio San Ramón y San Antonio (Agustinas Hermanas del Amparo), Rosa hizo mi razonamiento madurar, y José Luis, a través de su ejemplo, me animó a seguir su camino en la universidad. Específicamente, de José Luis aprendí que en matemáticas no solo importa la solución, sino el camino que lleva a ella y saber explicarlo en palabras.

Compartí clase y recreo en el "Sanra" con mis primos, José María y Carmen, y comedor con su madre, también llamada Carmen, que trabajaba en el colegio. Les estoy profundamente agradecido por lo bien que me acogieron y por hacer mi estancia tan feliz en una edad crucial, a la vez que convulsa, de mi vida. Fue importante también, en toda mi educación preuniversitaria, la formación espiritual recibida en estos tres colegios católicos concertados. Durante mucho tiempo mi corazón no estuvo abierto a ella, pero la llevaba muy dentro de mí. Hoy y en el futuro, doy y daré las gracias por tan poderosa arma.

Del claustro de la UAM, son tantos los profesores que han influido en mí que no podría nombrarlos a todos. No obstante, sí me gustaría referirme con mucho cariño, en primer lugar, a Ireneo Peral, fallecido en febrero de 2021. Jamás podré olvidar la empatía con la que me trató en una situación comprometida durante mi examen final de Cálculo III. Probablemente no reparó en la bondad de su gesto, pero para mí fue trascendental. Por otra parte, a pesar de que nunca me dio clase, quiero acordarme de Fernando Chamizo. Su encantadora página web alberga apuntes de lo más variados; auténticas obras confeccionadas con toda su ilusión, que han sido y serán referencia para muchísimos alumnos, entre los que me incluyo.

Casi la mitad de mi existencia ha estado ligada en exclusiva a la UAM, como institución educativa superior, desde que hice la selectividad en el Edificio de Biología de la Facultad de Ciencias, hasta ahora, que opto al grado de doctor. A la UAM debo una excelente formación

de base, al permitirnos compaginar los estudios de Licenciatura en Matemáticas con los de Ingeniería en Informática, en una época, además, en la que las dobles titulaciones estaban todavía muy poco extendidas. Uno de los artífices de aquel audaz plan de estudios simultáneos pionero, de singular belleza, fue José Ramón Dorronsoro. Promociones y promociones tuvimos el honor de recibir su magnífico magisterio en la Escuela Politécnica Superior.

Mientras terminaba los másteres en la UAM, pude disfrutar de una beca en una empresa emergente, donde aprendí mucho sobre el mundo laboral. Agradezco a su fundadora, Sonia Pacheco, que confiase en mí para aquella misión, aun cuando —ahora me doy cuenta— yo solo estaba en pañales. La tecnología aún no existía, pero ella soñaba a lo grande, y el tiempo acabó dando la razón a su visión de negocio. Siempre recordaré su afán por aprender y superarse. No en vano, ella me transmitió que la formación superior es siempre la inversión más rentable.

De mi trayectoria laboral posterior, en la gran empresa, quisiera mencionar como actor clave para mi doctorado a Mattia Chiodaroli. El lapso en que él fue mi responsable me brindó la paz necesaria para redescubrir, fuera del trabajo, mi faceta investigadora. Porque, en efecto, tuvo la opción de quemarme, y no lo hizo, lo que permitió que brotase mi creatividad. Así, gracias a él, pude desarrollar la prepublicación* que acabó sirviendo de carta de presentación ante mi director. Es por ello por lo que Mattia siempre será para mí un referente de jefe-líder moderno, que sabe conciliar la orientación a resultados y el respeto a las personas.

Ya en el capítulo de doctorado, estoy muy agradecido al Departamento de Matemáticas, entre otras cosas, por permitirme iniciar esta aventura. En concreto, debo referirme a los catedráticos de estadística José Ramón Berrendero y Antonio Cuevas, y a su inestimable labor al frente del programa. Asimismo, de vital importancia ha sido mi tutora, Amparo Baíllo, quien, además de haberme asesorado académicamente y ayudado con las gestiones, lidera nuestro grupo de investigación. Por último, quiero agradecer su simpatía a dos jóvenes colegas: Luis Alberto Rodríguez, hoy "postdoc" en Alemania, que me ofreció su consejo y cálida bienvenida, y Martín Sánchez Signorini, doctorando de primer año, al que deseo todo lo mejor.

Fuera de la UAM, agradezco la deferencia de Eduardo García Portugués y Andrea Meilán, de la Universidad Carlos III de Madrid, con quienes coincidí en múltiples ocasiones. Igualmente, agradezco su influencia en esta tesis a José Ameijeiras, de la Universidad de Santiago de Compostela, a quien también tuve el placer de conocer en persona. Asimismo, quiero reconocer su contribución a mi formación a los profesores y doctorandos encargados de la organización e impartición de seminarios, ya fueran en la UAM, la Universidad Complutense de Madrid o el Instituto de Ciencias Matemáticas (ICMAT). Finalmente, agradezco las valiosas aportaciones de los editores y revisores involucrados en la publicación de los capítulos de esta tesis.

Dejo como colofón al mayor impulsor de mi formación doctoral: mi director, José Enrique Chacón. Doy las gracias a la Universidad de Extremadura por habérmelo prestado estos años. Conocí a José Enrique en mi búsqueda de un autor que avalase mi prepublicación* en arXiv. Al final, de tan amable que se mostró, ese acercamiento devino en excusa para pedirle algo más: que se convirtiera en mi director. En aquel momento, torpe de mí, no fui plenamente consciente de lo eminente que era (hoy ya catedrático). Igualmente, en lo que se refiere a su calidad humana, que ya intuía, mis expectativas también se vieron ampliamente rebasadas. En ambos aspectos, la suerte —o la providencia— me sonrió más allá de lo que podía esperar. Ya no dudo de que la prepublicación, se publique o no, ha cumplido sobradamente su función.

En esta tesis y momento de mi vida, José Enrique ha sido el director perfecto. No solo por su extenso conocimiento y experiencia en la materia, fundamentales en toda la tesis, sino por su sabiduría, en general. Ha sabido en cada instante qué era importante y qué no, propiciando el desarrollo natural de la tesis. En particular, le agradezco que haya dado rienda suelta a mis cualidades y tenido en cuenta mis gustos personales, sin exponer innecesariamente mis debilidades. Mi experiencia a su lado resulta inmejorable, hasta tal punto que me he sentido más como su compañero de investigación que como un estudiante raso. Espero y deseo haberle podido corresponder del mismo modo.

*El *preprint* se encuentra disponible en arXiv (Fernández Serrano, 2021).


UAM
Universidad Autónoma
de Madrid


UAM
Universidad Autónoma
de Madrid

Resumen

El análisis de datos desempeña un papel crucial y transversal, potenciando cualquier rama del saber o actividad humana. A través de tres trabajos independientes, se presentan sendas propuestas metodológicas relativas a problemas clásicos en estadística de gran relevancia para el análisis de datos. El fundamento común a todos ellos es el estudio de la estructura poblacional subyacente a los datos, sustanciada en ciertas características de naturaleza diferencial y topológica de la función de densidad de probabilidad.

El primer trabajo aborda diversas propiedades de curvatura de la densidad, inéditas hasta ahora en estadística, con el fin de detectar subconjuntos significativos del espacio muestral. La investigación demuestra el buen comportamiento asintótico, tanto en consistencia como en inferencia, de los estimadores tipo núcleo de tales regiones de curvatura. Estas, asimismo, resultan de gran utilidad en aplicaciones del ámbito deportivo como herramienta de visualización para análisis exploratorio de datos multivariantes.

El segundo trabajo afronta el reto de estimar el número de modas de una densidad univariante, propiedad relacionada con la cantidad de subpoblaciones y, por tanto, con la complejidad de los datos. Marcando como meta la efectividad, se propone un nuevo método de estimación bayesiano que aúna múltiples perspectivas sobre la multimodalidad. El estudio de simulación llevado a cabo sitúa al nuevo método entre los más efectivos, a la vez que revela el mejorable rendimiento de algunas alternativas comúnmente empleadas.

El último trabajo revisa la estimación del centro de simetría de una densidad univariante, simétrica y unimodal, esto es, el valor poblacional más prominente. Los resultados asintóticos obtenidos, sobre el impacto del ancho de banda y la forma del núcleo en la eficiencia de la estimación, contribuyen a reducir la brecha entre las comunidades de estadística no paramétrica y estadística robusta. La propuesta subsiguiente de estimador adaptativo se contrasta con éxito en un estudio de simulación, destacando en escenarios con colas pesadas.

En conjunto, los tres trabajos ponen de relieve el potencial de estas características para extraer conocimiento de los datos, como evidencian los múltiples y variados casos de uso aportados. Los métodos propuestos suponen avances e innovaciones importantes respecto al estado del arte, tanto en el plano teórico como computacional, con especial énfasis en su uso práctico. En particular, se ahonda en la búsqueda de métodos flexibles en un contexto no paramétrico, ampliando el abanico de aplicaciones del estimador núcleo de la densidad.

Palabras y frases clave. densidad de probabilidad, estimación no paramétrica, estimador núcleo, curvatura, número de modas, centro de simetría.


UAM
Universidad Autónoma
de Madrid

Abstract

Data analysis plays a crucial and transversal role, enhancing any field of knowledge or human activity. Through three independent studies, methodological proposals are presented relating to classic problems in statistics of great relevance to data analysis. The common foundation of all three works is the study of the underlying population structure within the data, embodied in certain differential and topological features of the probability density function.

The first study addresses various curvature properties of the density, previously unpublished in statistics, to detect significant subsets within the sample space. The research demonstrates the sound asymptotic behavior, both in terms of consistency and inference, of kernel-type estimators of such curvature regions. Moreover, these are of great use in sports applications as a visualization tool for exploratory analysis of multivariate data.

The second study tackles the challenge of estimating the number of modes in a univariate density, a property connected to the number of subpopulations and, thus, to data complexity. With effectiveness as a goal, a new Bayesian estimation method is proposed, integrating multiple perspectives on multimodality. The simulation study conducted placed the new method among the most effective while revealing the improvable performance of some commonly used alternatives.

The final study revisits the estimation of the center of symmetry of a univariate, symmetric, and unimodal density, that is, the most prominent population value. The asymptotic results obtained, on the impact of the bandwidth and the shape of the kernel on estimation efficiency, contribute to bridging the gap between the nonparametric statistics and robust statistics communities. The subsequent proposal of an adaptive estimator is successfully tested in a simulation study, standing out in heavy-tailed scenarios.

Together, the three studies underscore the potential of these features to extract knowledge from data, as evidenced by the multiple and varied use cases provided. The proposed methods represent significant advances and innovations compared to the state of the art, both theoretically and computationally, with special emphasis on their practical use. In particular, the search for flexible methods in a nonparametric context is explored in depth, broadening the range of applications of kernel density estimators.

Key words and phrases. probability density, nonparametric estimation, kernel estimator, curvature, number of modes, center of symmetry.


UAM
Universidad Autónoma
de Madrid

Introducción

El análisis de datos se ha convertido en pieza fundamental para el desarrollo del conocimiento humano. Desde las disciplinas científicas más antiguas, centradas en la naturaleza, hasta las más recientes, que estudian diversos aspectos de la sociedad, pasando por las numerosas ramas de la técnica, los datos están presentes en la práctica totalidad de actividades humanas. Aunque no siempre ha contado con la formulación rigurosa conocida hoy, ni con las modernas herramientas computacionales, este análisis se encuentra íntimamente ligado a nuestra necesidad ancestral de lidiar con la incertidumbre para tomar las mejores decisiones posibles, cualesquiera que sean sus fines. Con todo, a pesar de tan poderoso e inmemorial origen, vivimos en la actualidad una auténtica revolución de la información, evidenciada por su explotación intensiva como motor económico y su popularización entre la población general, cada vez más consciente de su valor y dispuesta a guiarse mediante aplicaciones que se nutren masivamente de ella. En este sentido, la irrupción de una inteligencia artificial (IA) que rivaliza con la humana y la potencia de formas todavía desconocidas solo permite augurar una aceleración de esta transformación.

Paralelamente, la disciplina científica que se ocupa del análisis de datos, la estadística, ha disfrutado de un gran crecimiento en los últimos tiempos, fruto de la colaboración, no siempre estrecha, entre los mundos académico y empresarial. La industria, inmersa en un continuo proceso de digitalización, ofrece gran variedad de casos de uso a los investigadores, a la vez que demanda de estos nuevos y más potentes métodos, capaces de tratar grandes volúmenes de información heterogénea (*big data*) y de atender, además, a los particulares requisitos de los usuarios. Así, en lo referente a la complejidad de los datos, la estadística univariante clásica va dando paso progresivamente a la estadística multivariante y a la estadística infinitodimensional. Por otra parte, las hipótesis paramétricas tradicionales tienden a ser reemplazadas por otras no paramétricas, más flexibles y aptas para sacar el máximo partido a los mayores tamaños muestrales disponibles en la actualidad. Finalmente, la cada vez más apreciada interacción entre el usuario y el método (con el ordenador), a la hora de interpretar resultados e incorporar juicio experto, impulsa el florecimiento de las herramientas de visualización y del paradigma bayesiano.

La presente tesis doctoral, constituida como compendio de tres trabajos independientes —de los que se derivan un total de cuatro publicaciones—, se encuadra en varias de las tendencias mencionadas. Su eje central, nexo entre los distintos trabajos, es el estudio de propiedades geométrico-diferenciales y topológicas de las funciones de densidad de probabilidad. Dichas características son clave para analizar la estructura poblacional subyacente a conjuntos de datos univariantes o multivariantes. Para alcanzar el grado de aplicabilidad deseable en los métodos propuestos, se toma como base en todos los casos la técnica no paramétrica del estimador núcleo de la densidad (Chacón y Duong, 2018), siguiendo la estela de tesis doctorales recientes con objetivos afines como las de Ameijeiras-Alonso (2017) y Casa (2019). La presente investigación comparte también con ambas disertaciones el foco en el concepto de moda (Chacón, 2020), contemplado desde múltiples perspectivas: como regiones del espacio muestral, como subpoblaciones que contar o como medida de centralidad. La influencia de Good y Gaskins (1980) resulta notable en las contribuciones sobre los dos primeros puntos de vista. Por último, la obra posee un marcado carácter aplicado y computacional, coherente con la realidad actual de la estadística.

INTRODUCCIÓN

De acuerdo con la normativa de la UAM sobre tesis por compendio de publicaciones, las siguientes secciones presentan los tres trabajos realizados. A continuación, se incluyen en orden cronológico las versiones *postprint* de sus respectivos artículos, a razón de un capítulo por trabajo. Con ello se homogeneiza el formato, a la vez que se da cabida al material suplementario que, aunque revisado, quedó excluido de las versiones editoriales. Cada capítulo trata un tipo de propiedad de las funciones de densidad y cuenta con sus propias secciones de introducción, descripción del método, estudio de simulación o resultados teóricos, caso de uso, discusión, referencias y apéndices. De forma adicional, se incluye en el apéndice un artículo *rejoinder*, en versión *postprint*, derivado del tercer trabajo. Todos los artículos están firmados solo por el autor y su director, circunscribiéndose todas las aportaciones a este periodo doctoral. Antes del apéndice, la disertación concluye con un breve capítulo de síntesis y reflexión global sobre la investigación.

Bump hunting through density curvature features[1]

Chacón y Fernández Serrano (2023), en el transcurso de su primera tutela de doctorado (curso 2021-2022), se acercan al tema del *bump hunting*, una vertiente del análisis exploratorio de datos que no ha recibido suficiente atención, con contribuciones esporádicas y sin un marco conceptual común. Tomando como referencia fundamental a Good y Gaskins (1980), uno de los primeros trabajos en la materia, los *bumps* (del inglés, "bultos" o "protuberancias") se identifican con subconjuntos significativos del espacio muestral. A este respecto, la caracterización y posterior interpretación de tales conjuntos resulta difusa, y depende, en gran medida, del analista (el "cazador") y el caso de uso concreto. Suele entenderse el *bump hunting*, por tanto, como una tarea de estadística no supervisada sobre datos univariantes o multivariantes, típicamente en dimensiones pequeñas, que posibiliten la visualización. Sin embargo, enfoques radicalmente opuestos como el de Friedman y Fisher (1999) (supervisado, dimensiones altas, datos heterogéneos) evidencian la falta de consenso.

Good y Gaskins (1980) definieron los *bumps* en el caso univariante como regiones donde la densidad es cóncava, al igual que esbozaron una extensión al caso multivariante, sin dar una definición precisa. Así pues, el planteamiento inicial de Chacón y Fernández Serrano (2023) fue completar esta formalización pendiente. Pronto, no obstante, resultó claro que la concavidad es solo una de varias formas de entender la curvatura. Apoyándose en la geometría diferencial clásica, los autores establecen la conexión de la curvatura de la densidad con las derivadas parciales de orden dos de esta y, por tanto, con su matriz hessiana. Múltiples funcionales de la densidad se deducen de esta matriz, como el mayor de sus autovalores o la traza. Tras formular en abstracto el concepto conjuntista de *bump* —uno de los logros del trabajo—, el primero de estos funcionales proporciona la definición natural de *bumps* cóncavos buscada, mientras que el segundo da lugar a un nuevo e inesperado tipo de *bump*, bautizado como laplaciano. Asimilando la densidad a un paisaje montañoso, si los *bumps* cóncavos típicamente se forman en los "picos", los laplacianos abarcan "sierras" enteras.

Una vez definidos los distintos tipos de *bumps* en el ámbito poblacional, la labor se centró en verificar las buenas propiedades asintóticas de consistencia e inferencia del estimador *plug-in* basado en el estimador núcleo de las derivadas de la densidad. Esta parte teórica de la contribución se apoya en algunos de los últimos avances sobre estimación de conjuntos de nivel de la densidad, como Chen (2022) y Chen y col. (2017), con resultados sobre la estabilidad de variedades, los procesos gaussianos o el remuestreo (*bootstrap*, en inglés). En inferencia, el concepto de densidad suavizada, que volvería a aparecer en el desarrollo de la tesis en Chacón y Fernández Serrano (2025a), resultó clave para soslayar dificultades técnicas con el sesgo del estimador de la densidad.

[1] Su versión *postprint* conforma el Capítulo 1.

(A) Stephen Curry en 2016 (B) Shohei Ohtani en 2022

FIGURA 1. Jugadores protagonistas de los casos de uso con datos deportivos en Chacón y Fernández Serrano (2023, 2024). Ambas fotos están bajo una licencia Creative Commons "Atribución-CompartirIgual 2.0 Genérica", siendo simples recortes de originales de Keith Allison[2] y Mogami Kariya[3].

Al buen comportamiento teórico para muestras asintóticamente grandes de los nuevos *bumps* de curvatura, debe añadirse su uso destacado como herramienta de detección y visualización. Intuitivamente, en el grafo de la densidad pueden existir regiones que aglutinen poca masa de probabilidad y que tengan poca altura, pero que destaquen por su curvatura. Resulta esclarecedora la comparación de la nueva propuesta con las regiones de alta densidad introducidas por Hyndman (1996), técnica no siempre sencilla de ajustar y que tiende a pasar por alto detalles sutiles pero importantes. Una aplicación ilustrativa en dos dimensiones, empleada a lo largo del trabajo, es el análisis de las posiciones de lanzamientos a canasta en baloncesto, deporte también tratado como ejemplo en Chacón (2020). Merced a la nueva propuesta, es posible caracterizar el ADN de un lanzador de manera más sucinta y directa que con la alternativa clásica. Por ejemplo, los datos del célebre jugador Stephen Curry (Figura 1a) producen unos *bumps* de curvatura que resaltan las inmediaciones de la línea de tres puntos: toda una demostración gráfica del potencial del método.

Además de ampliar el repertorio de técnicas de *bump hunting*, siguiendo la línea marcada por Duong y col. (2008) y Godtliebsen y col. (2002), el estudio de tales propiedades geométrico-diferenciales supone una continuación de la tesis de Casa (2019), la cual se centraba en las derivadas parciales de orden uno (esto es, en el gradiente de la densidad). Por todo ello, tendría sentido incluir parte del contenido teórico de Chacón y Fernández Serrano (2023) en Chacón y Duong (2018, Sección 6.1.1), de cara a posteriores reediciones de la obra. Por otro lado, si bien el trabajo no se circunscribe exclusivamente al ámbito deportivo, ni viene motivado por este, los ejemplos aportados sobre baloncesto, béisbol y fútbol americano sugieren un uso práctico efectivo en el análisis de datos deportivos (*sports analytics*), un área que suscita mucho interés actualmente (Albert y col., 2005; Severini, 2020). Tales aplicaciones demuestran, asimismo, que los datos en dimensiones pequeñas todavía aportan valor, a pesar de la creciente tendencia a trabajar en dimensiones altas.

[2] `https://commons.wikimedia.org/wiki/File:Stephen_Curry_dribbling_2016.jpg`
[3] `https://commons.wikimedia.org/wiki/File:Shohei_Ohtani_(52251723213).jpg`

FIGURA 2. Cartel de la X Jornada Estadística UAM, organizada por el grupo de investigación en estadística infinito-dimensional del Departamento de Matemáticas de la UAM. Imagen original cortesía del profesor José Pedro Moreno, autor del diseño y de la foto de fondo.

El artículo se presentó oralmente durante la X Jornada Estadística UAM (véase Figura 2) en abril de 2023, coincidiendo con Berrendero y col. (2025). En junio del mismo año, se aceptó en la revista *TEST*. Posteriormente, fue citado por Zhang y Chen (2025).

(A) Sello de veinticinco centavos

(B) Sello de cien centavos

FIGURA 3. Ejemplos de la colección de sellos de Hidalgo de 1872, protagonista del problema estadístico clásico abordado en Chacón y Fernández Serrano (2024). Ambas fotos están en el dominio público.

Bayesian taut splines for estimating the number of modes[4]

El segundo trabajo de la tesis, Chacón y Fernández Serrano (2024), desarrollado a lo largo del curso 2022-2023, busca responder a la pregunta clásica de cuántas modas tiene una función de densidad desconocida. En el contexto del análisis de datos, este número entero representa la cantidad de subpoblaciones, entendidas cada una de estas como un fenómeno de interés diferenciado. Tal funcional de la densidad mide, por tanto, en cierto sentido, la complejidad de los datos. Se trata de un problema delicado, marcado por un resultado de Donoho (1988), ciertamente pesimista, que afirma que no es posible acotar superiormente el número de modas con cierto nivel de confianza a partir de muestras finitas si se asumen hipótesis completamente no paramétricas. Este escollo teórico se ve reforzado por la constatación en la práctica de una enorme disparidad de respuestas a problemas concretos largamente establecidos en la literatura, como el de los sellos de Hidalgo (véase Figura 3). Compárese, por ejemplo, la elección defendida en Marron y Dryden (2021, Figura 15.4) basada en el estimador núcleo de la densidad, que tiene siete modas, con la solución de Ameijeiras-Alonso y col. (2019, p. 917) basada en un test de multimodalidad, que estima cuatro.

El objetivo inicial fue el de arrojar luz sobre la efectividad de las propuestas de la literatura en la tarea de estimar el número de modas para el caso univariante. Según el leal saber y entender de los autores, a pesar de la relevancia del problema, tal estudio no se había realizado antes. Si bien se han estudiado ampliamente los métodos de selección de ancho de banda para el estimador núcleo de la densidad (Wand y Jones, 1995), siempre se ha hecho atendiendo a criterios globales, sin comprobar su eficacia respecto al número de modas. Por otra parte, aunque profusa la investigación sobre tests de multimodalidad para testar si el número de modas es menor o igual que una cierta cantidad, poca atención se ha prestado a cómo adaptarlos a la pregunta de cuántas modas hay. En tal caso, resulta natural emplearlos iterativamente, como hacen Ameijeiras-Alonso y col. (2019), tratando de descartar en primer lugar números de modas más pequeños. Sin embargo, como se aprecia en Ameijeiras-Alonso y col. (2019, Tabla 6), la evolución de los p-valores puede resultar errática y conducir a resultados poco intuitivos. Por último, métodos específicos como los de Davies y Kovac (2004) y Genovese y col. (2016) tampoco habían sido probados en un estudio de simulación exhaustivo.

Después de revisar las distintas alternativas disponibles, e inferir a partir de estas los puntos de vista más relevantes sobre la multimodalidad (tests, métodos gráficos, construcción de densidades), se consideró que el uso de una sola técnica *pura*, por eficaz que fuese,

[4]Su versión *postprint* conforma el Capítulo 2.

difícilmente permitiría alcanzar un entendimiento completo del problema. Por esta razón, los autores idearon una propuesta propia que, procurando la mayor eficacia posible, conjugase los distintos aspectos del problema de manera coherente y sencilla para los usuarios. La nueva herramienta debía ser flexible, permitiendo la exploración gráfica de manera similar a los árboles de modas (Minnotte y Scott, 1993) o SiZer (Marron y Dryden, 2021, Capítulo 15), pero sin permitir la explosión de complejidad típica de los métodos no paramétricos advertida por Donoho (1988). Asimismo, a la vez que ofreciese una respuesta directa sobre el número de modas, debía permitir cuantificar la incertidumbre de manera más consistente y aplicable que los p-valores. Finalmente, debía combinar el ajuste global de un modelo con la significatividad de las modas individuales. En este sentido, el paradigma bayesiano favorecería la interactividad con el juicio experto del analista y la interpretabilidad de los resultados. Fruto de esta reflexión y de un intenso proceso de experimentación, surge el método *Bayesian taut spline* (BTS).

En primer lugar, se validó BTS sobre conjuntos de datos reales. Además del ya mencionado de Hidalgo, para el que estimó concluyentemente las mismas siete modas que Marron y Dryden (2021), se planteó un nuevo caso de uso deportivo, como en Chacón y Fernández Serrano (2023). Tras obtener un resultado de cuatro modas similar a Chacón (2020, Figura 2) con datos baloncestísticos, se optó por analizar las velocidades de lanzamiento de un famoso jugador de béisbol, Shohei Ohtani (véase Figura 1b), en busca de sus distintos tipos característicos de picheo. Los resultados fueron satisfactorios, pues no solo la estimación del número de modas coincidió con la interpretación oficial del proveedor de los datos, sino que el análisis subyacente reflejó la incertidumbre del problema. Concretamente, en torno a un tipo de picheo conformado en realidad por dos subtipos englobados en la misma moda, difíciles de diferenciar atendiendo solo a la velocidad.

A continuación, el nuevo BTS y los métodos documentados en la literatura se probaron conjuntamente en un estudio de simulación exhaustivo. Como resultado, se obtuvo una clasificación de los distintos métodos, en la que se permitían posibles empates para descartar diferencias no significativas. El orden final resultó particularmente llamativo. BTS ocupaba la primera posición, empatado con varios métodos genéricos basados en el estimador núcleo de la densidad, quedando relegados a los últimos puestos todos los que habían sido diseñados específicamente para abordar problemas de multimodalidad. Tal hallazgo confirmaba la intuición inicial que había motivado el experimento, pero a la vez ponía en cuestión las técnicas de la competencia, especialmente en relación con su aplicabilidad práctica. En cuanto a la parte alta de la clasificación, aunque algunos estimadores núcleo ofrecían los mismos resultados que BTS con un menor coste computacional y, hasta cierto punto, mayores garantías teóricas, BTS tenía la ventaja de ofrecer mucha más información. Esto hizo pronosticar una buena aceptación de BTS por parte de analistas, siempre y cuando la implementación del método viniese acompañada de una interfaz gráfica acorde a sus posibilidades.

Además de proporcionar una nueva herramienta y de verificar experimentalmente el rendimiento de esta y de otras técnicas clásicas, las ideas presentadas en Chacón y Fernández Serrano (2024) abarcan ramas diversas de la estadística. Por una parte, BTS utiliza conceptos novedosos de estadística con datos funcionales. Por ejemplo, los espacios de Bayes y los *splines* composicionales (Machalová y col., 2020) dotan de estructura a las posibles soluciones durante la fase de exploración de BTS. Asimismo, se recurre al análisis de componentes principales funcional (Hron y col., 2016) de forma decisiva para simplificar dicho espacio de soluciones. Por otra parte, BTS supone un nuevo uso del enfoque bayesiano (Wasserman, 2000) en un contexto inédito, en consonancia con otros avances recientes (Berrendero y col., 2025). No obstante, ante todo, BTS recibe una influencia sustancial del método de verosimilitud penalizada de Good y Gaskins (1980), que, sin ser puramente bayesiano, ya recogía parte de ese espíritu. Concretamente, BTS toma de él su carácter aplicado, computacional, híbrido e interactivo. Por todo ello, podría entenderse que la contribución de BTS, aunque original, supone una evolución en la línea de Good y Gaskins (1980).

XL CONGRESO NACIONAL DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

XIV JORNADAS DE ESTADÍSTICA PÚBLICA

Dña. Mercedes Landete Ruiz y D. José Luis Ruiz Gómez, como presidentes del Comité Organizador del XL Congreso Nacional de Estadística e Investigación Operativa y de las XIV Jornadas de Estadística Pública (SEIO 2023) de la Sociedad de Estadística e Investigación Operativa, organizados por el Centro de Investigación Operativa de la Universidad Miguel Hernández de Elche, que se han celebrado del 7 al 10 de Noviembre de 2023 en Elche

CERTIFICAN QUE D./Dª **Javier Fernández Serrano** ha presentado en este congreso el trabajo:

Bayesian taut splines for estimating the number of modes

(Comunicación Oral)

Autores: **J. Fernández Serrano, J. E. Chacón**

Y para que conste y surta los efectos oportunos, se expide la presente certificación.

En Elche, a 13 de noviembre de 2023

Mercedes Landete Ruiz

José Luis Ruiz Gómez

PRESIDENTES COMITÉ ORGANIZADOR SEIO 2023

SEIO 2023

UNIVERSITAS Miguel Hernández

Seio Sociedad de Estadística e Investigación Operativa

Elche, 7-10 de Noviembre

FIGURA 4. Certificado de presentación de Chacón y Fernández Serrano (2024) durante el XL Congreso Nacional de Estadística e Investigación Operativa, organizado por la Sociedad de Estadística, Investigación Operativa y Ciencia de Datos (SEIO) y la Universidad Miguel Hernández de Elche.

El artículo se presentó oralmente en el XL Congreso Nacional de Estadística e Investigación Operativa (véase Figura 4) en noviembre de 2023, coincidiendo con Bolón (2024) y Chacón (2020). Se aceptó en *Computational Statistics & Data Analysis* (CSDA) en abril de 2024.

Mode-based estimation of the center of symmetry[5]

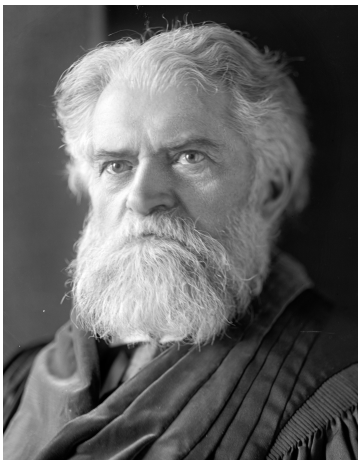
El tercer y último trabajo de la tesis, Chacón y Fernández Serrano (2025a), desarrollado durante el curso 2023-2024, subsana un déficit en la literatura en relación con la moda, como medida de centralidad alternativa a las clásicas media y mediana, en el caso de datos continuos, univariantes y simétricos. Las importantes comunidades de estadística robusta (Huber y Ronchetti, 2009; Maronna y col., 2019) y estadística no paramétrica (Devroye y Györfi, 1985; Tsybakov, 2009) han tratado este tema, en términos generales, solo de manera independiente y tangencial. Curiosamente, en sus respectivos inicios, se puede encontrar a Chernoff (1964) aludiendo tímidamente al caso simétrico en un contexto de estimación de la moda, y a Huber (1964), empleando implícitamente un tipo de estimador modal como caso particular de M-estimador. Sin embargo, puesto que cada comunidad persigue fines diferentes, la colaboración entre ambas ha sido virtualmente inexistente, desaprovechándose los avances que se realizaban desde la comunidad vecina.

La hipótesis de simetría, común en estadística robusta, tiene implicaciones importantes que han sido desatendidas formalmente desde la comunidad de estadística no paramétrica. Concretamente, asumiendo también unimodalidad, el estimador núcleo de la moda es un M-estimador del centro de simetría para cualquier ancho de banda, puesto que la moda de la densidad suavizada (concepto ya utilizado en Chacón y Fernández Serrano, 2023) coincide con la de la densidad original. En otras palabras: el sesgo desaparece. Ello provoca una situación paradójica en el contexto de estimadores núcleo, en la que anchos de banda asintóticamente pequeños producen un rápido deterioro del rendimiento, mientras que otros asintóticamente grandes pueden resultar relativamente eficientes. Además de formalizar estas afirmaciones, en Chacón y Fernández Serrano (2025a) se demuestra la existencia de un ancho de banda óptimo que permite al estimador núcleo de la moda superar en eficiencia a la media muestral cuando la densidad tiene colas pesadas, noción formalizada mediante el concepto de variación regular (Seneta, 1976).

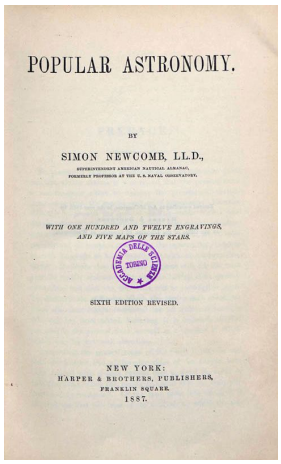
Los resultados teóricos aportados, bajo hipótesis no paramétricas, justifican una nueva propuesta de estimador núcleo de la moda que recomienda optimizar tanto el ancho de banda (la escala) como la forma del núcleo. La idea se concreta en una nueva familia paramétrica de núcleos cuyo parámetro óptimo guarda conexión —teórica y empírica— con el índice de variación regular de la densidad. Puesto que en estadística robusta se trabaja con la hipótesis de un modelo paramétrico contaminado, típicamente gaussiano, el ajuste del ancho de banda se reduce a una sencilla calibración respecto a la dispersión empírica de los datos. Además, desde la comunidad de estadística robusta tradicionalmente se ha argumentado que la forma del núcleo tiene un papel todavía más secundario, en comparación con la escala. A pesar de todo, en Chacón y Fernández Serrano (2025a) se evidencia a través de una serie de densidades clásicas en estadística robusta (como la t de Student) que forma y escala —y no solo esta última— son determinantes para obtener un estimador efectivo en la práctica.

La propuesta se probó empíricamente tanto en un caso de estudio con datos reales como en un estudio de simulación. A diferencia de Chacón y Fernández Serrano (2023, 2024), en esta ocasión los autores eligieron un ejemplo clásico perteneciente al ámbito de la física, donde la simetría constituye una hipótesis natural. Concretamente, se analizan los datos recogidos por Simon Newcomb (véase la Figura 5 y Stigler, 1977) en sus experimentos para determinar la velocidad de la luz, donde la moda permite aproximar el valor real tras los errores aleatorios en la medida. Los resultados sobre este conjunto de datos fueron reveladores, pues el nuevo método descartó automáticamente dos claros valores atípicos y seleccionó una forma de núcleo óptima para datos de naturaleza gaussiana, como era el caso. Por otra parte, el estudio de simulación demostró la gran eficiencia y versatilidad del método en comparación con las alternativas existentes. En particular, resultó sorprendente que el método optase por formas de núcleo inusitadas para ganar eficiencia en escenarios con colas pesadas.

[5]Su versión *postprint* conforma el Capítulo 3.



(A) Simon Newcomb



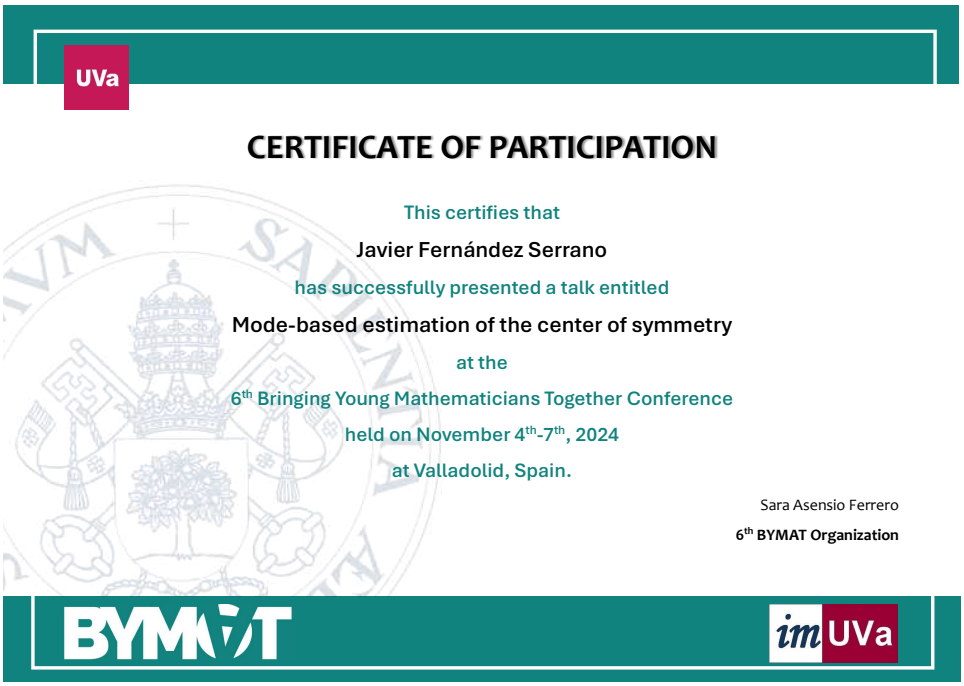
(B) *Popular astronomy*

FIGURA 5. A la izquierda, Simon Newcomb, astrónomo americano de origen canadiense, protagonista del caso de estudio con datos reales en Chacón y Fernández Serrano (2025a). A la derecha, la portada de una de las obras más influyentes de este autor. Ambas fotos están en el dominio público.

Con todo, uno de los méritos del trabajo consiste en reducir la distancia que existe actualmente entre las comunidades de estadística no paramétrica y estadística robusta. Lo hace a través de un tema, la moda, que —valga la redundancia— está muy de moda en investigación estadística (Chacón, 2020). La comunidad de no paramétrica avanza así en una tarea pendiente respecto a la estimación de la moda en el caso simétrico, mientras que la de robusta se enriquece con una visión complementaria. En particular, la estadística robusta incorpora con el estimador núcleo de la moda un nuevo ejemplo notable de M-estimador.

El trabajo se expuso oralmente en noviembre de 2024 en la VI Conferencia Internacional *Bringing Young Mathematicians Together* (BYMAT) (véase Figura 6). Al mes siguiente, en diciembre de 2024, el artículo se aceptó en *Annals of the Institute of Statistical Mathematics* (AISM), con el compromiso de editar una serie de artículos en torno a él, incluyendo dos discusiones (Hino, 2025; Pardo-Fernández, 2025), de sendos autores invitados, y una réplica opcional (Chacón y Fernández Serrano, 2025b). Más adelante, en abril de 2025, la revista ya había aceptado los artículos de discusión de Hino y Pardo-Fernández, y los puso a disposición de Chacón y Fernández Serrano para que elaborasen su contestación. Como colofón, el artículo de réplica fue aceptado por AISM en mayo de 2025.

La acogida de Chacón y Fernández Serrano (2025a) en las discusiones de Hino (2025) y Pardo-Fernández (2025) fue verdaderamente positiva, con sinceras palabras de reconocimiento y la aportación de ideas y comentarios útiles e interesantes. Hino (2025) hizo un gran resumen de las contribuciones esenciales del artículo, así como sobre los principales avances en torno a la moda como elemento de robustez. Igualmente, señaló las bondades de algunos estimadores alternativos de la moda, dignos de futuras investigaciones. Por su parte, Pardo-Fernández (2025) se animó a sugerir tres líneas de investigación muy pertinentes con las que continuar el trabajo: la adaptación del método a datos censurados y truncados, su extensión al caso multivariante y su uso dentro de los tests de simetría. Finalmente, la réplica de Chacón y Fernández Serrano (2025b), que puede consultarse en el apéndice, ahondó en todos estos temas, esbozando retos y oportunidades. En conjunto, esta serie de AISM demuestra el interés de la comunidad investigadora por la moda y contribuye a un diálogo constructivo desde ópticas diversas, uno de los objetivos del trabajo.


UVa

CERTIFICATE OF PARTICIPATION

This certifies that

Javier Fernández Serrano

has successfully presented a talk entitled

Mode-based estimation of the center of symmetry

at the

6th Bringing Young Mathematicians Together Conference

held on November 4th-7th, 2024

at Valladolid, Spain.

Sara Asensio Ferrero

6th BYMAT Organization

BYMAT

imUVa

FIGURA 6. Certificado de presentación de Chacón y Fernández Serrano (2025a) durante la VI Conferencia Internacional *Bringing Young Mathematicians Together* (BYMAT), organizada por la Universidad de Valladolid.

El proceso de publicación de Chacón y Fernández Serrano (2025a,b) se enmarcó en la cuarta[6] y última tutela de doctorado, durante el curso 2024-2025, tiempo que sirvió también para confeccionar la versión final de esta memoria de tesis.

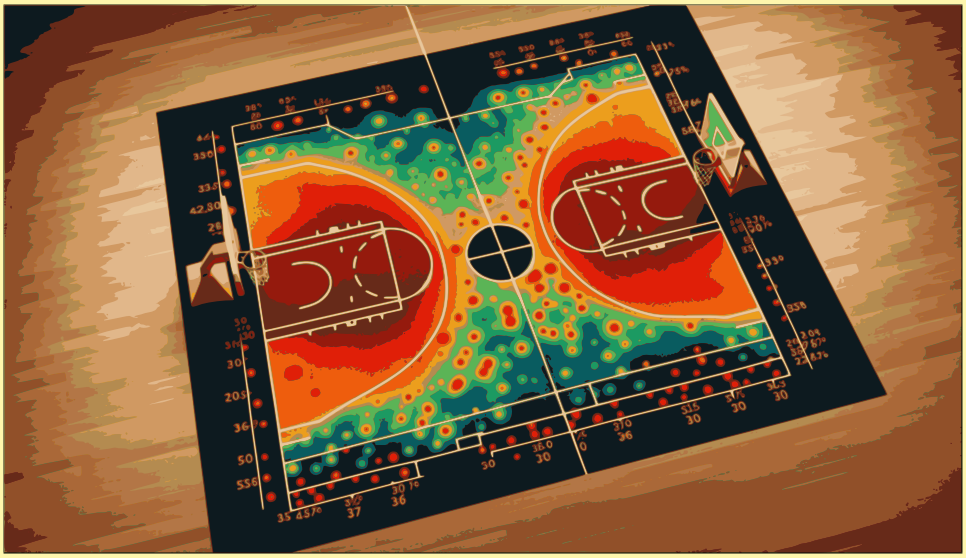
[6]El Real Decreto 576/2023 elevó la duración máxima de los estudios de doctorado de tres a cuatro años, sin duda un plazo más realista para alcanzar las tres publicaciones conducentes a una tesis por compendio en la UAM. Así, tal modificación de la ley se aplicó de oficio en la UAM a todos los doctorandos que habían comenzado sus estudios bajo el Real Decreto 99/2011.


UAM
Universidad Autónoma
de Madrid


UAM
Universidad Autónoma
de Madrid

Capítulo 1

Curvatura



TEST


UÁM
Universidad Autónoma
de Madrid

BUMP HUNTING THROUGH DENSITY CURVATURE FEATURES

JOSÉ E. CHACÓN[†] AND JAVIER FERNÁNDEZ SERRANO[‡]

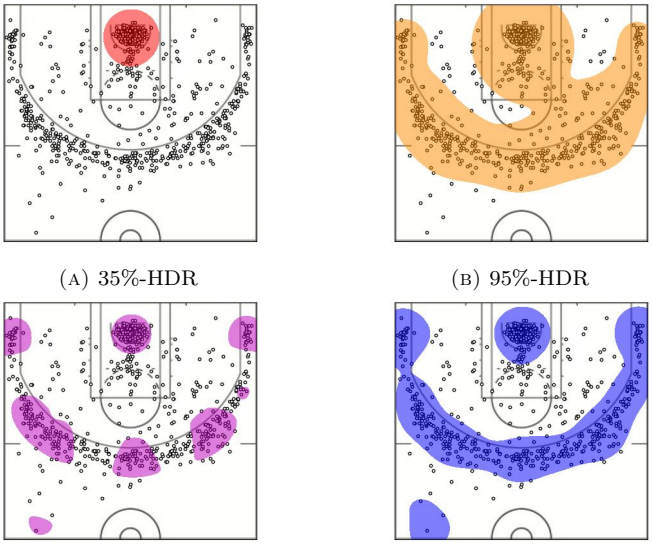
Abstract. Bump hunting deals with finding in sample spaces meaningful data subsets known as bumps. These have traditionally been conceived as modal or concave regions in the graph of the underlying density function. We define an abstract bump construct based on curvature functionals of the probability density. Then, we explore several alternative characterizations involving derivatives up to second order. In particular, a suitable implementation of Good and Gaskins' original concave bumps is proposed in the multivariate case. Moreover, we bring to exploratory data analysis concepts like the mean curvature and the Laplacian that have produced good results in applied domains. Our methodology addresses the approximation of the curvature functional with a plug-in kernel density estimator. We provide theoretical results that assure the asymptotic consistency of bump boundaries in the Hausdorff distance with affordable convergence rates. We also present asymptotically valid and consistent confidence regions bounding curvature bumps. The theory is illustrated through several use cases in sports analytics with datasets from the NBA, MLB and NFL. We conclude that the different curvature instances effectively combine to generate insightful visualizations.

1. Introduction

The subject of *bump hunting* (BH) refers to the set estimation task [2] of discovering meaningful data regions, called *bumps*, in a sample space [21]. The most representative example is the modal regions in a probability density function (pdf), which are literally bumps in its graph. Even though the concept has a broader scope, BH remains relatively unexplored.

Consider the problem of identifying made shots on a basketball court. Coaches, scouts and other personnel might be interested in extracting shooting patterns for adopting specific pre-game strategies, assessing talent or working on player development. Fig. 1 illustrates four different ways of constructing bumps with basketball shot data. Fig. 1a and Fig. 1b correspond to Hyndman's classical *highest density region* (HDR) configurations, while Fig. 1c and Fig. 1d follow our novel *curvature*-based characterizations. Each of them presents a distinctive perspective on the underlying shooting tendencies. Fig. 1a and Fig. 1c point at fine-grained locations, whereas Fig. 1b and Fig. 1d cover entire influence areas. Smaller regions suggest spots to prioritize in an offensive or defensive scheme. The larger ones *connect the dots*, revealing general trends. Both views complement each other to offer a complete picture.

[†]Departamento de Matemáticas, Universidad de Extremadura, Badajoz, Spain.
[‡]Departamento de Matemáticas, Universidad Autónoma de Madrid, Madrid, Spain.
E-mail addresses: [†]jechacon@unex.es ✉, [‡]javier.fernandezs01@estudiante.uam.es.
2020 *Mathematics Subject Classification.* 62G05 (Primary), 62G20, 60D05, 62R07.
Key words and phrases. bump hunting, concavity, Gaussian curvature, kernel density derivative estimation, Laplacian, mean curvature.
[†]https://orcid.org/0000-0002-3675-1960 .
[‡]https://orcid.org/0000-0001-5270-9941 .



(A) 35%-HDR (B) 95%-HDR

(C) Concave bumps (D) Laplacian bumps

FIG. 1. Four ways of constructing bumps for basketball converted shot data. The exact 804 made shot locations are scattered across each sub-figure. The top left and right bumps correspond to HDRs comprising 35% and 95% of all observations. The bottom left bumps highlight regions where the pdf subgraph is locally concave. The bottom right bumps comprise points where the Laplacian of the underlying pdf takes negative values.

1.1. **Goals.** We propose a new BH curvature-based methodology addressing some blind spots of classical methods. Fig. 1a and Fig. 1b either miss or mask relevant information. The finer-grained 35%-HDR does not include the perimeter concave bumps in Fig. 1c. Meanwhile, the 95%-HDR fails to keep the short, mid and long ranges well separated, as opposed to the Laplacian bumps in Fig. 1d.

Contributions. The main contributions of this paper are:

- Presenting a general set estimation framework for curvature-based BH.
- Extending concave bumps to the multivariate setting.
- Introducing mean curvature and Laplacian bumps.
- Deriving consistency convergence rates for curvature bump boundaries.
- Building valid and consistent confidence regions for curvature bumps.
- Showcasing the numerous applications of curvature-based BH.

1.2. **Related work.** One of the first BH references was due to Good and Gaskins in 1980 [21]. They offered a premier definition of a bump as the concave region delimited between two inflection points. Moreover, they suggested an extension to the multivariate case. Fig. 1c corresponds to our implementation of multivariate concave bumps.

In 1996, Hyndman introduced the concept of HDR, which he conceives as level sets of the pdf f that enclose a certain probability mass [23]. More formally, the $(1-\alpha)$-level HDR is defined as $R(f_\alpha) = \{\mathrm{x} : f(\mathrm{x}) \geq f_\alpha\}$, where f_α is the largest value such that $\mathbb{P}(\mathrm{X} \in R(f_\alpha)) \geq 1-\alpha$, and the random variable (rv) X is such that $\mathrm{X} \sim f$. HDRs satisfy the nice property of being the *smallest* sets with a given probability mass.

In 1999, Chaudhuri and Marron presented *SIgnificant ZERo crossings of derivatives* (SiZer), envisioning bumps as places where the first derivative becomes zero [9].

In 2002, Chaudhuri and Marron showcased the role of second derivatives in an unpublished manuscript [10]. Also in 2002, Godtliebsen, Marron, and Chaudhuri explored curvature features from a pointwise perspective by assessing Hessian eigenvalue sign combinations in the bivariate case [20]. A multivariate extension to [20] was formulated in 2008 by Duong et al., targeting the pointwise significance of nonzero Hessian determinants [18]. Lastly, in 2021, Marron and Dryden elaborate on second derivatives in their book *Object Oriented Data Analysis* [27].

1.3. **Outline.** The new methodology is presented in Section 2. The supplementary material (SM) [8] provides the necessary differential geometry foundations. In turn, Section 3 is entirely dedicated to asymptotic consistency and inference results. A sports analytics application is explored in Section 4. The SM [8] includes all the proofs and computational details. We reflect on the proposed methodology in Section 5.

2. Methods

Our methodology finds alternative ways of analyzing sample spaces by exploiting pdfs' curvature properties, adhering to Chaudhuri and Marron's defence of pdf derivatives. Considering Hyndman's approach a well-established tool, we believe there are still some blind spots to address with curvature.

Hyndman's HDRs have the advantage of always including *global* modes. However, they may generally miss *local* modes if small enough; lowering the threshold α might not capture them without obfuscating the HDR. On the other hand, when varying α works, questions remain on the specific value it should take. Moreover, sometimes it is necessary to explore the whole range of $\alpha \in (0,1)$ to recover all the relevant pdf features [31].

Consider a d-variate pdf $f : \mathbb{R}^d \to [0,\infty)$. We define bumps as subsets of $\mathbb{R}^d$ of the form

$$\mathcal{B}^{\phi} = \{\mathrm{x} \in \mathbb{R}^d : (-1)^s \phi[f](\mathrm{x}) \geq 0\}, \tag{1}$$

for some functional ϕ measuring the *curvature* of f at any point, and some sign selector $s \in \{0,1\}$ that will usually be kept implicit. If the gradient $\nabla\phi[f]$ does not vanish near the zero-level set of $\phi[f]$, the bump boundary $\partial\mathcal{B}^{\phi}$ is retrieved by substituting the inequality with an equality sign in (1) [29, Remark 3.1]; see Theorem 2 ahead for a formal condition [14, Assumption G] [11]. Contrary to HDRs, the idea behind (1) is that ϕ carries an implicit threshold, say zero, to determine if a point belongs to the bump, solving the arbitrariness of the choice of α in HDRs.

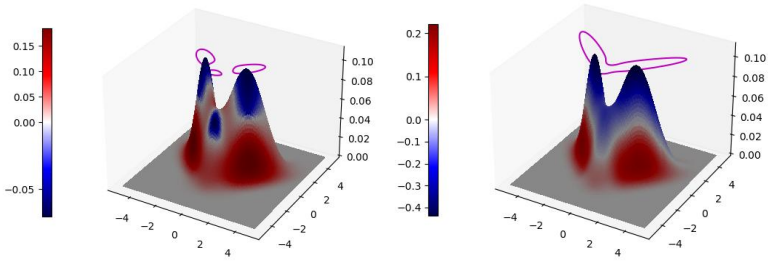
Once some curvature functional is chosen, we propose to employ a kernel plug-in estimator of $\mathcal{B}^{\phi}$, replacing f with its *kernel density estimator* (KDE) in (1). Thus, given a sample $\mathrm{X}_1, \ldots, \mathrm{X}_n$ of independent and identically distributed (i.i.d.) random variables with pdf f and a bandwidth $h > 0$, we consider the KDE of f as

$$\hat{f}_{n,h}(\mathrm{x}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\mathrm{x} - \mathrm{X}_i) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\mathrm{x} - \mathrm{X}_i}{h}\right), \tag{2}$$

for some kernel function K, typically a d-variate pdf. Using (2), we then define the plug-in estimator of (1) as

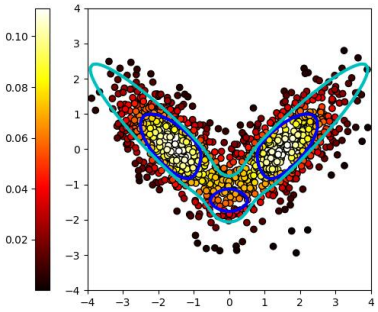
$$\tilde{\mathcal{B}}^{\phi}_{n,h} = \{\mathrm{x} \in \mathbb{R}^d : (-1)^s \phi[\hat{f}_{n,h}](\mathrm{x}) \geq 0\}. \tag{3}$$

To a first approximation, a *scalar* bandwidth is chosen for simplicity. Chacón and Duong demonstrated that, for $d > 1$, unconstrained bandwidth matrices produce significant performance gains, especially in *kernel density derivative estimation* (KDDE) [6, Section 5.2]. Preliminary experiments seem to support their recommendation also for curvature-based BH. Nonetheless, all the theoretical developments



(A) Concave bumps

(B) Mean curvature bump



(C) Both kinds of bumps

FIG. 2. Curvature bumps for a bivariate Gaussian mixture encompassing two equally-weighted components with means $\boldsymbol{\mu}_1 = [-3/2, 0]$, $\boldsymbol{\mu}_2 = [3/2, 0]$ and covariance matrices $\Sigma_1 = [1, -0.7; -0.7, 1]$, $\Sigma_2 = [1, 0.7; 0.7, 1]$. The top two sub-figures show the same graph of the pdf f. The area colours refer to the values taken by a specific curvature functional $\phi[f]$ at each point. For the left-hand picture, this function is the $\lambda_1[f]$ that defines concave bumps (5); on the right, it is the mean curvature $\operatorname{div}(\overline{\nabla} f)$ in (7). The magenta *halos* represent the zero level sets of those functionals and, thus, the corresponding bump boundaries. Concave and mean curvature bump boundaries show in blue and cyan in the bottom sub-figure, along with a 1,000-observation random sample from the mixture, where each point is coloured according to the value of f.

and, consequently, all the exhibition figures in this paper obey this simplification. On the other hand, the kernel K has a lower impact on the results. Most of the statements in Section 3 do not impose a particular choice. However, all of them are compatible with the Gaussian kernel (see [1, 12, 13]), which is almost universally preferred in a multivariate setting [6, p. 15].

For $d = 1$, Chaudhuri and Marron studied the functional $\phi[f] = f''$, which leads to *concave* bumps, if $s = 1$, or *convex* dips, if $s = 0$. Different alternatives arise in the multivariate case. The geometrical concepts in the SM [8] lay the grounds for characterizing bumps in alternative ways to HDRs. Considering pdfs as hypersurfaces, notions like the mean and Gaussian curvatures find new usages in statistics. Fig. 2 illustrates the two main kinds of curvature bumps in this paper. Even though ϕ may *a priori* depend on partial derivatives of f of arbitrary order, the theory of hypersurfaces in the SM [8] suggests that our quest for curvature features is essentially fulfilled with up to second derivatives of the pdf f.

Given the connection of curvature with second derivatives, we propose targeting $r = 2$ in one of the standard bandwidth selectors [5]. The same heuristic worked well for KDE-based applications such as mean shift clustering or feature significance testing [6, Chapter 6].

2.1. **Concavity and convexity.** Given a sufficiently smooth pdf f, let us define $\lambda_i[f]$, for $i \in \{1, 2, \ldots, d\}$, as the function mapping $\mathrm{x} \in \mathbb{R}^d$ to the i-th largest possibly repeated eigenvalue of $\mathrm{D}^2 f(\mathrm{x})$, the Hessian matrix of f at x, i.e.,

$$\lambda_1[f](\mathrm{x}) \geq \lambda_2[f](\mathrm{x}) \geq \cdots \geq \lambda_d[f](\mathrm{x}) \,, \tag{4}$$

for all $\mathrm{x} \in \mathbb{R}^d$. As mentioned in the SM [8], the eigenvalues of the Hessian (or the shape operator, equivalently) determine local concavity and convexity. Let us assume that $(-1)^s \lambda_i[f] > 0$, for all i on some subset $\mathcal{U} \subset \mathbb{R}^d$. If $s = 0$, f will be locally convex, whereas if $s = 1$, it will be locally concave on $\mathcal{U}$. Considering the ordering of functions (4), we can express the former concave and convex bumps in terms of a single functional, aligned with a specific sign s, as, respectively,

$$\mathcal{B}^{\lambda_1} = \{\mathrm{x} \in \mathbb{R}^d : \lambda_1[f](\mathrm{x}) \leq 0\} \,, \tag{5}$$

$$\mathcal{B}^{\lambda_d} = \{\mathrm{x} \in \mathbb{R}^d : \lambda_d[f](\mathrm{x}) \geq 0\} \,. \tag{6}$$

The concave region (5) yields the most recognizable flavour of bumps in the literature, this time in a multivariate setting. It is the method depicted in Fig. 1c. As for (6), they are actually not bumps but *dips*. Assuming non-degenerate Hessians, concave bumps typically delineate areas near local pdf modes, while convex dips do with local minima. Consequently, the former and the latter are known as *peaks* and *holes* [20, Table 1].

When concave bumps contain local modes, they make the most natural definition of a d-dimensional neighbourhood. Although straightforward, considering modal regions as ε-fattenings or enlargements (see Section 3.1.1 below) poses challenges regarding the choice of $\varepsilon > 0$, as similarly argued for α in HDRs. Besides, employing a single radius ε limits the overall expressiveness of the bump. On the other hand, if we saw modal regions as *basins of attraction* instead [3], despite ε disappearing and attaining more flexibility, we would not be pursuing a solution to a BH problem any more but a clustering one, giving up on the cohesive sense of bumps. In this respect, concave bumps provide us with an elegant compromise answer.

Moreover, this modal vicinity notion seamlessly incorporates the missing mode scenario. Concave bumps point out incipient modal regions as the central *mouth* in Fig. 2a and Fig. 2c, which does not contain a mode. Such *weak* modal regions are well-known in the context of univariate mode hunting as *shoulders*, representing complicated cases [15]. As for BH, d-dimensional shoulders deserve attention as evidence of hidden structure. See the NFL application in the SM [8] for an interpretable dynamic shoulder. In turn, the mouth in Fig. 2 is characteristic of mixtures whose components influence each other significantly. All in all, concave bumps subsume the modal regions, having a slightly broader reach.

2.2. **Gradient divergence.** Concave bumps may be too restrictive in some use cases. Imagine the pdf graph as a *landscape*, with mountains being local high-density regions. Concave bumps originate near mountain *peaks*, missing most of the *hillside*. Mean curvature allows the discovery of entire mountain chains.

The shape operator is a linear map of the tangent space that measures how a manifold bends in different directions (see the SM [8] for a formal definition). Let us consider its eigenvalues: the principal curvatures. Concavity requires all principal curvatures to be negative. By contrast, the mean curvature adds them all so that only the net sign matters. Computing curvature in this way fills the gaps between concave peaks in a *long ridge* [20, Table 1], as depicted in Fig. 2b and Fig. 2c in the form of a *boomerang*.

The SM [8] shows the connection between the mean curvature and divergence of the normalized version of the gradient $\bar{\nabla} f = \nabla f / \sqrt{1 + \|\nabla f\|^2}$. The divergence operator takes positive values when the argument field *diverges* from a point, whereas

the sign is negative when it *converges*. Therefore, we define the mean curvature bump as

$$\mathcal{B}^{\bar{\nabla}} = \{\mathrm{x} \in \mathbb{R}^d : \mathrm{div}(\bar{\nabla} f)(\mathrm{x}) \leq 0\}\,. \tag{7}$$

When the gradient is slight, as is usually the case for pdfs (one can even tweak the scale of the random variables to make $\|\nabla f\|$ small), the Laplacian $\Delta f = \mathrm{div}(\nabla f) = \sum_{i=1}^{d} \partial^2 f / \partial x_i^2$ roughly approximates the mean curvature (see [19, Equation 5.28]). Hence, we define the Laplacian bump as

$$\mathcal{B}^{\Delta} = \{\mathrm{x} \in \mathbb{R}^d : \Delta f(\mathrm{x}) \leq 0\}\,. \tag{8}$$

Note that $\mathcal{B}^{\lambda_1} \subset \mathcal{B}^{\Delta}$. Even though (8) may be less intrinsic than (7), it has a more straightforward form, for Δ is a second-order linear differential operator on f. A discretized version of the Laplacian operator has been used for contour detection in image processing through the *Laplacian-of-Gaussian* algorithm [22]. We have already seen an example of a Laplacian bump in Fig. 1d. The results would have been almost indistinguishable if the mean curvature had been employed.

The term *ridge* was used above to convey a mountain range covering several peaks following [20]. Ridges also refer in the statistical literature to a specific definition of higher-dimensional pdf modes [13]. This concept of ridge shares with Laplacian and mean curvature bumps the ability to unveil filament-like structures. However, ridges are intrinsically one-dimensional in their most typical form. For them to extend to $\mathbb{R}^d$, one would need to take an ε-enlargement, introducing some arbitrariness and rigidity with ε that gradient divergence bumps do not have. In our context, we will stick to the informal meaning of *ridge* in the following sections.

2.3. **Intrinsic curvature.** The Gaussian curvature is an intrinsic measure derived from the shape operator (see the SM [8] for a precise definition). This and the Hessian determinant provide alternative ways to detect warps. The analysis of these two notions is more subtle than in the previous sections: from the definition of Gaussian curvature in the SM [8], many sign combinations among the multiplied principal curvatures produce the same net sign.

The Gaussian curvature and the Hessian determinant differ by a positive factor; thus, if we set the bump detection threshold at zero, we can restrict our analysis to the latter. In the bivariate case, the bump

$$\mathcal{B}^{\det} = \{\mathrm{x} \in \mathbb{R}^2 : \det(\mathrm{D}^2 f)(\mathrm{x}) \geq 0\} \tag{9}$$

coincides with the union of (5) and (6). Therefore, (9) is helpful for detecting both concave bumps and convex dips simultaneously. We will refer to (9) as a *Gaussian bump*.

3. Asymptotics

This section will demonstrate the soundness of plug-in estimators in the asymptotic regime for curvature bumps.

3.1. **Consistency.** We rely on a recent result by Chen to prove consistency [11]. Let

$$\mathcal{M} = \{\mathrm{x} \in \mathbb{R}^d : \Psi(\mathrm{x}) = 0\}\,, \tag{10}$$

$$\tilde{\mathcal{M}} = \{\mathrm{x} \in \mathbb{R}^d : \tilde{\Psi}(\mathrm{x}) = 0\} \tag{11}$$

be two solution manifolds defined by their criterion functions $\Psi, \tilde{\Psi} : \mathbb{R}^d \to \mathbb{R}$, respectively. Chen's stability theorem shows that $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are near whenever the criterion functions and their derivatives are close. In our context, Ψ will represent a curvature measure and $\tilde{\Psi}$ the corresponding kernel plug-in estimator so that $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are the boundaries of the associated curvature bumps.

3.1.1. *Notational preliminaries.* The theory of convergence in the uniform norm for KDDE allows applying Chen's stability theorem to the curvature BH problem.

Vectors of nonnegative integers $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d) \in \mathbb{Z}_+^d$ shall represent partial derivatives through $\partial^{\boldsymbol{\beta}} f = \partial^{|\boldsymbol{\beta}|} f / \partial x_1^{\beta_1} \cdots \partial x_d^{\beta_d}$, where $|\boldsymbol{\beta}| = \sum_{i=1}^d \beta_i$. Let us call $\mathbb{Z}_+^d[k] = \{\boldsymbol{\beta} \in \mathbb{Z}_+^d : |\boldsymbol{\beta}| \leq k\}$. We also include the case $\boldsymbol{\beta} = \mathbf{0}$, which represents the identity. Let us also define, for any derivative index vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m \in \mathbb{Z}_+^d$, the function $\partial^{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m} f : \mathbb{R}^d \to \mathbb{R}^m$ as $\partial^{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m} f(\mathrm{x}) = (\partial^{\boldsymbol{\beta}_1} f(\mathrm{x}), \ldots, \partial^{\boldsymbol{\beta}_m} f(\mathrm{x}))$.

We will denote $\mathcal{C}^\ell(A)$ the class of functions $\varphi : A \subset \mathbb{R}^d \to \mathbb{R}$ with continuous partial derivatives up to ℓ-th order. Likewise, we will say that a function $\varphi : \mathbb{R}^d \to \mathbb{R}$ is Hölder continuous with exponent $\alpha \in (0, 1]$ if there exists $C \in (0, \infty)$ such that $|\varphi(\mathrm{x}) - \varphi(\mathrm{y})| \leq C\|\mathrm{x} - \mathrm{y}\|^\alpha$, for all $\mathrm{x}, \mathrm{y} \in \mathbb{R}^d$ [24]. By convention, we include the case $\alpha = 0$ when Hölder continuity does not hold for any positive exponent.

For any $\varphi : \mathbb{R}^d \to \mathbb{R}$ and some $A \subset \mathbb{R}^d$, we denote $\|\varphi\|_\infty = \sup_{\mathrm{x} \in A} |\varphi(\mathrm{x})|$, and we will indicate that the supremum is over A by explicitly stating that $\|\varphi\|_\infty$ satisfies some property *on* A. Also, write $\|\varphi\|_{\infty,k} = \max\left\{\|\partial^{\boldsymbol{\beta}}\varphi\|_\infty : \boldsymbol{\beta} \in \mathbb{Z}_+^d, |\boldsymbol{\beta}| = k\right\}$. All these norms will formalize how close the criterion functions and their respective derivatives are.

On the other hand, the stability theorem invokes some other concepts related to sets. Let us define the distance from a point $x \in \mathbb{R}^d$ to some subset $A \subset \mathbb{R}^d$ as $d(\mathrm{x}, A) = \inf_{\mathrm{y} \in A} \|\mathrm{x} - \mathrm{y}\|$, and the ε-fattening of a set $A \subset \mathbb{R}^d$, where $\varepsilon > 0$, as $A \oplus \varepsilon = \{\mathrm{x} \in \mathbb{R}^d : d(\mathrm{x}, A) \leq \varepsilon\}$. Finally, the *Hausdorff distance* between two subsets $A, B \subset \mathbb{R}^d$ is $\mathrm{Haus}(A, B) = \max\{\sup_{\mathrm{x} \in B} d(\mathrm{x}, A), \sup_{\mathrm{x} \in A} d(\mathrm{x}, B)\}$.

The problem of uniformly bounding the KDDE error refers to finding an infinitesimal bound for $\sup_{\mathrm{x} \in \mathbb{R}^d} |\partial^{\boldsymbol{\beta}} \hat{f}_{n,h}(\mathrm{x}) - \partial^{\boldsymbol{\beta}} f(\mathrm{x})|$. Note that the latter is bounded by the *bias* $\sup_{\mathrm{x} \in \mathbb{R}^d} |\mathbb{E}[\partial^{\boldsymbol{\beta}} \hat{f}_{n,h}(\mathrm{x})] - \partial^{\boldsymbol{\beta}} f(\mathrm{x})|$ plus the *stochastic error* $\sup_{\mathrm{x} \in \mathbb{R}^d} |\partial^{\boldsymbol{\beta}} \hat{f}_{n,h}(\mathrm{x}) - \mathbb{E}[\partial^{\boldsymbol{\beta}} \hat{f}_{n,h}(\mathrm{x})]|$. We will analyze both terms separately.

3.1.2. *Bias analysis.* Lemma 1 is an extended version of [1, Lemma 2] with alternative hypotheses to ensure consistency under less stringent differentiability assumptions. Namely, we resort to Hölder and uniform continuity, following the example of [24] and [28, Theorem 1.1, p. 42].

Lemma 1. *Let $\boldsymbol{\beta} \in \mathbb{Z}_+^d$ be a partial derivative index vector. Let f be a pdf in $\mathcal{C}^{|\boldsymbol{\beta}|+r}(\mathbb{R}^d)$, for some $r \in \mathbb{Z}_+ \cup \{\infty\}$, with all partial derivatives bounded up to $(|\boldsymbol{\beta}| + r)$-th order. Assume that $\partial^{\boldsymbol{\beta}} f$ is Hölder continuous on $\mathbb{R}^d$ with exponent $\alpha \in [0, 1]$. If the exponent is $\alpha = 0$, then ultimately assume that $\partial^{\boldsymbol{\beta}} f$ is uniformly continuous. Finally, let $\hat{f}_{n,h}$ be the KDE of f based on a true pdf kernel K vanishing at infinity and satisfying the moment constraints*

$$\int_{\mathbb{R}^d} \mathrm{x}\ K(\mathrm{x})\ d\mathrm{x} = \mathbf{0}\,, \qquad\qquad \int_{\mathbb{R}^d} |\mathrm{x}_i \mathrm{x}_j|\ K(\mathrm{x})\ d\mathrm{x} < \infty\,,$$

for all $i, j \in \{1, \ldots, d\}$. Then,

$$\sup_{\mathrm{x} \in \mathbb{R}^d} |\mathbb{E}[\partial^{\boldsymbol{\beta}} \hat{f}_{n,h}(\mathrm{x})] - \partial^{\boldsymbol{\beta}} f(\mathrm{x})| = \begin{cases} O(h^s), & \text{if } \max\{r, \alpha\} > 0 \\ o(1) \text{ as } h \to 0, & \text{otherwise} \end{cases},$$

where $s = \max\{\alpha, \min\{r, 2\}\}$.

3.1.3. *Stochastic error analysis.* Lemma 2 below appears as an auxiliary result in [1] in the case $\ell = 3$, but the proof works for an arbitrary ℓ.

Lemma 2 (Arias-Castro, Mason, and Pelletier, 2016 [1])**.** *Let f be a bounded pdf in $\mathbb{R}^d$ and let $\hat{f}_{n,h}$ be the KDE of f. Fix a nonnegative integer ℓ as the maximum partial derivative order. Assume that K is a product kernel of the form $K(x_1, \ldots, x_d) =$*

$\prod_{i=1}^{d} \kappa_i(x_i)$, *where each* κ_i *is a univariate pdf of class* $\mathcal{C}^{\ell}(\mathbb{R})$. *Further, assume that all the partial derivatives up to* ℓ*-th order of* K *are of bounded variation and integrable on* $\mathbb{R}^d$. *Then, there exists* $b \in (0,1)$ *such that, if* $h \equiv h_n$ *is a sequence satisfying* $\log n \leq nh^d \leq bn$, *then*

$$\sup_{\mathrm{x}\in\mathbb{R}^d} |\partial^{\boldsymbol{\beta}} \hat{f}_{n,h}(\mathrm{x}) - \mathbb{E}[\partial^{\boldsymbol{\beta}} \hat{f}_{n,h}(\mathrm{x})]| = O\left(\sqrt{\frac{\log n}{nh^{d+2|\boldsymbol{\beta}|}}}\right),$$

almost surely (a.s.) for all $\boldsymbol{\beta} \in \mathbb{Z}_+^d[\ell]$.

Finally, note that Lemma 2 also holds for a sufficiently small but constant h.

3.1.4. *Total error analysis.* Combining Lemma 1 and Lemma 2, we obtain a general consistency result in the supremum norm for KDDE. We will focus on the Gaussian kernel for simplicity, but any other satisfying the conditions in both Lemma 1 and Lemma 2 would do.

Theorem 1. *Let* $\boldsymbol{\beta} \in \mathbb{Z}_+^d$ *be a partial derivative index vector. Let* f *be a pdf in* $\mathcal{C}^{|\boldsymbol{\beta}|+r}(\mathbb{R}^d)$, *for some* $r \in \mathbb{Z}_+ \cup \{\infty\}$, *with all partial derivatives bounded up to* $(|\boldsymbol{\beta}| + r)$*-th order. Assume that* $\partial^{\boldsymbol{\beta}} f$ *is Hölder continuous on* $\mathbb{R}^d$ *with exponent* $\alpha \in [0,1]$. *If the exponent is* $\alpha = 0$, *then ultimately assume that* $\partial^{\boldsymbol{\beta}} f$ *is uniformly continuous. Let* $\hat{f}_{n,h}$ *be the KDE of* f *based on the Gaussian kernel. Finally, let* $h \equiv h_n$ *be a sequence converging to zero as* $n \to \infty$ *and satisfying* $nh^d \geq \log n$. *Then,*

$$\sup_{\mathrm{x}\in\mathbb{R}^d} |\partial^{\boldsymbol{\beta}} \hat{f}_{n,h}(\mathrm{x}) - \partial^{\boldsymbol{\beta}} f(\mathrm{x})| = \begin{cases} O\left(h^s + \sqrt{\frac{\log n}{nh^{d+2|\boldsymbol{\beta}|}}}\right), & \text{if } \max\{r,\alpha\} > 0 \\ o(1) + O\left(\sqrt{\frac{\log n}{nh^{d+2|\boldsymbol{\beta}|}}}\right), & \text{otherwise} \end{cases},$$

a.s. as $n \to \infty$, *where* $s = \max\{\alpha, \min\{r, 2\}\}$. *In particular,*

$$\sup_{\mathrm{x}\in\mathbb{R}^d} |\partial^{\boldsymbol{\beta}} \hat{f}_{n,h}(\mathrm{x}) - \partial^{\boldsymbol{\beta}} f(\mathrm{x})| \xrightarrow[n\to\infty]{a.s.} 0 \text{ , if } \frac{\log n}{nh^{d+2|\boldsymbol{\beta}|}} \xrightarrow[n\to\infty]{} 0\,.$$

3.1.5. *Manifold stability.* Theorem 2 gathers the essential elements of Chen's stability theorem needed in our context.

Theorem 2 (Chen, 2022 [11])**.** *Let* $\Psi, \tilde{\Psi} : \mathbb{R}^d \to \mathbb{R}$ *and let* $\mathcal{M}$ *and* $\tilde{\mathcal{M}}$ *be as defined in* (10) *and* (11), *respectively. Assume that:*

A1. There exists $\delta > 0$ *such that* Ψ *has bounded first-order derivatives on* $\mathcal{M} \oplus \delta$.
A2. There exists $\lambda > 0$ *such that* $\|\nabla\Psi(\mathrm{x})\| > \lambda$, *for all* $\mathrm{x} \in \mathcal{M} \oplus \delta$.
A3. $\|\tilde{\Psi} - \Psi\|_\infty$ *is sufficiently small on* $\mathbb{R}^d$.

Moreover, suppose that:

B1. $\tilde{\Psi}$ *has bounded first-order derivatives on* $\mathcal{M} \oplus \delta$.
B2. $\|\tilde{\Psi} - \Psi\|_{\infty,1}$ *is sufficiently small on* $\mathcal{M} \oplus \delta$.

Then, $\mathrm{Haus}(\tilde{\mathcal{M}}, \mathcal{M}) = O(\|\tilde{\Psi} - \Psi\|_\infty)$.

We have introduced in Theorem 2 a slight relaxation on the differentiability constraint for $\tilde{\Psi}$. Chen supposes differentiability and bounds on $\mathbb{R}^d$, whereas we allow for a narrower domain $\mathcal{M} \oplus \delta$. This deviation is justified since hypotheses (A) imply $\tilde{\mathcal{M}} \subset \mathcal{M} \oplus \varepsilon \subset \mathcal{M} \oplus \delta$, where $\varepsilon < \delta$. Since pdfs typically vanish at infinity, it might be unfeasible to ask $\tilde{\Psi} = \phi[\hat{f}_{n,h}]$ to be differentiable everywhere. This is the case for the eigenvalues (4) in Proposition 1 below, where condition (12) would not hold if the infimum were taken over $\mathbb{R}^d$.

Finally, putting all the pieces together, we get the following main result.

Theorem 3. *Assume the following:*

- *Let ϕ be a curvature functional defined over d-variate pdfs depending on their partial derivatives up to ℓ-th order. More formally, given a pdf p, we have $\phi[p] = \varphi \circ \partial^{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_m} p$, for some $\varphi : \mathbb{R}^m \to \mathbb{R}$ and derivative index vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m \in \mathbb{Z}_+^d[\ell]$.*
- *Let f be a pdf in $\mathcal{C}^{\ell+r}(\mathbb{R}^d)$, for some $r \in \{1, 2, \ldots, \infty\}$, with all partial derivatives bounded up to $(\ell+r)$-th order. If $r = 1$, further assume that the $(\ell+1)$-th partial derivatives of f are either Hölder continuous with exponent $\alpha \in (0, 1]$ or uniformly continuous.*
- *Let $\hat{f}_{n,h}$ be the KDE of f based on the Gaussian kernel.*
- *Let $h \equiv h_n$ converge to zero and satisfy $\lim_{n\to\infty} n^{-1}h^{-(d+2\ell+2)} \log n = 0$.*

Let the curvature bump boundary and its plug-in estimator, respectively, be

$$\partial\mathcal{B}^{\phi} = \{\mathrm{x} \in \mathbb{R}^d : \phi[f](\mathrm{x}) = 0\}\,, \qquad \partial\tilde{\mathcal{B}}^{\phi}_{n,h} = \{\mathrm{x} \in \mathbb{R}^d : \phi[\hat{f}_{n,h}](\mathrm{x}) = 0\}\,.$$

Further, suppose that:

- *There exists $\delta > 0$ such that $\varphi \in \mathcal{C}^1(\mathcal{U})$, for some open set $\mathcal{U} \subset \mathbb{R}^m$ containing the images of $\partial\mathcal{B}^{\phi} \oplus \delta$ under both $\partial^{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_m} f$ and $\partial^{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_m} \hat{f}_{n,h}$ a.s.*
- *There exists $\lambda > 0$ such that $\|\nabla\phi[f](\mathrm{x})\| > \lambda$, for all $\mathrm{x} \in \partial\mathcal{B}^{\phi} \oplus \delta$.*

Then,

$$\mathrm{Haus}(\partial\tilde{\mathcal{B}}^{\phi}_{n,h}, \partial\mathcal{B}^{\phi}) = O\left(h^{\min\{r,2\}} + \sqrt{\frac{\log n}{nh^{d+2\ell}}}\right) \xrightarrow[n\to\infty]{a.s.} 0\,.$$

The optimal bound is $\mathrm{Haus}(\partial\tilde{\mathcal{B}}^{\phi}_{n,h}, \partial\mathcal{B}^{\phi}) = O([n^{-1}\log n]^{2/(d+2\ell+4)})$, achieved with $h \asymp [n^{-1}\log n]^{1/(d+2\ell+4)}$ $(r \geq 2)$. The former coincides up to a logarithmic term with the optimum in KDDE for ℓ-th order partial derivatives according to the *root mean integrated square error* criterion, which is $O(n^{-2/(d+2\ell+4)})$ [7].

Theorem 3 straightforwardly leads to bump boundary convergence results for the determinants and traces of the shape operator and the Hessian matrix.

Example 1. Consider the Laplacian and Gaussian bumps (8) and (9) for a bivariate pdf $f : \mathbb{R}^2 \to [0, \infty)$, with $\phi[f]$ equal to, respectively,

$$\mathrm{tr}(\mathrm{D}^2 f) \equiv \Delta f = \frac{\partial^2 f}{\partial x_1^2} + \frac{\partial^2 f}{\partial x_2^2}\,, \qquad \det(\mathrm{D}^2 f) = \frac{\partial^2 f}{\partial x_1^2}\frac{\partial^2 f}{\partial x_2^2} - \left(\frac{\partial^2 f}{\partial x_1 \partial x_2}\right)^2.$$

For the trace, the underlying derivative functional is $\varphi(a_1, a_2) = a_1 + a_2$, considering $\boldsymbol{\beta}_1 = (2, 0)$ and $\boldsymbol{\beta}_2 = (0, 2)$. In turn, the functional is $\varphi(a_1, a_2, a_3) = a_1 a_2 - a_3^2$ for the determinant, taking $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ as before plus $\boldsymbol{\beta}_3 = (1, 1)$. In both cases, φ is an infinitely smooth function over $\mathcal{U} = \mathbb{R}^m$, making every $\delta > 0$ satisfy the requirement in Theorem 3 without imposing additional hypotheses on the original pdf and its KDE.

The case for the Hessian eigenvalues is more involved. The functions $\lambda_i[f]$ in (4) are not generally $\mathbb{R}^d$-differentiable. To solve this differentiability issue, we will follow the standard assumption in Kato's book that, for every $\mathrm{x} \in \mathbb{R}^d$, all the eigenvalues of $\mathrm{D}^2 f(\mathrm{x})$ have multiplicity one [25, Theorem 5.16, p. 119]. We will ask for an even stronger hypothesis to ensure that all plug-in estimators $\lambda_i[\hat{f}_{n,h}]$ are eventually distinct everywhere for large n a.s.

Proposition 1. *Let f be a pdf and let $\hat{f}_{n,h}$ be its KDE. Let us assume that f and $\hat{f}_{n,h}$ satisfy all the conditions in Theorem 1 so that the second-order partial derivatives of f are consistently approximated with plug-in estimators. Let us call $\partial\mathcal{B}^{\phi}$*

the bump boundary for the criterion function $\phi \equiv \lambda_j[f]$, for some $j \in \{1, \dots, d\}$. If there exists $\delta > 0$ such that

$$\inf_{\mathrm{x} \in \partial \mathcal{B}^\phi \oplus \delta} \{\lambda_i[f](\mathrm{x}) - \lambda_{i+1}[f](\mathrm{x})\} > 0\,, \tag{12}$$

for all $i \in \{1, \dots, d-1\}$, then (12) also holds a.s. for n sufficiently large if we replace f by $\hat{f}_{n,h}$. In particular, both $\lambda_j[f]$ and $\lambda_j[\hat{f}_{n,h}]$ are infinitely differentiable functions of the second-order partial derivatives of f and $\hat{f}_{n,h}$, respectively, on some neighbourhood $\partial\mathcal{B}^\phi \oplus \delta$ a.s. for n sufficiently large.

3.2. **Inference.** In this section, we derive bootstrap inference for curvature bumps, following similar steps as in the scheme developed by Chen, Genovese, and Wasserman for pdf level sets [14]. To accommodate the required techniques, we will exclusively focus on curvature functionals ϕ deriving from the pdf Hessian $\mathrm{D}^2 f$.

3.2.1. *Inference scheme.* We will simplify the inference problem by targeting $f_h : \mathbb{R}^d \to [0, \infty)$, given by $f_h(\mathrm{x}) = \mathbb{E}[\hat{f}_{n,h}(\mathrm{x})]$, instead of f, considering the bias negligible for a small h. There are compelling arguments favouring f_h against f for inference purposes (see [14, Section 2.2] for a thorough discussion).

Let us call $\mathcal{B}_h^\phi$ the smoothed version of (1) derived by replacing f with f_h. We will assume that $\mathcal{B}_h^\phi \subset \Theta$, for some $\Theta \subset \mathbb{R}^d$, or at least that the inferential procedure focuses on $\mathcal{B}_h^\phi \cap \Theta$. Ideally, Θ should be as *small* as possible (hopefully $\Theta \neq \mathbb{R}^d$) so that the resulting confidence regions are *efficient*.

Given $\alpha \in (0, 1)$, a path for narrowing down a $(1-\alpha)$-level confidence region for $\mathcal{B}_h^\phi$ is constructing two sets

$$\begin{aligned} \bar{\mathcal{B}}_{n,h}^\phi(\zeta_{n,h}^\alpha) &= \{\mathrm{x} \in \Theta : (-1)^s \phi[\hat{f}_{n,h}](\mathrm{x}) \geq -\zeta_{n,h}^\alpha\} \\ \underline{\mathcal{B}}_{n,h}^\phi(\zeta_{n,h}^\alpha) &= \{\mathrm{x} \in \Theta : (-1)^s \phi[\hat{f}_{n,h}](\mathrm{x}) \geq \zeta_{n,h}^\alpha\} \end{aligned}\,, \tag{13}$$

for some margin $\zeta_{n,h}^\alpha \in [0, \infty)$. Note that $\underline{\mathcal{B}}_{n,h}^\phi(\zeta_{n,h}^\alpha) \subset \tilde{\mathcal{B}}_{n,h}^\phi \subset \bar{\mathcal{B}}_{n,h}^\phi(\zeta_{n,h}^\alpha)$, thus (13) are set bounds for the $\tilde{\mathcal{B}}_{n,h}^\phi$ in (3) approximating $\mathcal{B}_h^\phi$. This *vertical* scheme is similar to Chen, Genovese, and Wasserman's second method for pdf level set inference [14] and a particular case of Mammen and Polonik's universal approach [26].

Our inference results will establish conditions to ensure the previous set inequality eventually holds too with probability $1 - \alpha$ when replacing $\tilde{\mathcal{B}}_{n,h}^\phi$ with $\mathcal{B}_h^\phi$ while the set bounds (13) draw nearer $\mathcal{B}_h^\phi$, namely

$$\begin{cases} \mathbb{P}\left(\underline{\mathcal{B}}_{n,h}^\phi(\tilde{\zeta}_{n,h}^\alpha) \subset \mathcal{B}_h^\phi \subset \bar{\mathcal{B}}_{n,h}^\phi(\tilde{\zeta}_{n,h}^\alpha)\right) \geq 1 - \alpha + o(1) \\ \tilde{\zeta}_{n,h}^\alpha = o(1) \end{cases}\,, \tag{14}$$

as $n \to \infty$, for some sequence $\{\tilde{\zeta}_{n,h}^\alpha\}_{n=1}^\infty$. The inference scheme (14) can be proven for all curvature bumps using Theorem 4. From Section 3.1, it is an exercise to realize that, under the conditions in which (14) will hold, and with a few mild additional assumptions, the boundaries of the set bounds (13) converge in the Hausdorff distance to $\partial\mathcal{B}_h^\phi$.

In what follows, we will equivalently denote $\mathcal{Q}_p\{X\} \equiv \mathcal{Q}_X(p)$ the p-th quantile, $p \in (0, 1)$, of the rv X [32, p. 304].

Theorem 4. *In the context described above, assume the following:*

I There exists a sequence of random variables $\{Z_{n,h}\}_{n=1}^\infty$ such that, for sufficiently large $n \in \mathbb{N}$, $\mathcal{S}_{n,h}[\phi] \equiv \sup_{\mathrm{x} \in \Theta} |\phi[\hat{f}_{n,h}](\mathrm{x}) - \phi[f_h](\mathrm{x})| \leq Z_{n,h}$ a.s. Let us further assume that $\sqrt{n} Z_{n,h}$ converges weakly [33] to some rv $\mathcal{Z}$ as $n \to \infty$, denoted by $\sqrt{n} Z_{n,h} \rightsquigarrow \mathcal{Z}$. Suppose that $\mathcal{Z}$ has a continuous and strictly increasing cumulative distribution function (cdf).

II For each $\alpha \in (0,1)$*, there is* $\{\zeta^{\alpha}_{n,h}\}_{n=1}^{\infty}$ *satisfying* $\zeta^{\alpha}_{n,h} \geq \mathcal{Q}_{1-\alpha}\{Z_{n,h}\}$*, for all* $n \in \mathbb{N}$*, and* $\lim_{n\to\infty} \zeta^{\alpha}_{n,h} = 0$*.*

III For each $\alpha \in (0,1)$*, there is* $\{\tilde{\zeta}^{\alpha}_{n,h}\}_{n=1}^{\infty}$ *satisfying* $|\tilde{\zeta}^{\alpha}_{n,h} - \zeta^{\alpha}_{n,h}| = o(n^{-1/2})$ *as* $n \to \infty$*.*

Then, for all $\alpha \in (0,1)$*, the asymptotic validity of the inference scheme* (14) *holds.*

The following sections will introduce theoretical results leading to bootstrap estimates $\tilde{\zeta}^{\alpha}_{n,h}$ that can be feasibly computed in practice.

Mammen and Polonik's approach [26] achieves a sharp asymptotic coverage probability $1-\alpha$ in (14). A key difference separating their proposal from Chen, Genovese, and Wasserman's and ours is that they manage to bootstrap from an rv that is a supremum over a neighbourhood of the level set, unlike $\mathcal{S}_{n,h}[\phi]$ in Theorem 4, which considers the whole Θ. See [30] for an overview of similar local strategies for level sets. Based on that, Mammen and Polonik's method will generally be less conservative.

3.2.2. *Bootstrap outline.* The main point to fill the Theorem 4 template is approximating the stochastic errors for second-order linear differential operators $\mathcal{D}$

$$\mathcal{E}_{n,h}[\mathcal{D}] = \sup_{\mathrm{x}\in\Theta} |\mathcal{D}\hat{f}_{n,h}(\mathrm{x}) - \mathcal{D}f_h(\mathrm{x})|\,, \tag{15}$$

using bootstrap estimates

$$\mathcal{E}^{*}_{n,h}[\mathcal{D}|\mathfrak{X}_n] = \sup_{\mathrm{x}\in\Theta} |\mathcal{D}\hat{f}^{*}_{n,h}(\mathrm{x}|\mathfrak{X}_n) - \mathcal{D}\hat{f}_{n,h}(\mathrm{x}|\mathfrak{X}_n)|\,, \tag{16}$$

where $\hat{f}^{*}_{n,h}(\cdot|\mathfrak{X}_n)$ denotes the KDE based on n i.i.d. random variables $\mathrm{X}^{*}_1,\ldots,\mathrm{X}^{*}_n \sim \mathbb{P}^{*}_n\{\mathfrak{X}_n\}$ of the empirical bootstrap probability measure $\mathbb{P}^{*}_n\{\mathfrak{X}_n\}$ assigning equal masses $1/n$ to each component $\mathrm{x}_i \in \mathbb{R}^d$ of a particular n-size i.i.d. realization $\mathfrak{X}_n = \{\mathrm{x}_1,\ldots,\mathrm{x}_n\}$ from f, and $\hat{f}_{n,h}(\cdot|\mathfrak{X}_n)$ is the realization of the KDE based on $\mathfrak{X}_n$, i.e.,

$$\hat{f}^{*}_{n,h}(\mathrm{x}|\mathfrak{X}_n) = \frac{1}{n}\sum_{i=1}^{n} K_h(\mathrm{x}-\mathrm{X}^{*}_i)\,, \qquad \hat{f}_{n,h}(\mathrm{x}|\mathfrak{X}_n) = \frac{1}{n}\sum_{i=1}^{n} K_h(\mathrm{x}-\mathrm{x}_i)\,.$$

Assume that both (15) and (16) use the same kernel K everywhere. Estimating confidence regions for curvature bumps will go through, directly or indirectly, approximating the cdf of (15) with that of (16).

3.2.3. *Gaussian process approximation.* Lemma 3 below allows a *Gaussian process* (GP) approximation between the suprema (15) and (16). See [33] for further knowledge about GPs. The *empirical process* [33] on a sample $\mathrm{X}_1,\ldots,\mathrm{X}_n$ of i.i.d. d-dimensional random variables indexed by a class $\mathcal{F}$ of measurable functions $\varphi : \mathbb{R}^d \to \mathbb{R}$ is defined as the functional $\mathbb{G}_n$ mapping a function $\varphi \in \mathcal{F}$ to the rv

$$\mathbb{G}_n(\varphi) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \left(\varphi(\mathrm{X}_i) - \mathbb{E}\left[\varphi(\mathrm{X}_i)\right]\right)\,.$$

Lemma 3 invokes the pointwise measurable (PM) and Vapnik–Chervonenkis (VC)-type classes of functions. We refer the reader to [33] for the former and briefly define the latter, including the auxiliary Definition 1.

Definition 1 (Covering number [33])**.** Let $(\mathcal{V}, \|\cdot\|)$ be a vector space with a seminorm and let $\mathcal{F} \subset \mathcal{V}$. We define the ϵ-*covering number* of $\mathcal{F}$, denoted by $\mathcal{N}(\mathcal{F},\mathcal{V},\epsilon)$, as the minimum number of ϵ-balls of the form $\{\mathrm{x}\in\mathcal{V} : \|\mathrm{x}-\mathrm{y}\| < \epsilon\}$, where $\mathrm{y}\in\mathcal{V}$, needed to cover $\mathcal{F}$.

Definition 2 (VC-type class of functions [16])**.** Let $\mathcal{F}$ be a class of measurable functions $\varphi : \mathbb{R}^d \to \mathbb{R}$. Let Ψ be an *envelope* function for $\mathcal{F}$, i.e., $\Psi : \mathbb{R}^d \to \mathbb{R}$ measurable such that $\sup_{\varphi\in\mathcal{F}} |\varphi(\mathrm{x})| \leq \Psi(\mathrm{x})$ for all $\mathrm{x} \in \mathbb{R}^d$. An $\mathcal{F}$ class equipped with an envelope Ψ is called a VC-type class if there exist $A, \nu \in (0,\infty)$ such that, for all $\epsilon \in (0,1)$,

$$\sup_{\mathbb{Q}} \mathcal{N}(\mathcal{F}, \mathcal{L}^2(\mathbb{R}^d;\mathbb{Q}), \epsilon\|\Psi\|_{2,\mathbb{Q}}) \leq \left(\frac{A}{\epsilon}\right)^{\nu},$$

where the supremum is taken over all finitely discrete probability measures $\mathbb{Q}$ defined on $\mathbb{R}^d$ and $\|\Psi\|_{2,\mathbb{Q}} = (\int_{\mathbb{R}^d} |\Psi|^2 \, d\mathbb{Q})^{1/2}$ is the seminorm of $\mathcal{L}^2(\mathbb{R}^d;\mathbb{Q})$.

We will denote the Kolmogorov distance as $\rho_{\mathrm{cdf}}(X,Y) = \sup_{t\in\mathbb{R}} |F_X(t) - F_Y(t)|$, where F_X is the cdf of the rv X. Likewise, $X \overset{d}{=} Y$ will denote equality in distribution between the random variables.

Lemma 3 (Chernozhukov, Chetverikov, and Kato, 2014 [12, 13, 16])**.** *Consider a sample $\mathrm{X}_1, \ldots, \mathrm{X}_n$ of i.i.d. random variables. Let $\mathcal{F}$ be a PM and VC-type class of functions with constant envelope $b \in (0,\infty)$. Let $\sigma \in (0,\infty)$ be such that $\sup_{\varphi\in\mathcal{F}} \mathbb{E}\left[\varphi(\mathrm{X}_1)^2\right] \leq \sigma^2 \leq b^2$. Let $\mathbb{B}$ be a centred tight GP with sample paths on the space of bounded functions $\ell^\infty(\mathcal{F})$, and with covariance function*

$$\mathrm{Cov}(\mathbb{B}(\varphi_1), \mathbb{B}(\varphi_2)) = \mathbb{E}\left[\varphi_1(\mathrm{X}_1)\varphi_2(\mathrm{X}_1)\right] - \mathbb{E}\left[\varphi_1(\mathrm{X}_1)\right]\mathbb{E}\left[\varphi_2(\mathrm{X}_1)\right], \tag{17}$$

for $\varphi_1, \varphi_2 \in \mathcal{F}$. Then, there exists an rv $\mathbf{B} \overset{d}{=} \sup_{\varphi\in\mathcal{F}} |\mathbb{B}(\varphi)|$ such that, for all $\gamma \in (0,1)$ and n sufficiently large,

$$\mathbb{P}\left(\left|\sup_{\varphi\in\mathcal{F}} |\mathbb{G}_n(\varphi)| - \mathbf{B}\right| > A_1 \frac{b^{1/3}\sigma^{2/3}\log^{2/3} n}{\gamma^{1/3} n^{1/6}}\right) \leq A_2\gamma,$$

where $\mathbb{G}_n$ is based on $\mathrm{X}_1, \ldots, \mathrm{X}_n$, and A_1, A_2 are universal constants.

If we apply Lemma 3 to (15), we get the following result.

Theorem 5. *Let $\mathcal{D}$ denote any linear ℓ-th order differential operator. Let $K \in \mathcal{C}^\ell(\mathbb{R}^d)$ be a kernel with bounded ℓ-th derivatives. Further, suppose that the class*

$$\mathcal{K} = \left\{\mathrm{y} \in \mathbb{R}^d \mapsto \partial^{\boldsymbol{\beta}} K\left(\frac{\mathrm{x}-\mathrm{y}}{h}\right) : \mathrm{x} \in \Theta, h > 0, \boldsymbol{\beta} \in \mathbb{Z}_+^d, |\boldsymbol{\beta}| = \ell\right\} \tag{18}$$

is VC-type. Let $h \equiv h_n$ be a sequence with $h \in (0,1)$ and $h^{-(d+\ell)} = O(\log n)$. Moreover, let $\mathbb{B}$ be a GP with the same properties as in Lemma 3 and indexed by

$$\mathcal{F}_h = \left\{\mathrm{y} \in \mathbb{R}^d \mapsto \frac{1}{\sqrt{h^{d+\ell}}} \mathcal{D}K\left(\frac{\mathrm{x}-\mathrm{y}}{h}\right) : \mathrm{x} \in \Theta\right\}. \tag{19}$$

Then, there exists $\mathbf{B}_h \overset{d}{=} \sup_{\varphi\in\mathcal{F}_h} |\mathbb{B}(\varphi)|$ such that, for n sufficiently large,

$$\rho_{\mathrm{cdf}}\left(\sqrt{nh^{d+\ell}}\mathcal{E}_{n,h}[\mathcal{D}], \mathbf{B}_h\right) = O\left(\left[\frac{\log^7 n}{nh^{d+\ell}}\right]^{1/8}\right) \xrightarrow[n\to\infty]{} 0.$$

Moreover, if we fix $h \in (0,1)$ and define $\bar{\mathbf{B}}_h = \mathbf{B}_h/\sqrt{h^{d+\ell}}$, then $\sqrt{n}\ \mathcal{E}_{n,h}[\mathcal{D}]$ converges in probability to $\bar{\mathbf{B}}_h$, denoted $\sqrt{n}\ \mathcal{E}_{n,h}[\mathcal{D}] \xrightarrow{\mathbb{P}} \bar{\mathbf{B}}_h$, as $n \to \infty$.

A similar result establishes the asymptotic distribution for (16).

Theorem 6. *Let $\mathcal{D}$ denote any linear ℓ-th order differential operator. Let $K \in \mathcal{C}^\ell(\mathbb{R}^d)$ be a kernel with bounded ℓ-th derivatives. Further, suppose that the class*

$\mathcal{K}$ in (18) *is VC-type. Moreover, let $\mathbb{B}_{\mathfrak{X}_n}$ be a GP with the same properties as in Lemma 3, indexed by $\mathcal{F}_h$ as in* (19)*, and with covariance*

$$\mathrm{Cov}(\mathbb{B}_{\mathfrak{X}_n}(\varphi_1), \mathbb{B}_{\mathfrak{X}_n}(\varphi_2)) = \frac{1}{n}\sum_{i=1}^{n} \varphi_1(\mathrm{x}_i)\varphi_2(\mathrm{x}_i) - \frac{1}{n^2}\prod_{j=1}^{2}\left(\sum_{i=1}^{n}\varphi_j(\mathrm{x}_i)\right),$$

where x_i is the i-th observation in $\mathfrak{X}_n$. If $h \equiv h_n$ is a sequence with $h \in (0,1)$ and $h^{-(d+\ell)} = O(\log n)$, then there exists $\mathbf{B}_{n,h}\{\mathfrak{X}_n\} \overset{d}{=} \sup_{\varphi\in\mathcal{F}_h} |\mathbb{B}_{\mathfrak{X}_n}(\varphi)|$ such that, for n sufficiently large,

$$\rho_{\mathrm{cdf}}\left(\sqrt{nh^{d+\ell}}\mathcal{E}^*_{n,h}[\mathcal{D}|\mathfrak{X}_n], \mathbf{B}_{n,h}\{\mathfrak{X}_n\}\right) = O\left(\left[\frac{\log^7 n}{nh^{d+\ell}}\right]^{1/8}\right) \xrightarrow[n\to\infty]{} 0\,.$$

Theorem 6 holds for *any* observations $\mathfrak{X}_n$. The applicability of this theorem relies on the assumption that $\mathbf{B}_{n,h}\{\mathfrak{X}_n\} \rightsquigarrow \mathbf{B}_h$ a.s. This connection crystallises in the following result, which can be straightly derived from Theorem 5 and Theorem 6.

Theorem 7. *Let $\mathcal{D}$ denote any linear ℓ-th order differential operator. Let $K \in \mathcal{C}^\ell(\mathbb{R}^d)$ be a kernel with bounded ℓ-th derivatives. Further, suppose that the class $\mathcal{K}$ in* (18) *is VC-type. Let $h \equiv h_n$ be a sequence with $h \in (0,1)$ and $h^{-(d+\ell)} = O(\log n)$. Moreover, let us write $\Omega_{n,h}(\mathfrak{X}_n) \equiv \rho_{\mathrm{cdf}}\left(\mathbf{B}_{n,h}\{\mathfrak{X}_n\}, \mathbf{B}_h\right)$, where $\mathbf{B}_h$ and $\mathbf{B}_{n,h}\{\mathfrak{X}_n\}$ are as in Theorem 5 and Theorem 6, respectively. Let us allow $\mathfrak{X}_n$ to vary as a random sample from the pdf f underlying the covariance structure* (17) *of $\mathbf{B}_h$. Further, suppose that $\Omega_{n,h}(\mathfrak{X}_n) = o(1)$ a.s. under the previous hypotheses on h. Then, for n sufficiently large,*

$$\rho_{\mathrm{cdf}}\left(\sqrt{nh^{d+\ell}}\mathcal{E}^*_{n,h}[\mathcal{D}|\mathfrak{X}_n], \mathbf{B}_h\right) = O\left(\Omega_{n,h}(\mathfrak{X}_n) + \left[\frac{\log^7 n}{nh^{d+\ell}}\right]^{1/8}\right) \xrightarrow[n\to\infty]{a.s.} 0\,.$$

We can state sufficient conditions under which $\Omega_{n,h}(\mathfrak{X}_n)$ would converge to zero a.s. Corollary 1 gathers all the previous findings in an easy, ready-to-use form.

Corollary 1. *In the hypotheses of Theorem 7, if we further take a constant h and define $\bar{\mathbf{B}}_h = \mathbf{B}_h/\sqrt{h^{d+\ell}}$, then*

$$\begin{aligned}\rho_{\mathrm{cdf}}\left(\sqrt{n}\ \mathcal{E}_{n,h}[\mathcal{D}], \bar{\mathbf{B}}_h\right) &\xrightarrow[n\to\infty]{} 0\\ \rho_{\mathrm{cdf}}\left(\sqrt{n}\ \mathcal{E}^*_{n,h}[\mathcal{D}|\mathfrak{X}_n], \bar{\mathbf{B}}_h\right) &\xrightarrow[n\to\infty]{a.s.} 0\end{aligned}.$$

In particular, $\sqrt{n}\ \mathcal{E}^*_{n,h}[\mathcal{D}|\mathfrak{X}_n] \rightsquigarrow \bar{\mathbf{B}}_h$ a.s. Moreover, $\sqrt{n}\ \mathcal{E}_{n,h}[\mathcal{D}] \xrightarrow{\mathbb{P}} \bar{\mathbf{B}}_h$. Finally, $\bar{\mathbf{B}}_h$ has a continuous and strictly increasing cdf.

3.2.4. *Inference for curvature bumps.* The results from the previous section hold the key to ensuring (14) for curvature bumps.

Laplacian bumps. Theorem 8 straightly follows from Corollary 1 and Theorem 4.

Theorem 8. *Let us fix $h \in (0,1)$. Let $\mathcal{E}^*_{n,h}[\cdot|\mathfrak{X}_n]$ be as defined in* (16) *with KDE based on a kernel $K \in \mathcal{C}^2(\mathbb{R}^d)$ with bounded second derivatives. Taking $\ell = 2$, suppose that the class $\mathcal{K}$ in* (18) *is VC-type. For any $\alpha \in (0,1)$, define the margin $\tilde{\zeta}^{\alpha}_{n,h} = \mathcal{Q}_{1-\alpha}\{\mathcal{E}^*_{n,h}[\Delta|\mathfrak{X}_n]\}$. Then, for all $\alpha \in (0,1)$, the asymptotic validity of the inference scheme* (14) *holds a.s. for the smoothed version of the Laplacian bump* (8)*.*

Concave bumps & convex dips. Concave bumps and convex dips are more involved. To obtain a parallel result to Theorem 8, we will borrow the *Tail Value at Risk* (TVaR) concept from financial risk management [17]. The TVaR at level $p \in (0,1)$ of an rv X is defined as

$$\mathrm{TVaR}_p\{X\} = \frac{1}{1-p}\int_p^1 \mathcal{Q}_X(q)\, dq\,.$$

The TVaR is utilised to *aggregate* risks governed by an *unknown* dependence structure, for it satisfies $\mathrm{TVaR}_p\{X\} \geq \mathcal{Q}_p\{X\}$ and is sub-additive [17]. Contrary to quantiles, weak convergence does not guarantee TVaR convergence. Lemma 4 requires the random variables to be *asymptotically uniformly integrable* (a.u.i.) [32, p. 17].

Lemma 4. *Let $\{X_n\}_{n=1}^{\infty}$ be an a.u.i. sequence of random variables satisfying $X_n \rightsquigarrow X$ for some rv X with a strictly increasing cdf. Then, $\lim_{n\to\infty}\mathrm{TVaR}_p\{X_n\} = \mathrm{TVaR}_p\{X\}$ for all $p \in (0,1)$, being the limit finite.*

Then, Lemma 4 allows proving the main result.

Theorem 9. *Let us fix $h \in (0,1)$. Let $\mathcal{E}_{n,h}[\cdot]$ and $\mathcal{E}^*_{n,h}[\cdot|\mathfrak{X}_n]$ be as defined in* (15) *and* (16) *with KDE based on the same kernel $K \in \mathcal{C}^2(\mathbb{R}^d)$ with bounded second derivatives. Taking $\ell = 2$, suppose that the class $\mathcal{K}$ in* (18) *is VC-type. For any $\alpha \in (0,1)$, define the margin*

$$\tilde{\zeta}^{\alpha}_{n,h} = \sum_{i=1}^{d}\sum_{j=1}^{d}\mathrm{TVaR}_{1-\alpha}\left\{\mathcal{E}^*_{n,h}[\mathrm{D}_{ij}|\mathfrak{X}_n]\right\}\,,$$

where D_{ij} denotes second-order partial differentiation in the i and j variables. Moreover, let us assume the following:

(1) Letting $\bar{\mathbf{B}}_h[\mathrm{D}_{ij}]$ be the rv such that $\sqrt{n}\ \mathcal{E}_{n,h}[\mathrm{D}_{ij}] \xrightarrow{\mathbb{P}} \bar{\mathbf{B}}_h[\mathrm{D}_{ij}]$, the sum rv $\mathcal{Z} = \sum_{i=1}^{d}\sum_{j=1}^{d}\bar{\mathbf{B}}_h[\mathrm{D}_{ij}]$ has a continuous and strictly increasing cdf.

(2) For each pair (i,j), we have:

- *$\{\sqrt{n}\ \mathcal{E}_{n,h}[\mathrm{D}_{ij}]\}_{n=1}^{\infty}$ is a.u.i.*
- *$\{\sqrt{n}\ \mathcal{E}^*_{n,h}[\mathrm{D}_{ij}|\mathfrak{X}_n]\}_{n=1}^{\infty}$ is a.u.i. a.s.*

Then, for all $\alpha \in (0,1)$, the asymptotic validity of the inference scheme (14) *holds a.s. for the smoothed version of the concave bump* (5) *and the convex dip* (6).

The assumptions (1) and (2) seem natural. Hypothesis (1) asks a sum of non-negative random variables with continuous and strictly increasing cdfs to have a continuous and strictly increasing cdf too, which should be valid except in pathological cases. Similarly, knowing both sequences in hypothesis (2) converge weakly, being a.u.i. amounts to the convergence of their expectations [32, Theorem 2.20].

Gaussian bumps. A similar result to Theorem 9 holds for Gaussian bumps (9).

Theorem 10. *Consider the same hypotheses in Theorem 9 in the case $d = 2$. Assume a Gaussian kernel K. Further, assume that the true pdf f is bounded. Let C be a constant such that $C > (\pi h^4)^{-1}$. For any $\alpha \in (0,1)$, define the margin*

$$\tilde{\zeta}^{\alpha}_{n,h} = C\ \sum_{i=1}^{2}\sum_{j=1}^{2}\mathrm{TVaR}_{1-\alpha}\left\{\mathcal{E}^*_{n,h}[\mathrm{D}_{ij}|\mathfrak{X}_n]\right\}\,.$$

Then, for all $\alpha \in (0,1)$, the asymptotic validity of the inference scheme (14) *holds a.s. for the smoothed version of the Gaussian bump* (9).

4. Application

We will explore a *sports analytics* application for $d = 2$ in the *National Basketball Association* (NBA). See the SM [8] for additional applications with $d \in \{1, 3\}$ in two American leagues: the *National Football League* (NFL) and the *Major League Baseball* (MLB). Each player and team has its own style, a form of DNA. Following the biological analogy, if a single gene activates a trait in natural DNA, even minor bumps in data may reveal essential features.

All three sports applications are representative of the use of kernel methods for *exploratory data analysis* (EDA). Moreover, our proposal has a marked visual intent, thus excelling in low dimensions. In this context, the *curse of dimensionality* that harms kernel methods, demanding larger sample sizes to retain precision, becomes less relevant [6, Section 2.8].

Bivariate made shots in the NBA. Most people are familiar with basketball's *three-point line* (3PL), behind which a made shot earns not two but three points. Sports analytics have demonstrated that attempting more of these shots is well worth the risk, given the increased efficiency of three-point shooters. This trend has recently changed the basketball landscape, especially in the NBA.

Chacón exemplified univariate multimodality with shooting distances to the basket in the NBA [4]. We could see that the highest mode in a pdf model of all shots for the 2014-2015 season peaked beyond the 3PL. Looking at shots from a bivariate perspective will reveal the 3PL not as two separate modes but as a *ridge* [6].

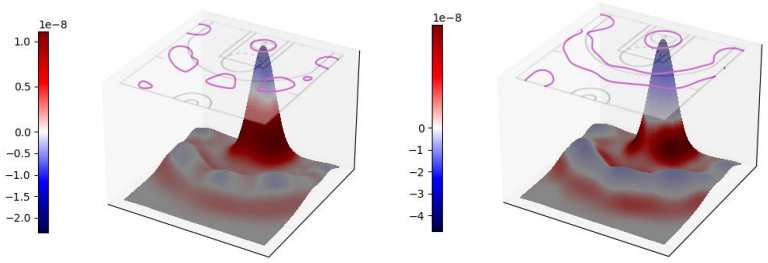
We will examine bumps from shot data by the three best scorers in the 2015-2016 NBA season: Stephen Curry, James Harden and Kevin Durant. Fig. 3 and Fig. 4 present different perspectives on concave and Laplacian bumps. Setting the near-the-rim shots aside, the three players have different shooting DNAs. Stephen Curry (Fig. 4a) operates beyond the 3PL, covering the entire ridge. He also demonstrates good range with even some half-court shots. However, he barely uses the mid-range area. His shooting patterns are mostly symmetrical. James Harden (Fig. 4b) has similar trends to Curry's. He almost covers the 3PL while leaning towards some mid-range areas without half-court shots. Some notable asymmetries are present. Kevin Durant (Fig. 4c) has a more balanced game between mid and long shots. He shoots facing the basket mainly, with lower usage of lateral shots.

Fig. 5 complements the previous figures with confidence sets. As refers to concave bumps, a wholly or partially ring-shaped area around the basket can be excluded with confidence for the three players. Apart from the shots near the rim, we cannot find other spots likely contained in the concave bumps. Regarding Laplacian bumps, the lower bound confidence sets become more relevant, even far apart from the rim. For Curry, up to four high-confidence spots appear beyond the 3PL, including the left-field corner; for Harden, the number of outside high-confidence spots decreases to two, while for Durant, there is only one.

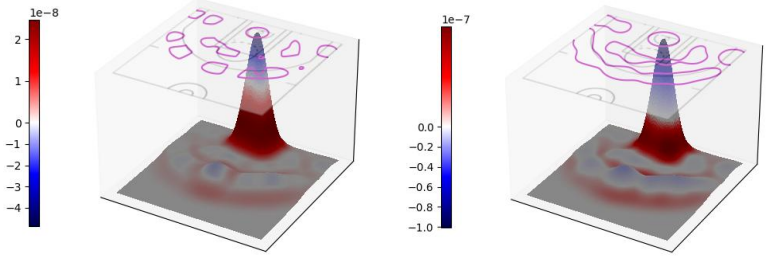
5. Discussion

Our curvature BH methodology represents the next step in density BH techniques, a path opened by Good and Gaskins [21] and consolidated with Hyndman [23] and Chaudhuri and Marron [9]. Rather than sticking to a purely probabilistic view on pdfs, our proposal thrives on sound geometry principles that have produced good results in applied areas like image processing [22].

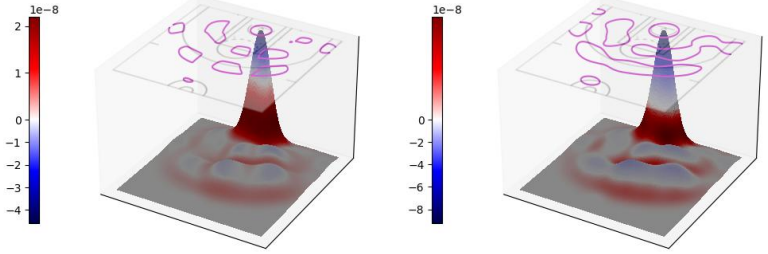
Our work strongly relies on KDDE, continuing the exploration of applications for higher-order partial derivatives of the pdf [5]. On the other hand, we bring



(A) Stephen Curry



(B) James Harden



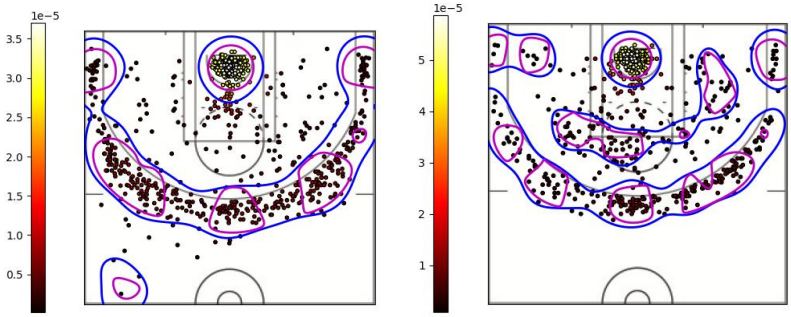
(C) Kevin Durant

FIG. 3. Concave and Laplacian bumps for Stephen Curry, James Harden and Kevin Durant. The three sub-figures have the same structure. On the left are concave bumps (5); on the right are Laplacian bumps (8). On either side, the two-dimensional surface is the fitted KDE pdf. The area colour refers to the curvature functional value at each point. The bump boundaries appear as lines on a flat basketball court at the top.

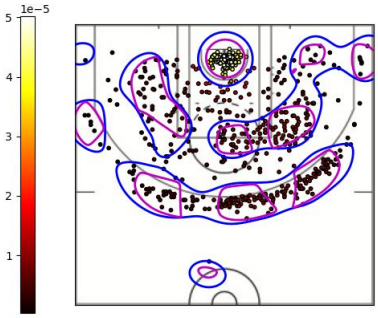
to curvature BH some cutting-edge techniques for level set estimation and inference that extend the pointwise-oriented initial works by Godtliebsen, Marron, and Chaudhuri [20] and Duong et al. [18].

The presented curvature framework shows great applicability from a theoretical standpoint. Under mild assumptions, the mean curvature, Laplacian and Gaussian bumps are consistent with affordable convergence rates. The confidence regions for Laplacian bumps are also asymptotically valid and consistent. The cases for Gaussian bumps (inference), concave bumps and convex dips (consistency and inference) are slightly more technical. Notwithstanding, pathological cases should not often appear in practice.

The NBA application shows promise for EDA and *clustering*. Fig. 4a presents a most pleasing result, identifying the 3PL area and the most relevant shooting



(A) Stephen Curry (B) James Harden



(C) Kevin Durant

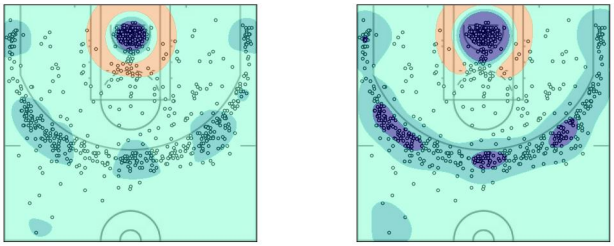
FIG. 4. Shot scatter data with concave and Laplacian bumps for Stephen Curry, James Harden and Kevin Durant. The three sub-figures have the same structure. Each point corresponds to a made shot location. The number of observations is 804, 710 and 698 for Stephen Curry, James Harden and Kevin Durant. The lines represent bump boundaries: magenta for concave bumps (5); blue, Laplacian bumps (8). The colour of the dots in the scatter plot conveys the value of the KDE pdf at each point.

spots. Both bumps are valuable and combine to produce insightful visualizations. Comparing the pictures in Fig. 4, we see that curvature bumps capture the players' rich shooting DNAs. Despite the ultimately unavoidable threat of the *curse of dimensionality* in KDE settings [6], the relatively small sample sizes did not detract from the accuracy of the results.

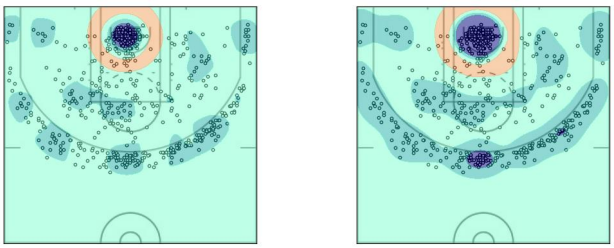
Our methodology's apparent least impressive achievement is confidence regions despite asymptotic guarantees. In Fig. 5, the upper-bound confidence sets tend to be conservative. This was not wholly unexpected, as Chen, Genovese, and Wasserman warned [14]. The margin is especially coarse for the concave bumps. In practice, we can mitigate this effect by splitting the bump and calculating the margin over smaller domains, employing a pilot estimation for guidance. Nonetheless, further research following Mammen and Polonik's universal approach [26] should yield even better results.

ACKNOWLEDGEMENTS

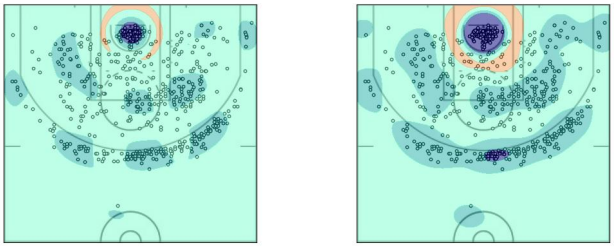
The first author's research has been supported by the MICINN grant PID2019-109387GB-I00 and the Junta de Extremadura grant GR21044. The second author would like to thank Amparo Baíllo Moreno for her advice as a doctoral counsellor at the Autonomous University of Madrid. Finally, we thank two anonymous reviewers for their helpful comments.



(A) Stephen Curry



(B) James Harden



(C) Kevin Durant

FIG. 5. Confidence sets for Stephen Curry, James Harden and Kevin Durant's bumps. The three sub-figures have the same structure. On the left, 90%-confidence sets for concave bumps (5); on the right, 90%-confidence sets for Laplacian bumps (8). The confidence margins are based on 200 bootstrap samples, each with the same resample size as the original one. On either side, the area colours convey the same meaning. The non-blue *sandy* areas fall *outside* the confidence set bounds; the blue-coloured areas lie *inside* the confidence region. The darkest blue corresponds to the lower bound confidence set: a set that is likely contained in the bump. The remaining blue areas cover the upper bound confidence set: a set that likely contains the bump. Finally, the mid-light blue colour points out the estimated bump.

REFERENCES

[1] ARIAS-CASTRO, E., MASON, D., and PELLETIER, B. (2016). On the Estimation of the Gradient Lines of a Density and the Consistency of the Mean-Shift Algorithm. *Journal of Machine Learning Research* **17**, 1–28.

[2] BAÍLLO, A., CUEVAS, A., and JUSTEL, A. (2000). Set estimation and nonparametric detection. *Canadian Journal of Statistics* **28**, 765–82.

[3] CHACÓN, J. E. (2015). A Population Background for Nonparametric Density-Based Clustering. *Statistical Science* **30**, 518–32.

[4] CHACÓN, J. E. (2020). The Modal Age of Statistics. *International Statistical Review* **88**, 122–41.

[5] CHACÓN, J. E. and DUONG, T. (2013). Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electronic Journal of Statistics* **7**, 499–532.

[6] CHACÓN, J. E. and DUONG, T. (2018). *Multivariate Kernel Smoothing and its Applications*. Chapman and Hall/CRC.

[7] Chacón, J. E., Duong, T., and Wand, M. P. (2011). Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica* **21**, 807–40.

[8] Chacón, J. E. and Fernández Serrano, J. (2023). *Supplementary material to "Bump hunting through density curvature features"*.

[9] Chaudhuri, P. and Marron, J. S. (1999). SiZer for Exploration of Structures in Curves. *Journal of the American Statistical Association* **94**, 807–23.

[10] Chaudhuri, P. and Marron, J. S. (2002). Curvature vs. Slope Inference for Features in Nonparametric Curve Estimates. *Unpublished manuscript*.

[11] Chen, Y.-C. (2022). Solution manifold and its statistical applications. *Electronic Journal of Statistics* **16**, 408–50.

[12] Chen, Y.-C., Genovese, C. R., Tibshirani, R. J., and Wasserman, L. (2016). Nonparametric modal regression. *The Annals of Statistics* **44**, 489–514.

[13] Chen, Y.-C., Genovese, C. R., and Wasserman, L. (2015). Asymptotic theory for density ridges. *The Annals of Statistics* **43**, 1896–928.

[14] Chen, Y.-C., Genovese, C. R., and Wasserman, L. (2017). Density Level Sets: Asymptotics, Inference, and Visualization. *Journal of the American Statistical Association* **112**, 1684–96.

[15] Cheng, M.-Y. and Hall, P. (1999). Mode testing in difficult cases. *The Annals of Statistics* **27**, 1294–315.

[16] Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics* **42**, 1564–97.

[17] Dhaene, J., Vanduffel, S., Goovaerts, M. J., Kaas, R., Tang, Q., and Vyncke, D. (2006). Risk Measures and Comonotonicity: A Review. *Stochastic Models* **22**, 573–606.

[18] Duong, T., Cowling, A., Koch, I., and Wand, M. P. (2008). Feature significance for multivariate kernel density estimation. *Computational Statistics & Data Analysis* **52**, 4225–42.

[19] Folland, G. B. (2002). *Advanced Calculus*. Pearson.

[20] Godtliebsen, F., Marron, J. S., and Chaudhuri, P. (2002). Significance in Scale Space for Bivariate Density Estimation. *Journal of Computational and Graphical Statistics* **11**, 1–21.

[21] Good, I. J. and Gaskins, R. A. (1980). Density Estimation and Bump-Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data. *Journal of the American Statistical Association* **75**, 42–56.

[22] Haralick, R. M. and Shapiro, L. G. (1992). *Computer and Robot Vision*. USA: Addison-Wesley Longman Publishing Co., Inc.

[23] Hyndman, R. J. (1996). Computing and Graphing Highest Density Regions. *The American Statistician* **50**, 120–6.

[24] Jiang, H. (2017). "Uniform Convergence Rates for Kernel Density Estimation". *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, 1694–703.

[25] Kato, T. (1995). *Perturbation Theory for Linear Operators*. Vol. 132. Springer Berlin Heidelberg.

[26] Mammen, E. and Polonik, W. (2013). Confidence regions for level sets. *Journal of Multivariate Analysis* **122**, 202–14.

[27] Marron, J. S. and Dryden, I. L. (2021). *Object Oriented Data Analysis*. Chapman and Hall/CRC.

[28] Nadaraya, E. A. (1989). *Nonparametric Estimation of Probability Densities and Regression Curves*. Kluwer Academic Publishers.

[29] Qiao, W. (2020). Asymptotics and optimal bandwidth for nonparametric estimation of density level sets. *Electronic Journal of Statistics* **14**, 302–44.

[30] Qiao, W. and Polonik, W. (2019). Nonparametric confidence regions for level sets: Statistical properties and geometry. *Electronic Journal of Statistics* **13**, 985–1030.

[31] Stuetzle, W. (2003). Estimating the Cluster Tree of a Density by Analyzing the Minimal Spanning Tree of a Sample. *Journal of Classification* **20**, 25–47.

[32] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

[33] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer New York.

Supplementary material

The following *appendices* are provided as *supplementary material* to the manuscript *Bump hunting through density curvature features*. Theorem and equation numbers refer to the manuscript. The numbers for new equations are prefixed. References are included at the end. Notation and acronyms are reused from the manuscript.

Appendix A contains proof of the results in the manuscript. Appendix B introduces and motivates the geometrical concepts behind the curvature bumps in Section 2. Appendix C illustrates the BH methodology for $d \in \{1, 3\}$ with data from the NFL and the MLB. Finally, Appendix D adds some comments on the underlying algorithms and data and acknowledges computational resources.

Appendix A. Proofs

A.1. **Consistency.**

Proof of Lemma 1. Since the partial derivatives of f and K are bounded and vanish at infinity, respectively, integration by parts and a change of variables yields

$$\begin{aligned}
\mathbb{E}[\partial^{\boldsymbol{\beta}} \hat{f}_{n,h}(\mathrm{x})] &= \int_{\mathbb{R}^d} \partial^{\boldsymbol{\beta}}[K_h(\cdot - \mathrm{y})](\mathrm{x})\ f(\mathrm{y})\ d\mathrm{y} \\
&= \int_{\mathbb{R}^d} (-1)^{|\boldsymbol{\beta}|}\ \partial^{\boldsymbol{\beta}}[K_h(\mathrm{x} - \cdot)](\mathrm{y})\ f(\mathrm{y})\ d\mathrm{y} \\
&= (-1)^{|\boldsymbol{\beta}|} \int_{\mathbb{R}^d} (-1)^{|\boldsymbol{\beta}|}\ K_h(\mathrm{x} - \mathrm{y})\ \partial^{\boldsymbol{\beta}} f(\mathrm{y})\ d\mathrm{y} \\
&= \frac{1}{h^d} \int_{\mathbb{R}^d} K\left(\frac{\mathrm{x} - \mathrm{y}}{h}\right)\ \partial^{\boldsymbol{\beta}} f(\mathrm{y})\ d\mathrm{y} \\
&= \int_{\mathbb{R}^d} \partial^{\boldsymbol{\beta}} f(\mathrm{x} - h\mathrm{z})\ K(\mathrm{z})\ d\mathrm{z}\,.
\end{aligned} \tag{S1}$$

Therefore,

$$\mathbb{E}[\partial^{\boldsymbol{\beta}} \hat{f}_{n,h}(\mathrm{x})] - \partial^{\boldsymbol{\beta}} f(\mathrm{x}) = \int_{\mathbb{R}^d} [\partial^{\boldsymbol{\beta}} f(\mathrm{x} - h\mathrm{z}) - \partial^{\boldsymbol{\beta}} f(\mathrm{x})]\ K(\mathrm{z})\ d\mathrm{z}\,. \tag{S2}$$

Now, depending on the differentiability of f, we have one of the following Taylor expansions, using the integral form of the remainder [10, Theorem 2.68]:

$$\partial^{\boldsymbol{\beta}} f(\mathrm{x} - h\mathrm{z}) - \partial^{\boldsymbol{\beta}} f(\mathrm{x}) = \begin{cases} -\nabla \partial^{\boldsymbol{\beta}} f(\mathrm{x})\mathrm{z}\ h + \mathcal{R}_2(\mathrm{x}, \mathrm{z}, h)\ h^2, & \text{if } r \geq 2 \\ \mathcal{R}_1(\mathrm{x}, \mathrm{z}, h)\ h, & \text{if } r = 1 \end{cases}, \tag{S3}$$

where

$$\mathcal{R}_s(\mathrm{x}, \mathrm{z}, h) = \begin{cases} \int_0^1 (1 - t)\ \mathrm{z}^\top \mathrm{D}^2 \partial^{\boldsymbol{\beta}} f(\mathrm{x} - th\mathrm{z})\mathrm{z}\ dt, & \text{if } s = 2 \\ -\int_0^1 \nabla \partial^{\boldsymbol{\beta}} f(\mathrm{x} - th\mathrm{z})\mathrm{z}\ dt, & \text{if } s = 1 \end{cases}. \tag{S4}$$

The z term vanishes after plugging (S3) into (S2) because of the zero-mean constraint on K, yielding the following explicit expression for the bias:

$$\mathbb{E}[\partial^{\boldsymbol{\beta}} \hat{f}_{n,h}(\mathrm{x})] - \partial^{\boldsymbol{\beta}} f(\mathrm{x}) = \left(\int_{\mathbb{R}^d} \mathcal{R}_s(\mathrm{x}, \mathrm{z}, h)\ K(\mathrm{z})\ d\mathrm{z} \right) h^s\,, \tag{S5}$$

where $s = \min\{r, 2\}$.

From (S5), there only remains to bound (S4) to finish the proof in the differentiable case. For $r \geq 2$, let $C_2 = \sup_{\mathrm{x} \in \mathbb{R}^d} \|\mathrm{vec}\ \mathrm{D}^2 \partial^{\boldsymbol{\beta}} f(\mathrm{x})\|_\infty$, where $\|\cdot\|_\infty$ denotes

the maximum absolute value of a vector's components. Likewise, for $r = 1$, let $C_1 = \sup_{\mathrm{x}\in\mathbb{R}^d} \|\nabla \partial^{\boldsymbol{\beta}} f(\mathrm{x})\|_\infty$. Then, it is straightforward to see

$$|\mathbb{E}[\partial^{\boldsymbol{\beta}} \hat{f}_{n,h}(\mathrm{x})] - \partial^{\boldsymbol{\beta}} f(\mathrm{x})| \leq \begin{cases} \frac{1}{2} C_2 \left(\sum_{i=1}^d \sum_{j=1}^d \int_{\mathbb{R}^d} |\mathrm{z}_i \mathrm{z}_j| \; K(\mathrm{z}) \; d\mathrm{z} \right) h^2, & \text{if } r \geq 2 \\ C_1 \left(\sum_{i=1}^d \int_{\mathbb{R}^d} |\mathrm{z}_i| \; K(\mathrm{z}) \; d\mathrm{z} \right) h, & \text{if } r = 1 \end{cases},$$

which yields the desired orders after considering the moment constraints on K.

In turn, when $r = 0$, we can resort to Hölder continuity, if $\alpha > 0$. Considering the $\|\cdot\|_1$ norm without loss of generality and letting $C > 0$ be the corresponding Hölder constant, from (S2) follows $|\mathbb{E}[\partial^{\boldsymbol{\beta}} \hat{f}_{n,h}(\mathrm{x})] - \partial^{\boldsymbol{\beta}} f(\mathrm{x})| \leq C \left(\int_{\mathbb{R}^d} \|\mathrm{z}\|_1^\alpha K(\mathrm{z}) \; d\mathrm{z} \right) h^\alpha$. Then, the integral is finite because of the moment constraints on K after applying Jensen's inequality with $\|\mathrm{z}\|_1^\alpha \mapsto (\|\mathrm{z}\|_1^\alpha)^{1/\alpha} = \|\mathrm{z}\|_1 = \sum_{i=1}^d |\mathrm{z}_i|$, which is convex for $\alpha \in (0, 1]$. This leads to a bound similar to the case $r = 1$, proving the order $O(h^\alpha)$.

Finally, if $\alpha = 0$, we apply uniform continuity. Branching from (S1) with a change of variables, we get $\mathbb{E}[\partial^{\boldsymbol{\beta}} \hat{f}_{n,h}(\mathrm{x})] = \int_{\mathbb{R}^d} \partial^{\boldsymbol{\beta}} f(\mathrm{x} - \mathrm{y}) \; K_h(\mathrm{y}) \; d\mathrm{y}$. Hence, letting $C > 0$ be an upper bound for $|\partial^{\boldsymbol{\beta}} f|$ on $\mathbb{R}^d$ and letting $\delta > 0$,

$$\begin{aligned} |\mathbb{E}[\partial^{\boldsymbol{\beta}} \hat{f}_{n,h}(\mathrm{x})] - \partial^{\boldsymbol{\beta}} f(\mathrm{x})| &\leq \int_{\mathbb{R}^d} |\partial^{\boldsymbol{\beta}} f(\mathrm{x} - \mathrm{y}) - \partial^{\boldsymbol{\beta}} f(\mathrm{x})| \; K_h(\mathrm{y}) \; d\mathrm{y} \\ &\leq \sup_{\|\mathrm{y}\| \leq \delta} |\partial^{\boldsymbol{\beta}} f(\mathrm{x} - \mathrm{y}) - \partial^{\boldsymbol{\beta}} f(\mathrm{x})| + 2C \int_{\|\mathrm{y}\| > \delta} K_h(\mathrm{y}) \; d\mathrm{y} \, . \end{aligned}$$

By uniform continuity, the supremum term will be small as long as δ is small regardless of x. Once fixed a sufficiently small δ, the integral term approaches zero as h does. In conclusion, the bias is $o(1)$ as $h \to 0$. ■

Proof of Theorem 3. The result follows after applying Theorem 2 with $\Psi = \phi[f]$ and $\tilde{\Psi} = \phi[\hat{f}_{n,h}]$. We must check that $\|\phi[\hat{f}_{n,h}] - \phi[f]\|_{\infty,k}$, $k \in \{0, 1\}$, can be made arbitrarily small.

From Theorem 1, uniform convergence for all the KDE's derivatives is ensured, except for a finite union of zero-probability events. Then, since partial derivatives of f are bounded, those of $\hat{f}_{n,h}$ are also eventually bounded a.s. Therefore, we can restrict the domain of the function φ to a compact set $\mathcal{U}$. This claim implies that both $\phi[p]$ and $\nabla\phi[p]$ are uniformly continuous functions of partial derivatives of p, which proves that $\|\phi[\hat{f}_{n,h}] - \phi[f]\|_{\infty,k}$, for $k \in \{0, 1\}$, can be made arbitrarily small. In particular, φ is Lipschitz continuous, which yields the desired convergence order for the Hausdorff distance. ■

Proof of Proposition 1. From [27, Wielandt-Hoffman theorem], any ordered eigenvalue function is Lipschitz continous. In particular, plug-in estimators of the eigenvalues are consistent in the uniform norm a.s. Therefore, using the triangle inequality, one arrives at

$$\begin{aligned} \inf_{\mathrm{x}\in\partial\mathcal{B}^\phi\oplus\delta} \{\lambda_i[\hat{f}_{n,h}](\mathrm{x}) - \lambda_{i+1}[\hat{f}_{n,h}](\mathrm{x})\} \geq & \inf_{\mathrm{x}\in\partial\mathcal{B}^\phi\oplus\delta} \{\lambda_i[f](\mathrm{x}) - \lambda_{i+1}[f](\mathrm{x})\} \\ & - \sup_{\mathrm{x}\in\mathbb{R}^d} |\lambda_i[f](\mathrm{x}) - \lambda_i[\hat{f}_{n,h}](\mathrm{x})| \\ & - \sup_{\mathrm{x}\in\mathbb{R}^d} |\lambda_{i+1}[f](\mathrm{x}) - \lambda_{i+1}[\hat{f}_{n,h}](\mathrm{x})| > 0 \, , \end{aligned}$$

a.s. for n sufficiently large, which proves (12) for $\hat{f}_{n,h}$. Let $\lambda_j : \mathbb{R}^{d^2} \to \mathbb{R}$ be the j-th eigenvalue function defined over d-dimensional square matrices. From [18, Theorem 5.16, p. 119], for every $\mathrm{x} \in \partial\mathcal{B}^\phi \oplus \delta$, λ_j is infinitely differentiable in some neighbourhoods of $\mathrm{D}^2 f(\mathrm{x})$ and $\mathrm{D}^2 \hat{f}_{n,h}(\mathrm{x})$, since all eigenvalues have multiplicity one in both cases. ■

A.2. **Inference.**

Proof of Theorem 4. It is an exercise to show that $\sup_{\mathrm{x}\in\Theta}|\phi[\hat{f}_{n,h}](\mathrm{x})-\phi[f_h](\mathrm{x})|\leq\tilde{\zeta}^{\alpha}_{n,h}$ implies $\underline{\mathcal{B}}^{\phi}_{n,h}(\tilde{\zeta}^{\alpha}_{n,h})\subset\mathcal{B}^{\phi}_{h}\subset\bar{\mathcal{B}}^{\phi}_{n,h}(\tilde{\zeta}^{\alpha}_{n,h})$. Therefore, for sufficiently large n,

$$\begin{aligned}
\mathbb{P}\left(\underline{\mathcal{B}}^{\phi}_{n,h}(\tilde{\zeta}^{\alpha}_{n,h})\subset\mathcal{B}^{\phi}_{h}\subset\bar{\mathcal{B}}^{\phi}_{n,h}(\tilde{\zeta}^{\alpha}_{n,h})\right)&\geq\mathbb{P}\left(\sup_{\mathrm{x}\in\Theta}|\phi[\hat{f}_{n,h}](\mathrm{x})-\phi[f_h](\mathrm{x})|\leq\tilde{\zeta}^{\alpha}_{n,h}\right)\\
&\geq\mathbb{P}\left(Z_{n,h}\leq\tilde{\zeta}^{\alpha}_{n,h}\right)\\
&=\mathbb{P}\left(Z_{n,h}\leq\zeta^{\alpha}_{n,h}+o(n^{-1/2})\right)\\
&\geq\mathbb{P}\left(Z_{n,h}\leq\mathcal{Q}_{1-\alpha}\{Z_{n,h}\}+o(n^{-1/2})\right)\\
&=\mathbb{P}\left(\sqrt{n}Z_{n,h}\leq\mathcal{Q}_{1-\alpha}\{\sqrt{n}Z_{n,h}\}+o(1)\right)\\
&=\mathbb{P}\left(\mathcal{Z}\leq\mathcal{Q}_{1-\alpha}\{\sqrt{n}Z_{n,h}\}+o(1)\right)+o(1)\\
&=1-\alpha+o(1)\,,
\end{aligned}$$

where we have used that the weak convergence to a continuous rv implies convergence in the Kolmogorov distance [23, Lemma 2.11], and $\mathcal{Q}_{1-\alpha}\{\sqrt{n}Z_{n,h}\}$ converges to $\mathcal{Q}_{1-\alpha}\{\mathcal{Z}\}$ for all $\alpha\in(0,1)$ [23, Lemma 21.2], since having $\mathcal{Z}$ a strictly increasing cdf implies its quantile function is continuous [23, p. 305]. Finally, $\lim_{n\to\infty}\tilde{\zeta}^{\alpha}_{n,h}=0$ follows from $\lim_{n\to\infty}\zeta^{\alpha}_{n,h}=0$ and $\tilde{\zeta}^{\alpha}_{n,h}=\zeta^{\alpha}_{n,h}+o(n^{-1/2})$. ■

Proof of Theorem 5. One can easily check that $\mathcal{D}f_h(\mathrm{x})=\mathbb{E}[\mathcal{D}K_h(\mathrm{x}-\mathrm{X}_i)]$, for all i. This leads to $\sqrt{nh^{d+\ell}}\mathcal{E}_{n,h}[\mathcal{D}]=\sup_{\varphi\in\mathcal{F}_h}|\mathbb{G}_n(\varphi)|$, for $\mathcal{F}_h$ as in (19). On the other hand, the class

$$\mathcal{F}'_h=\left\{\mathrm{y}\in\mathbb{R}^d\mapsto\mathcal{D}K\left(\frac{\mathrm{x}-\mathrm{y}}{h}\right):\mathrm{x}\in\Theta\right\}=\{\sqrt{h^{d+\ell}}\varphi:\varphi\in\mathcal{F}_h\}\,,$$

for any $h>0$, is VC-type by our assumption on the kernel class $\mathcal{K}$ in (18), since VC-type classes are closed under summation and product [5]. Also, $\mathcal{F}'_h$ is clearly PM, as the continuity of $\mathcal{D}K$ allows for a countable subset indexed by a rational d-tuple x. Therefore, we can apply Lemma 3 to $\mathcal{F}'_h$, yielding the existence of $\mathbf{B}'_h\stackrel{d}{=}\sup_{\varphi\in\mathcal{F}'_h}|\mathbb{B}(\varphi)|$ near $\sup_{\varphi\in\mathcal{F}'_h}|\mathbb{G}_n(\varphi)|$ with high probability. Now, by our boundedness assumption on the ℓ-th derivatives of K, letting $C>0$ be an envelope for $\mathcal{F}_h$, and assumming $h\in(0,1)$, the same C will be an envelope for $\mathcal{F}'_h$, yielding $\sup_{\varphi\in\mathcal{F}'_h}\mathbb{E}\left[\varphi(\mathrm{X}_1)^2\right]\leq h^{d+\ell}C^2\leq C^2$, whence we can take $b=C$ and $\sigma=C\sqrt{h^{d+\ell}}$ in Lemma 3. Dividing both sides of the inequality inside $\mathbb{P}$ in Lemma 3 by $\sqrt{h^{d+\ell}}$, we arrive at

$$\mathbb{P}\left(\left|\sup_{\varphi\in\mathcal{F}_h}|\mathbb{G}_n(\varphi)|-\frac{1}{\sqrt{h^{d+\ell}}}\mathbf{B}'_h\right|>A_1\frac{C\log^{2/3}n}{\gamma^{1/3}(nh^{d+\ell})^{1/6}}\right)\leq A_2\gamma\,,$$

for n sufficiently large. Letting $\mathbf{B}_h=\mathbf{B}'_h/\sqrt{h^{d+\ell}}$, $\tilde{A}_1=CA_1$, and bypassing the empirical process, we get

$$\mathbb{P}\left(\left|\sqrt{nh^{d+\ell}}\mathcal{E}_{n,h}[\mathcal{D}]-\mathbf{B}_h\right|>\tilde{A}_1\frac{\log^{2/3}n}{\gamma^{1/3}(nh^{d+\ell})^{1/6}}\right)\leq A_2\gamma\,.\tag{S6}$$

Note that $\mathbf{B}_h\stackrel{d}{=}\sup_{\varphi\in\mathcal{F}_h}|\mathbb{B}(\varphi)|$ [3, Supplementary material].

To go from (S6) to the desired result, we apply [4, Lemma 10 – supplementary material], which in turn derives from [5, Lemma 2.3]. As in [4], the assumptions (A1)-(A3) hold under our hypotheses. (A1) is a PM requirement, which again stands valid for $\mathcal{F}_h$. Then, (A2) is satisfied with the bound C, and we can use [5,

Lemma 2.1] to infer the pre-Gaussian requirement (A3) from (A2) and the fact that VC-type classes have finite *uniform entropy integral* [5, 23]. Therefore,

$$\sup_{t\geq 0}\left|\mathbb{P}\left(\sqrt{nh^{d+\ell}}\mathcal{E}_{n,h}[\mathcal{D}]<t\right)-\mathbb{P}\left(\mathbf{B}_h<t\right)\right|\leq \mu\mathbb{E}\left[\mathbf{B}_h\right]\left(\tilde{A}_1\frac{\log^{2/3}n}{\gamma^{1/3}(nh^{d+\ell})^{1/6}}\right)+A_2\gamma$$
$$\leq \mu'\tilde{A}_1\frac{\log^{7/6}n}{\gamma^{1/3}(nh^{d+\ell})^{1/6}}+A_2\gamma\,,$$

for positive constants μ,μ', where it has been used that $\mathbb{E}\left[\mathbf{B}_h\right]=O(\sqrt{\log n})$. The order of $\mathbb{E}\left[\mathbf{B}_h\right]$ is obtained noting that, by the coupling hypothesis between h and n, $\mathbb{E}\left[\mathbf{B}_h\right]=O(1/\sqrt{h^{d+\ell}})=O(\sqrt{\log n})$, where we have used that $\mathbb{E}\left[\mathbf{B}'_h\right]=O(1)$ uniformly in h due to Dudley's inequality [24, Corollary 2.2.8] applied to the VC-type class $\bigcup_{\delta>0}\mathcal{F}'_\delta$, which is larger than $\mathcal{F}'_h$. Differentiating the right-hand side of the last inequality with respect to γ yields the optimal order $O[\log^{7/8}n/(nh^{d+\ell})^{1/8}]$, which also happens to be that of γ.

Finally, plugging the optimal γ in (S6), given $\epsilon>0$, there exists $n_0\in\mathbb{N}$ such that, for $n\geq n_0$, we have $\mathbb{P}\left(\left|\sqrt{n}\;\mathcal{E}_{n,h}[\mathcal{D}]-\bar{\mathbf{B}}_h\right|>\epsilon\right)=O([n^{-1}\log^7 n]^{1/8})$, meaning $\sqrt{n}\;\mathcal{E}_{n,h}[\mathcal{D}]\xrightarrow{\mathbb{P}}\bar{\mathbf{B}}_h$. ■

Proof of Theorem 6. Let us start by decoupling the sample sizes in (16). Let us fix some observed values $\mathfrak{X}_n=\{\mathrm{x}_1,\ldots,\mathrm{x}_n\}$ and take $\mathrm{X}_1^*,\ldots,\mathrm{X}_m^*$ i.i.d. random variables from $\mathbb{P}_n^*\{\mathfrak{X}_n\}$, for m independent of n. Keeping n fixed, let us consider the decoupled version of (16) $\mathcal{E}^*_{m,h}[\mathcal{D}|\mathfrak{X}_n]=\sup_{\mathrm{x}\in\Theta}|\mathcal{D}\hat{f}^*_{m,h}(\mathrm{x}|\mathfrak{X}_n)-\mathcal{D}\hat{f}_{n,h}(\mathrm{x}|\mathfrak{X}_n)|$. Note that, for $i\in\{1,\ldots,m\}$, we have $\mathbb{E}\left[K_h(\mathrm{x}-\mathrm{X}_i^*)|\mathfrak{X}_n\right]=\hat{f}_{n,h}(\mathrm{x}|\mathfrak{X}_n)$. This leads to

$$\mathcal{D}\hat{f}^*_{m,h}(\mathrm{x}|\mathfrak{X}_n)-\mathcal{D}\hat{f}_{n,h}(\mathrm{x}|\mathfrak{X}_n)=\frac{1}{mh^{d+\ell}}\sum_{i=1}^{m}\mathcal{D}K\left(\frac{\mathrm{x}-\mathrm{X}_i^*}{h}\right)-\mathbb{E}\left[\mathcal{D}K\left(\frac{\mathrm{x}-\mathrm{X}_i^*}{h}\right)\middle|\mathfrak{X}_n\right],$$

and, subsequently, $\sqrt{mh^{d+\ell}}\mathcal{E}^*_{m,h}[\mathcal{D}|\mathfrak{X}_n]=\sup_{\varphi\in\mathcal{F}_h}|\mathbb{G}_m(\varphi)|$, for $\mathcal{F}_h$ as in (19). From this point on, one can apply exactly the same steps in the proof of Theorem 5 to obtain an approximation for $\mathcal{E}^*_{m,h}[\mathcal{D}|\mathfrak{X}_n]$ in the form of the supremum $\mathbf{B}_{n,h}\{\mathfrak{X}_n\}$ of a GP. The covariance structure for this $\mathbf{B}_{n,h}\{\mathfrak{X}_n\}$ immediately follows by considering X_1^* and $\mathbb{P}_n^*\{\mathfrak{X}_n\}$ in (17).

A question remains to be solved to finish the proof: undo the decoupling of m from n and the original sample $\mathfrak{X}_n$. We have proved that, given $\mathfrak{X}_n$, the Gaussian approximation works for sufficiently large m, but we have no guarantee the same would work if we took $m=n\to\infty$. A close look at Lemma 3 and Theorem 5 shows all the constants involved are universal or depend only on $\mathcal{F}_h$, which remains unchanged for every $\mathfrak{X}_n$. Therefore, we can take $m=n$ and conclude that the result holds for a sufficiently large n. ■

Proof of Corollary 1. It suffices to check that $\Omega_{n,h}(\mathfrak{X}_n)$ converges to zero. First, note that the GP $\mathbb{B}_{\mathfrak{X}_n}$ converges weakly to the GP $\mathbb{B}$ since the sample covariances converge to the population covariances (17) a.s. due to the law of large numbers. Therefore, applying the *continuous mapping theorem* [24, Theorem 1.3.6], we assure the suprema of the GPs also converge weakly. Let us assume for a moment that $\mathbf{B}_h$ has a continuous cdf. Then, the uniform convergence in distribution over all $t\in\mathbb{R}$ would follow using [23, Lemma 2.11], finishing the proof. Now, $\mathbf{B}_h$ has a continuous and strictly increasing cdf following [11, Corollary 1.3 + Remark 4.1] since we automatically rule out the case $\mathcal{K}$ in (18) is degenerate by the regularity assumptions on the kernel K. That is, we have $\mathcal{K}\neq\{\mathrm{y}\mapsto 0\}$. Therefore, $\mathbf{B}_h$

cannot be the Dirac distribution concentrated at zero. Meanwhile, the properties of $\bar{\mathbf{B}}_h$ derive from those of $\mathbf{B}_h$. ■

Proof of Lemma 4. From [6, Theorem 2.1], we have, for all $p \in (0,1)$,

$$\mathrm{TVaR}_p\{X_n\} = \mathcal{Q}_p\{X_n\} + \frac{1}{1-p}\mathbb{E}\left[(X_n - \mathcal{Q}_p\{X_n\})_+\right] ,$$

where $(\cdot)_+ \equiv \max\{\cdot, 0\}$. Now, since X has a strictly increasing cdf, its quantile function is continuous at every $p \in (0,1)$. Following [23, Lemma 21.2], we thus get $\lim_{n\to\infty} \mathcal{Q}_p\{X_n\} = \mathcal{Q}_p\{X\}$. On the other hand, using [23, Theorem 2.20] and Slutsky's [23, Lemma 2.8 (i)], it is not difficult to see that $\{X_n - \mathcal{Q}_p\{X_n\}\}_{n=1}^{\infty}$ is a.u.i. Next, noting that $(X_n - \mathcal{Q}_p\{X_n\})_+ \leq |X_n - \mathcal{Q}_p\{X_n\}|$ it easily follows that $\{(X_n - \mathcal{Q}_p\{X_n\})_+\}_{n=1}^{\infty}$ is a.u.i. too. Then, using again [23, Theorem 2.20], the latter implies $\lim_{n\to\infty} \mathbb{E}\left[(X_n - \mathcal{Q}_p\{X_n\})_+\right] = \mathbb{E}\left[(X - \mathcal{Q}_p\{X\})_+\right]$ after applying the *continuous mapping theorem* [23, Theorem 2.3] on $(\cdot)_+$. Finally, from the previous derivations and [6, Theorem 2.1], $\lim_{n\to\infty} \mathrm{TVaR}_p\{X_n\} = \mathrm{TVaR}_p\{X\}$. ■

Proof of Theorem 9. The eigenvalues are Lipschitz continuous as functions of matrix entries. As shown in [27, Equation 48.1, p. 104], the absolute difference between eigenvalues at distinct matrices is bounded by the Euclidean 2-norm [27, Equation 54.1, p. 57] of the difference between the corresponding matrices. Let $\lambda[\cdot]$ be either $\lambda_1[\cdot]$ or $\lambda_d[\cdot]$ in (4). Taking suprema at both sides of the inequality, and noting that $\|\cdot\|_2 \leq \|\cdot\|_1$, we get

$$\mathcal{S}_{n,h}[\lambda] \leq \sup_{\mathrm{x}\in\mathbb{R}^d} \sum_{i=1}^{d}\sum_{j=1}^{d} |\mathrm{D}_{ij}\hat{f}_{n,h}(\mathrm{x}) - \mathrm{D}_{ij}f_h(\mathrm{x})| \leq \sum_{i=1}^{d}\sum_{j=1}^{d} \mathcal{E}_{n,h}[\mathrm{D}_{ij}] \equiv Z_{n,h} .$$

Note that $\sqrt{n}\ Z_{n,h} \xrightarrow{\mathbb{P}} \mathcal{Z} = \sum_{i=1}^{d}\sum_{j=1}^{d} \bar{\mathbf{B}}_h[\mathrm{D}_{ij}]$ since the terms in the sum jointly converge [23, Theorem 2.7 (vi)], and then we can apply the *continuous mapping theorem* [23, Theorem 2.3]. This fulfils the pillar (I) in Theorem 4.

Now, $Z_{n,h}$ is the sum of random variables whose cdfs can be approximated via bootstrap. The obstacle here lies in the dependencies between the terms in $Z_{n,h}$; Corollary 1 only applies to the margin cdfs, not the joint cdf. Notwithstanding, we only need some $\zeta_{n,h}^{\alpha}$ satisfying $\zeta_{n,h}^{\alpha} \geq \mathcal{Q}_{1-\alpha}\{Z_{n,h}\}$. This $\zeta_{n,h}^{\alpha}$ can be obtained by resorting to the TVaR concept, which bounds from above the corresponding quantile at the same confidence level while sub-additive [6, Equation 37]. Interestingly, the TVaR is the lowest *concave distortion risk measure* satisfying both requirements [6, Theorem 5.2.2]. Therefore,

$$\mathcal{Q}_{1-\alpha}\{Z_{n,h}\} \leq \mathrm{TVaR}_{1-\alpha}\{Z_{n,h}\} \leq \sum_{i=1}^{d}\sum_{j=1}^{d} \mathrm{TVaR}_{1-\alpha}\{\mathcal{E}_{n,h}[\mathrm{D}_{ij}]\} \equiv \zeta_{n,h}^{\alpha} .$$

Note that $\zeta_{n,h}^{\alpha}$ converges to zero because, by Lemma 4 and assumption (2), for every pair (i,j), $\lim_{n\to\infty} \sqrt{n}\ \mathrm{TVaR}_{1-\alpha}\{\mathcal{E}_{n,h}[\mathrm{D}_{ij}]\} = \mathrm{TVaR}_{1-\alpha}\left\{\bar{\mathbf{B}}_h[\mathrm{D}_{ij}]\right\} < \infty$, where we have used Corollary 1 and $\sqrt{n}\ \mathrm{TVaR}_p\{X\} = \mathrm{TVaR}_p\{\sqrt{n}X\}$, for [6, Lemma 2.1]. This completes hypothesis (II) in Theorem 4.

Next, consider the bootstrap margin $\tilde{\zeta}_{n,h}^{\alpha}$. There only remains to verify (III) in Theorem 4, i.e., $|\tilde{\zeta}_{n,h}^{\alpha} - \zeta_{n,h}^{\alpha}| = o(n^{-1/2})$, to finish the proof. In turn, this follows using again [6, Lemma 2.1] after checking, for every pair (i,j),

$$\left|\mathrm{TVaR}_{1-\alpha}\left\{\sqrt{n}\ \mathcal{E}_{n,h}^{*}[\mathrm{D}_{ij}|\mathfrak{X}_n]\right\} - \mathrm{TVaR}_{1-\alpha}\left\{\sqrt{n}\ \mathcal{E}_{n,h}[\mathrm{D}_{ij}]\right\}\right| \xrightarrow[n\to\infty]{a.s.} 0 . \tag{S7}$$

Finally, (S7) is a direct consequence of Lemma 4 and the assumption (2), considering the two sequences have the same weak limit $\bar{\mathbf{B}}_h[\mathrm{D}_{ij}]$, by Corollary 1. ■

Proof of Theorem 10. First, it is an exercise to check that, for the Gaussian kernel, $\|K_h\|_{\infty,2} = (2\pi h^4)^{-1}$ on $\mathbb{R}^2$ and, thus, $\|K_h\|_{\infty,2} < C/2$ on $\mathbb{R}^2$. Then, one can quickly check that $\|f_h\|_{\infty,2} < C/2$ on Θ, and, from Lemma 2, $\|\hat{f}_{n,h}\|_{\infty,2} < C/2$ on Θ for sufficiently large n a.s. Therefore, we can use [17, Theorem 2.12, p. 768] along with [27, Equation 54.5, p. 57] and the fact that $\|\cdot\|_2 \leq \|\cdot\|_1$ to obtain

$$\begin{aligned}\mathcal{S}_{n,h}[\det(\mathrm{D}^2\cdot)] &\leq 2\max\{\|f_h\|_{\infty,2}, \|\hat{f}_{n,h}\|_{\infty,2}\} \sup_{\mathrm{x}\in\Theta} \sum_{i=1}^{2}\sum_{j=1}^{2} |\mathrm{D}_{ij}\hat{f}_{n,h}(\mathrm{x}) - \mathrm{D}_{ij}f_h(\mathrm{x})| \\ &\leq C \sum_{i=1}^{2}\sum_{j=1}^{2} \mathcal{E}_{n,h}[\mathrm{D}_{ij}] \equiv Z_{n,h}\,,\end{aligned}$$

a.s. for sufficiently large n. From this point on, the proof follows the same steps as that of Theorem 9. ■

Appendix B. Curvature

The rationale behind Section 2 lies in classical geometry. We refer the reader to do Carmo's books for a comprehensive introduction to differential and Riemannian geometry [7, 8] and Grinfeld's for an operational perspective on tensor calculus [13]. Then, Petersen's book has a complete chapter devoted to the central topic of hypersurfaces [22, Chapter 4]. The appendix of Ecker's book also includes some key concepts [9]. Finally, we recommend Gray, Abbena, and Salamon's book to build up some intuition in $\mathbb{R}^3$ [12, Chapter 13].

B.1. **Outline.** After recalling some basic notions, we will address the concept of principal curvature. Concavity and convexity are defined through specific sign configurations among the principal curvatures. On the other hand, the *mean curvature* and the *Gaussian curvature* are alternative scalar summaries of principal curvatures. All these measures involve first and second-order derivatives of the pdf. Theoretical considerations below will discard the need for higher-order derivatives beyond those. Meanwhile, combining derivatives of different orders is a major inconvenience in a KDE setting. Hence, we present similarly-intended features that rely only on the Hessian matrix.

B.2. **Fundamental concepts.** Let $f : \mathbb{R}^d \to [0,\infty)$ be a d-dimensional pdf. Let us assume that f is as smooth as needed. The graph of f, that is, $\mathcal{G} = \{(\mathrm{x}, f(\mathrm{x})) \in \mathbb{R}^{d+1} : \mathrm{x} \in \mathbb{R}^d\}$, is a hypersurface embedded in $\mathbb{R}^{d+1}$ and a d-dimensional smooth manifold, in particular. Let $\mathcal{X} : \mathcal{G} \to \mathbb{R}^d$ be the global coordinates chart of $\mathcal{G}$ mapping every $(\mathrm{x}, f(\mathrm{x})) \in \mathcal{G}$ to x. Let also $\mathcal{F} : \mathbb{R}^d \to \mathcal{G}$ be the global parameterization of $\mathcal{G}$, defined as $\mathcal{F}(\mathrm{x}) = \mathcal{X}^{-1}(\mathrm{x}) = (\mathrm{x}, f(\mathrm{x}))$.

At every $p \in \mathcal{G}$, we can define a d-dimensional tangent vector space $T_p\mathcal{G}$ comprising different ways of deriving smooth functions from $\mathcal{G}$ to $\mathbb{R}$ in a neighbourhood of p [8]. In particular, the i-th element of the canonical basis determined by the chart is the i-th partial derivative $\partial_i|_p$ acting on functions $\varphi : \mathcal{G} \to \mathbb{R}$ as $\partial_i|_p(\varphi) = [\mathrm{D}_i(\varphi \circ \mathcal{F}) \circ \mathcal{X}](p)$, where D_i takes the i-th partial derivative of $\varphi \circ \mathcal{F} : \mathbb{R}^d \to \mathbb{R}$.

The union of all tangent spaces across $p \in \mathcal{G}$ is also a smooth manifold, known as the tangent bundle $T\mathcal{G}$ [8]. Elements in $T\mathcal{G}$ are called vector fields. A vector field can be thought of as a function mapping each point $p \in \mathcal{G}$ to a tangent vector in $T_p\mathcal{G}$. Every vector field X can be expressed, using Einstein's summation convention [13], as $X|_\cdot = X^i(\cdot)\partial_i|_\cdot$, where every X^i is a smooth function $\mathcal{G} \to \mathbb{R}$.

Let us now build on the previous concepts considering $\mathcal{G}$ is a hypersurface.

B.2.1. *Tangent vector fields.* Being $\mathcal{G}$ immersed in $\mathbb{R}^{d+1}$, there exists a canonical identification of basis vector fields ∂_i with real functions $\mathbb{R}^d \to \mathbb{R}^{d+1}$ [9]. To see this, consider the canonical identity immersion $\iota : \mathcal{G} \to \mathbb{R}^{d+1}$. The differential $d\iota$ transforms vector fields in $T\mathcal{G}$ into vector fields in $\mathbb{R}^{d+1}$. Given a smooth function $\varphi : \mathbb{R}^{d+1} \to \mathbb{R}$, we have $[d\iota(\partial_i)](\varphi) = \mathrm{D}_i(\varphi \circ \mathcal{F}^\iota)$, where $\mathcal{F}^\iota \equiv \iota \circ \mathcal{F} : \mathbb{R}^d \to \mathbb{R}^{d+1}$ is the immersed parameterization of $\mathcal{G}$. Now, using the chain rule, for any $\mathrm{x} \in \mathbb{R}^d$, $\mathrm{D}_i(\varphi \circ \mathcal{F}^\iota)(\mathrm{x}) = \langle \nabla\varphi(\mathrm{x}, f(\mathrm{x})), \mathrm{D}_i\mathcal{F}^\iota(\mathrm{x})\rangle$, where $\mathrm{D}_i\mathcal{F}^\iota$ is the i-th column vector field in the Jacobian matrix of $\mathcal{F}^\iota$. Combining the last two equations, we see that $d\iota(\partial_i)$ *derives* the function φ in the *direction* of the vector field $\mathrm{D}_i\mathcal{F}^\iota$. Therefore, there is a canonical isomorphism

$$d\iota(\partial_i) \cong \mathrm{D}_i\mathcal{F}^\iota = \mathbf{e}_i + \mathrm{D}_i f \ \mathbf{e}_{d+1}\,, \tag{S8}$$

where $\mathbf{e}_i$ denotes the i-th canonical basis vector in the Euclidean space $\mathbb{R}^{d+1}$.

The identification (S8) unlocks both intrinsic and extrinsic properties of $\mathcal{G}$. On the one hand, we have an inner product in $\mathbb{R}^{d+1}$ that can be brought into $T\mathcal{G}$ via (S8). On the other hand, we have d tangent vectors in $T\mathcal{G}$ and $d+1$ linearly independent vector fields in $\mathbb{R}^{d+1}$, leaving space for an orthogonal complement *normal* to the hypersurface.

B.2.2. *Normal vector field.* At any given $\mathrm{x} \in \mathbb{R}^d$, the vectors $\mathrm{D}_i\mathcal{F}^\iota(\mathrm{x})$ form a basis for the d-dimensional tangent hyperplane $\pi_\mathrm{x} \subset \mathbb{R}^{d+1}$ at $\mathcal{F}(\mathrm{x})$. Its orthogonal complement $\pi_\mathrm{x}^\perp$ is a straight line passing through $\mathcal{F}(\mathrm{x})$. One can check that

$$\check{N}(\mathrm{x}) = \frac{(\nabla f(\mathrm{x}), -1)}{\sqrt{1 + \|\nabla f(\mathrm{x})\|^2}} \tag{S9}$$

corresponds to the downward unit normal vector to π_x, perpendicular to vectors (S8), thus generating $\pi_\mathrm{x}^\perp$ [9, Equation A.2].

Studying how the normal vector field (S9) changes along all directions over $\mathcal{G}$ provides a means to characterize curvature.

B.2.3. *First fundamental form.* Utilizing (S8), the Euclidean inner product in $\mathbb{R}^{d+1}$ induces a metric field g on $\mathcal{G}$ with components [9]

$$g_{ij} \equiv g(\partial_i, \partial_j) \coloneqq \langle \mathrm{D}_i\mathcal{F}^\iota \circ \mathcal{X}, \mathrm{D}_j\mathcal{F}^\iota \circ \mathcal{X}\rangle = \delta_{ij} + (\mathrm{D}_i f \circ \mathcal{X})(\mathrm{D}_j f \circ \mathcal{X})\,, \tag{S10}$$

where δ_{ij} denotes the Kronecker delta twice covariant tensor. The metric tensor g is known as the *first fundamental form*. It allows measuring angles and lengths on $T\mathcal{G}$ *intrinsically*. In particular, it defines at each point a norm $\|X\| = \sqrt{g(X,X)}$, for $X \in T\mathcal{G}$.

The metric tensor has an inverse g^{ij} satisfying $\delta^i_k = g^{ij}g_{jk}$, where δ^i_k is the Kronecker delta $(1,1)$-tensor. From the previous equation and (S10), using the Sherman–Morrison formula [15, Corollary 18.2.11], one can check that

$$g^{ij} = \delta_{ij} - \frac{(\mathrm{D}_i f \circ \mathcal{X})(\mathrm{D}_j f \circ \mathcal{X})}{1 + \|\nabla f \circ \mathcal{X}\|^2}\,.$$

A simple expression for the metric determinant follows from [15, Corollary 18.1.3] when applied to (S10), yielding $\det(g) = 1 + \|\nabla f \circ \mathcal{X}\|^2$.

In our context, compatibility with the metric tensor leads to a *natural* definition of derivatives for tangent vector fields over $\mathcal{G}$.

B.2.4. *Field divergence.* The metric tensor allows defining *covariant derivatives* over vector fields through the *Levi-Civita connection* $\nabla^\mathcal{G}$ [8], with components given by the Christoffel symbols Γ^k_{ij} [13, Equation 5.66]. Letting $X, Y \in T\mathcal{G}$, the covariant derivative of $\nabla^\mathcal{G}_Y X$ is a vector field in $T\mathcal{G}$. See [13, Equation 8.9] for an explicit expression of $\nabla^\mathcal{G}_i X \equiv \nabla^\mathcal{G}_{\partial_i} X$ involving the Christoffel symbols, the components of X and the basis vector fields.

The covariant derivative then can be used to define the *divergence* operator [13, Equation 8.2] on $X \in T\mathcal{G}$ as $\mathrm{div}_{\mathcal{G}}(X) = (\nabla^{\mathcal{G}}_i X)^i$, which is a function $\mathcal{G} \to \mathbb{R}$. As it turns out, the divergence can be extended to *extrinsic* vector fields $\vec{V} : \mathbb{R}^d \to \mathbb{R}^{d+1}$ living in the ambient space through [9]

$$\mathrm{div}^{\nmid}_{\mathcal{G}}(\vec{V}) = g^{ij} \langle \mathrm{D}_i \vec{V} \circ \mathcal{X}, \mathrm{D}_j \mathcal{F}^\iota \circ \mathcal{X} \rangle \,, \tag{S11}$$

meaning $\mathrm{div}^{\nmid}_{\mathcal{G}}(\cdot)$ coincides with $\mathrm{div}_{\mathcal{G}}(\cdot)$ over tangent vector fields, i.e., for all j, $\mathrm{div}^{\nmid}_{\mathcal{G}}(\mathrm{D}_j \mathcal{F}^\iota) = \mathrm{div}_{\mathcal{G}}(\partial_j)$. To see this, note that $\mathrm{D}_i \mathrm{D}_j \mathcal{F}^\iota = (\Gamma^k_{ij} \circ \mathcal{F})\ \mathrm{D}_k \mathcal{F}^\iota + N$, where N is a normal vector field [9].

The divergence operator (S11) appears in an alternative expression for the mean curvature defined below.

B.2.5. *Second fundamental form.* Similarly to how the first fundamental form (S10) measures change intrinsically, we can define another twice covariant tensor measuring changes in the normal vector field along different tangent vector fields. This new tensor A is the *second fundamental form* [9]. It has components

$$A_{ij} \equiv A(\partial_i, \partial_j) \coloneqq \langle \mathrm{D}_i \check{N} \circ \mathcal{X}, \mathrm{D}_j \mathcal{F}^\iota \circ \mathcal{X} \rangle = \frac{\mathrm{D}_{ij} f \circ \mathcal{X}}{\sqrt{1 + \|\nabla f \circ \mathcal{X}\|^2}} \,, \tag{S12}$$

where D_{ij} denotes second-order partial differentiation in the i and j variables [9]. The second fundamental form is *extrinsically* defined since it contains the normal vector, which *lives* outside the tangent space.

B.2.6. *Shape operator.* Having introduced the first and second fundamental forms, the shape operator $S : T\mathcal{G} \to T\mathcal{G}$ connects the two via $g(S(\partial_i), \partial_j) = A_{ij}$ [22]. One can easily verify that the shape operator has components $S^i_k = g^{ij} A_{jk}$ [9].

Since A is symmetric, i.e., $A_{ij} = A_{ji}$, one can check that S is a self-adjoint operator, meaning $g(S(X), Y) = g(X, S(Y))$, for $X, Y \in T\mathcal{G}$. This implies that, at any point $p \in \mathcal{G}$, all the eigenvalues of S_p are real and there exists an associated orthonormal basis consisting of eigenvectors of S_p [22, Section 2.4.6]. These eigenvalues correspond to some curvature measure along the direction of the corresponding eigenvectors.

The shape operator is easily confused with the second fundamental form. Petersen refers to this issue as a matter of perspective [22, Section 2.3.1]. Ecker even employs the same letter A instead of S to refer to the shape operator, relying on the position of the indices to distinguish between them (S is a $(1,1)$-tensor, while A is $(2,0)$). Meanwhile, Gray, Abbena, and Salamon equivalently define A from S and g [12, Definition 13.28].

The *curvature tensor*, pervasive in Riemannian geometry, takes a simple form for hypersurfaces, exclusively involving g and S [22, Section 4.2]. This virtually ensures that curvature can be apprehended by inspecting only the first and second derivatives of the pdf f.

B.2.7. *Normal and principal curvatures.* The definition of the *normal curvature* [7, 12] of $\mathcal{G}$ in the non-null direction $X = X^i \partial_i \in T\mathcal{G}$ also applies to hypersurfaces:

$$\kappa(X) = \frac{g(S(X), X)}{g(X, X)} = \frac{A_{ij} X^i X^j}{g_{ij} X^i X^j} \,, \tag{S13}$$

where the denominator ensures $\kappa(\lambda X) = \kappa(X)$, for all $\lambda \neq 0$.

Using the Cauchy–Schwarz inequality, for all non-null $X \in T\mathcal{G}$, we have at any point $|\kappa(X)| \leq \|S(\hat{X})\|$, where $\hat{X} = X/\|X\|$. Moreover, the equality holds if and only if (iff) X is an eigenvector of S. In that case, $\kappa(X)$ is equal to the respective eigenvalue of X. Given an orthonormal basis of eigenvectors X_i of S, we say that X_i is the i-th *principal direction* of $\mathcal{G}$ and its corresponding eigenvalue κ_i is known

as the i-th *principal curvature* [8]. The *minimax principle* establishes a variational connection between principal curvatures and the quotient (S13) [18, Section 6.10]. In particular, for any non-null $X \in T\mathcal{G}$, $\min_i \kappa_i \leq \kappa(X) \leq \max_i \kappa_i$.

The principal curvatures can be summarized into a single value according to several criteria, giving rise to the mean and Gaussian curvatures.

B.2.8. *Mean curvature.* At any point in $\mathcal{G}$, the mean curvature $\mathfrak{D}$ is defined as the sum of all principal curvatures, that is, the trace of the diagonalized matrix of S [9]. Since the trace is a matrix invariant, we have $\mathfrak{D} = \sum_{i=1}^{d} \kappa_i = \operatorname{tr}(S) = S_i^i$. Equivalently, using (S11) we get (S14), while extra calculations yield (S15):

$$\mathfrak{D} = \operatorname{div}_{\mathcal{G}}^{\sharp}(\check{N})\,, \qquad \text{(S14)} \qquad\qquad \mathfrak{D} = \operatorname{div}_{\mathbb{R}^d}\left(\bar{\nabla} f\right) \circ \mathcal{X}\,, \qquad \text{(S15)}$$

where $\bar{\nabla} f = \nabla f / \sqrt{1 + \|\nabla f\|^2}$ [9, Equation A.3]. Both expressions convey the same idea: $\mathfrak{D}$ is the divergence of a *normalized* vector field. The mean curvature will take negative values when $\check{N}$ and $\bar{\nabla} f$ locally converge towards a point.

B.2.9. *Gaussian curvature.* The shape operator contains another way of summarizing principal curvatures, this time by its determinant:

$$\mathfrak{B} \coloneqq \prod_{i=1}^{d} \kappa_i = \det(S) = \frac{\det(A)}{\det(g)} = \frac{\det(\mathrm{D}^2 f)}{(1 + \|\nabla f\|^2)^{1+\frac{d}{2}}} \circ \mathcal{X}\,. \qquad \text{(S16)}$$

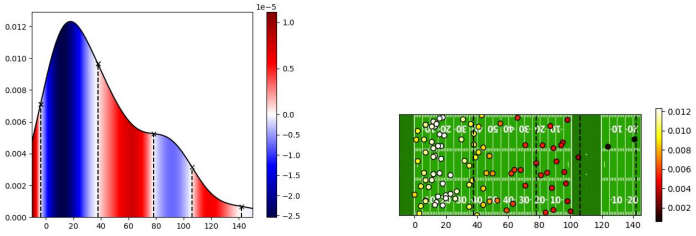
The Gaussian curvature $\mathfrak{B}$ owns outstanding geometrical and topological properties. Gauss' *Theorema Egregium* states that $\mathfrak{B}$ and $|\mathfrak{B}|$ are intrinsic invariants of a hypersurface, respectively, if d is even or odd [22, Lemma 3.1, p. 96]. Interestingly, when $d = 2$, $\mathfrak{B}$ is precisely half the *scalar curvature*, the scalar contraction of the curvature tensor (see [22, Section 2.2.5] and [22, Proposition 2.1, p. 92]).

B.2.10. *Hessian matrix.* The Hessian matrix is present in the previous curvature concepts through the second fundamental form (S12). The Hessian $\mathrm{D}^2 f$ corresponds to the non-normalized version of A, replacing $\check{N}$ by $N = (\nabla f, -1)$ in (S12). The eigenvalues of $\mathrm{D}^2 f$ are scalar multiples of those of A. Moreover, expressing the shape operator in matrix form as $S = G^{-1}A$, and using Sylvester's law of inertia [15, Exercise 12 (c), p. 275], we can infer that S coincides in *signature* with A and, consequently, with $\mathrm{D}^2 f$. On the other hand, it is well-known that the eigenvalues of the Hessian $\mathrm{D}^2 f$ relate to the concavity and convexity of f. Namely, if all the eigenvalues at a point are negative (alternatively, positive), i.e., $\mathrm{D}^2 f$ is negative (positive) definite, then f is locally strictly concave (convex) around that point. Hence, the shape operator also determines local concavity and convexity because of the previous argument [22, p. 91].

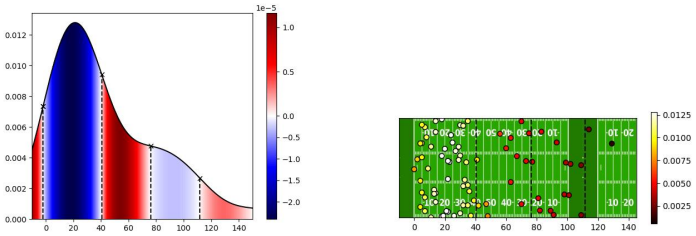
The shape operator and the Hessian are equal iff G is the identity matrix, or, equivalently, iff $\|\nabla f\| = 0$. The connection between Gaussian curvature and the Hessian determinant is evident from (S16). Furthermore, the trace of the Hessian is equal to the divergence of the gradient, meaning $\operatorname{tr}(\mathrm{D}^2 f) = \operatorname{div}_{\mathbb{R}^d}(\nabla f)$. Therefore, attending to (S15), the mean curvature differs from the Hessian trace by a normalizing term acting on the gradient argument (see [10, Equation 5.28] for an explicit link between the two). In conclusion, the essence of curvature lies in the Hessian $\mathrm{D}^2 f$.

Appendix C. Extra applications

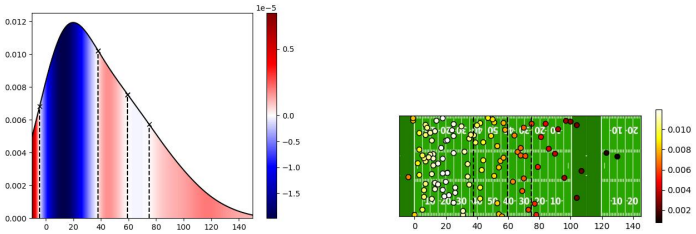
C.1. **Univariate receiving yards in the NFL.** American football is inherently one-dimensional: only distances projected on the sidelines matter, despite the game



(A) 2014



(B) 2016



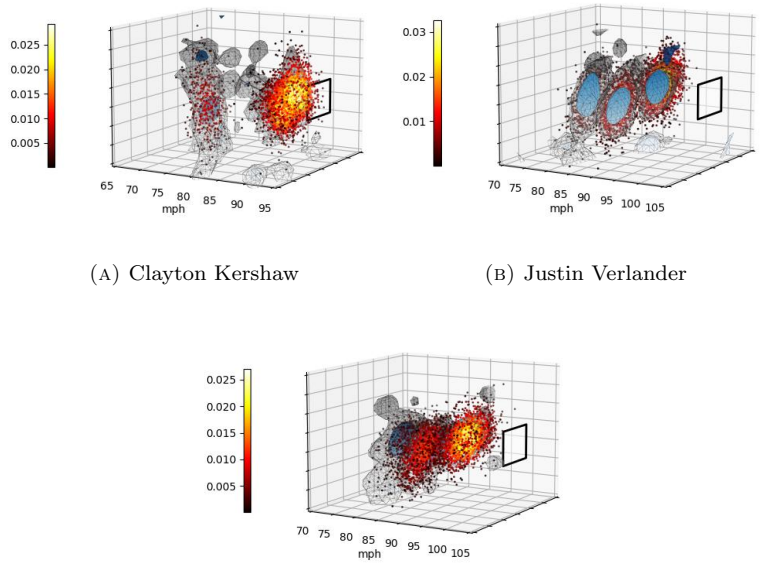
(C) 2018

FIG. 6. New England Patriots' receiving yards in the 2014, 2016 and 2018 seasons. The three sub-figures have the same structure. On the left is the underlying KDE pdf graph; on the right is a one-dimensional scatter plot of the original observations with random jitter for better resolution. Each point in the scatter plot corresponds to the receiving yards from Tom Brady by some player at some game. The number of observations is 98, 82 and 111 for the 2014, 2016 and 2018 seasons. The coloured areas under the pdf's curve represent the sign of the second derivative: red means convex; blue, concave. The colour of the dots in the scatter plot conveys the value of the KDE pdf at each point. Inflection points show up as vertical dotted lines on both sides.

taking place on a rectangular field. Hence, we propose a case study about the NFL to illustrate our methods on univariate data.

The New England Patriots from the NFL experienced noticeable changes in their passing DNA while winning the *Super Bowl* in 2014, 2016 and 2018. We will analyze the bumps in their receiving profiles in those three championship seasons. Fig. 6 shows the original data and the estimated curvature bumps. All three sub-figures have a primary bump between 0 and 40 yards. In 2014, this bump is just slightly narrower than in 2016 and 2018. What differentiates all three seasons is the existence or absence of a secondary bump. In 2014, the second bump ranges between 80 and 100 yards and is perfectly visible. Two years later, this bump still exists but has become smoother. Finally, in 2018, the bump has almost been wholly *ironed*. This evolution reflects the transition from a team with a few go-to players to a more numerous receiver squad sharing responsibilities.

C.2. **Trivariate pitches in MLB.** The pitching event lives at least in two dimensions. The batter needs to anticipate the arrival coordinates of the ball. KDE



(A) Clayton Kershaw

(B) Justin Verlander

(C) Max Scherzer

FIG. 7. Pitching scatter data for Clayton Kershaw, Justin Verlander and Max Scherzer with 2680, 3839 and 3284 pitches, respectively. Concave bumps (5) and Laplacian bumps (8) show in blue and grey, respectively. Speeds are reported in mph. The arrival coordinates at the home plate for each speed value should be interpreted relative to an *averaged* strike zone, which shows at the front. The colour of each point corresponds to the value of the underlying KDE pdf.

methods have shortly become standard for these bivariate samples. However, the pitched ball has other attributes that will make the batter's job difficult. A first approximation suggests considering pitches as a pair of location coordinates with a *speed* value measured from the release point.

We propose studying trivariate pitching samples from three top MLB pitchers in the 2019 season: Clayton Kershaw, Justin Verlander and Max Scherzer. Fig. 7 shows pitching scatter data and bumps. In general, the three pitching patterns resemble Gaussian mixtures. It seems that each player chooses his pitches from a finite arsenal. Then, each pitch type has its location and speed variability. Most pitchers employ three mechanisms: the *fastball*, which relies on sheer speed; the *slider*, which moves sideways at relatively high speeds; and the *curveball*, which sinks at low speeds.

In Fig. 7a, Kershaw's fastballs and sliders have speeds ranging between 85 and 90 *miles per hour* (mph). Fastballs usually range beyond 95 mph, as for Verlander (Fig. 7b) and Scherzer (Fig. 7c), who have a defined separation between both pitch types. As for curveballs, Kershaw's have low speed and high vertical variability. Verlander and Scherzer's curveball clouds have similar compact shapes, but the former usually aims for the right side of the strike zone and the latter for the left. Scherzer's usage of curveballs is the lowest of all three.

Appendix D. Computational details

D.1. **Source code.** We provide this project's source code for reproducibility purposes. This paper's figures can be safely generated from scratch in any environment by leveraging Docker's containerization [20]. See the `README.md` and `LICENSE.txt` files for further details [2].

D.2. **Scientific computing.** We employed Python's *NumPy* [14], *SciPy* [26] and *pandas* [19] packages for array and data manipulation, optimization, and linear algebra operations, among other uses.

D.3. **Data.** The three main American professional sports leagues maintain vast data collections and publicly expose them through web APIs. NBA shot records are accessible via the `nba_api` Python package [30]. NFL passing data can be retrieved through the `nflgame-redux` Python package [28]. MLB pitching samples are available from the `pybaseball` Python package [29]. To ease reproducibility, we include the downloaded data as CSV files and the parameterized scripts to call the APIs. See the `README.md` file for more information [2].

D.4. **Bandwidth selection.** The bandwidth selection process targeted the second derivatives of the pdf in all the examples. We employed the smoothed cross-validation criterion for the NFL and NBA applications, beginning with a Gaussian pilot bandwidth [1]; three and two stages were used, respectively. In the NBA case, we limited the optimization to 50 iterations. For the MLB application, we selected the bandwidth assuming the true pdf was a Gaussian mixture [1] with three components, following the three main pitching mechanisms discussed. The Gaussian mixture was fit using Python's package *scikit-learn* [21].

D.5. **Graphing.** Visualizing results is crucial in our proposal, as demonstrated in many examples. Up to three dimensions, computing and picturing bump boundaries as level sets is workable. In the univariate case, we have used conventional root-finding algorithms for f''. In dimension two, we employed the graphical routine `contour` from Python's package *Matplotlib* [16], which uses the algorithm *marching squares*. Finally, for $d = 3$, we resorted to the routine `marching_cubes` from Python's package *scikit-image* [25], implementing the homonym algorithm for dimension three.

References

[1] Chacón, J. E. and Duong, T. (2018). *Multivariate Kernel Smoothing and its Applications*. Chapman and Hall/CRC.

[2] Chacón, J. E. and Fernández Serrano, J. (2023). *Source code for "Bump hunting through density curvature features"*.

[3] Chen, Y.-C., Genovese, C. R., Tibshirani, R. J., and Wasserman, L. (2016). Nonparametric modal regression. *The Annals of Statistics* **44**, 489–514.

[4] Chen, Y.-C., Genovese, C. R., and Wasserman, L. (2017). Density Level Sets: Asymptotics, Inference, and Visualization. *Journal of the American Statistical Association* **112**, 1684–96.

[5] Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics* **42**, 1564–97.

[6] Dhaene, J., Vanduffel, S., Goovaerts, M. J., Kaas, R., Tang, Q., and Vyncke, D. (2006). Risk Measures and Comonotonicity: A Review. *Stochastic Models* **22**, 573–606.

[7] do Carmo, M. P. (1976). *Differential Geometry of Curves and Surfaces*. Prentice-Hall.

[8] do Carmo, M. P. (1992). *Riemannian Geometry*. Mathematics: Theory & Applications. Birkhäuser.

[9] Ecker, K. (2004). *Regularity Theory for Mean Curvature Flow*. Birkhäuser Boston.

[10] Folland, G. B. (2002). *Advanced Calculus*. Pearson.

[11] Gaenssler, P., Molnár, P., and Rost, D. (2007). On Continuity and Strict Increase of the CDF for the Sup-Functional of a Gaussian Process with Applications to Statistics. *Results in Mathematics* **51**, 51–60.
[12] Gray, A., Abbena, E., and Salamon, S. (2006). *Modern Differential Geometry of Curves and Surfaces with Mathematica, Third Edition*. Chapman and Hall/CRC.
[13] Grinfeld, P. (2013). *Introduction to Tensor Analysis and the Calculus of Moving Surfaces*. Springer New York.
[14] Harris, C. R. et al. (2020). Array programming with NumPy. *Nature* **585**, 357–62.
[15] Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective*. Springer New York.
[16] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **9**, 90–5.
[17] Ipsen, I. C. F. and Rehman, R. (2008). Perturbation Bounds for Determinants and Characteristic Polynomials. *SIAM Journal on Matrix Analysis and Applications* **30**, 762–76.
[18] Kato, T. (1995). *Perturbation Theory for Linear Operators*. Vol. 132. Springer Berlin Heidelberg.
[19] McKinney, W. (2010). "Data Structures for Statistical Computing in Python". *Proceedings of the Python in Science Conference*. SciPy.
[20] Merkel, D. (2014). Docker: Lightweight Linux Containers for Consistent Development and Deployment. *Linux J.* **2014**.
[21] Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–30.
[22] Petersen, P. (1998). *Riemannian Geometry*. Springer New York.
[23] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.
[24] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer New York.
[25] van der Walt, S. et al. (2014). scikit-image: image processing in Python. *PeerJ* **2**, e453.
[26] Virtanen, P. et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–72.
[27] Wilkinson, J. H. (1988). *The Algebraic Eigenvalue Problem*. OUP Oxford.

Software

[28] Gallant, A. (2020). *nflgame-redux. An API to retrieve and read NFL Game Center JSON data. It can work with real-time data, which can be used for fantasy football*. Version 3.0.0. url: `https://pypi.org/project/nflgame-redux/`.
[29] LeDoux, J. (2021). *pybaseball. Retrieve baseball data in Python*. Version 2.2.1. url: `https://pypi.org/project/pybaseball/`.
[30] Patel, S. (2021). *nba_api. An API Client package to access the APIs for NBA.com*. Version 1.1.11. url: `https://pypi.org/project/nba-api/`.

Authors: José E. Chacón[†] and Javier Fernández Serrano[‡].

[†]Departamento de Matemáticas, Universidad de Extremadura, Badajoz, Spain.
[‡]Departamento de Matemáticas, Universidad Autónoma de Madrid, Madrid, Spain.
E-mail addresses: [†]`jechacon@unex.es` ✉, [‡]`javier.fernandezs01@estudiante.uam.es`.
[†]`https://orcid.org/0000-0002-3675-1960` .
[‡]`https://orcid.org/0000-0001-5270-9941` .


UÁM
Universidad Autónoma
de Madrid

https://doi.org/10.1007/s11749-023-00872-z


UAM
Universidad Autónoma
de Madrid


UAM
Universidad Autónoma
de Madrid

Capítulo 2

Número de modas

DECLASSIFIED

Computational Statistics & Data Analysis


UÁM
Universidad Autónoma
de Madrid

BAYESIAN TAUT SPLINES FOR ESTIMATING THE NUMBER OF MODES

JOSÉ E. CHACÓN[†] AND JAVIER FERNÁNDEZ SERRANO[‡]

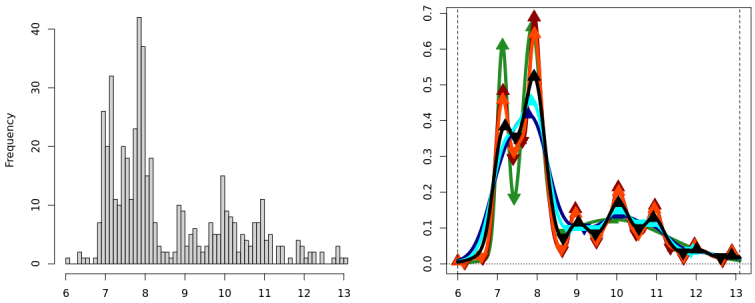
Abstract. The number of modes in a probability density function is representative of the complexity of a model and can also be viewed as the number of subpopulations. Despite its relevance, there has been limited research in this area. A novel approach to estimating the number of modes in the univariate setting is presented, focusing on prediction accuracy and inspired by some overlooked aspects of the problem: the need for structure in the solutions, the subjective and uncertain nature of modes, and the convenience of a holistic view that blends local and global density properties. The technique combines flexible kernel estimators and parsimonious compositional splines in the Bayesian inference paradigm, providing soft solutions and incorporating expert judgment. The procedure includes feature exploration, model selection, and mode testing, illustrated in a sports analytics case study showcasing multiple companion visualisation tools. A thorough simulation study also demonstrates that traditional modality-driven approaches paradoxically struggle to provide accurate results. In this context, the new method emerges as a top-tier alternative, offering innovative solutions for analysts.

1. Introduction

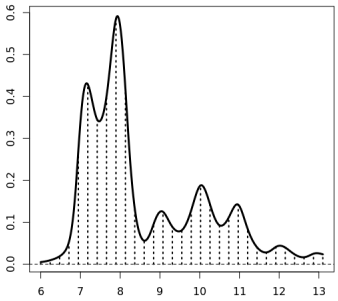
The concept of *mode* has regained the attention of the research community in recent years (Chacón, 2020). Density modes are defined as local maxima of a probability density function (pdf). As such, they mark regions of relatively high concentration of probability mass. Consequently, they are essential features in *exploratory data analysis*, pointing out new phenomena (Ameijeiras-Alonso et al., 2018; Arias-Castro and Jiang, 2022). Notable application examples include the silica composition of meteors in geology (Good and Gaskins, 1980), the distribution of net income in econometrics (Marron and Schmitz, 1992), and the thickness of stamps in philately (Izenman and Sommer, 1988). Even so, since the true pdf is unknown, estimating it from possibly scarce data is prone to generating *spurious* modes that can be confused with actual discoveries (Good and Gaskins, 1980; Minnotte, 1997).

Concerning modes, two primary considerations emerge: the determination of their quantity and their spatial locations. The latter subsumes the former, but the techniques and hypotheses vary (Chacón, 2020). The number of modes (NoM) is an integer-valued statistical functional, deceptively leading one to believe that *counting* them is more straightforward and less compelling than *locating* them. However, a result by Donoho (1988) states the impossibility of bounding from above this quantity with certain confidence if the underlying finite sample comes from a *genuinely*

[†]Departamento de Matemáticas, Universidad de Extremadura, Badajoz, Spain.
[‡]Departamento de Matemáticas, Universidad Autónoma de Madrid, Madrid, Spain.
E-mail addresses: [†]jechacon@unex.es ✉, [‡]javier.fernandezs01@estudiante.uam.es.
2020 *Mathematics Subject Classification.* 62G05 (Primary), 62G07, 62F15, 62C10, 62C86.
Key words and phrases. number of modes, Bayesian inference, compositional spline, kernel density estimation, model selection, mode testing.
[†]https://orcid.org/0000-0002-3675-1960 .
[‡]https://orcid.org/0000-0001-5270-9941 .



(A) Hidalgo sample (B) Several pdfs



(C) BTS solution

FIG. 1. The Hidalgo stamps data (Izenman and Sommer, 1988). The bar chart on the left displays the sample, which comprises 485 measurements of stamp thickness in hundredths of a millimetre. Several pdfs for that sample are shown on the right. There are some kernel density estimators with different bandwidth selectors, such as PI, LSCV and STE, the first two of which count on variations targeting the r-th pdf derivative (Chacón and Duong, 2013): PI_r and LSCV_r. Namely, the pdfs are PI_0 (black, 7 modes), PI_1 (cyan, 5 modes), PI_2 (dark blue, 2 modes), STE (orange, 9 modes), LSCV_0 (red, 11 modes), and a Gaussian mixture (green, 3 modes). The bottom picture shows our BTS solution with seven modes based on 32 spline basis functions.

nonparametric distribution. Moreover, from a practical standpoint, the NoM corresponds to the number of groups in many clustering methodologies (Chacón, 2018; Cuevas et al., 2000).

Consider the Hidalgo stamps data, consisting of 485 measurements of stamp thicknesses (Izenman and Sommer, 1988). The raw data histogram is depicted in Fig. 1a. The issue from 1872 comprises a mix of thicknesses deriving from several extinct paper types and manufacturing processes in various factories (Fisher and Marron, 2001). Philatelists are willing to trace the number of sources of that collection to measure the value of the stamp, which calls for estimating the NoM. Interestingly, the problem is open to this day. As Fig. 1b shows, many sensible pdf estimators, such as the *kernel density estimator* (KDE), provide different solutions. Although each new modality approach is traditionally tested on this dataset, no consensus answer exists (Ameijeiras-Alonso et al., 2018). The *Bayesian taut spline* (BTS) method introduced in this paper outputs the pdf in Fig. 1c with seven modes, quantifying the uncertainty of several alternatives.

Goals. This paper addresses the classical problem of estimating the NoM in the univariate setting. We advance the development of new ideas targeting some overlooked themes in modality. Namely, we stress the convenience of structure in the solutions, the uncertain and subjective nature of modes, and the desirability of a holistic view combining global and local properties.

The resulting novel BTS proposal allows incorporating prior knowledge regarding the number and significance of modes while balancing data fitting and model complexity in *soft* solutions, gaining valuable insights along the way. The method will be tested and compared with existing alternatives from the literature in a thorough simulation study based on accuracy.

Highlights.

- Combining kernel estimators and compositional splines benefits mode exploration
- Dimensionality reduction with one principal component captures the essential modes
- Bayesian inference admits expert knowledge regarding modality in soft solutions
- The new proposal and other generic methods outperform classic modality approaches

Related work. Modality research has evolved along several branches. Although BTS is influenced by many of the approaches in this section, its philosophy is reminiscent of the *penalised likelihood* method by Good and Gaskins (1980), which arguably had no straight continuation.

One of the all-time classics is the *critical bandwidth* test by Silverman (1981). The critical bandwidth h_{crit} is the smallest h at which a KDE with a Gaussian kernel has at most k modes. If h_{crit} turns out large after *bootstrapping*, the null hypothesis of k modes is rejected in favour of more than k. Mammen et al. (1992) refined some asymptotic results by Silverman, adjusted the scaling of the bootstrap to increase the power of the test (see also Fisher et al., 1994), and later provided an extension (Fisher and Marron, 2001). Lastly, Hall and York (2001) also suggested a recalibration of the p-values.

An alternative perspective based on the *mass* of the mode was presented by Muller and Sawitzki (1991). Assuming k modes, the *excess mass* $E_k(\lambda)$ is the maximal sum of deviations between the empirical distribution and λ times the Lebesgue measure, where the maximum is taken over all sequences of k disjoint sets (Cheng and Hall, 1998). A large value of $\sup_{\lambda>0} E_k(\lambda) - E_{k-1}(\lambda)$ rejects $k-1$ modes in favour of k. This test statistic for $k = 2$ is equivalent in the univariate case to the *dip* test by Hartigan and Hartigan (1985) (see also Cheng and Hall, 1998, 1999; Hall and York, 2001). Polonik (1995a) and Polonik (1995b) further developed the idea of excess mass. Since then, several calibration methods have been proposed (Ameijeiras-Alonso et al., 2018, 2021; Cheng and Hall, 1998, 1999).

Minnotte and Scott (1993) introduced *mode trees*. Using a Gaussian kernel, the sequence of KDEs as h decreases depicts the *branching* of minor modes from larger ones, forming a tree-like diagram. The original version can be extended with visual hints of the location of antimodes and *bumps*, and the mass and significance of each mode (Minnotte, 1997). Minnotte et al. (1998) investigated the use of jittering and bootstrapping techniques to build a *mode forest*, which helps to overcome sampling variability.

Chaudhuri and Marron (1999) developed SiZer, a graphical tool to examine KDEs at different scales. SiZer explores modes attending to statistically significant

sign changes in the first derivative of the KDE at each h. Chaudhuri and Marron (2002) later promoted evaluating second derivatives to complement vanilla SiZer. The multi-scale view was also explored by Dümbgen and Walther (2008). Genovese et al. (2016) introduced a mode testing technique, which estimates the NoM as the maximum number of significant ones at any scale h. Finally, Sommerfeld et al. (2017) used *topological data analysis* to assess modes from a twofold scale-*lifetime* perspective.

Following Hartigan and Hartigan (1985, p. 79), Davies and Kovac (2004) proposed the *taut string*, a piecewise-linear cumulative distribution function (cdf) confined on an ε-radius tube surrounding the empirical cdf (see also Davies et al., 2009). The hyperparameter $\varepsilon > 0$ controls the amount of smoothing, with the NoM monotonically increasing as ε decreases. Interestingly, the taut string minimises the NoM under the ε-radius constraint. A more recent proposal fitting pdfs with a fixed NoM in a dynamical programming scheme is developed in Arias-Castro and Jiang (2022).

Outline. Section 2 provides some technical preliminaries. Then, our novel proposal is presented in Section 3. A real-world sports analytics application is explored in Section 4 for illustrative purposes. The simulation study testing and comparing the performance of the new method to existing alternatives comes in Section 5. Finally, Section 6 discusses the achievements, points of improvement and possibilities of the new method.

2. Preliminaries

The following lines define concepts such as modes, antimodes and modal regions, making our assumptions explicit. We will also introduce the Bayesian inference notation used across all the stages of BTS. Finally, we will recall some elements of KDEs, Bayes spaces and compositional splines necessary for our construction.

In what follows, assume a non-empty and finite dataset $\mathcal{D} = \{x_1, \ldots, x_n\} \subset \mathbb{R}$ consisting of n independent and identically distributed (i.i.d.) random variable realisations.

Modes. All intermediate and final univariate pdfs of BTS are assumed to be *Morse functions*, i.e., functions whose critical points are nondegenerate (Chacón, 2015), avoiding pdfs with flat parts where defining modes is more involved (Donoho, 1988, Section 4.1). We will also suppose that all the pdfs f are as smooth as required and have compact support $[a, b]$, i.e., $f(x) > 0$ if $x \in [a, b]$, and $f(x) = 0$ for all $x \notin [a, b]$. Consequently, f will have finitely many critical points (Chacón, 2015, p. 530).

We define *modes* as *local maxima* of the pdf, i.e., points $\hat{\mathsf{m}} \in [a, b]$ such that $f(\hat{\mathsf{m}}) > f(x)$ for all $x \in [\hat{\mathsf{m}} - \epsilon, \hat{\mathsf{m}} + \epsilon] \setminus \{\hat{\mathsf{m}}\}$, for some $\epsilon > 0$. Such definition includes, but is not limited to, critical points x where $f'(x) = 0$ and $f''(x) < 0$. In particular, the boundaries a and b could be modes despite not obeying the latter derivative constraints (remember that f is defined over $\mathbb{R}$). Similarly, we define *antimodes* as *local minima* of the pdf, i.e., points $\check{\mathsf{m}} \in [a, b]$ such that $f(\check{\mathsf{m}}) < f(x)$ for all $x \in [\check{\mathsf{m}} - \epsilon, \check{\mathsf{m}} + \epsilon] \setminus \{\check{\mathsf{m}}\}$, for some $\epsilon > 0$.

Under the previous assumptions, the minimum NoM is one. In general, if there are $k \geq 1$ modes, the number of antimodes will be $k - 1$. For $k > 1$, the modes and antimodes alternate as $a \leq \hat{\mathsf{m}}_1 < \check{\mathsf{m}}_1 < \hat{\mathsf{m}}_2 < \cdots < \hat{\mathsf{m}}_{k-1} < \check{\mathsf{m}}_{k-1} < \hat{\mathsf{m}}_k \leq b$. Hence, we see that $[a, b]$ can be expressed as a union of intervals containing exactly one mode: the *modal regions*. More formally, if we write $\check{\mathsf{m}}_0 = a$ and $\check{\mathsf{m}}_k = b$, the i-th modal region containing $\hat{\mathsf{m}}_i$ is $[\check{\mathsf{m}}_{i-1}, \check{\mathsf{m}}_i]$, where $\hat{\mathsf{m}}_i$ is allowed to coincide with

one of the boundaries only if $i = 1$ or $i = k$. With that notation, when $k = 1$, the unique modal region is $[\check{\mathsf{m}}_0, \check{\mathsf{m}}_1] = [a, b]$.

Bayesian inference. Continuous and discrete parameters and hypotheses are treated as random variables in Bayesian inference. We will use the same notation $\Pr(\cdot)$ for pdfs and probability mass functions, inferring the continuous/discrete nature from the context.

Consider a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m) \in \mathbb{R}^m$. We will refer to its Bayesian prior as $\Pr(\theta_1, \ldots, \theta_m)$. The model pdf conditioning on $\boldsymbol{\theta}$ will be denoted as $\Pr(\cdot|\theta_1, \ldots, \theta_m)$. Then, $\Pr(\mathcal{D}|\theta_1, \ldots, \theta_m) = \prod_{x\in\mathcal{D}} \Pr(x|\theta_1, \ldots, \theta_m)$ is the likelihood of $\boldsymbol{\theta}$ given data $\mathcal{D}$, and $\Pr(\theta_1, \ldots, \theta_m|\mathcal{D}) \propto \Pr(\mathcal{D}|\theta_1, \ldots, \theta_m) \times \Pr(\theta_1, \ldots, \theta_m)$ is the posterior of $\boldsymbol{\theta}$ given $\mathcal{D}$, where "$\propto$" indicates proportionality, i.e., equality except for a normalising constant making the left-hand side a pdf or a probability mass function. To specify a value $\vartheta \in \mathbb{R}$ for some parameter θ_i, we will replace the parameter "θ_i" with "$\theta_i = \vartheta$" in the notations above.

Kernel density estimation. The KDE for the dataset $\mathcal{D}$ based on a Gaussian kernel ϕ and bandwidth $h > 0$ is the function $\mathbb{R} \to (0, \infty)$ defined by

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \phi\left(\frac{x - x_i}{h}\right). \tag{1}$$

The parameter h controls the amount of smoothing. Fixing $\mathcal{D}$, a large value of h hides prominent features, while a small one produces spurious *wiggles*. A Gaussian kernel is needed to ensure continuous tracking of modes (Minnotte and Scott, 1993; Silverman, 1981).

Bayes spaces. The *Bayes space* $\mathcal{B}[a, b]$ of positive pdfs with bounded support $[a, b]$ and square-integrable logarithm has a Hilbert space structure (Machalová et al., 2020) originating from mapping $f \in \mathcal{B}[a, b]$ to the household $L^2([a, b])$ via the *centred log-ratio* (CLR) transformation

$$\operatorname{clr}[f](x) = \log f(x) - \frac{1}{b-a} \int_a^b \log f(y)\, dy\,. \tag{2}$$

Since $\int_a^b \operatorname{clr}[f](x)\, dx = 0$, if we consider the subset complying with that zero-integral constraint, $L_0^2([a, b]) = \{p \in L^2([a, b]) : \int_a^b p(x)\, dx = 0\}$, then $\mathcal{B}[a, b]$ is isometric to $L_0^2([a, b])$ with inverse $\operatorname{clr}^{-1}[p](x) = \exp p(x) / \int_a^b \exp p(y)\, dy$. The vector space operations of sum (perturbation) and scalar multiplication (powering) are, respectively, $(f \oplus g)(x) = f(x)g(x) / \int_a^b f(y)g(y)dy$, and $(\gamma \odot f)(x) = f(x)^\gamma / \int_a^b f(y)^\gamma dy$, where $f, g \in \mathcal{B}[a, b]$ and $\gamma \in \mathbb{R}$. Let us denote $f \ominus g = f \oplus (-1) \odot g$. Similarly, the inner product between $f, g \in \mathcal{B}[a, b]$ is given by

$$\langle f, g\rangle_{\mathcal{B}} = \frac{1}{2(b-a)} \int_a^b \int_a^b \log \frac{f(x)}{f(y)} \log \frac{g(x)}{g(y)}\, dx\, dy\,, \tag{3}$$

that is, $\langle f, g\rangle_{\mathcal{B}} = \int_a^b \operatorname{clr}[f](x)\operatorname{clr}[g](x)\, dx = \langle \operatorname{clr}[f], \operatorname{clr}[g]\rangle_2$.

Compositional splines. Let us fix a polynomial degree $r \geq 3$, enabling a non-flat second derivative. Given $d \geq r$, we can construct a d-dimensional vector subspace $\mathcal{Z}_d[a, b] \subset \mathcal{B}[a, b]$ consisting of pdfs whose CLR transformations (2) are r-th degree spline functions $s : [a, b] \to \mathbb{R}$ such that $\int_a^b s(x)\, dx = 0$. The elements in $\mathcal{Z}_d[a, b]$ are known as *compositional splines*.

Explicitly constructing compositional splines relies on a variant of the usual B-splines, known as ZB-splines, obeying the zero-integral constraint (Machalová et al., 2020). Let us split $[a, b]$ through arbitrary knots $a = \kappa_0 < \cdots < \kappa_{d-r+1} = b$. Then,

we can define d ZB-spline basis functions $Z_1, \ldots, Z_d : [a, b] \to \mathbb{R}$ built up from r-th degree polynomials joining at the knots with maximal $\mathcal{C}^{r-1}([a, b])$ smoothness.

Let us define the ZB-spline $z_{\boldsymbol{\theta}} = \sum_{i=1}^{d} \theta_i Z_i$, for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$. The CLR back-transformed pdf $\zeta_{\boldsymbol{\theta}} = \mathrm{clr}^{-1}[z_{\boldsymbol{\theta}}] = \bigoplus_{i=1}^{d} (\theta_i \odot \mathrm{clr}^{-1}[Z_i])$ is a compositional spline. Therefore, we can define $\mathcal{Z}_d[a, b] = \{\zeta_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^d\}$. Note that $\mathcal{Z}_d[a, b]$ is isometric to $\mathbb{R}^d$ via $\zeta_{\boldsymbol{\theta}} \mapsto \boldsymbol{\theta}$ with inner product $\langle \zeta_{\boldsymbol{\theta}_1}, \zeta_{\boldsymbol{\theta}_2} \rangle_{\mathcal{B}} = \boldsymbol{\theta}_1^\top \mathbf{M} \boldsymbol{\theta}_2$, for $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$, where the symmetric matrix $\mathbf{M}$ has entries $\mathbf{M}_{ij} = \langle Z_i, Z_j \rangle_2$.

Every $f \in \mathcal{B}[a, b]$ can be approximated in the least squares sense within $\mathcal{Z}_d[a, b]$ provided equidistant knots and a sufficiently large d are used. The value r is less critical, making $r = 3$ a widespread choice. Let us fix a fine grid with evenly spaced points $a = t_1 < \cdots < t_m = b$. For all $f \in \mathcal{B}[a, b]$ and every smoothing penalty factor $\alpha \in (0, 1)$, we define the curvature-penalised quadratic loss of a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$ as

$$\mathcal{L}(\boldsymbol{\theta}; f, \alpha) = \alpha \sum_{i=1}^{m} \{\mathrm{clr}[f](t_i) - z_{\boldsymbol{\theta}}(t_i)\}^2 + (1 - \alpha) \int_a^b z''_{\boldsymbol{\theta}}(x)^2 \, dx \,. \tag{4}$$

The sum term relates to data fitting, whereas the integral value amounts to the *total curvature* of the spline, representing its complexity. Then, the compositional spline $[f]_\alpha$ approximating $f \in \mathcal{B}[a, b]$ with smoothing penalty factor $\alpha \in (0, 1)$ is

$$[f]_\alpha = \zeta_{\hat{\boldsymbol{\theta}}(f,\alpha)} \,, \text{ where } \hat{\boldsymbol{\theta}}(f, \alpha) = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}; f, \alpha) \,. \tag{5}$$

Problem (5) has a straightforward linear algebra solution (Machalová et al., 2020). The larger α, the closer (5) is to f. In contrast, smaller values of α produce smoother solutions. If f is positive but does not have bounded support $[a, b]$, we will assume that $[f]_\alpha \equiv [f^\star]_\alpha$, where the pdf $f^\star$ satisfies $f^\star(x) \propto f(x) \cdot \mathbb{1}_{[a,b]}(x)$, since $\mathrm{clr}[f]$ and $\mathrm{clr}[f^\star]$ coincide over $[a, b]$.

3. The Bayesian taut spline (BTS) method

This section introduces the new BTS method, encompassing several steps. First, an *exploration* phase scans the KDE mode tree while building spline models. Next, an *analysis* phase summarises the splines into a linear one-parameter model. Then, a *selection* phase obtains probabilities and pdfs for each k-mode hypothesis. Finally, a *testing* phase evaluates the modes of each representative k-modal pdf.

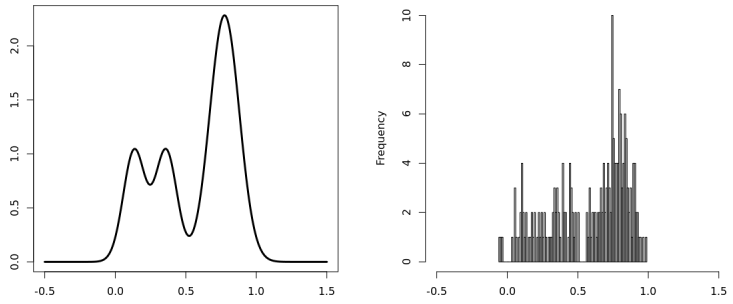
We refer the reader to the supplementary material (SM) (Chacón and Fernández Serrano, 2023) for implementation details and discussing the pillars of the proposal: model structure, Bayesian inference and a holistic view joining global and local density properties.

We will illustrate the different stages through the example in Fig. 2. Fig. 2a depicts a complicated case with three modes, where two are little pronounced. The NoM in the resulting random sample in Fig. 2b is debatable, for one of the two apparent modes on the left block seems isolated and weak. We shall see how BTS handles the situation.

3.1. Exploration.

BTS starts off exploring modes through a combination of KDEs and compositional splines. Informally speaking, for every KDE bandwidth $h > 0$ and smoothing penalty factor $\alpha \in (0, 1)$, we first build the KDE $\hat{f}_h$ from $\mathcal{D}$ and then fit a compositional spline to $\hat{f}_h$ using least squares with penalty α. More formally, for every pair of smoothing parameters $(h, \alpha) \in (0, \infty) \times (0, 1)$, we define a pdf model for $\mathcal{D}$ conditioning on an uncertain pair (h, α) as

$$\Pr(\cdot | h, \alpha) = [\hat{f}_h]_\alpha \,. \tag{6}$$



(A) Theoretical pdf (B) Random sample

FIG. 2. An illustrative test-bed for estimating the NoM. The `M25` mixture model pdf from Ameijeiras-Alonso et al. (2018) is shown on the left. A histogram of a random sample of size 200 from that model is displayed on the right.

Since $\hat{f}_h$ in (6) has unbounded support, we will set $a = \min(\mathcal{D})$ and $b = \max(\mathcal{D})$ in practice. In this context, let us abbreviate for later use $\hat{\boldsymbol{\theta}}(h, \alpha) \equiv \hat{\boldsymbol{\theta}}(\hat{f}_h, \alpha)$, where the right-hand side is as in (5).

The KDE is known to be sensitive to outliers in the tails (Chacón and Duong, 2013). Spurious modes, almost invisible to the human eye, might appear despite the curvature penalisation of BTS in cases with severely isolated points. We propose removing the outliers before building the KDE in (6), similarly to Chacón and Duong (2013, Section 5.2). Using (1) as a pilot, one can calculate the probability masses of all the modal regions and remove those $x \in \mathcal{D}$ linked to underrepresented modes. After removal, the resulting subsample is again fed to (1). This preprocessing step shall be implicitly assumed in (6) in the upcoming sections.

Model (6) is not a typical parametric pdf, for (h, α) are a *proxy* of the underlying spline coordinates $\boldsymbol{\theta}$. Since direct likelihood-based estimation tends to *overfit* data, especially with discretised data, we suggest imposing strong regularity conditions through a prior $\Pr(h, \alpha)$. Then, exploring the spline space and the subsequent modes comes down to evaluating the posterior $\Pr(h, \alpha | \mathcal{D})$. This can be achieved via Markov chain Monte Carlo (MCMC) simulation (Kass and Raftery, 1995), securing a posterior sample $\mathcal{S} = \{(h_i, \alpha_i)\}_{i=1}^{\nu}$.

We propose an *improper* prior $\Pr(h, \alpha) \equiv \Pr(h) \times \Pr(\alpha) \times \Pr(k) \times \Pr(\xi)$ factoring the shape parameters alongside quantities representing the complexity of the true pdf, such as $k \equiv k(h, \alpha)$, the NoM of $\Pr(\cdot | h, \alpha)$, and $\xi \equiv \xi(h, \alpha)$, the total curvature term in (4) with $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(h, \alpha)$. See Good and Gaskins (1980, p. 45) for a similar use of an "improper" prior in regularisation. A design with well-known parametric distributions is

$$\begin{aligned} h &\sim \text{LogNormal}(\mu = \mu_h, \sigma = \sigma_h)\,, \\ 1 - \alpha &\sim \text{Beta}(\alpha = 1, \beta = \beta_{1-\alpha})\,, \\ k &\sim \text{Poisson}(\lambda = 1)\,, \\ \xi &\sim \text{Exponential}(\lambda = \lambda_\xi)\,. \end{aligned} \tag{7}$$

See the SM (Chacón and Fernández Serrano, 2023) for justification of the previous scheme and the selection of μ_h, σ_h, $\beta_{1-\alpha}$ and λ_ξ, depending on the case.

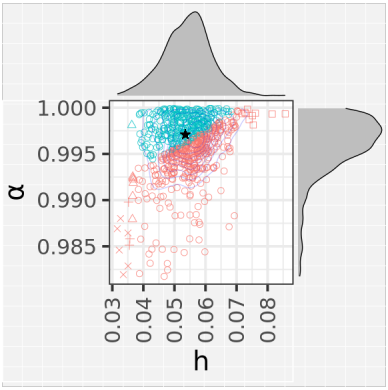


FIG. 3. MCMC sample from $\Pr(h, \alpha|\mathcal{D})$ consisting of 700 observations. The horizontal and vertical axes represent the h and α components, respectively. The points are coloured according to the NoM of model (6): red (57% of the total, 2 modes) and blue (43% of the total, 3 modes). The shape of each point represents the NoM of the underlying KDE, thus only depending on h: square (2 modes), circle (3 modes) and triangle (4 modes), among others. The average point is the black star in the middle of the point cloud. KDEs for the margins are also provided.

The first BTS estimator for the NoM shall be called the *raw* variant of BTS. It is the most frequent modality in $\mathcal{S}$, i.e.,

$$\hat{k}_{\mathrm{BTS},0} = \arg\max_{k\in\mathbb{N}} \left|\{i \in \{1,\dots,\nu\} : \Pr(\cdot|h = h_i, \alpha = \alpha_i) \text{ has } k \text{ modes}\}\right|, \quad (8)$$

where the lower k should be taken for parsimony in case of a draw. The same criterion shall also apply to the upcoming estimators. The counts in (8) can also be normalised to obtain probabilities for each k hypothesis, gaining more insight. We will further process $\mathcal{S}$ in the analysis phase.

Fig. 3 shows the results from the exploration phase of BTS on the random sample in Fig. 2b. The two main scenarios, of two and three modes, are considered in the MCMC sample. The former hypothesis has an inconclusive advantage with 57% of the *ballots*. Because of the intended conservatism of this stage, BTS sticks to two modes without overlooking three. Fig. 3 displays some compelling features of BTS. The spline structure and regularisation leave a fourth and a fifth mode in the underlying KDE with no effect, restraining excessive complexity. Similarly, many points correspond to partially *ironed* three-mode KDEs. We can graphically see a neat oblique line separating the two modality hypotheses, meaning KDEs and smoothing splines benefit from each other to test modes.

Revisiting the classic Hidalgo example in Fig. 1, the above prior design oriented to discrete data leads in (8) to $\hat{k}_{\mathrm{BTS},0} = 7$ with 100% probability. This unanimity ensures that this seven-mode result stands through the upcoming stages.

3.2. **Analysis.** The exploration phase of BTS could benefit from some improvement. Since $\Pr(h, \alpha|\mathcal{D})$ is not genuinely parametric, we depend on finite bivariate MCMC samples. Robust approximations require long MCMC chains, and solving (6) at each iteration is time-consuming. On the other hand, although $\Pr(h, \alpha|\mathcal{D})$ captures modes thoroughly, mode trees count on just one parameter, h, making them easier to interpret (Minnotte and Scott, 1993). The analysis stage of BTS will address both issues.

Analysing $\mathcal{S}$ is problematic, for (h, α) does not reflect the spline structure. Instead, let us define $\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}(h_i, \alpha_i)$ for every $(h_i, \alpha_i) \in \mathcal{S}$. We propose a dimensionality reduction on the pdf sample $\{\zeta_{\boldsymbol{\theta}_i}\}_{i=1}^{\nu} \subset \mathcal{Z}_d[a, b]$ using *simplicial functional*

principal component analysis (SFPCA) (Hron et al., 2016). SFPCA works similarly to its non-functional counterpart, expressing the compositional spline space in terms of orthonormal axes $\mathrm{PC}_1, \ldots, \mathrm{PC}_d \in \mathcal{Z}_d[a,b]$ (each one, a principal component (PC)) with respective variances $\lambda_1 \geq \cdots \geq \lambda_d$. Then, the usual next step is to simplify a sample by projecting each centred functional datum over a few PCs that retain a considerable proportion of the original variability. See the SM (Chacón and Fernández Serrano, 2023) for a brief note on SFPCA in our context.

Beyond convenience and simplicity, a single PC provides enough power in the BTS context. See the SM (Chacón and Fernández Serrano, 2023) for justification. We propose a refined pdf model for $\mathcal{D}$ depending on a single parameter δ through

$$\Pr(\cdot|\delta) = \mu \oplus \delta \odot \sigma\,, \tag{9}$$

where $\mu = (1/\nu) \odot \bigoplus_{i=1}^{\nu} \zeta_{\boldsymbol{\theta}_i}$ is a sample average pdf, and $\sigma = \sqrt{\lambda_1} \odot \mathrm{PC}_1$ is a standard deviation pdf along the first PC. Even though δ could range over the whole $\mathbb{R}$, extrapolation beyond certain limits leads to nonsensical solutions. We can assess sensible values for δ by inspecting the centred projections along PC_1, i.e., the scores $s_i = \langle \zeta_{\boldsymbol{\theta}_i} \ominus \mu, \mathrm{PC}_1 \rangle_{\mathcal{B}}$. If we write $\delta_i = s_i/\sqrt{\lambda_1}$, we can build a *support* for δ as $\Delta = [\delta_{\min}, \delta_{\max}]$, where $\delta_{\min} = \min_{1\leq i\leq \nu} \delta_i$ and $\delta_{\max} = \max_{1\leq i\leq \nu} \delta_i$. Provided $\delta \in \Delta$, (9) represents the modal features in $\Pr(h, \alpha|\mathcal{D})$.

Since (9) captures the essential parts of (6) by construction, one can adopt a more *objective* approach for the prior distribution. The Jeffreys prior is a standard alternative in a univariate setting (Bernardo, 1994). It assigns equal probabilities to regions of the statistical manifold (9) with the same *volume*, conveying the *principle of indifference*. Additionally, the Jeffreys prior is invariant to reparametrisation. Also, note that since Δ is bounded, the Jeffreys prior is *proper*, integrating up to one. Finally, straightforward calculations for this type of prior yield, for $\delta \in \Delta$,

$$\Pr(\delta) \propto \sqrt{\mathrm{Var}(\log \sigma(X_\delta))}\,, \text{ where } X_\delta \sim \Pr(\cdot|\delta)\,, \tag{10}$$

and $\mathrm{Var}(\cdot)$ denotes the variance of a random variable. We can see that $\Pr(\delta) > 0$ for all $\delta \in \Delta$, since σ can never be constant and $\Pr(x|\delta) > 0$ for all $x \in [a,b]$.

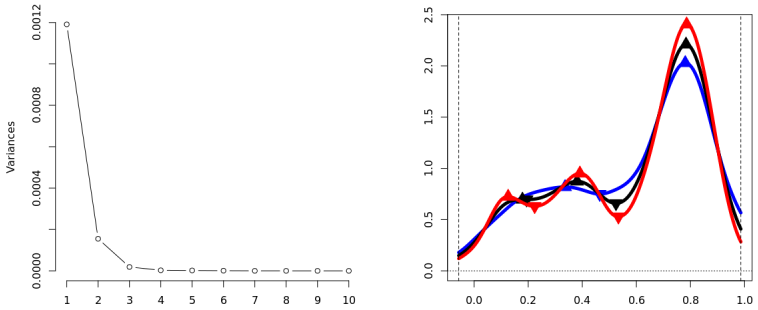
We will refer to the combination of (9) and (10) as the SFPCA model of BTS. BTS considers a second Bayesian inference on the SFPCA model to estimate the NoM.

Fig. 4 shows the results of the analysis stage after the exploration in Fig. 3. First, Fig. 4a plots the PCs against their variances. The sudden drop is characteristic, justifying keeping one dimension. Then, Fig. 4b sheds some light on the shape of the splines in the sample in Fig. 3. All the uncertainty gravitates around whether the left block splits into two modes. The mean μ has three modes, one of which is barely noticeable as an incipient *shoulder* that finally emerges as the left-most mode of $\mu \oplus \delta_{\max} \odot \sigma$.

3.3. **Selection.** The SFPCA model encompasses several hypotheses about the NoM of the pdf. Indeed, the prior (10) is an *encompassing prior* (Klugkist et al., 2005), expressing the relative uncertainty of δ under different domain restrictions. More precisely, let us define, for $k \in \mathbb{N}$, $\Delta_k = \{\delta \in \Delta : \Pr(\cdot|\delta) \text{ has } k \text{ modes}\}$, i.e., the set of parameters δ producing k modes, and let $\mathcal{K} = \{k \in \mathbb{N} : \Delta_k \neq \emptyset\}$ be the set of all the reachable modality numbers. For every $k \in \mathcal{K}$, we can define a prior ensuring k modes as

$$\Pr(\delta|k) \propto \Pr(\delta) \cdot \mathbb{1}_{\Delta_k}(\delta)\,. \tag{11}$$

Combining the new priors (11) with the original model (9), we obtain a class of SFPCA k-modal pdfs. This way, BTS translates estimating the NoM into a Bayesian model selection problem.



(A) SFPCA variances scree plot (B) Mode of variation of PC_1

FIG. 4. SFPCA analysis phase results. The scree plot of the ordered PCs against their variances is presented on the left. Some representative pdfs in the SFPCA model (9) are displayed on the right: the mean μ (black, 3 modes), the lower bound $\mu \oplus \delta_{\min} \odot \sigma$ (blue, 2 modes) and the upper bound $\mu \oplus \delta_{\max} \odot \sigma$ (red, 3 modes).

Note that $\mathcal{K}$ is finite since the pdfs (9) are derived from splines. From our experience, we can virtually assure that every $k \in \mathcal{K}$ shows up as the modality of some pdf in the exploratory sample $\mathcal{S}$. However, it is not uncommon for some modalities observed in the exploration phase to be absent in $\mathcal{K}$. This loss of information, which is welcomed for parsimony, roots in dimensionality reduction.

The *marginal likelihood* of the k-modal prior, $k \in \mathcal{K}$, and the total prior, respectively, are $\Pr(\mathcal{D}|k) = \int_\Delta \Pr(\mathcal{D}|\delta) \times \Pr(\delta|k)\, d\delta$ and $\Pr(\mathcal{D}) = \int_\Delta \Pr(\mathcal{D}|\delta) \times \Pr(\delta)\, d\delta$. After assigning prior probabilities $\Pr(k)$ to each hypothesis $k \in \mathbb{N}$, we can calculate the posterior probabilities (Wasserman, 2000) as $\Pr(k|\mathcal{D}) \propto \Pr(\mathcal{D}|k) \times \Pr(k)$, assuming that $\Pr(k|\mathcal{D}) = 0$ whenever $k \notin \mathcal{K}$.

Calculating $\Pr(\mathcal{D}|k)$ is generally intractable (Kass and Raftery, 1995). Instead, we can use the fact that, in the case of an encompassing prior (Klugkist et al., 2005, p. 60), for every $k \in \mathcal{K}$, we have $\Pr(\mathcal{D}|k)/\Pr(\mathcal{D}) = \int_{\Delta_k} \Pr(\delta|\mathcal{D})\, d\delta / \int_{\Delta_k} \Pr(\delta)\, d\delta$, where the denominator on the right-hand side is positive. This ratio of marginal likelihoods is the *Bayes factor*. Plugging it into the equation for $\Pr(k|\mathcal{D})$ above, we get

$$\Pr(k|\mathcal{D}) \propto \frac{\int_{\Delta_k} \Pr(\delta|\mathcal{D})\, d\delta}{\int_{\Delta_k} \Pr(\delta)\, d\delta} \times \Pr(k)\,. \tag{12}$$

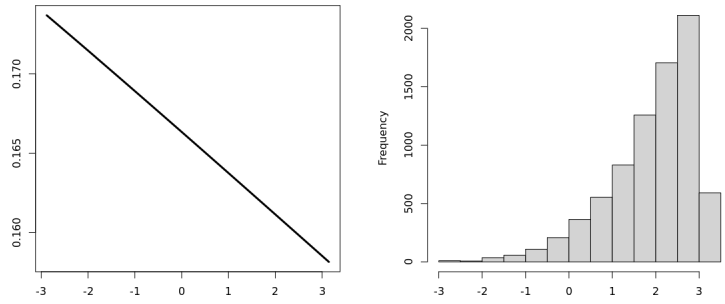
Based on (12), a sensible choice for the prior is $\Pr(k) = \int_{\Delta_k} \Pr(\delta)\, d\delta$. However, $\Pr(k) \propto 1$ is usually preferred (Kass and Raftery, 1995; Klugkist et al., 2005). Furthermore, we can consider the prior probabilities from the posterior sample $\mathcal{S}$ in the exploration phase.

Maximising the posterior probability (12), we finally get the *processed* estimator variant of BTS:

$$\hat{k}_{\mathrm{BTS},1} = \arg\max_{k \in \mathbb{N}} \Pr(k|\mathcal{D})\,. \tag{13}$$

In addition to the new insights gained through the process, the latter estimator will generally produce better results than (8), as demonstrated later in Section 5. The testing phase of BTS, introduced in Section 3.4 below, will further refine (13), providing yet more information.

One can easily verify that, for all $k \in \mathcal{K}$, the Bayesian update to the k-modality prior (11) is $\Pr(\delta|k, \mathcal{D}) \propto \Pr(\delta|\mathcal{D}) \cdot \mathbb{1}_{\Delta_k}(\delta)$. Hence, the prior $\Pr(\delta|\mathcal{D})$ in (12) can



(A) Jeffreys prior (B) Posterior sample

FIG. 5. Second Bayesian inference on the SFPCA model. The Jeffreys prior pdf (10) is shown on the left. A histogram of a sample from $\Pr(\delta|\mathcal{D})$ consisting of 7,840 observations is shown on the right.

be used to obtain a representative pdf of the k-modality hypothesis, typically summarising $\Pr(\delta|k,\mathcal{D})$ (Bernardo, 1994). However, neither the *posterior predictive distribution* nor the *posterior mean* preserves the NoM. A common choice in such circumstances is the *posterior median*, which belongs to Δ_k. The median Bayes estimator can then be interpreted as the minimiser of the average distance induced by (3).

Fig. 5 shows the before and after of the second inference. Fig. 5a depicts the Jeffreys prior (10), which exhibits a slight slope, mildly penalising the more complex pdfs. Then, Fig. 5b shows a histogram of the posterior sample displaying a varying slope in the opposite direction. The Bayesian update for the SFPCA model favours the three-mode hypothesis more than the exploratory inference, as confirmed by Fig. 6. Regardless of the choice of $\Pr(k)$ (among those proposed here), BTS correctly selects three modes based on posterior probabilities ranging between 0.92 and 0.94.

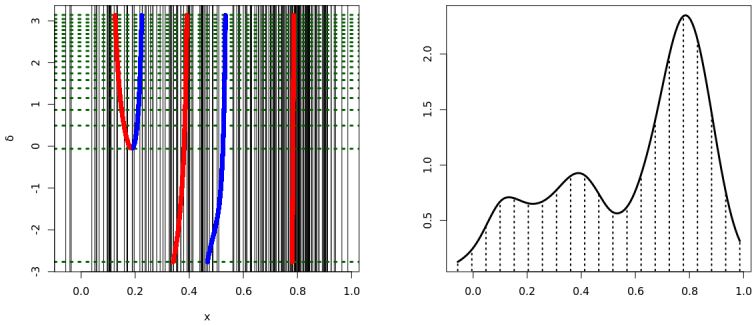
3.4. **Testing.** The NoM in BTS measured *model complexity* in Section 3.1 and was a *feature* encompassing a range of parameter values in Section 3.3. Both correspond to *global* views. However, modes are defined locally, which calls for testing their significance with nearby data. We thus invoke the concept of *excess mass region*.

For every $k \in \mathcal{K}$ modality hypothesis, let us call $\tilde{f}_k$ the median k-modal pdf selected in Section 3.3 with probability $\Pr(k|\mathcal{D})$. Then, for the i-th mode $\hat{\mathsf{m}}_i$ of the model $\tilde{f}_k$, we define the i-th excess mass region as $\mathcal{M}_{k,i} = \{x \in [\check{\mathsf{m}}_{i-1}, \check{\mathsf{m}}_i] : \tilde{f}_k(x) \geq \eta_i\}$, where $\check{\mathsf{m}}_{i-1}$ and $\check{\mathsf{m}}_i$ are the minimum and the maximum, respectively, of the modal region corresponding to $\hat{\mathsf{m}}_i$, and $\eta_i = \max\{\tilde{f}_k(x) : x \in \{\check{\mathsf{m}}_{i-1}, \check{\mathsf{m}}_i\} \setminus \{\hat{\mathsf{m}}_i\}\}$. Naturally, $\hat{\mathsf{m}}_i \in \mathcal{M}_{k,i}$. These excess mass regions are a slight variation of those considered in Minnotte and Scott (1993) to allow modes to appear at modal region boundaries.

The rationale behind excess mass regions is as follows. Let $d(f_1, f_2)$ denote the distance between two functions $f_1, f_2 \in L^1([a,b])$. Then, define, for $x \in [a,b]$,

$$g_{k,i}(x) = \begin{cases} \eta_i, & \text{if } x \in \mathcal{M}_{k,i} \\ \tilde{f}_k(x), & \text{if } x \in [a,b] \setminus \mathcal{M}_{k,i} \end{cases}.$$

As mentioned in Minnotte and Scott (1993), for any continuous f without local maxima in $[\check{\mathsf{m}}_{i-1}, \check{\mathsf{m}}_i]$, we have $d(\tilde{f}_k, g_{k,i}) \leq d(\tilde{f}_k, f)$. In other words, $g_{k,i}$ is the



(A) Mode tree (B) Posterior median spline

FIG. 6. Results of the BTS method. The left-hand side picture shows a mode tree with δ values on the vertical axis. Red points correspond to modes, whereas the blue ones are antimodes. The black vertical lines highlight the original sample observations in Fig. 2b. The dashed green horizontal lines represent uniform quantiles of the posterior sample in Fig. 5b. The figure on the right displays the posterior median model, which has three modes and a CLR dimension of 22.

least distinguishable function from $\tilde{f}_k$ with such constraints. This suggests that, to test the existence of $\hat{\mathsf{m}}_i$, assuming $\tilde{f}_k$ is true, the most conservative null hypothesis of non-existence should be based on $g_{k,i}$. Since $g_{k,i}$ only differs from $\tilde{f}_k$ in being constant over $\mathcal{M}_{k,i}$, we see an identification between removing $\hat{\mathsf{m}}_i$ and being *uniform* over $\mathcal{M}_{k,i}$.

Given a k-modality hypothesis, we can *zoom in* on the i-th excess mass region to test the strength of evidence in favour of $\hat{\mathsf{m}}_i$ locally. Let us define a one-parameter power variation of $\tilde{f}_k$ around $\mathcal{M}_{k,i}$ through

$$\Pr(x|\tau,k,i) \propto (\tau \odot \tilde{f}_k)(x) \cdot \mathbb{1}_{\mathcal{M}_{k,i}}(x)\,, \tag{14}$$

where $\tau \in [0,\infty)$. We note that (14) coincides with the τ-powering of the restriction of $\tilde{f}_k$ to $\mathcal{M}_{k,i}$, which is retrieved for $\tau = 1$. Powering has the effect of intensifying or weakening the mode at $\hat{\mathsf{m}}_i$. For a large τ, the probability mass concentrates around $\hat{\mathsf{m}}_i$, while for a small τ, the mass is spread over $\mathcal{M}_{k,i}$, reaching uniformity for $\tau = 0$. This observation motivates a mode significance test in which the null hypothesis $\mathcal{H}_0$ of the non-existence of the mode is paired with $\tau = 0$, while $\tau \neq 0$ represents the alternative hypothesis $\mathcal{H}_1$ of existence. This way, testing the mode becomes a Bayesian single-parameter value testing problem.

The Savage-Dickey method (Wagenmakers et al., 2010) assigns probabilities to $\mathcal{H}_0$ and $\mathcal{H}_1$, weighing the evidence in data $\mathcal{D}$ with prior knowledge. Namely, we can transform prior odds into posterior odds via

$$\frac{\Pr(\mathcal{H}_1|\mathcal{D},k,i)}{\Pr(\mathcal{H}_0|\mathcal{D},k,i)} = \frac{\Pr(\mathcal{D}|\mathcal{H}_1,k,i)}{\Pr(\mathcal{D}|\mathcal{H}_0,k,i)} \times \frac{\Pr(\mathcal{H}_1|k,i)}{\Pr(\mathcal{H}_0|k,i)}\,. \tag{15}$$

We will denote the prior odds for $\mathcal{H}_1$ as $\mathrm{Odds}(\mathcal{H}_1|k,i) = \Pr(\mathcal{H}_1|k,i)/\Pr(\mathcal{H}_0|k,i)$, and those in favour of $\mathcal{H}_0$ as $\mathrm{Odds}(\mathcal{H}_0|k,i) = \mathrm{Odds}(\mathcal{H}_1|k,i)^{-1}$. If $\mathrm{Odds}(\mathcal{H}_1|k,i) = 0$, we shall assume $\Pr(\mathcal{H}_1|\mathcal{D},k,i) = 0$. At this point, one usually takes $\Pr(\mathcal{H}_0|k,i) = \Pr(\mathcal{H}_1|k,i)$, making the posterior odds equal a ratio of marginal likelihoods, the *Bayes factor*, where

$$\Pr(\mathcal{D}|\mathcal{H}_1,k,i) = \int_0^\infty \Pr(\mathcal{D}|\tau,k,i) \times \Pr(\tau|k,i)\; d\tau\,, \tag{16}$$

and $\Pr(\mathcal{D}|\mathcal{H}_0, k, i) = \Pr(\mathcal{D}|\tau = 0, k, i)$. The likelihood $\Pr(\mathcal{D}|\tau, k, i)$ is evaluated only with data in the excess mass region, i.e., $\Pr(\mathcal{D}|\tau, k, i) = \prod_{x \in \mathcal{D} \cap \mathcal{M}_{k,i}} \Pr(x|\tau, k, i)$. Whenever $\mathcal{D} \cap \mathcal{M}_{k,i} = \emptyset$, we can assume $\Pr(\mathcal{H}_1|\mathcal{D}, k, i) = 0$. A natural prior density choice is $\tau|k, i \sim \text{Exponential}(\lambda = 1)$, regardless of k and i. Hence, τ has mean one and a global mode at zero, favouring the null hypothesis and leaning the average scenario towards $\tilde{f}_k$.

Computing (16) is generally complicated (Kass and Raftery, 1995). However, defining the posterior $\Pr(\tau|\mathcal{D}, k, i) \propto \Pr(\mathcal{D}|\tau, k, i) \times \Pr(\tau|k, i)$, the ratio of marginal likelihoods can be expressed as the ratio between the prior and the posterior at $\tau = 0$, i.e.,

$$\frac{\Pr(\mathcal{D}|\mathcal{H}_1, k, i)}{\Pr(\mathcal{D}|\mathcal{H}_0, k, i)} = \frac{\Pr(\tau = 0|k, i)}{\Pr(\tau = 0|\mathcal{D}, k, i)} = \Pr(\tau = 0|\mathcal{D}, k, i)^{-1}\,, \tag{17}$$

assuming an exponential prior in the numerator. Plugging (17) into (15) and solving, we finally get

$$\Pr(\mathcal{H}_1|\mathcal{D}, k, i) = \left[1 + \frac{\Pr(\tau = 0|\mathcal{D}, k, i)}{\text{Odds}(\mathcal{H}_1|k, i)}\right]^{-1}. \tag{18}$$

The probability (18) expresses the significance of the mode $\hat{\mathsf{m}}_i$ in the k-modal pdf $\tilde{f}_k$. The classical theory of Bayes factors for hypothesis testing establishes some reference values to interpret (18). See Kass and Raftery (1995, Section 3.2) for a scale in terms of (17). This methodology gives the benefit of the doubt to the null hypothesis $\mathcal{H}_0$, requiring a value of (18) well above 0.50 to reject it. Such conservatism would also be justified in our case for parsimony.

The probabilities (18) for the left, centre and right modes in Fig. 6b are 0.69, 0.74 and 0.99, respectively, assuming all the prior odds equal one. These results agree with our intuition that the left mode is relatively weak, and only that to the right is beyond doubt.

Applying the same procedure on the pdf in Fig. 1c of the Hidalgo problem, we obtain probabilities from left to right: 0.63, 0.99, 0.83, 0.96, 0.91, 0.51 and 0.75. According to the scale in Kass and Raftery (1995), only the second to fifth modes would be significant, yielding four modes instead of the original seven. It is easy to see why by looking at Fig. 1a. The modes at 7 and 12 are *fractured*, while few data points support the one at 13. From this one-dimensional mode testing perspective, BTS would agree with the four-mode solution by Ameijeiras-Alonso et al. (2018).

Notwithstanding, limiting our analysis to the probabilities (18) has several drawbacks. Setting a decision threshold might be too rigid and arbitrary, and, more importantly, discarding modes on the grounds of significance equates to forgetting the global results making the local analysis (18) possible. In fact, for the Hidalgo problem, the previous stages of BTS led unequivocally to seven modes.

We propose to round BTS off by combining the *global* probabilities obtained in Section 3.3 with new *scores* for each k based on (18). Let $\Pr(\mathcal{H}_1|k, \mathcal{D})$ be the new score for the k hypothesis, representing its *overall* significance. If the latter plays the part of the likelihood and we take $\Pr(k|\mathcal{D})$ as prior probabilities, the "Bayesian update" yields the posterior $\Pr(k|\mathcal{H}_1, \mathcal{D}) \propto \Pr(\mathcal{H}_1|k, \mathcal{D}) \times \Pr(k|\mathcal{D})$, assuming that $\Pr(k|\mathcal{H}_1, \mathcal{D}) = 0$ whenever $\Pr(k|\mathcal{D}) = 0$. The latter posterior can be seen as the sum of two compositional vectors. If $\Pr(\mathcal{H}_1|k, \mathcal{D}) = 0$ for all k, we should take $\Pr(k|\mathcal{H}_1, \mathcal{D}) = \Pr(k|\mathcal{D})$. See Egozcue and Pawlowsky-Glahn (2018) for a similar treatment of likelihoods as general *evidence functions*.

Finally, the *refined* estimator variant of BTS is defined analogously to (13) as

$$\hat{k}_{\text{BTS},2} = \arg\max_{k \in \mathbb{N}} \Pr(k|\mathcal{H}_1, \mathcal{D})\,. \tag{19}$$

If there is only one $k \in \mathbb{N}$ such that $\Pr(k|\mathcal{D}) > 0$, the estimators (13) and (19) coincide. Therefore, in the case of the Hidalgo problem, despite some modes being non-significant, we have $\hat{k}_{\mathrm{BTS},2} = \hat{k}_{\mathrm{BTS},1} = 7$.

There are many ways to aggregate all the probabilities (18) into a single score. We propose taking the *harmonic mean*. Amongst the Pythagorean means, it is the most sensitive to the lower outliers while stable against the higher ones. Let us define $\mathrm{Odds}(\mathcal{H}_1|k)$ as the harmonic mean of the prior odds for each of the k modes in $\tilde{f}_k$, i.e., $\mathrm{Odds}(\mathcal{H}_1|k) = k/\sum_{i=1}^{k} \mathrm{Odds}(\mathcal{H}_1|k,i)^{-1}$. Also, define the mixture

$$\Pr(\tau|k,\mathcal{D}) = \sum_{i=1}^{k} w_i \Pr(\tau|\mathcal{D},k,i)\,, \text{ where } w_i = \frac{\mathrm{Odds}(\mathcal{H}_0|k,i)}{\sum_{j=1}^{k}\mathrm{Odds}(\mathcal{H}_0|k,j)}\,.$$

Then, the harmonic mean $\Pr(\mathcal{H}_1|k,\mathcal{D}) = k/\sum_{i=1}^{k}\Pr(\mathcal{H}_1|\mathcal{D},k,i)^{-1}$ satisfies

$$\Pr(\mathcal{H}_1|k,\mathcal{D}) = \left[1 + \frac{\Pr(\tau=0|k,\mathcal{D})}{\mathrm{Odds}(\mathcal{H}_1|k)}\right]^{-1}\,,$$

making it a generalisation of (18) for k higher than one.

4. Case study

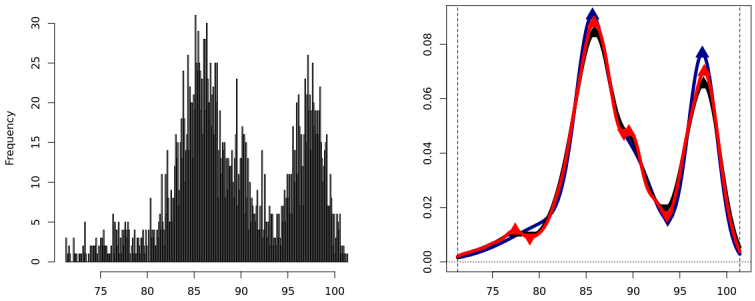
Pitchers are a central part of baseball. They develop a comprehensive throw repertoire that varies in speed, spin and target. Thus, classifying a pitch is challenging, even for a well-trained eye. Nonetheless, to a first approximation, speed takes considerable variability while widely recognised as the main asset of a pitcher. Therefore, let us study pitching from the univariate perspective of speed.

Knowing the *arsenal* of the opposing pitcher increases the chances of a batter hitting the ball. In particular, we can gain valuable knowledge from the speed modes. The larger the NoM, the greater the unpredictability of the pitcher. A scouting report could advise that batters focus their pre-game training and in-game strategy on specific ball speeds corresponding to the modes.

We will look at the pitching speeds of *Major League Baseball* (MLB) top player Shohei Ohtani in the 2022 season. See the SM (Chacón and Fernández Serrano, 2023) for further details on the data. Fig. 7a is a bar chart of the underlying pitches, omitting some outliers below 70 miles per hour (mph). The 2,626-point sample consists of speed values reported up to 0.1 mph, making up a discretised dataset. Hence, extra smoothing will be induced in BTS, as mentioned in the SM (Chacón and Fernández Serrano, 2023). A large compositional spline space dimension $d = 32$ will be employed, as well.

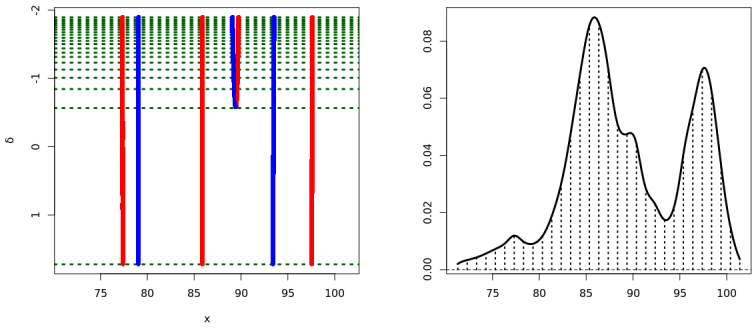
The official interpretation of the pitches by the MLB comprises four modes. Three can be identified with specific pitch types: the *curveball* (with an average of 78 mph), the *slider* (85 mph) and the *fastball* (97 mph). The last mode, located at approximately 90 mph, emerges from a mix of *changeups* and *cutters*. Several preliminary pdf models are depicted in Fig. 7b. The PI_0-variant of the KDE points out four modes, whereas PI_1 eludes the 90-mph mode. The parametric Gaussian mixture model is the least expressive with two modes, missing the latter plus the curveball.

The intermediate results of BTS are qualitatively very similar to the examples in Section 3. We will go straight to the final results from the selection and testing stages and refer the reader to the SM (Chacón and Fernández Serrano, 2023) for the rest. Depending on the prior probabilities $\Pr(k)$, we have both $\Pr(k=4|\mathcal{D})$ and $\Pr(k=4|\mathcal{H}_1,\mathcal{D})$ ranging between 0.93 and 0.98. The mode tree weighted with the posterior sample is in Fig. 8a, while the posterior median spline for $k=4$ is in



(A) Ohtani's 2022 pitches (B) Several pdfs

FIG. 7. Shohei Ohtani's 2022 pitching season data. The left-hand side picture shows a bar chart of the sample, consisting of 2,626 pitch speeds greater than or equal to 70 mph. Several pdfs for that sample are shown on the right. Namely, the pdfs are PI_0 (red, 4 modes), PI_1 (black, 3 modes), and a Gaussian mixture (blue, 2 modes).



(A) Mode tree (B) Posterior median spline

FIG. 8. Results of the BTS method for the MLB case study with the same structure as Fig. 6. The black vertical lines are omitted from the mode tree on the left to enhance readability under a larger sample size. The dimension of the posterior median model with four modes is 32.

Fig. 8b. The modes of the latter pdf are located at 77.4, 85.9, 89.7 and 97.6 mph, having significance probabilities of 0.85, 0.99, 0.62 and 0.99, respectively.

5. Simulation study

This section demonstrates the effectiveness of our proposal in a thorough comparison with other well-established techniques. As we will see, BTS is a top-tier method according to an overall ranking aggregating results from a broad array of test-beds. We will also examine under what circumstances each procedure performs best and worst and analyse the distribution of the predictions. The reader is referred to the SM (Chacón and Fernández Serrano, 2023) for extra auxiliary results and setup details.

5.1. **Setup.** The experimental design is fully described in the following lines.

Methods. Seven variants of BTS are tested in this simulation study:

- `BTS0`: The *raw* BTS estimator (8).
- `BTS1S`: The *processed* BTS estimator (13) with $\Pr(k)$ estimated from $\mathcal{S}$.
- `BTS1J`: The *processed* BTS estimator (13) with $\Pr(k) = \int_{\Delta_k} \Pr(\delta)\ d\delta$.
- `BTS1U`: The *processed* BTS estimator (13) with $\Pr(k) \propto 1$.
- `BTS2S`: The *refined* BTS estimator (19) with $\Pr(k)$ estimated from $\mathcal{S}$.
- `BTS2J`: The *refined* BTS estimator (19) with $\Pr(k) = \int_{\Delta_k} \Pr(\delta)\ d\delta$.
- `BTS2U`: The *refined* BTS estimator (19) with $\Pr(k) \propto 1$.

When comparing alternative methods, we will focus on `BTS2U`, which encompasses all four BTS stages and provides competitive results. We refer the reader to the SM (Chacón and Fernández Serrano, 2023) to compare all the different BTS variants. Standard BTS configurations with $d = 22$ basis functions were used in all cases.

The alternative methods considered in this simulation study, which are described and justified in the SM (Chacón and Fernández Serrano, 2023), are the following:

- `PI0`: The NoM of the KDE with *plug-in* (PI) bandwidth for $r = 0$.
- `PI1`: The NoM of the KDE with PI bandwidth for $r = 1$.
- `PI2`: The NoM of the KDE with PI bandwidth for $r = 2$.
- `SCV`: The NoM of the KDE with *smoothed cross-validation* (SCV) bandwidth.
- `STE`: The NoM of the KDE with *solve-the-equation* (STE) bandwidth.
- `LSCV0`: The NoM of the KDE with *least-squares cross-validation* (LSCV) bandwidth for $r = 0$.
- `LSCV1`: The NoM of the KDE with LSCV bandwidth for $r = 1$.
- `LSCV2`: The NoM of the KDE with LSCV bandwidth for $r = 2$.
- `GM`: The NoM of a Gaussian mixture model selected via *Bayesian information criterion*.
- `TS`: The number of *peaks* of a *taut string* fitted based on Kuiper metrics (Davies and Kovac, 2004).
- `SI`: The highest NoM that the critical bandwidth test by Silverman (1981) cannot reject at a 0.05 significance level.
- `FM`: The highest NoM that the critical bandwidth test by Fisher and Marron (2001) cannot reject at a 0.05 significance level.
- `EIG`: The maximum number of significant KDE modes at a 0.10 significance level, according to Genovese et al. (2016).

Test-beds. We propose as test-beds the $T = 5$ three-modal Gaussian mixtures considered in Ameijeiras-Alonso (2017, pp. 124-125): `M21`, `M22`, `M23`, `M24` and `M25`. They are the most complex models in Ameijeiras-Alonso (2017) and Ameijeiras-Alonso et al. (2018). We refer the reader to the SM (Chacón and Fernández Serrano, 2023) for the definition and plotting of these pdfs.

Experimental design. Each test-bed will be paired with $S = 3$ sample sizes: $n = 100$ (small), $n = 400$ (medium-sized), and $n = 1600$ (large), adding up to $T \times S = 15$ sampling configurations. Then, $m = 200$ replications of the experiment, drawing i.i.d. observations from each combination of test-bed and sample size, will be carried out to obtain significant results, as in Lee (2003).

The results from each sampling configuration will be analysed separately first. For each method $\mathtt{M}_j$, $j \in \{1, \ldots, M\}$, and sample $\mathcal{D}_i$, $i \in \{1, \ldots, m\}$, an estimate $\hat{k}_{j,i} \in \mathbb{N}$ of the NoM will be recorded. Comparing the BTS variants, we have $M = 7$, whereas comparing `BTS2U` with the alternative methods yields $M = 14$. Assuming a ground truth target NoM $k = 3$ for every test-bed model, an outcome $\omega_{j,i} \in \{0, 1\}$

PI0	PI1	PI2	SCV	STE	LSCV0	LSCV1	LSCV2	GM	TS	SI	FM	EIG	BTS2U
1	2	5	**1**	**1**	2	2	3	**1**	6	6	4	6	**1**

TABLE 1. Global ranking.

is derived taking $\omega_{j,i} = 1$, if $\hat{k}_{j,i} = k$, and $\omega_{j,i} = 0$, otherwise. Then, the *accuracy* of the j-th method is $\varpi_j = m^{-1} \sum_{i=1}^{m} \omega_{j,i}$.

Two methods $\mathtt{M}_\mu$ and $\mathtt{M}_\nu$ are confronted by comparing the outcome sequences $(\omega_{\mu,i})_{i=1}^m$ and $(\omega_{\nu,i})_{i=1}^m$. One method is ranked ahead if (i) its accuracy is greater than that of the other *and* (ii) both methods provide significantly different performances according to McNemar's test (McNemar, 1947) using a standard significance level $\alpha = 0.01$. Otherwise, both methods shall be ranked the same. The $M(M-1)/2$ pairwise orderings are then aggregated using the Kemeny distance approach in Amodio et al. (2016), which computes a *median* ranking, possibly including ties among the individual method ranks. If there is no unique solution, we propose aggregating all the resulting rankings via simple component-wise rank averaging, a typical default procedure mentioned in Amodio et al. (2016).

After computing the intermediate rankings $\mathcal{R}_1, \ldots, \mathcal{R}_{T \times S}$, a final aggregation using the previous procedure, consisting of a median consensus ranking and averaging in case of multiple solutions, yields a unique global ranking $\mathcal{R}$, possibly with ties.

5.2. **Results.** We will now look at the results following the previous setup and methodology. By convention, the best position is 1 in all rankings, as in Amodio et al. (2016).

Before diving into the results, we urge the reader to take them cautiously. As warned by Lee (2003), these are only simulations, spanning a small and simplified piece of the problem. Moreover, there is no canonical way of analysing and aggregating the results (see, for instance, Cao et al., 1994, Section 3.2).

Global ranking. The global ranking $\mathcal{R}$ is presented in Table 1. As we can see, `BTS2U` belongs in the same top-ranked group as the parametric `GM` and the KDEs `PI0`, `SCV` and `STE`, all of them generic direct approaches. In the second place, we find two LSCV variants, `LSCV0` and `LSCV1`, and the first-order PI version, `PI1`. Next, `LSCV2`, the remaining LSCV, holds the third rank. The bottom three positions gather all the methods tailored explicitly for mode estimation. Coming fourth and fifth, we have the best performer of the two test-based approaches, `FM`, and the last member of the PI family, `PI2`. Finally, ranking at the bottom appears the other test, `SI`, alongside two methods genuinely related to modes, `TS` and `EIG`.

The same exact final ranking is obtained for various significance levels in McNemar's test, aside from our reference value $\alpha = 0.01$, evidencing the robustness of our conclusions.

The results in Table 1 may look counterintuitive. Direct non-specific methods perform better than those specifically designed for modality assessment. In that sense, customary KDE bandwidth selectors for the pdf usually provide the correct answer, matching the performance of `GM`, which, given the nature of the considered test-beds, can be seen as an upper limit in performance.

The `FM` method by Fisher and Marron (2001) comes as the top-qualified classic approach, outmatching `SI`, the original critical bandwidth proposal by Silverman (1981). In turn, the poor results of `TS` and `EIG` are especially striking, demonstrating that too much parsimony might not work well in practice.

		PI0	PI1	PI2	SCV	STE	LSCV0	LSCV1	LSCV2	GM	TS	SI	FM	EIG	BTS2U
M21	100	2	3	3	2	**1**	**1**	**1**	2	2	3	3	3	3	2
M21	400	2	4	6	2	**1**	**1**	2	3	2	6	6	5	6	2
M21	1600	2	4	6	2	2	3	4	5	**1**	9	7	5	8	2
M22	100	3	7	8	5	2	4	6	7	6	8	8	7	8	**1**
M22	400	**1**	**1**	6	**1**	2	5	3	4	**1**	7	2	**1**	7	2
M22	1600	2	**1**	**1**	2	4	8	6	6	**1**	7	**1**	**1**	5	3
M23	100	2	4	6	2	**1**	2	2	3	**1**	7	7	5	7	2
M23	400	2	**1**	3	2	5	7	6	6	**1**	6	6	4	7	4
M23	1600	4	2	**1**	5	7	8	8	7	**1**	3	**1**	3	6	6
M24	100	3	4	5	3	**1**	3	3	3	5	5	5	5	3	2
M24	400	**1**	2	3	**1**	**1**	2	2	2	3	3	3	3	2	**1**
M24	1600	**1**	**1**	4	**1**	2	5	2	2	5	5	7	6	3	**1**
M25	100	2	5	6	2	**1**	2	3	4	6	8	8	5	7	3
M25	400	**1**	3	8	**1**	**1**	2	4	6	7	10	9	5	10	**1**
M25	1600	2	**1**	2	2	2	3	3	4	**1**	6	4	2	5	2

TABLE 2. Intermediate rankings by test-bed and sample size configuration.

Intermediate rankings. The rankings $\mathcal{R}_1, \ldots, \mathcal{R}_{T\times S}$ are reported in Table 2. See the SM (Chacón and Fernández Serrano, 2023) for an in-depth commentary with plots.

As we see, no method performs uniformly better than the rest across all settings. Nonetheless, the intermediate rankings confirm the superiority of the top-ranked methods in Table 1: `GM`, `PI0`, `SCV`, `STE` and our `BTS2U`. They rank high most of the time and, more importantly, consistently escape from the bottom position. Indeed, at least one of these methods ranks first in each considered case. Especially worth mentioning is the `M22`-100 setting, where `BTS2U` ranks ahead of the rest.

Beyond the top-tier methods, Table 2 shows that the global ranking is quite unfair with `PI1`, as `PI1` lands the first position once more than `PI0` and `SCV`. The reason why `PI1` ranks globally worse lies in its uneven performance. Small datasets particularly harm `PI1`. Also, `PI1` struggles with `M21`, regardless of the sample size. As we can see in the SM (Chacón and Fernández Serrano, 2023), the slight middle mode of `M21` poses serious problems for the over-smoothing strategy of `PI1`.

6. DISCUSSION

BTS faces estimating the NoM inspired by some little-known aspects of the problem. The need for structure diagnosed by Donoho (1988) is implemented through a combination of KDEs (Wand and Jones, 1995) and compositional splines (Machalová et al., 2020). Subsequently, the Bayesian inference machinery (Bernardo, 1994) offers standard procedures to balance data fitting and model complexity, incorporate expert knowledge and assess the uncertainty of the results. Finally, global model selection (Wasserman, 2000) and local testing (Minnotte and Scott, 1993) allow for a holistic view of modality.

We follow a strategic *divide-and-conquer* approach similar to that suggested by Good and Gaskins (1980). An exploration phase via MCMC enforces regularity in the solutions, penalising curvature and multimodality. A numerous sample of candidate compositional splines is summarised using SFPCA (Hron et al., 2016), obtaining a one-parameter model that retains the essential modal features. Then, the Bayesian update weighs the probabilities of each k-modality hypothesis in an encompassing prior (Klugkist et al., 2005), producing a characteristic median compositional spline for each k. Next, each representative pdf has its k modes tested using a Bayesian Savage-Dickey scheme (Wagenmakers et al., 2010) with data in the excess mass region of the mode. Lastly, the global and local scores are *summed* to determine the most likely k.

Traditional approaches for modality either offer a lot of sparse information, such as mode trees (Minnotte and Scott, 1993) or SiZer (Chaudhuri and Marron, 1999), or act as a *black box* outputting an actionable but cryptic p-value, such as the critical bandwidth test (Fisher and Marron, 2001; Silverman, 1981). BTS delivers traceable and interpretable intermediate products, providing a fuzzy but straightforward result. The h-α scatter plot in Fig. 3 studying modality from a twofold perspective represents a legitimate innovation for mode exploration. Then, Fig. 6a and Fig. 8a are upgraded mode trees with several advantages: sensible prescribed upper and lower bounds, a well-behaved natural parameter without resorting to the logarithmic scale, and a posterior pdf that allows placing cue quantiles in the parameter axis. Additionally, the final decision for k is supported on a tangible pdf, as in Fig. 8b.

Several practical experiences support BTS. Our seven-mode result for the Hidalgo problem is well-grounded on likelihood. There are also qualitative reasons, such as the modes approximately appearing at regular integer values, a much human-like trait. On the other hand, the four-mode solution of Ameijeiras-Alonso et al. (2018), for instance, demands a similar smoothing to `PI1`, which we know does not perform well on medium-sized samples with subtle features, according to Section 5. Also, such a smoothing level would imply that the stamp manufacturing process had high variability, including a strange *shoulder* in the density of the thickness distribution. As for the MLB case study, our four-mode solution agrees with the official interpretation of the MLB. Interestingly, our results grant a small probability of having only three modes, with a weak significance score for the dubious changeup-cutter mode. Unlike our method, the Gaussian mixture approach failed conservatively in both real settings since, contrary to the simulation study, non-Gaussianity and discretisation were present.

The simulation study unveils highly unexpected findings. The traditional methods for modality, e.g., `FM`, `SI` or `TS`, fall deeply in the rankings behind generic plug-in KDE-based estimators such as `PI0` and `STE`. Hypothesis testing methods underperform if evaluated far from their restricted theoretical framework, which leads us to question their actual value in practice when they are sequentially employed to estimate the NoM. Moreover, we have not observed the typical asymptotic accuracy boost of nonparametric methods. In turn, the taut string `TS` is conservatively biased, recalling the approach by Donoho (1988) for calculating a lower confidence bound for the NoM. As for our proposal, even though the `BTS2U` variant of BTS is a top-tier method, there is still room for accuracy improvement, perhaps using auxiliary techniques such as ensembles or bootstrapping.

If the user has to make a *blindfold* decision, our overall recommendation for estimating the NoM is `PI0` based on accuracy, robustness, versatility, simplicity and low computational cost. The empirical evidence generally supports the sound theory behind it. With large samples, looking at `PI1` should pay off, as well. Nevertheless, analysts usually prefer analysing data rather than simply *hitting a button* and reporting some output. For them, BTS offers valuable resources with similar accuracy to `PI0`. In this respect, BTS would benefit from a graphical interface for the analysts to work interactively and efficiently access all the available information. Additionally, beyond the NoM, BTS could serve as a more elaborate and dependable version of `PI0` for pdf estimation in the context of bounded data.

Acknowledgements

The research of the first author has been supported by the MICINN grants PID2019-109387GB-I00 and PID2021-124051NB-I00. The second author would like to thank Professor Amparo Baíllo Moreno for her advice as a doctoral counsellor at

the Autonomous University of Madrid. Finally, we thank an associate editor and two anonymous reviewers for their helpful comments.

References

Ameijeiras-Alonso, J. (2017). "Assessing simplifying hypotheses in density estimation". PhD thesis. Universidade de Santiago de Compostela.

Ameijeiras-Alonso, J., Crujeiras, R. M., and Rodríguez-Casal, A. (2018). Mode testing, critical bandwidth and excess mass. *TEST* **28**, 900–19.

Ameijeiras-Alonso, J., Crujeiras, R. M., and Rodríguez-Casal, A. (2021). Multimode: an R package for mode assessment. *Journal of Statistical Software* **97**.

Amodio, S., D'Ambrosio, A., and Siciliano, R. (2016). Accurate algorithms for identifying the median ranking when dealing with weak and partial rankings under the Kemeny axiomatic approach. *European Journal of Operational Research* **249**, 667–76.

Arias-Castro, E. and Jiang, H. (2022). Fitting a multi-modal density by dynamic programming. *arXiv preprint*.

Bernardo, J. M. (1994). Bayesian statistics. *Probability and statistics*. Vol. 2, 345–407.

Cao, R., Cuevas, A., and González-Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation. *Computational Statistics & Data Analysis* **17**, 153–76.

Chacón, J. E. (2015). A population background for nonparametric density-based clustering. *Statistical Science* **30**, 518–32.

Chacón, J. E. (2018). Mixture model modal clustering. *Advances in Data Analysis and Classification* **13**, 379–404.

Chacón, J. E. (2020). The modal age of statistics. *International Statistical Review* **88**, 122–41.

Chacón, J. E. and Duong, T. (2013). Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electronic Journal of Statistics* **7**, 499–532.

Chacón, J. E. and Fernández Serrano, J. (2023). *Supplementary material to "Bayesian taut splines for estimating the number of modes"*.

Chaudhuri, P. and Marron, J. S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association* **94**, 807–23.

Chaudhuri, P. and Marron, J. S. (2002). Curvature vs. slope inference for features in nonparametric curve estimates. *Unpublished manuscript*.

Cheng, M.-Y. and Hall, P. (1998). Calibrating the excess mass and dip tests of modality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**, 579–89.

Cheng, M.-Y. and Hall, P. (1999). Mode testing in difficult cases. *The Annals of Statistics* **27**, 1294–315.

Cuevas, A., Febrero, M., and Fraiman, R. (2000). Estimating the number of clusters. *Canadian Journal of Statistics* **28**, 367–82.

Davies, L., Gather, U., Nordman, D., and Weinert, H. (2009). A comparison of automatic histogram constructions. *ESAIM: Probability and Statistics* **13**, 181–96.

Davies, L. and Kovac, A. (2004). Densities, spectral densities and modality. *The Annals of Statistics* **32**, 1093–136.

Donoho, D. L. (1988). One-sided inference about functionals of a density. *The Annals of Statistics* **16**, 1390–420.

Dümbgen, L. and Walther, G. (2008). Multiscale inference about a density. *The Annals of Statistics* **36**, 1758–85.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2018). Evidence functions: a compositional approach to information. *SORT. Statistics and Operations Research Transactions*, 101–24.

Fisher, N. I., Mammen, E., and Marron, J. S. (1994). Testing for multimodality. *Computational Statistics & Data Analysis* **18**, 499–512.

Fisher, N. I. and Marron, J. S. (2001). Mode testing via the excess mass estimate. *Biometrika* **88**, 499–517.

Genovese, C., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2016). Non-parametric inference for density modes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**, 99–126.

Good, I. J. and Gaskins, R. A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association* **75**, 42–56.

Hall, P. and York, M. (2001). On the calibration of Silverman's test for multimodality. *Statistica Sinica* **11**, 515–36.

Hartigan, J. A. and Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics* **13**, 70–84.

Hron, K., Menafoglio, A., Templ, M., Hrůzová, K., and Filzmoser, P. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics & Data Analysis* **94**, 330–50.

Izenman, A. J. and Sommer, C. J. (1988). Philatelic mixtures and multimodal densities. *Journal of the American Statistical Association* **83**, 941–53.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–95.

Klugkist, I., Kato, B., and Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica* **59**, 57–69.

Lee, T. C. M. (2003). Smoothing parameter selection for smoothing splines: a simulation study. *Computational Statistics & Data Analysis* **42**, 139–48.

Machalová, J., Talská, R., Hron, K., and Gába, A. (2020). Compositional splines for representation of density functions. *Computational Statistics* **36**, 1031–64.

Mammen, E., Marron, J. S., and Fisher, N. I. (1992). Some asymptotics for multimodality tests based on kernel density estimates. *Probability Theory and Related Fields* **91**, 115–32.

Marron, J. S. and Schmitz, H.-P. (1992). Simultaneous density estimation of several income distributions. *Econometric Theory* **8**, 476–88.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–7.

Minnotte, M. C. (1997). Nonparametric testing of the existence of modes. *The Annals of Statistics* **25**, 1646–60.

Minnotte, M. C., Marchette, D. J., and Wegman, E. J. (1998). The bumpy road to the mode forest. *Journal of Computational and Graphical Statistics* **7**, 239–51.

Minnotte, M. C. and Scott, D. W. (1993). The mode tree: a tool for visualization of nonparametric density features. *Journal of Computational and Graphical Statistics* **2**, 51–68.

Muller, D. W. and Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association* **86**, 738–46.

Polonik, W. (1995a). Density estimation under qualitative assumptions in higher dimensions. *Journal of Multivariate Analysis* **55**, 61–81.

Polonik, W. (1995b). Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *The Annals of Statistics* **23**, 855–81.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **43**, 97–9.

Sommerfeld, M., Heo, G., Kim, P., Rush, S. T., and Marron, J. S. (2017). Bump hunting by topological data analysis. *Stat* **6**, 462–71.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., and Grasman, R. (2010). Bayesian hypothesis testing for psychologists: a tutorial on the Savage–Dickey method. *Cognitive Psychology* **60**, 158–89.

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*. Springer US.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology* **44**, 92–107.

Supplementary material

The following *appendices* are provided as *supplementary material* to the manuscript *Bayesian taut splines for estimating the number of modes*. Equation numbers refer to the manuscript. The numbers for new equations are prefixed. References are included at the end. Notation and acronyms are reused from the manuscript.

Appendix A and Appendix B explain the philosophy and implementation details of BTS. Then, Appendix C, Appendix D, and Appendix E expand on the case and simulation studies. Finally, Appendix F and Appendix G discuss relevant theoretical and practical issues.

Appendix A. Motivation

The following lines motivate the design of BTS in Section 3 along three axes.

Structure. Researchers have mainly tackled modality through nonparametric approaches, for parametric ones are deemed too rigid (Ameijeiras-Alonso, 2017). However, the work by Donoho (1988) suggests that modality is impossible to *tame* under weak nonparametric regularity assumptions. For that matter, the quest for modes has the existence of a pdf as a prerequisite, something we cannot test empirically (Donoho, 1988).

Consequently, some structure seems due in the study of modes, balancing data fitting and model complexity. This policy is exhibited by Davies et al. (2009) in the context of histograms or Good and Gaskins (1980) with Fourier series. In this respect, smoothing splines are the right tool, allowing restrained flexibility (Eilers and Marx, 1996; Hron et al., 2016; Machalová et al., 2020). BTS thus targets a compromise between both dimensions by blending KDEs and compositional splines.

On the one hand, KDEs serve as a *scaffold* for splines, guiding high-dimensional fitting more efficiently than customary histograms (Eilers and Marx, 1996; Machalová et al., 2020). On the other, the comparatively rigid structure of splines and their built-in curvature penalisation (Machalová et al., 2020) represent a *harness* for KDEs, preventing spurious modes. Additionally, curvature penalisation mitigates the effect of the KDE having a single global bandwidth h (Minnotte and Scott, 1993). Finally, splines allow for a deeper analysis and simplification of the modal structure via dimensionality reduction (Hron et al., 2016).

Bayesian inference. Mode estimation reveals philosophical issues excellently handled by Bayesian inference (Bernardo, 1994). Emphasising *frequentist* population-wide properties over actual data is unrealistic in some cases. The Hidalgo problem is a paradigmatic example since the stamps are no longer issued. In turn, the Bayesian approach is the right choice when data is scarce, and any information, such as philately expert knowledge, could be helpful.

Bayesian inference provides *soft* solutions, quantifying their uncertainty (Bernardo, 1994). Current methods cannot assign a probability to each k-modality hypothesis. For instance, hypothesis testing procedures report p-values limited to the null hypothesis (Wagenmakers et al., 2010). Meanwhile, mode trees (Minnotte, 1997; Minnotte et al., 1998; Minnotte and Scott, 1993) and SiZer (Chaudhuri and Marron, 1999), though incorporating mechanisms to assess the uncertainty, fall short of providing actionable answers. By contrast, the Bayesian framework excels at operating with probabilities, offering standard hypothesis selection tools (Kass and Raftery, 1995; Klugkist et al., 2005; Spiegelhalter et al., 2014; Wagenmakers et al., 2010; Wasserman, 2000).

The subjective nature of modes particularly suits Bayesian methods. The wide array of graphical methods (Chaudhuri and Marron, 1999, 2002; Minnotte, 1997; Minnotte et al., 1998; Minnotte and Scott, 1993) evidence that the human eye,

aided by a computer, better appreciates such features (Good and Gaskins, 1980). Consequently, the usual criticism that Bayesian inference is not objective loses force (Bernardo, 1994). Moreover, Bayesian tools are less prone to overfitting since they examine a range of plausible outcomes rather than isolated optima. This inherent *parsimony* (Wagenmakers et al., 2010) will be valuable against spurious modes.

Holism. Modes are challenging for their dual local and global nature. They are defined via a local property of the pdf, but, at the same time, that pdf is built from disconnected data. In that sense, modes are *emergent* phenomena.

BTS aims to combine both perspectives. During the first three stages, the mode concept helps build candidate pdfs. Then, at the fourth stage, the representative pdfs of each modality hypothesis have their modes tested individually using neighbouring data, yielding *significance* scores.

The penalised likelihood approach by Good and Gaskins (1980) and mode trees (Minnotte and Scott, 1993) also include local testing mechanisms after global fitting. Our proposal merges the global and local probabilities into a single result, obtaining a holistic view of modality.

Appendix B. Implementation

The following lines discuss the implementation of the BTS method in Section 3.

B.1. **Hyperparameter tuning.** BTS requires setting several configurations. We comment here on how this can be done in practice with attention to the data.

Prior design. The parametric design (7) of the prior in the exploration stage of BTS was left unexplained. Let us now go deeper into the underlying principles and experiences.

The choice of beta distribution for $1-\alpha$ is conventional when $\alpha \in (0,1)$. Fixing $\alpha = 1$ makes the pdf diverge at zero, while $\beta_{1-\alpha}$ controls its expected value through $\mathbb{E}[1-\alpha] = (1+\beta_{1-\alpha})^{-1}$. Next, the rationale behind the bandwidth distribution lies in the logarithmic scale, known for improving the appreciation of the KDE changes in mode trees (Minnotte and Scott, 1993). Assuming a normal distribution for $\log h$ with location μ_h and scale σ_h yields maximum entropy and allows focusing on a suitable region of the mode tree. Moreover, preliminary simulations studying critical bandwidths and the posterior $\Pr(h, \alpha|\mathcal{D})$ confirm that the log-normal provides a good approximation. On the other hand, the distributions of k and ξ have been selected to penalise complexity. In the case of k, the Poisson distribution with mean one favours unimodality. For ξ, the exponential pulls the curvature towards zero, leaving control over the mean via $\mathbb{E}[\xi] = \lambda_\xi^{-1}$.

Imposing hyperpriors on (7) would make the MCMC heavier. We propose choosing the hyperparameters empirically. Taking $\beta_{1-\alpha} = 99$ yields $\mathbb{E}[\alpha] = 0.99$, which works well in practice. In turn, for μ_h, σ_h and λ_ξ, we first recommend estimating two tentative values, say $h_1 < h_2$, from distinct bandwidth selectors. Imposing $\log h$ to enclose a central probability $\Phi(\sigma) - \Phi(-\sigma)$ between $\log h_1$ and $\log h_2$, where Φ is the standard univariate Gaussian cdf and $\sigma > 0$, implies $\mu_h = \log\sqrt{h_1 h_2}$ and $\sigma_h = \sigma^{-1}\log\sqrt{h_2/h_1}$. On the other hand, if ξ_1 and ξ_2 are the curvatures of $\hat{f}_{h_1}$ and $\hat{f}_{h_2}$ in the sense of (4), respectively, taking $\lambda_\xi^{-1} = (\xi_1 + \xi_2)/2$ produces a λ_ξ that is the harmonic mean of the λ parameters corresponding to ξ_1 and ξ_2.

The recommended bandwidth selectors for calculating h_1 and h_2 belong to the PI family of methods PI_r, targeting the r-th derivative for $r = 0, 1, 2$ (Chacón and Duong, 2013). These are robust in an asymptotic sense, avoiding overfitting to $\mathcal{D}$. Namely, we propose taking $(h_1, h_2) = (h_{\mathrm{PI}_0}, h_{\mathrm{PI}_1})$ in a general setting. For

severely discretised data, $(h_1, h_2) = (h_{\mathrm{PI}_1}, h_{\mathrm{PI}_2})$ offers an extra *shield* against spurious modes. In both cases, we propose $\sigma = 1$ to leave room for exploration beyond (h_1, h_2). All in all, the previous configurations are somewhat conservative but conform to the regularising goal of the prior $\Pr(h, \alpha)$.

Other configurations. We recommend the *deviance information criterion* (DIC) (Spiegelhalter et al., 2014) to assess the optimal spline dimension d and knot placement strategy, fixing the spline degree to $r = 3$. Too small d will produce too low likelihood values, whereas a too large d will increase the *effective* number of parameters, i.e., the model complexity. The DIC will generally advise against both extremes. Typical values for d are 22 or 32 (i.e., 21 or 31 knots), depending on the intricacies of the data, far from the hundreds of Fourier series terms in Good and Gaskins (1980).

Generally, the grid size m is far less critical than d and can be held to a constant value such as $m = 1001$. The larger the m, the more accurate the spline approximation but the higher the computational cost.

B.2. **Simulation.** BTS strongly relies on MCMC (Bernardo, 1994). In all three Bayesian inference steps in Section 3, the updated parameters have different support than $\mathbb{R}$. Sampling directly from those posteriors would lead to abnormally low acceptance rates in MCMC.

In the case of $\Pr(h, \alpha|\mathcal{D})$, we recommend applying the change of variables $\bar{h} = \log h$, $\bar{\alpha} = \Phi^{-1}(\alpha)$ to obtain the posterior $\Pr(\bar{h}, \bar{\alpha}|\mathcal{D}) = \Pr(h, \alpha|\mathcal{D}) \cdot e^{\bar{h}} \cdot \phi(\bar{\alpha})$. Then, we can sample from $(\bar{h}, \bar{\alpha})$ and obtain $\mathcal{S}$ after undoing the change of variables. Similarly, for $\Pr(\delta|\mathcal{D})$, we propose taking $\bar{\delta} = \Phi^{-1}[(\delta - \delta_{\min})/(\delta_{\max} - \delta_{\min})]$, which has pdf $\Pr(\bar{\delta}|\mathcal{D}) \propto \Pr(\delta|\mathcal{D}) \cdot \phi(\bar{\delta})$. Finally, for $\Pr(\tau|\mathcal{D}, k, i)$, we suggest $\bar{\tau} = \log \tau$, yielding $\Pr(\bar{\tau}|\mathcal{D}, k, i) = \Pr(\tau|\mathcal{D}, k, i) \cdot e^{\bar{\tau}}$.

Finding a good initial state for MCMC by calculating the *maximum a posteriori* estimator through a small optimisation will ensure the proper behaviour of the posterior. This will prevent MCMC from including outliers in the posterior sample: the so-called *burn-in* period observations that are usually removed (Wagenmakers et al., 2010).

B.3. **Time complexity.** BTS is a compound method. It comprises several stages with algorithms of varied nature, including estimation, optimisation, and simulation. In addition to the intrinsic complexity and size of the input data, each procedure has its configuration options, affecting both the execution time and the precision of the results. On the other hand, most of these algorithms are readily implemented as packages, which eases building BTS from scratch but may lead to small inefficiencies. Considering the above, making a comprehensive time complexity analysis of BTS that honours reality is not easy.

Regarding pdf evaluations, all the Bayesian inferences have complexity $\mathcal{O}(Ns)$, where N is the number of data points, and s is the number of MCMC steps. Except for the testing stage, where N is the number of points in the excess mass region, N coincides with the total sample size n. Then, all the pdfs are compositional splines, which can be efficiently evaluated at a single point using B-splines with complexity $\mathcal{O}(d + r^2)$ (Boor, 1972). However, building the model sometimes has a different cost. In the selection and testing stages, the pdfs (9) and (14) belong to a one-dimensional parametric family, so instantiating the model for each δ and τ is almost immediate. By contrast, in the exploration phase, we must build (6) for each (h, α), consuming much time. The latter step involves several hyperparameters, deserving careful examination.

First, building the CLR grid of the KDE in (4) depends on the sample size n. Naively evaluating (1) at m points has complexity $\mathcal{O}(nm)$ regarding Gaussian

kernel ϕ evaluations. Even if m is held to a moderate constant value, as indicated above, if n is sufficiently large, the execution time may be unaffordable. Using the concept of *binning*, implemented in the package *ks* (Duong, 2022), we can build a discrete approximation to (1) in $\mathcal{O}(n)$ steps that can be evaluated over the grid in $\mathcal{O}(m \log m)$ steps via the FFT algorithm (Gramacki and Gramacki, 2017). Assuming $m \ll n$, the dominant term is $\mathcal{O}(n)$, vastly improving the naive algorithm.

Secondly, solving for the d compositional spline coordinates in (5) requires computations that no longer depend on the sample size once the m-size grid is formed. See Machalová et al. (2020) for further details. Constructing the final linear system matrix has complexity $\mathcal{O}(md)$ in terms of B-spline evaluations, plus $\mathcal{O}(d^2)$ integral calculations for the total curvature penalty component in (4). Finally, using standard implementations, finding the solution to the $d \times d$ linear system has complexity $\mathcal{O}(d^3)$.

Apart from the Bayesian inferences, there is an analysis phase with SFPCA between the exploration and selection stages. See Section F.1 below for further details on SFPCA. Roughly speaking, SFPCA reduces to vanilla PCA and, subsequently, to solving an eigenvalue problem followed by a linear system over $d \times d$ matrices, yielding a complexity $\mathcal{O}(d^3)$ in both cases.

Beyond the sample size and the rest of the hyperparameters, BTS hides a strong dependency regarding the intricacies of the data. The more *ambiguous* the NoM in the data and the higher the maximum predicted NoM, the more calculations are necessary. In some cases, the computational cost increase is negligible. For instance, if many k values have non-null posterior probability (12), it is more expensive to compute the maximum (13). However, letting K be the maximum $k \in \mathbb{N}$ with non-zero probability after the selection phase, the testing stage consists of $K(K+1)/2$ inferences in the worst-case scenario, a complexity $\mathcal{O}(K^2)$, corresponding to testing the k modes of $\tilde{f}_k$ for $k \in \{1, \ldots, K\}$.

As the last remark reminds us, BTS is especially suited for data exploration in an interactive environment, possibly with the assistance of a graphical tool. Employing BTS from start to finish in *batch* mode, as we have done in this paper, evenly demonstrates the value of the proposal compared to other alternatives but misses much of its power. Many of the time complexity issues above can be solved with an expert *hand* making good decisions behind the algorithms. For instance, a low-value d might suffice if spline knots are intelligently placed by hand, dramatically reducing the computational cost. Also, running MCMC chains with as many steps s as we have used here may be unnecessary in practice. Moreover, of course, if analysts are overwhelmed by the complexity of the data, they may use the graphical tool to omit some stages of BTS or discard on the run some modality hypotheses based on expert knowledge.

Appendix C. Extra case study

This annexe completes the exposition of the case study in Section 4.

Data. The pitching data has been retrieved using the R package *baseballr* (Petti and Gilani, 2022). Fig. 9 is taken from the MLB-supported advanced metrics website *BaseballSavant* (Willman, 2023). It shows a mixture pdf model of the pitching speeds by Shohei Ohtani with the four modes advanced in Section 4.

Intermediate results. The results of the exploration phase of BTS are shown in Fig. 10, which is similar to Fig. 3 in the synthetic mixture example from Section 3. Up to 77% of the odds favour three modes, while the remaining 23% correspond to four. The subsequent analysis phase results are gathered in Fig. 11. On one side,

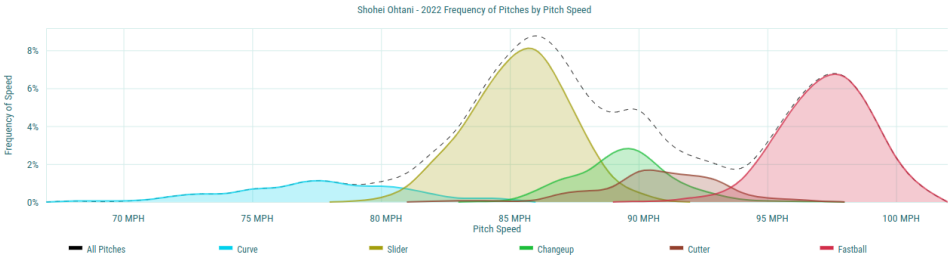


FIG. 9. Mixture pdf model of the pitching speeds by Shohei Ohtani in the 2022 season taken from *BaseballSavant* (Willman, 2023). Pitches of different types are modelled in separate mixture components: *curveballs* (blue), *sliders* (yellow), *changeups* (green), *cutters* (brown), and *fastballs* (red). The overall pdf shows up as the black dashed line.

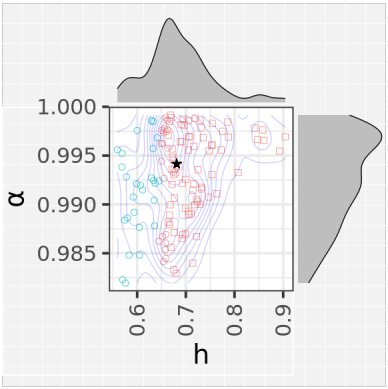


FIG. 10. MCMC sample from the exploration phase of BTS consisting of 120 observations for the MLB case study. The figure follows the structure of Fig. 3. Blue corresponds to four modes (23% of all points) and red to three (77% of the total). Squares and circles refer to three and four modes, respectively.

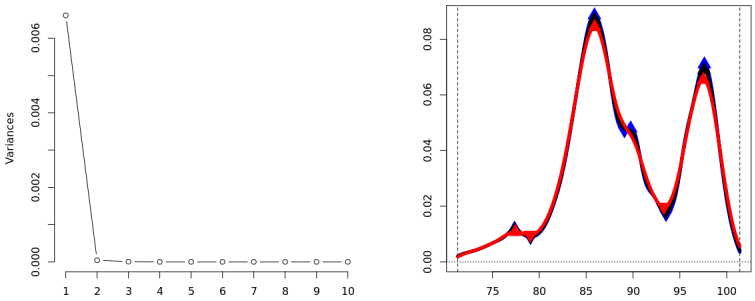
Fig. 11a shows an even more prominent first PC than Fig. 4a. In turn, Fig. 11b is similar to Fig. 7b in that only $\mu \oplus \delta_{\min} \odot \sigma$ captures the elusive 90-mph mode. In Fig. 12a, the Jeffreys prior for the SFPCA model behind Fig. 11b is qualitatively very similar to Fig. 5a, displaying a mild uniform slope that slightly penalises the fourth mode. The posterior sample for the selection phase appears in Fig. 12b, also having a very similar look to Fig. 5b, definitely leaning the odds towards four modes.

APPENDIX D. EXTRA SETUP

This appendix expands on the simulation study setup in Section 5.1.

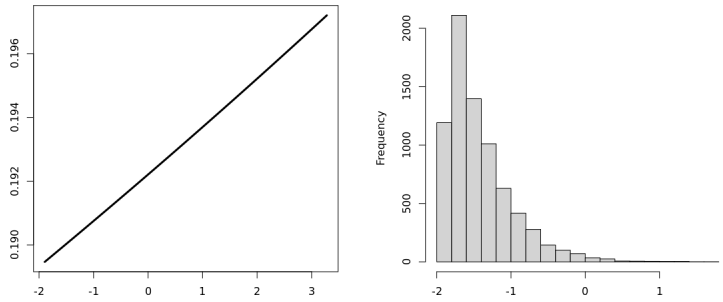
D.1. **Methods.** The following lines justify and explain the methods compared in Section 5.1.

Theoretical grounds. The KDE-based methods `PI0`, `PI1`, `PI2`, `SCV`, `STE`, `LSCV0`, `LSCV1`, `LSCV2` and the mixture-based `GM` are direct *plug-in* approaches, meaning the NoM derives from counting the modes in a fitted pdf model. To do so, we apply the definition in Section 2 of modes as local maxima, restricting the number of pdf evaluations to a sufficiently fine grid over $[a, b]$. Among these direct approaches, `GM` is parametric, whereas the rest are nonparametric. None of them is tailored explicitly for mode estimation.



(A) SFPCA variances scree plot (B) Mode of variation of PC_1

FIG. 11. SFPCA analysis phase results for the MLB case study with the same structure as Fig. 4. The upper bound (red) and the mean (black) pdfs have three modes, while the lower bound (blue) has four.



(A) Jeffreys prior (B) Posterior sample

FIG. 12. Second Bayesian inference on the SFPCA model for the MLB case study with the same structure as Fig. 5. The posterior sample to the right comprises 7,437 observations.

The taut string `TS` method, nonparametric and centred on modality, is also a direct approach but with some additional peculiarities. The continuous and non-uniform definition of mode does not translate well to histogram-like pdfs. In this case, modes should be counted as peaks: the midpoints of bins *taller* than their immediate neighbours. Also, we will interpret a flat taut string, which arguably has no modes, as having modality one, so the minimum NoM will be the same across all direct methods.

Some previous KDEs have been portrayed in Fig. 1b and Fig. 7b. The theory behind the different bandwidth selectors targeting the pdf ($r = 0$) can be found in Wand and Jones (1995). The STE selector is closely related to PI_0, minimising the same loss function but with a slightly different solution scheme (Wand and Jones, 1995, p. 74). The corresponding extensions for density derivative estimation ($r > 0$) are described in Chacón and Duong (2018). Those were included because of the connection of the first and second derivatives with local maxima (Chaudhuri and Marron, 1999, 2002).

To make a fair comparison with BTS, all the KDEs are equipped with the same outlier-filtering preprocessing step described for BTS in Section 3.1, employing an almost negligible mass threshold of 0.001. Data points belonging to modal regions below that mass are discarded when building the final pdf. The preliminary findings leading to the design of BTS evidenced the extreme sensitivity of KDEs to isolated points when estimating the NoM, a rare risk but with a high impact on the results. The vast majority of such spurious modes generated nearly imperceptible modal regions, easily ignored by the human eye. Therefore, we decided to expand all the outlier-prone methods to match the performance a human would get from them under regular operation. This extra help is unnecessary for the rest.

A second group comprises the mode hypothesis testing methods `SI` and `FM`. To obtain the NoM estimate, we iteratively test the null hypothesis that the NoM is less than or equal to k against the alternative of being greater than k, beginning with $k = 1$ and stopping the first time the null hypothesis cannot be rejected. Such an iterative process is customarily used to convert hypothesis testing procedures into estimation ones (Ameijeiras-Alonso et al., 2018, p. 917). The same intermediate significance level $\alpha = 0.05$ is used at every iteration, while the number of bootstrap samples is $B = 500$, being both settings as in Ameijeiras-Alonso et al. (2018).

The excess mass approach is a notable absence among the studied methods because of its high computational cost. Calculating the excess mass statistic has a high asymptotic complexity as the sample size n and the tested k grow. Despite approximations, the execution time under the recommended number of bootstrap samples ($B = 500$) was unworkable for reasonable values $n \geq 1000$ and $k > 1$. Moreover, the risks implied are magnified by the iterative use of the test, making it difficult to limit the total time per task. Therefore, we finally refrained from an exhaustive comparison. Nonetheless, preliminary experiments suggested a similar performance to `FM` in terms of accuracy.

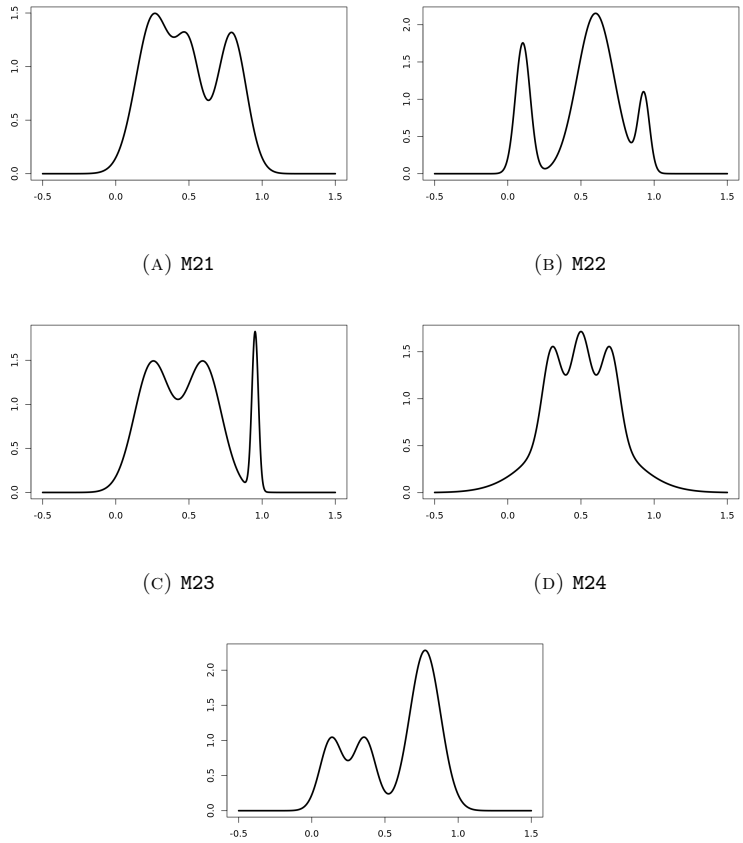
Last but not least, the `EIG` method by Genovese et al. (2016), which we named after the role of the eigenvalues in its multivariate version, is based on very different principles, such as testing modes locally and splitting data into train and test sets. The confidence level for `EIG` was the same as in Genovese et al. (2016).

Software implementation. All the methods under study are coded in the R programming language. The source code for BTS, an R package, shall be distributed under licence and on demand. Our implementation relies on the package *robCompositions* for compositional data analysis (Templ et al., 2009). Our package will also include the source code for `EIG` by Genovese et al. (2016) and the data from the case study.

The rest of the methods are publicly available. `GM` is based on the classical package *mclust* (Scrucca et al., 1999, 2016). `TS` is implemented in *ftnonpar* (Davies and Kovac, 2012). The test-based approaches are provided in the package *multimode* by Ameijeiras-Alonso et al. (2021a) (see also Ameijeiras-Alonso et al., 2021b). Except for `STE`, which corresponds to the default routine `bw.SJ` in R with `method = "ste"`, all the bandwidth selectors are in the package *ks* (Duong, 2022).

D.2. **Test-beds.** The test-bed pdfs in Section 5.1 condense many different shapes, as seen in Fig. 13. In particular, Fig. 13e coincides with the example pdf in Fig. 2a. The five pdfs take the form $x \mapsto \sum_{i=1}^{K} w_i \, \phi((x - \mu_i)/\sigma_i)$, where ϕ is the standard Gaussian pdf, K is the number of mixture components and μ_i, σ_i and w_i are, respectively, the mean, the standard deviation and the weight of the i-th mixture component. Collecting all the mixture model parameters in vectors $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_K^2)$ and $\boldsymbol{w} = (w_1, \ldots, w_K)$, the five mixtures are defined in Table 3.

D.3. **Rankings.** The global and intermediate rankings in Table 1 and Table 2, respectively, are computed with the assistance of the R package *ConsRank* (D'Ambrosio



(A) `M21` (B) `M22`

(C) `M23` (D) `M24`

(E) `M25`

FIG. 13. Test-bed Gaussian mixture model pdfs.

	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}^2$	$\boldsymbol{w}$
`M21`	$(0.26, 0.79145, 0.5)$	$(0.01476, 0.01, 0.007)$	$(0.45, 0.33, 0.22)$
`M22`	$(0.6, 0.10245, 0.93)$	$(0.01588, 0.0025, 0.0015)$	$(0.68, 0.22, 0.1)$
`M23`	$(0.25, 0.6, 0.95222)$	$(0.015, 0.015, 0.00049)$	$(0.45, 0.45, 0.1)$
`M24`	$(0.5, 0.3, 0.5, 0.7)$	$(0.08425, 0.004, 0.004, 0.004)$	$(0.55, 0.15, 0.15, 0.15)$
`M25`	$(0.7749, 0.1345, 0.36)$	$(0.011, 0.006, 0.006)$	$(0.6, 0.2, 0.2)$

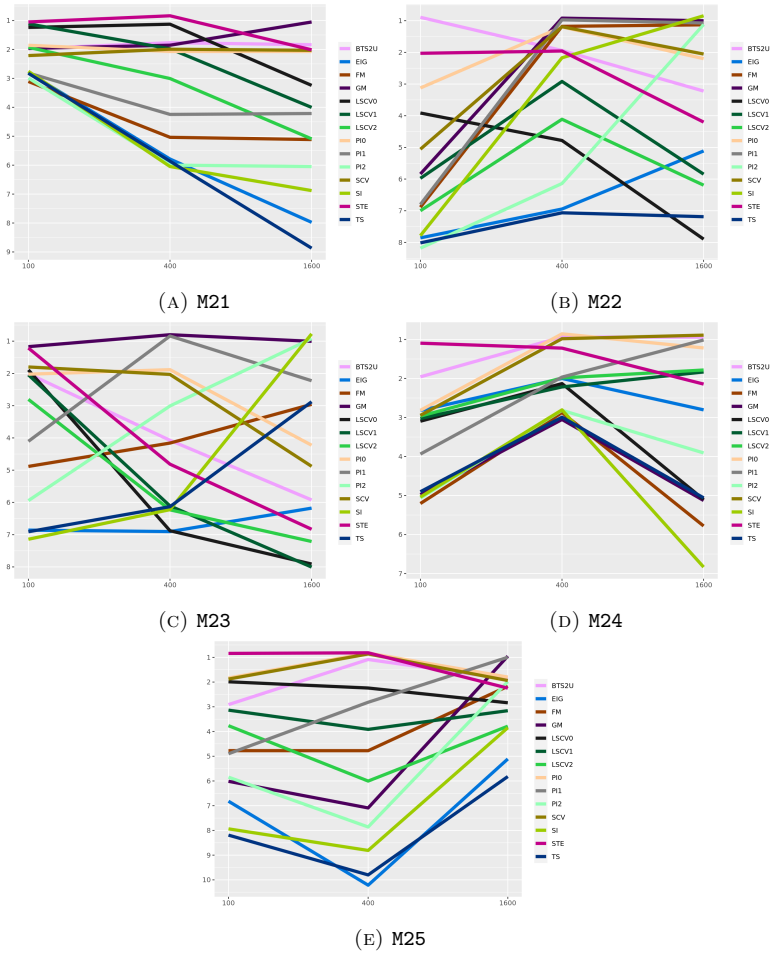
TABLE 3. Test-bed Gaussian mixture model parameters.

et al., 2015). Namely, we used the routine `consrank` with all the default parameters except one for suppressing screen output. In particular, we selected `algorithm = "BB"`, corresponding to the *branch-and-bound* algorithm (Amodio et al., 2016), and `full = FALSE`, meaning ties were allowed among ranks.

APPENDIX E. EXTRA RESULTS

This section is a follow-up of the simulation study in Section 5.2 of the manuscript, including further results and comments.

E.1. **Intermediate rankings.** The rankings $\mathcal{R}_1, \ldots, \mathcal{R}_{T \times S}$ in Table 2 are depicted in Fig. 14. The Kendall correlations between ranks are presented in Fig. 15. With



(A) M21 (B) M22

(C) M23 (D) M24

(E) M25

FIG. 14. Intermediate rankings in Table 2 by test-bed. In every subfigure, the horizontal and vertical axes correspond to the sample size and the rank of the method, respectively. A small amount of jitter has been added to the ranks to appreciate overlapping trajectories better.

these auxiliary representations, let us further analyse the results in Section 5.2 focusing on the less successful methods.

The two second-ranked methods apart from `PI1`, i.e., `LSCV0` and `LSCV1`, also have some success. Contrary to `PI1`, the LSCV methods tend to under-smooth, helping to detect the short-lived mode in Fig. 13a. However, under-smoothing is counter-productive most of the time, leading to spurious modes as n grows. The last LSCV member, the third-ranked `LSCV2`, underperforms across all settings, securing zero first ranks.

The fifth-ranked `PI2` goes deeper into regularisation than `PI1`, becoming too insensitive for most cases (see Fig. 14a, Fig. 14d and Fig. 14e). There are two first ranks in `M22` and `M23`, though, always with $n = 1600$. Especially in `M22`, all the modes are long-lived and well-separated. Hence, the risk of not detecting them is outweighed by that of making up spurious ones. The fourth-placed method, `FM`, also shows top performance in `M22`-400 and `M22`-1600 for the same reason.

Among the sixth-ranked methods, we still see a top performance by `SI` in `M22`-1600 and `M23`-1600. Even though `SI` and `FM` appear highly correlated in Fig. 15, the former is dominated by the latter across most configurations. Nonetheless, both

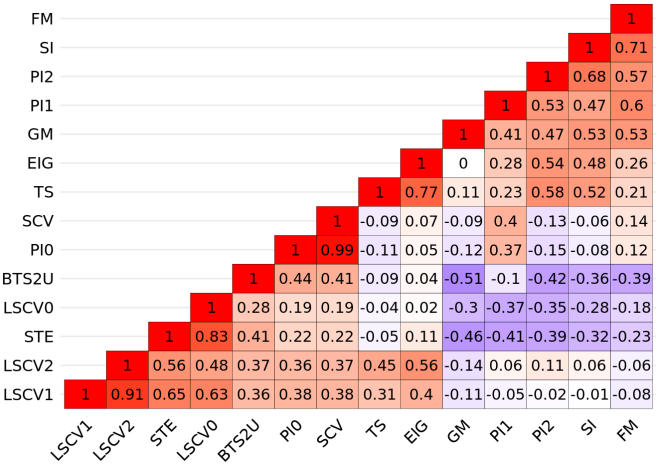


FIG. 15. Kendall correlations between the intermediate ranks in Table 2.

are *niche* methods that excel in some straightforward cases, with large sample sizes and well-separated modes. The same cannot be said about `TS` and `EIG`, which have no first positions and usually close the ranking, as depicted in Fig. 14a, Fig. 14b and Fig. 14e.

Table 2 also reveals noticeable differences among the first-ranked methods. `GM` negatively correlates with the other four, which are positively correlated among themselves. For instance, `GM` stands out with `M23` while the rest struggle (see Fig. 14c) and the opposite happens with `M24`. On the other hand, `PI0` and `SCV` usually perform best with medium to large sample sizes (see Fig. 14b), whereas `STE` operates better with small to medium ones (see Fig. 14a). The three KDEs are more or less even in the middle (Fig. 14d and Fig. 14e). The performance of `BTS2U` is mainly correlated with that of `PI0` and, to a lesser extent, `SCV` and `STE`.

Some curiosities are found in Table 2. `BTS2U` is one of only three methods, alongside `GM` and `STE`, with an unmatched first rank, the `M22-100` mentioned in the manuscript. Additionally, `BTS2U` is the only method not named `GM` capable of sustaining the first rank at least once in each of the three sample sizes. On the other hand, no method ranks first at least once in each of the five test-beds. The most versatile methods in that sense are `GM`, `STE` and `PI1`, with four.

E.2. **Distribution.** We will now look at the distribution of the predicted NoM in each of the $T \times S = 15$ sampling configurations through Fig. 16, Fig. 17, Fig. 18, Fig. 19, and Fig. 20. Here, the reader will see the *raw* accuracies and variabilities of each method in each scenario. We shall concentrate our commentary on some selected cases as the conclusions for the rest are qualitatively very similar.

Fig. 17a shows the top performance mentioned above for `BTS2U`, with a small sample size. `BTS2U` passes `STE` based on the significance of the results. The broad middle block of `M22` traps `STE` and `LSCV0` in over-predicting four modes. Meanwhile, `GM` misses the mode to the right many times. In turn, `PI1`, `FM` and `EIG` behave even more conservatively, while `SI` and `PI2` get stuck at one and two modes, respectively. Interestingly, `FM` extends beyond three modes more than `BTS2U`.

A more complicated scenario is in Fig. 19c despite the large sample size. The model `M24` is like the classic *claw* density used by Davies and Kovac (2004) but with three short-lived *fingers* instead of five. `TS` performs very well with the *claw* but not with `M24`, almost exclusively predicting one mode, similar to `FM`. Overall,

BTS0	BTS1S	BTS1J	BTS1U	BTS2S	BTS2J	BTS2U
2	**1**	**1**	**1**	**1**	**1**	**1**

TABLE 4. Global ranking among BTS variants.

M24 produces high variability, and almost all the methods widely spread their predictions over the one to five range. Having a lower accuracy, BTS2U ranks first, tied with three methods (PI0, PI1 and SCV) based on statistical significance. Very surprisingly, EIG manages to rank third ahead of GM and LSCV0. The latter is *off the chart*, as most of its predictions are above five. Lastly, SI lies at the bottom, predicting one mode 100% of the time.

The variability in Fig. 20b is the lowest of all the three cases considered. On the one hand, M22 is less complex than M25, but data is more scarce in Fig. 17a. On the other, M24 surpasses the complexity of M25, but with larger samples in Fig. 19c. In this case, BTS2U, PI0, SCV and STE rank at first, having very similar profiles between two and four modes. Only the three LSCV methods have predictions over four, while SI, TS and EIG are the only ones betting on one mode. Finally, FM surprisingly outperforms GM, which is excessively conservative.

The previous figures suggest many relevant correlations between the contending methods, as depicted in Fig. 21. We see several groups. BTS2U is similar to PI0, SCV and STE, while all the LSCV variants are highly concordant. A third group, loosely coupled, gathers TS, SI, GM, PI1, PI2 and FM, although PI1 also correlates with PI0. The fourth and final group comprises the outlying EIG, the only method with all its correlations below 0.5.

E.3. **BTS variants.** In Section 5.1, we chose BTS2U as our reference BTS variant to compare with the alternative traditional methods. Here, we will compare all the BTS variants in a similar exercise.

Table 4 shows the global ranking for the BTS variants. BTS0 ranks behind the rest, including BTS2U. The intermediate rankings are reported in Table 5 and depicted in Fig. 22. The rank correlations appear in Fig. 23.

The problem with BTS0 is the sample size: with large samples, it improves considerably, just like PI0 and PI1. On the other hand, the refined versions BTS2S, BTS2J and BTS2U are not significantly better than their respective predecessors, BTS1S, BTS1J and BTS1U.

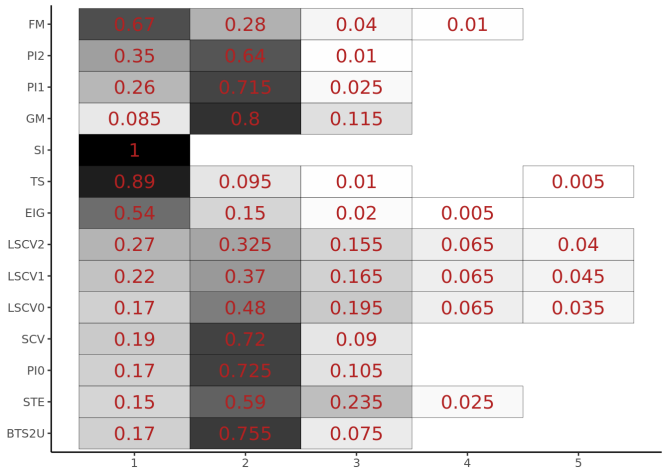
Fig. 24 shows the correlations between the predictions of the methods. The refined versions are highly correlated with their respective base methods. In turn, we observe that the more biased the prior probabilities $\Pr(k)$ are, the higher the correlation with BTS0.

APPENDIX F. DIMENSIONALITY REDUCTION

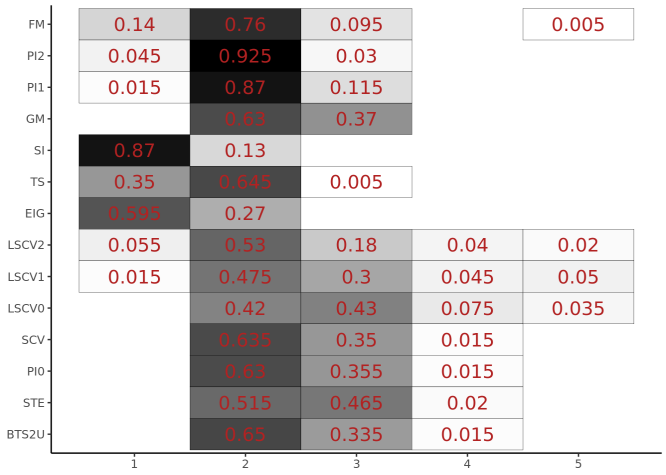
This appendix expands on the dimensionality reduction in the analysis phase of BTS. First, we outline the technicalities of SFPCA. Then, we address the requests of two reviewers for further justification on picking just the first PC.

F.1. **SFPCA.** The following lines discuss SFPCA in our context for $\mathcal{Z}_d[a, b]$. We translate the original theory in Hron et al. (2016) to our notation so the reader can more easily identify the elements involved.

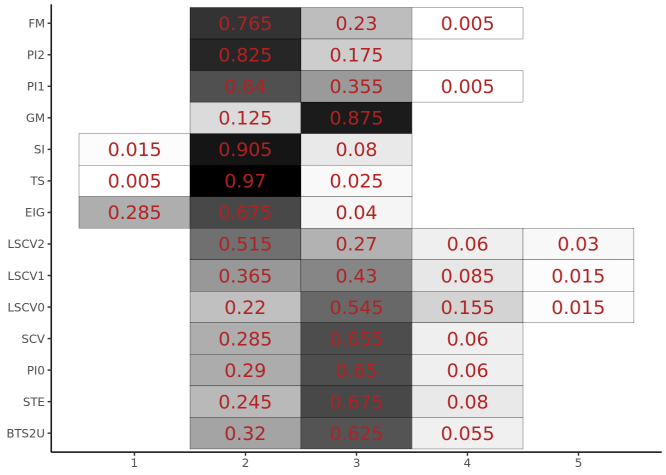
Let us express the CLR of each of the d PCs in terms of ZB-spline basis functions as $\operatorname{clr}[\mathrm{PC}_i] = \sum_{j=1}^{d} \mathbf{b}_i^j Z_j$. That is, each PC_i corresponds to the column vector $\mathbf{b}_i = (\mathbf{b}_i^1, \ldots, \mathbf{b}_i^d) \in \mathbb{R}^d$. Similarly, let us also expand the centred functional data



(A) $n = 100$

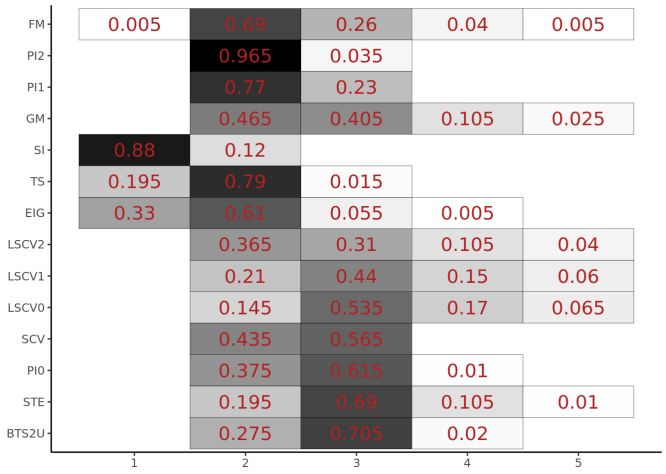


(B) $n = 400$

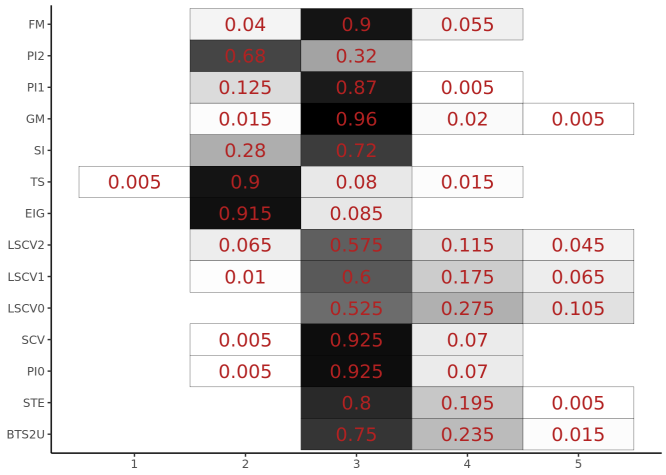


(C) $n = 1600$

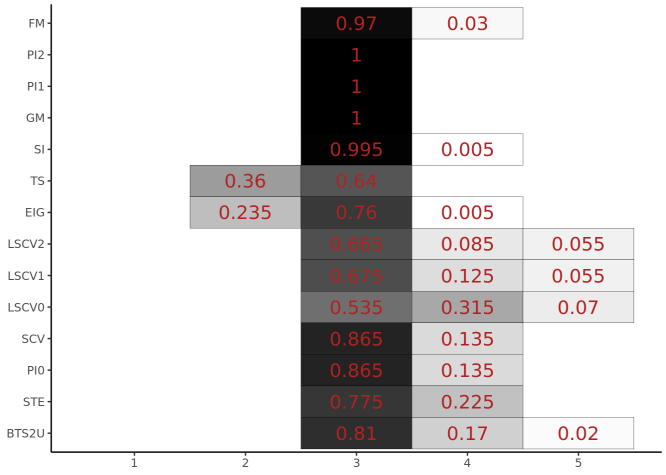
FIG. 16. Distribution of the predicted NoM for the `M21` test-bed for several sample sizes n.



(A) $n = 100$

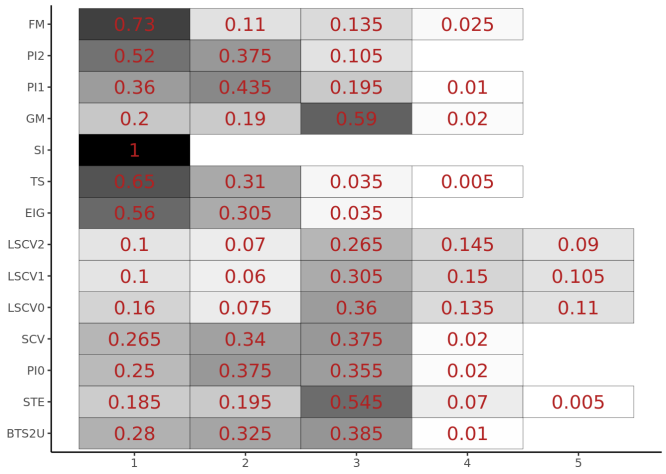


(B) $n = 400$

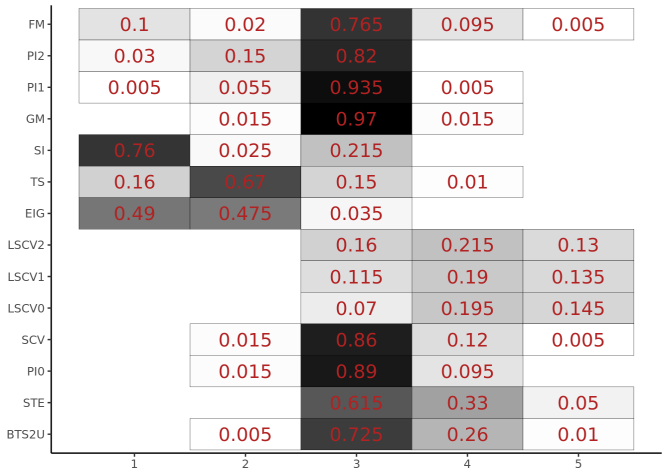


(C) $n = 1600$

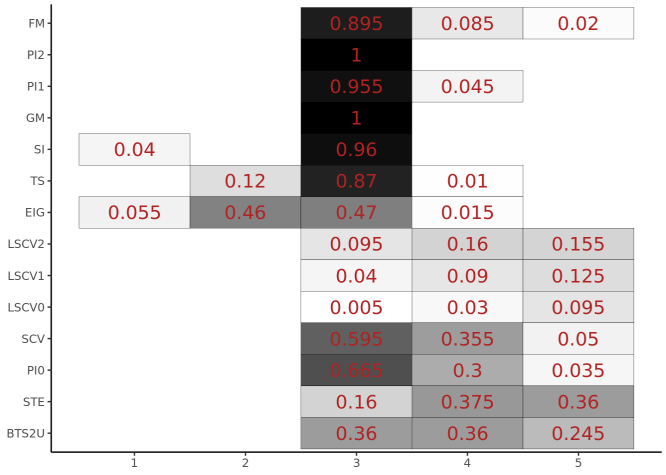
FIG. 17. Distribution of the predicted NoM for the `M22` test-bed for several sample sizes n.



(A) $n = 100$

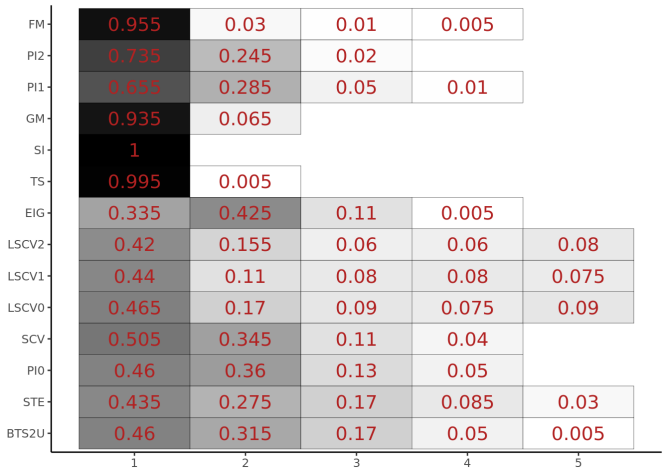


(B) $n = 400$

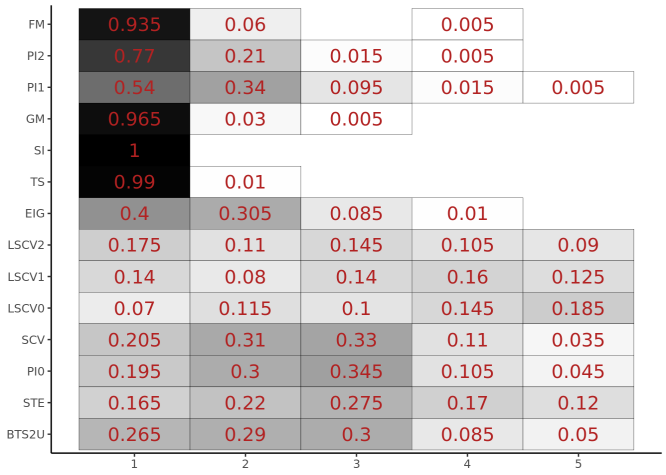


(C) $n = 1600$

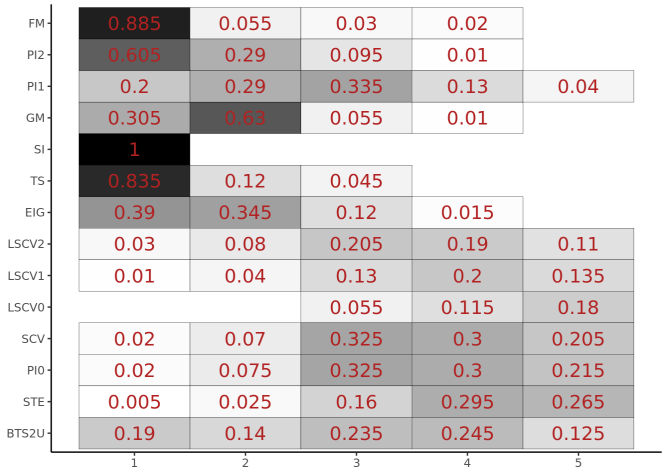
FIG. 18. Distribution of the predicted NoM for the `M23` test-bed for several sample sizes n.



(A) $n = 100$

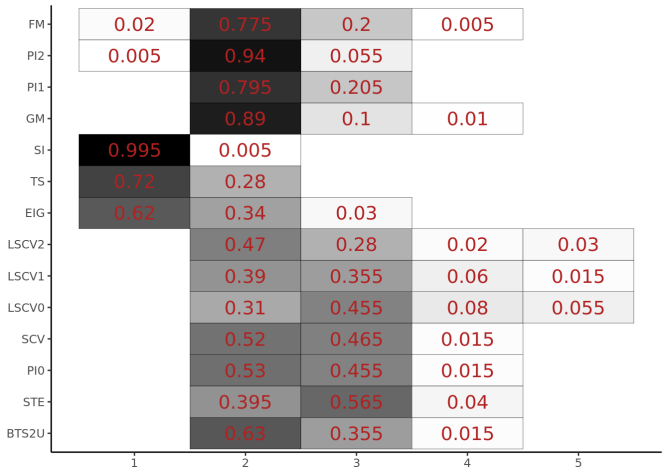


(B) $n = 400$

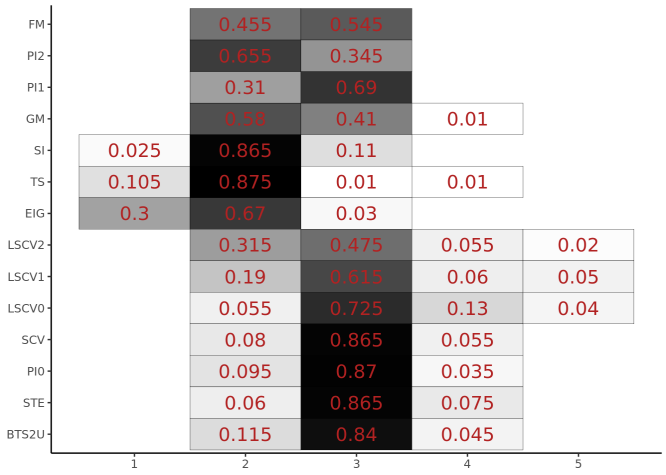


(C) $n = 1600$

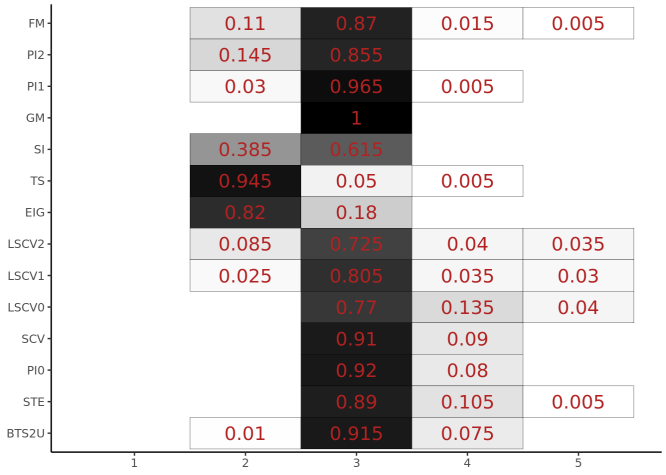
FIG. 19. Distribution of the predicted NoM for the `M24` test-bed for several sample sizes n.



(A) $n = 100$



(B) $n = 400$



(C) $n = 1600$

FIG. 20. Distribution of the predicted NoM for the `M25` test-bed for several sample sizes n.

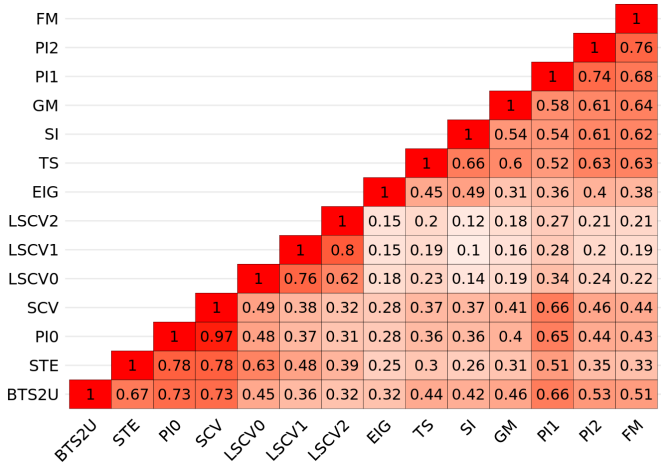


FIG. 21. Kendall correlations between predictions along the $T \times S \times m = 3000$ samples.

		BTS0	BTS1S	BTS1J	BTS1U	BTS2S	BTS2J	BTS2U
M21	100	2	2	**1**	**1**	2	**1**	**1**
M21	400	3	2	2	**1**	2	2	**1**
M21	1600	3	2	2	**1**	2	2	**1**
M22	100	3	2	2	**1**	2	2	**1**
M22	400	**1**	2	2	3	2	2	3
M22	1600	**1**	2	2	3	2	2	3
M23	100	4	3	2	**1**	3	2	**1**
M23	400	**1**	2	2	3	2	2	3
M23	1600	**1**	2	2	2	2	2	2
M24	100	2	2	**1**	**1**	2	**1**	**1**
M24	400	4	3	2	**1**	3	2	**1**
M24	1600	**1**	**1**	**1**	**1**	**1**	**1**	**1**
M25	100	4	3	2	**1**	3	2	**1**
M25	400	2	**1**	**1**	**1**	**1**	**1**	**1**
M25	1600	**1**	2	2	2	2	2	2

TABLE 5. Intermediate rankings by test-bed and sample size configuration among BTS variants.

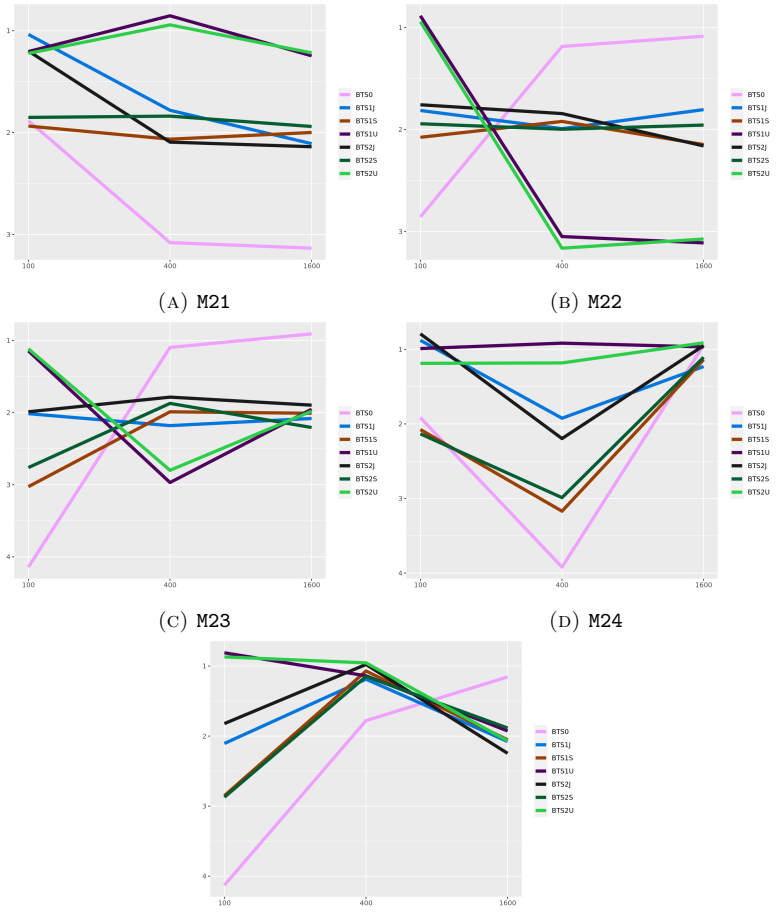
as $\mathrm{clr}[\zeta_{\boldsymbol{\theta}_i} \ominus \mu] = \sum_{j=1}^{d} \mathbf{C}_{ij} Z_j$, i.e., the coordinates forming the rows of the matrix $\mathbf{C} \in \mathbb{R}^{\nu \times d}$.

To obtain $\mathbf{b}_i$, we *first* have to solve the i-th largest eigenvalue λ_i problem

$$\frac{1}{\nu} \mathbf{M}^{1/2} \mathbf{C}^\top \mathbf{C} \mathbf{M}^{1/2} \mathbf{u}_i = \lambda_i \mathbf{u}_i \,, \tag{S1}$$

where $\mathbf{u}_i \in \mathbb{R}^d$ has Euclidean norm one, i.e., $\mathbf{u}_i^\top \mathbf{u}_i = 1$, and $\mathbf{M}^{1/2}$ is the square root of the ZB-spline inner product matrix $\mathbf{M}$ defined in Section 2. Problem (S1) corresponds to the usual principal component analysis in $\mathbb{R}^d$ for the transformed data matrix $\mathbf{C}\mathbf{M}^{1/2}$ (Hron et al., 2016, p. 5) and can be solved directly using the routine `eigen` in R. *Then*, finally, one takes $\mathbf{b}_i$ satisfying $\mathbf{M}^{1/2} \mathbf{b}_i = \mathbf{u}_i$, while the corresponding eigenvalue is λ_i.

Considering the solutions of all the λ_i-problems (S1), it can be easily checked that $\langle \mathrm{PC}_i, \mathrm{PC}_j \rangle_{\mathcal{B}} = \mathbf{b}_i^\top \mathbf{M} \mathbf{b}_j = \mathbf{u}_i^\top \mathbf{u}_j = \delta_{ij}$, the Kronecker delta, meaning the



(A) M21 (B) M22

(C) M23 (D) M24

(E) M25

FIG. 22. Intermediate rankings by test-bed among BTS variants.

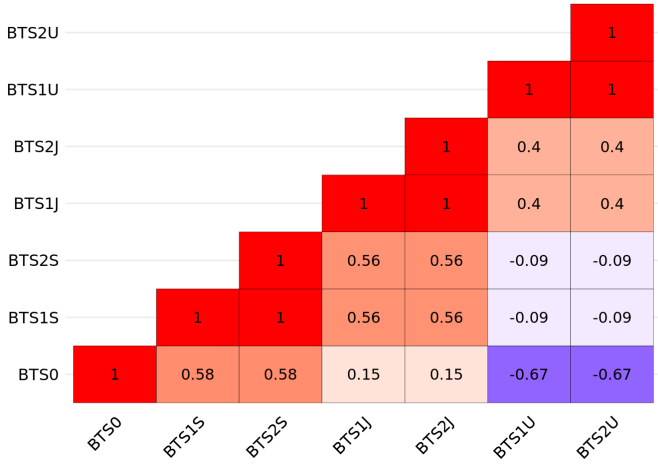


FIG. 23. Kendall correlations between the intermediate ranks in Table 5.

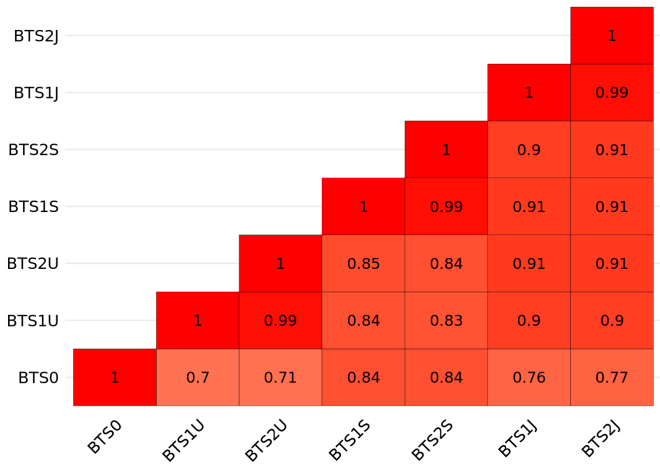


FIG. 24. Kendall correlations between predictions for the NoM along the $T \times S \times m = 3000$ samples for the BTS variants.

PCs form an orthonormal basis. To grasp the true meaning of PCs, specifically their variability-maximising property, we refer the reader to Hron et al. (2016). In particular, the sum of squared scores $\sum_{i=1}^{\nu} s_i^2$ (see Section 3.2) is maximal when a unitary PC_1 is the first PC (Hron et al., 2016, Equation 7).

F.2. **Dimension justification.** In Section 3.2 of the manuscript, we sketched some convenience and simplicity reasons behind keeping just one dimension coincidental with the goals of the analysis phase. Namely, we alluded to making robust inferences and improving interpretability. We also claimed that one PC provided enough power in the BTS context but gave no justification beyond the scree plot in Fig. 4a, complemented in this SM with Fig. 11a. Let us now imagine the consequences of including several PCs.

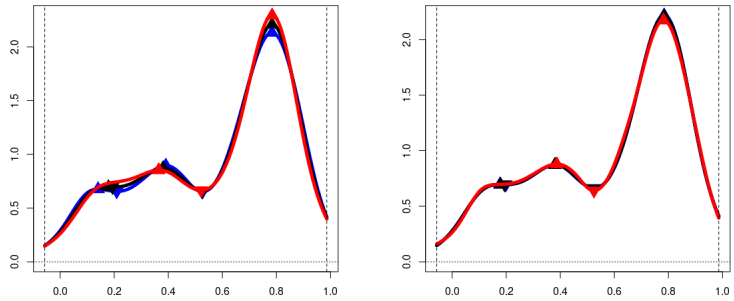
First, the multiparameter generalisation of (9) is

$$\Pr(\cdot|\delta_1, \ldots, \delta_\ell) = \mu \oplus \bigoplus_{i=1}^{\ell} (\delta_i \odot \sigma_i)\,, \tag{S2}$$

where μ stays the same as in (9), $\sigma_i = \sqrt{\lambda_i} \odot \mathrm{PC}_i$, and typically $\ell \ll d$. A similar procedure based on centred projections allows prescribing a rectangular support for the parameter vector $(\delta_1, \ldots, \delta_\ell)$.

Using (S2), new information would enter into our analysis. Though valuable for other purposes, such information would be noisy and redundant for estimating the NoM. By *debugging* samples like that in Fig. 3, we observed that many different but similar pdfs lead to the same NoM. The second and subsequent PCs generally add variability regarding other pdf properties. For instance, in some scenarios, mode A is higher than mode B, while in others, it is the other way around. Similarly, modes A and B appear in slightly different positions in some scenarios. Taking all into account, we would end up with a complex model like (S2), where several δ_i have to be tuned, increasing the computational cost and reducing the robustness of the inference process. Moreover, making sense of such an analysis would be challenging compared to the simplicity of Fig. 6a. In particular, note that the mode tree in Fig. 6a does not say anything about the height of a mode.

The above explanation about redundancy and noise will be more apparent after examining the modes of variation for the subsequent PCs in Fig. 25. First, note that the amplitude of the *oscillations* around the mean gets increasingly weak from



(A) Mode of variation of PC_2 (B) Mode of variation of PC_3

FIG. 25. Modes of variation for subsequent PCs following the example in Fig. 4. Both subfigures have the same structure and colouring conventions as Fig. 4b, but replacing PC_1 with PC_2 and PC_3, respectively. Also, the black curve in both cases is the same as in Fig. 4b, corresponding to the mean μ. The support bounds $\delta_{\min}$ and $\delta_{\max}$ are calculated analogously as for (10), considering PC_2 or PC_3 instead of PC_1. On the left, the lower bound (blue) has three modes, while the upper bound (red) has two. By contrast, the situation in the right-hand side subfigure is unclear, given the small amplitude of the *oscillation* around the mean.

Fig. 4b to Fig. 25a and then Fig. 25b. In fact, the *oscillation* for PC_3 is already virtually imperceptible in Fig. 25b, being PC_3 arguably noisy. The case of Fig. 25a is far more interesting, however. There, we see that the pdf in which the two left-most modes are most pronounced (the blue curve) is also the one in which the right-most mode has the lowest height. That is precisely the opposite case of Fig. 4b, where the three modes grow simultaneously. This means that PC_2 is associated with calibrating the relative height of the modes, which is redundant for estimating the NoM.

There is a natural reason why one PC works so well. One has to consider that the set of functions $\{\zeta_{\boldsymbol{\theta}_i}\}_{i=1}^{\nu}$ we aim to summarise with SFPCA has a relatively low complexity compared to arbitrary datasets in functional data analysis. The random mechanism generating them is comparatively straightforward: a joint use of KDE and compositional splines. The complementary interaction between h and α in Fig. 3 should be relatively simple to capture with a single PC. In this respect, we note that an oblique straight boundary, like in Fig. 3, is standard but not universal, as shown in Fig. 10. In the latter case, the boundary is vertical, indicating an even simpler scenario dominated by h.

Finally, adding more PCs would also be inconvenient for two reasons. First, we would no longer be able to pick a median spline representative $\tilde{f}_k$ of each k-modality hypothesis for the testing phase, at least not directly and efficiently, as in the one-dimensional case. Also, the Jeffreys prior in (10) in the one-dimensional case has the desirable property of being a *reference prior*, which does not happen in the multiparameter setting (Bernardo, 1994).

APPENDIX G. COMPUTING

This section comments on the computational environment used throughout our investigation. Emerging technologies such as *containerisation* and *cloud computing*

are increasingly drawing the attention of the research community as a means to enhance reproducibility.

Containerisation. All our research outputs have been produced within a *Docker* container (Merkel, 2014). The figures and tables were generated upon building a Docker image as part of the installed vignettes of an R package. The `Dockerfile` with the specification of that image will be distributed along the BTS package for R. Next, the built image was made available for the simulation study in a private *cloud* environment through the *Docker Hub* container registry.

Cloud computing. The simulation study was carried out with the help of Microsoft's *Azure Databricks* service (Etaati, 2019). The whole experiment was scheduled as a Databricks *workflow*, where each of the $T \times S = 15$ sampling configurations was a Databricks *task* executing a Databricks R *notebook* with distinct parameters. Then, the $m = 200$ replications took the form of parallelisable Spark *tasks* (Karau and Warren, 2017). The required R files and a JSON file specifying the workflow will be distributed along the BTS package. Finally, the workflow job was assigned to a Databricks cluster with 25 type `Standard_DS3_v2` workers, each counting on four cores. Therefore, 100 threads ran in parallel at any given time for a total execution time of approximately 20 hours.

References

Ameijeiras-Alonso, J. (2017). "Assessing simplifying hypotheses in density estimation". PhD thesis. Universidade de Santiago de Compostela.

Ameijeiras-Alonso, J., Crujeiras, R. M., and Rodríguez-Casal, A. (2018). Mode testing, critical bandwidth and excess mass. *TEST* **28**, 900–19.

Ameijeiras-Alonso, J., Crujeiras, R. M., and Rodríguez-Casal, A. (2021b). Multimode: an R package for mode assessment. *Journal of Statistical Software* **97**.

Amodio, S., D'Ambrosio, A., and Siciliano, R. (2016). Accurate algorithms for identifying the median ranking when dealing with weak and partial rankings under the Kemeny axiomatic approach. *European Journal of Operational Research* **249**, 667–76.

Bernardo, J. M. (1994). Bayesian statistics. *Probability and statistics*. Vol. 2, 345–407.

Boor, C. de (1972). On calculating with B-splines. *Journal of Approximation Theory* **6**, 50–62.

Chacón, J. E. and Duong, T. (2013). Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electronic Journal of Statistics* **7**, 499–532.

Chacón, J. E. and Duong, T. (2018). *Multivariate kernel smoothing and its applications*. Chapman and Hall/CRC.

Chaudhuri, P. and Marron, J. S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association* **94**, 807–23.

Chaudhuri, P. and Marron, J. S. (2002). Curvature vs. slope inference for features in nonparametric curve estimates. *Unpublished manuscript*.

Davies, L., Gather, U., Nordman, D., and Weinert, H. (2009). A comparison of automatic histogram constructions. *ESAIM: Probability and Statistics* **13**, 181–96.

Davies, L. and Kovac, A. (2004). Densities, spectral densities and modality. *The Annals of Statistics* **32**, 1093–136.

Donoho, D. L. (1988). One-sided inference about functionals of a density. *The Annals of Statistics* **16**, 1390–420.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89–121.

Etaati, L. (2019). "Azure Databricks". *Machine learning with Microsoft technologies*. Apress, 159–71.

Genovese, C., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2016). Non-parametric inference for density modes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**, 99–126.

Good, I. J. and Gaskins, R. A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association* **75**, 42–56.

Gramacki, A. and Gramacki, J. (2017). FFT-based fast bandwidth selector for multivariate kernel density estimation. *Computational Statistics & Data Analysis* **106**, 27–45.

Hron, K., Menafoglio, A., Templ, M., Hrůzová, K., and Filzmoser, P. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics & Data Analysis* **94**, 330–50.

Karau, H. and Warren, R. (2017). *High performance spark: best practices for scaling and optimizing Apache Spark*. O'Reilly Media, Incorporated.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–95.

Klugkist, I., Kato, B., and Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica* **59**, 57–69.

Machalová, J., Talská, R., Hron, K., and Gába, A. (2020). Compositional splines for representation of density functions. *Computational Statistics* **36**, 1031–64.

Merkel, D. (2014). Docker: lightweight Linux containers for consistent development and deployment. *Linux J.* **2014**.

Minnotte, M. C. (1997). Nonparametric testing of the existence of modes. *The Annals of Statistics* **25**, 1646–60.

Minnotte, M. C., Marchette, D. J., and Wegman, E. J. (1998). The bumpy road to the mode forest. *Journal of Computational and Graphical Statistics* **7**, 239–51.

Minnotte, M. C. and Scott, D. W. (1993). The mode tree: a tool for visualization of nonparametric density features. *Journal of Computational and Graphical Statistics* **2**, 51–68.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal* **8**, 289–317.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Linde, A. van der (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 485–93.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., and Grasman, R. (2010). Bayesian hypothesis testing for psychologists: a tutorial on the Savage–Dickey method. *Cognitive Psychology* **60**, 158–89.

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*. Springer US.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology* **44**, 92–107.

Resources

Ameijeiras-Alonso, J., Crujeiras, R. M., and Rodríguez-Casal, A. (2021a). *multimode. Mode testing and exploring*. Version 1.5. URL: https://cran.r-project.org/package=multimode.

D'Ambrosio, A., Amodio, S., Mazzeo, G., Albano, A., and Plaia, A. (2015). *ConsRank. Compute the median ranking(s) according to the Kemeny's axiomatic approach*. Version 2.1.0. URL: https://cran.r-project.org/package=ConsRank.

Davies, L. and Kovac, A. (2012). *ftnonpar. Features and strings for nonparametric regression*. Version 0.1-88. URL: https://cran.r-project.org/package=ftnonpar.

Duong, T. (2022). *ks. Kernel smoothing*. Version 1.13.5. URL: https://cran.r-project.org/package=ks.

Petti, B. and Gilani, S. (2022). *baseballr. Acquiring and analyzing baseball data*. Version 1.5.0. URL: https://cran.r-project.org/web/packages/baseballr/index.html.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (1999). *mclust. Gaussian mixture modelling for model-based clustering, classification, and density estimation*. URL: https://cran.r-project.org/package=mclust.

Templ, M., Hron, K., and Filzmoser, P. (2009). *robCompositions. Compositional data analysis*. Version 2.3.1. URL: https://cran.r-project.org/package=robCompositions.

Willman, D. (2023). *BaseballSavant*. URL: https://baseballsavant.mlb.com/.

Authors: José E. Chacón[†] and Javier Fernández Serrano[‡].

[†]Departamento de Matemáticas, Universidad de Extremadura, Badajoz, Spain.
[‡]Departamento de Matemáticas, Universidad Autónoma de Madrid, Madrid, Spain.
E-mail addresses: [†]jechacon@unex.es ✉, [‡]javier.fernandezs01@estudiante.uam.es.
[†]https://orcid.org/0000-0002-3675-1960 .
[‡]https://orcid.org/0000-0001-5270-9941 .

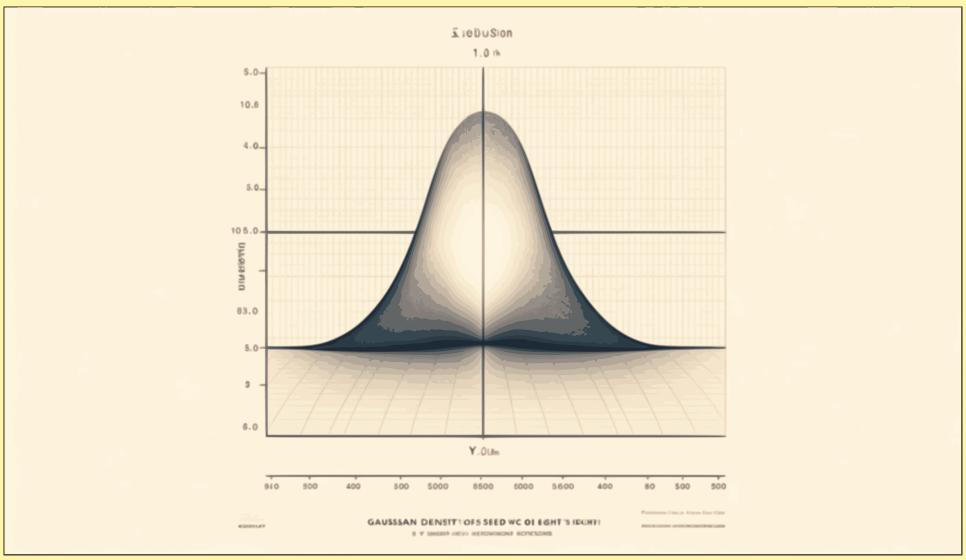

UAM
Universidad Autónoma
de Madrid


UÁM
Universidad Autónoma
de Madrid

https://doi.org/10.1016/j.csda.2024.107961


UAM
Universidad Autónoma
de Madrid


UAM
Universidad Autónoma
de Madrid



Capítulo 3

Centro de simetría

Annals of the Institute of Statistical Mathematics


UÁM
Universidad Autónoma
de Madrid

MODE-BASED ESTIMATION OF THE CENTER OF SYMMETRY

JOSÉ E. CHACÓN[†] AND JAVIER FERNÁNDEZ SERRANO[‡]

Abstract. In the mean-median-mode triad of univariate centrality measures, the mode has been overlooked for estimating the center of symmetry in continuous and unimodal settings. This paper expands on the connection between kernel mode estimators and M-estimators for location, bridging the gap between the nonparametrics and robust statistics communities. The variance of modal estimators is studied in terms of a bandwidth parameter, establishing conditions for an optimal solution that outperforms the household sample mean. A purely nonparametric approach is adopted, modeling heavy-tailedness through regular variation. The results lead to an estimator proposal that includes a novel one-parameter family of kernels with compact support, offering extra robustness and efficiency. The effectiveness and versatility of the new method are demonstrated in a real-world case study and a thorough simulation study, comparing favorably to traditional and more competitive alternatives. Several myths about the mode are clarified along the way, reopening the quest for flexible and efficient nonparametric estimators.

1. Introduction

The mean-median-mode triad is well-rooted in statistical *folklore*. In general, these three summaries depict different notions of centrality and have different properties, so the relative preference for either of them is dependent on the context. Even so, the mode has certainly received less attention in the literature, perhaps because it poses a more challenging estimation problem and also due to definition technicalities in the continuous setting. In any case, the concept of mode has recently emerged as a solution to several seemingly unrelated statistical problems. See Chacón (2020) for a comprehensive overview.

It is well known that these three centrality measures coincide when assuming symmetry and unimodality, in which case they all represent the so-called *center of symmetry* of the distribution. Indeed, under symmetry, the mean and the median have long been studied as location statistics, and several attempts have been made to find robust intermediate estimators (Huber, 1964). Nevertheless, although the mean and the median are known to underperform in fat-tailed scenarios (Lai et al., 1983), the mode has been overlooked for estimating the center of symmetry.

On the other hand, there is abundant literature on nonparametric estimation of the mode (Chacón, 2020), particularly in kernel density estimation (Parzen, 1962). Very accurate results describing the asymptotic properties of the kernel-based mode estimator can be found in the exhaustive studies of Romano (1986, 1988a,b). However, symmetry has relevant implications in the analysis of this estimator, which have not been exploited before. Interestingly, symmetry makes

[†]Departamento de Matemáticas, Universidad de Extremadura, Badajoz, Spain.
[‡]Departamento de Matemáticas, Universidad Autónoma de Madrid, Madrid, Spain.
E-mail addresses: [†]jechacon@unex.es ✉, [‡]javier.fernandezs01@estudiante.uam.es.
2020 *Mathematics Subject Classification.* 62G05 (Primary), 62G07, 62G35.
Key words and phrases. kernel mode estimator, center of symmetry, unimodality, regularly varying density, redescending M-estimator, efficient nonparametric estimation.
[†]https://orcid.org/0000-0002-3675-1960 .
[‡]https://orcid.org/0000-0001-5270-9941 .

the bias of the kernel-based mode estimator vanish (Chernoff, 1964, p. 41; Eddy, 1982, Section 4), allowing us to focus on reducing variance.

This paper aims to bridge the gap between robust location statistics and kernel density estimation theory for mode estimation in the symmetric, unimodal case. Theoretical asymptotic results and finite-sample simulations will demonstrate the effectiveness of the modal approach. Our findings also expand on the connection noted by Eddy (1982, Section 4) with the theory of M-estimators (Huber, 1964; Huber and Ronchetti, 2009; Lehmann and Casella, 1998, p. 484), which elegantly subsumes the three classic centrality points of view.

The rest of the paper is organized as follows. The links between the two communities (robust statistics and nonparametrics) are briefly reviewed in Section 2, where the fundamental notation is also introduced. Section 3 presents original theoretical results and a novel proposal for mode-based estimation of the center of symmetry. Proofs are given in Appendix A. The practical effectiveness of our method is demonstrated through a case study with real-world data in Section 4 and a thorough simulation study in Section 5. Finally, Section 6 summarizes and discusses the relevance of our research.

2. Background

Estimating the mode of an unknown univariate probability density function (pdf) f has arguably been a foundational goal for nonparametrics, being the kernel density estimator (KDE) one of the most popular tools (Grenander, 1965; Parzen, 1962). Given a sample of independent and identically distributed (i.i.d.) absolutely continuous random variables $X_1, \ldots, X_n$ with common pdf f, a *bandwidth* $h \in (0, \infty)$, and a *kernel* function $K : \mathbb{R} \to \mathbb{R}$ (typically a symmetric, unimodal pdf), the KDE is defined as

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

The KDE converges locally and globally to f as $n \to \infty$ if $h \equiv h_n \to 0^+$ and $nh \to \infty$, provided certain mild assumptions are satisfied (Chacón and Duong, 2018).

In our context, the mode will be the unique maximizer $\mathsf{m} \in \mathbb{R}$ of f, i.e.,

$$\mathsf{m} = \arg\max_{x \in \mathbb{R}} f(x). \tag{2}$$

Assuming natural conditions on K, there is a random variable $\hat{\mathsf{m}}_{n,h}$, which Parzen (1962) calls the *sample mode* and which we shall refer to as the *kernel mode estimator* (KME), such that

$$\hat{f}_{n,h}(\hat{\mathsf{m}}_{n,h}) = \max_{x \in \mathbb{R}} \hat{f}_{n,h}(x). \tag{3}$$

In general, there are several versions of $\hat{\mathsf{m}}_{n,h}$. Eddy (1980) and Romano (1988b) pick the smallest argument that maximizes the KDE when the optimum is not unique. Romano claims all his results hold for any random variable complying with (3). In turn, Grund and Hall (1995) directly state their results by *breaking* ties at random.

On the other hand, M-estimators (Huber, 1964; Huber and Ronchetti, 2009) of the location parameter $\theta \equiv \theta_F \in \mathbb{R}$ of a cumulative distribution function F are defined as sample random variables $\hat{\theta}_n$ minimizing

$$\sum_{i=1}^{n} \rho(X_i - \hat{\theta}_n), \tag{4}$$

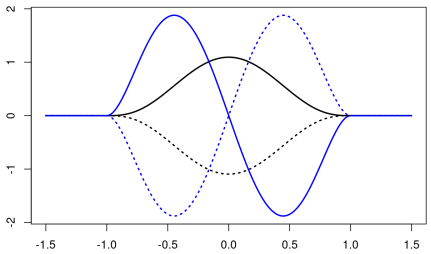


FIG. 1. The M-estimator perspective of KMEs. A kernel function K and its derivative are shown in solid black and blue, respectively. The dashed black and blue lines are, respectively, $\rho(x) = -K(x)$, and $\psi(x) = \rho'(x) = -K'(x)$.

for some fixed function $\rho : \mathbb{R} \to \mathbb{R}$. Or equivalently, if ρ is differentiable with $\psi = \rho'$, through

$$\sum_{i=1}^{n} \psi(X_i - \hat{\theta}_n) = 0\,. \tag{5}$$

For instance, the sample mean $\hat{\mu}_n$ and the sample median $\hat{M}_n$ are the M-estimators corresponding to $\rho(x) = x^2$ and $\rho(x) = |x|$, respectively. Moreover, the population target θ—the mean μ and the median M in the previous examples—can also be expressed in terms of ρ and ψ by cautiously replacing sums with expectations (Huber and Ronchetti, 2009, pp. 46–47).

Regarding the mode, Eddy (1982) mentioned the link between the KME and M-estimators in his concluding remarks, but he did not further develop the idea. Indeed, given (1) and (3), for a symmetric kernel, the KME $\hat{\mathsf{m}}_{n,h}$ can be seen as the M-estimator corresponding to $\rho(x) = -K_h(x) = -K(x/h)/h$. For fixed $h > 0$, the population target of the KME would be the value $\bar{\mathsf{m}}_h \in \mathbb{R}$ maximizing $\bar{f}_h(x) = \mathbb{E}[\hat{f}_{n,h}(x)]$. In general, $\bar{\mathsf{m}}_h \neq \mathsf{m}$, and only as $h \to 0^+$ the bias goes to zero. However, under the symmetry of K (about zero) and f (about m), there is no bias effect (Chernoff, 1964), so that $\bar{\mathsf{m}}_h = \mathsf{m}$ (this is rigorously stated and proved in Theorem 1 below). Hence, the KME would effectively be an M-estimator of the mode. See in Fig. 1 the M-estimator perspective of KMEs.

For most common kernels, KMEs are members of the class of *redescending* M-estimators, which are those characterized by a function ψ that tends to zero as $|x| \to \infty$ (Huber and Ronchetti, 2009; Maronna et al., 2019). Surprisingly, neither of these celebrated robust statistics books mentions KMEs. The importance of redescending M-estimators lies in their resilience to *outliers*, especially when ψ has compact support (in the case of KMEs, when K has compact support). See, for instance, the concepts of *influence function* and *breakdown point* in the former references. Nevertheless, despite their compelling features, redescending M-estimators have also received some criticism. Huber and Ronchetti (2009, p. 101) point out an efficiency decrease and a higher sensitivity to wrong *scaling*, which in the case of KMEs translates into the importance of a good choice of the bandwidth h.

Consequently, one of the main goals of this paper is to study how the bandwidth affects the performance of the KME in this scenario and propose a practical data-driven choice.

3. Method

Section 3.1 introduces the main theoretical results. Section 3.2 and Section 3.3 address the two steps of our KME proposal. Finally, Section 3.4 illustrates our method with a *toy* example.

3.1. **Theoretical results.** For a fixed bandwidth $h > 0$, the population-wide modal location statistic corresponding to (1) as $n \to \infty$, denoted $\bar{\mathsf{m}}_h$, is obtained by maximizing the *smoothed pdf* $\bar{f}_h$ (Wand and Jones, 1995, Eq. 2.4) defined by

$$\bar{f}_h(x) = \mathbb{E}[\hat{f}_{n,h}(x)] = \int_{-\infty}^{\infty} K_h(x-y) f(y)\, dy = (K_h * f)(x)\,, \tag{6}$$

which does not depend on the sample size n. If K is a proper pdf, so it is (6).

As noted in Section 2, the *smoothed* mode $\bar{\mathsf{m}}_h$ does not generally coincide with the true mode m of f. To change that situation, we need to assume the following two hypotheses.

Definition 1 (Symmetry). An absolutely continuous random variable X with pdf f is *symmetric* about a center of symmetry $\theta \in \mathbb{R}$ if $X - \theta$ and $\theta - X$ have the same distribution, or, equivalently, if $f(\theta + x) = f(\theta - x)$ for all $x \in \mathbb{R}$.

Definition 2 (Unimodality). An absolutely continuous random variable X with pdf f is *unimodal* about a mode $\mathsf{m} \in \mathbb{R}$ if $f_0(x) = f(\mathsf{m} + x)$ is a strictly decreasing function of $|x|$ over its support.

The definitions of symmetry and unimodality can be found in p. 22 and p. 228 of Maronna et al. (2019), respectively. In both cases, we shall extend these terms to pdfs rather than only random variables. In particular, the kernel K in (1) will typically be, in our context, a symmetric, unimodal pdf about zero. There is a wide array of symmetric, unimodal pdfs, as shown in Fig. 2. Indeed, symmetry arises as a natural assumption in the location model $X = \theta + \varepsilon$, where ε represents the error variable, to formalize the idea that there are no systematic errors (Maronna et al., 2019, p. 17). Moreover, symmetry and unimodality of the error term appear as key assumptions in modern modal regression (Wang, 2024).

The symmetry assumption implies $\mu = M = \theta$ (Maronna et al., 2019, p. 22). However, note that μ is not always defined, whereas M is. This happens, for instance, in the case of the Cauchy distribution, that is, Student's t with one degree of freedom. If, in addition to symmetry, the maximizer (2) is assumed to be unique, then it is easy to see that $\mathsf{m} = \theta$ as well. Next, we show that the stronger condition of unimodality ensures that $\bar{\mathsf{m}}_h = \theta$ for every $h > 0$, making the KME $\hat{\mathsf{m}}_{n,h}$ an M-estimator of θ, as suggested by Chernoff (1964) and Eddy (1982).

Theorem 1. *Assume that a pdf f is symmetric and unimodal about θ. Also, consider a differentiable pdf kernel K with bounded K' that is symmetric and unimodal about zero. Then, the smoothed pdf $\bar{f}_h$ is also symmetric and unimodal about θ for any $h > 0$.*

The fact that the *convolution* of two symmetric and unimodal distributions—the case of (6)—is also symmetric and unimodal is well known (Purkayastha, 1998). However, the concept of unimodality employed in that reference slightly differs from the usual one in robust statistics (see Maronna et al., 2019, Theorem 10.2), covering a broader range of distributions but producing a weaker result than required. Theorem 1 is tailored to our needs and has a more straightforward elementary proof.

The role of the bandwidth h in Theorem 1 seems minor, permanently attached to the kernel in K_h. One could think of h as incorporated *inside* K. Carrying

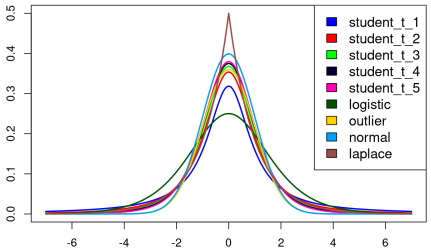


FIG. 2. Several examples of symmetric, unimodal pdfs (about zero). Instances of the Student's t family (Maronna et al., 2019) with $\nu \in \{1, ..., 5\}$ appear in dark blue, red, light green, black, and magenta, respectively. The *standard* normal, Laplace, and logistic pdfs appear in light blue, brown, and dark green, respectively. See Maronna et al. (2019) for the former two and Huber and Ronchetti (2009, Example 3.13) for the latter. Finally, the *outlier* pdf, shown in yellow, is defined as a Gaussian mixture with two components having the same mean (zero) but very different variances. For further details, see Section 5 below, where all these pdfs will be used as test-beds in a simulation study.

h responds to historical reasons in the kernel smoothing literature. However, as we shall see, fine-tuning h will make the difference in obtaining efficient kernel estimators of the center of symmetry.

Theorem 1 ensures that the population-wide target maximizing (1) for a fixed bandwidth h is unique and equal to the center of symmetry θ. According to Maronna et al. (2019, Theorem 10.5), the latter guarantees consistency, namely convergence in probability of the KME $\hat{\mathsf{m}}_{n,h}$ to θ as $n \to \infty$. The following intuitive result shows that the KME is also *on target* for every finite sample size n.

Theorem 2. *In the hypotheses of Theorem 1, if we assume that ties are broken uniformly at random in the condition (3), then $\hat{\mathsf{m}}_{n,h}$ is symmetric about θ for every $n \in \mathbb{N}$ and $h > 0$. Consequently, if $\hat{\mathsf{m}}_{n,h}$ is integrable, then it is an unbiased estimator of θ, i.e., $\mathbb{E}[\hat{\mathsf{m}}_{n,h}] = \theta$.*

Let us now turn to the asymptotic distribution of the KME. To make arguments shorter and simpler, we shall work with *bell-shaped* kernels.

Definition 3 (Bell-shaped kernel). Consider a twice continuously differentiable pdf kernel K that is symmetric and unimodal about zero, with bounded K' and K''. We say that K is *bell-shaped* if K has exactly two inflection points at $\pm a$, for some $a > 0$. That is, $\pm a$ are the only points where K'' changes its sign.

Definition 3 adds further regularity conditions to the ones imposed in Theorem 1. We require additional smoothness and assume that K is concave over $|x| < a$ (i.e., $K''(x) \leq 0$) and convex over $|x| > a$ (i.e., $K''(x) \geq 0$). One can easily see that the convexity and unimodality of K forces $\lim_{x\to\infty} K'(\pm x) = 0$, making the KME with a bell-shaped kernel a redescending M-estimator. The most common smooth kernels, such as the Gaussian, are bell-shaped, so Definition 3 is not too restrictive. Also, note that if K is bell-shaped, so it is K_h for any $h > 0$.

Using Definition 3, we can state the following asymptotic normality result. It can be seen as a particular case of a general asymptotic distribution result for M-estimators (see Maronna et al., 2019, Theorem 10.7), adapted for KMEs in their specific setting. The notations K_h' and K_h'' should be interpreted as successive derivatives of K_h.

Theorem 3. *Assume that $X \sim f$ is symmetric and unimodal about θ. Consider a bell-shaped kernel K. Then, for any $h > 0$, $\sqrt{n}(\hat{\mathsf{m}}_{n,h} - \theta) \rightsquigarrow \mathcal{N}(0, \sigma^2_{f,K,h})$ as $n \to \infty$, with variance*

$$\sigma^2_{f,K,h} = \frac{\mathbb{E}[K_h'(X-\theta)^2]}{\mathbb{E}[K_h''(X-\theta)]^2} = \frac{\int_{-\infty}^{\infty} K_h'(x-\theta)^2 f(x)\, dx}{[\int_{-\infty}^{\infty} K_h''(x-\theta) f(x)\, dx]^2} = \frac{[(K_h')^2 * f](\theta)}{(K_h'' * f)(\theta)^2}\,. \tag{7}$$

Although the result of Theorem 3 is not particularly surprising, its implications are significant in understanding and optimizing the performance of the KME. The variance (7) has an interesting behavior as $h \to 0^+$ and $h \to \infty$ that has not been addressed in the current robust statistics literature.

Theorem 4. *In what follows, assume the hypotheses of Theorem 3.*

(1) Let f be twice continuously differentiable with bounded f' and f'' integrable. Let K be such that $\lim_{x\to\infty} xK(x) = 0$ and $\mathcal{R}(K') = \int_{-\infty}^{\infty} K'(x)^2\, dx$ is finite. Then, $\lim_{h\to 0^+} \sigma^2_{f,K,h} = \infty$. Moreover, if $f''(\theta) < 0$, and if we define $\sigma^2_{f,K} = |f''(\theta)|^{-2} f(\theta)\mathcal{R}(K')$, then $\sigma^2_{f,K,h} \sim \sigma^2_{f,K} h^{-3}$ as $h \to 0^+$.

(2) If f has finite variance σ^2, and $K''(0) < 0$, then $\lim_{h\to\infty} \sigma^2_{f,K,h} = \sigma^2$.

Part 1 of Theorem 4 is compatible with the observation by Huber and Ronchetti (2009) that redescending M-estimators are more sensitive to wrong *scaling*, as a too-small h dangerously increases the variance. Part 2 of Theorem 4 shows that KMEs are asymptotically nearly as efficient as the sample mean if h is sufficiently large.

Theorem 3 is a fixed-bandwidth version of Romano (1988a, Theorem 2.1). He states, roughly speaking, that $\sqrt{nh^3}(\hat{\mathsf{m}}_{n,h} - \bar{\mathsf{m}}_h) \rightsquigarrow \mathcal{N}(0, \sigma^2_{f,K})$ as $n \to \infty$ and $h \to 0^+$, where $\sigma^2_{f,K}$ is as in Theorem 4 but replacing θ with m. The connection between our result and Romano's derives from the fact that $\bar{\mathsf{m}}_h = \mathsf{m} = \theta$ under symmetry and unimodality.

Heavy-tailed distributions exist in which the KME is guaranteed to outperform the sample mean. To model the tail behavior of the distribution we rely on the concept of regularly varying functions (Seneta, 1976).

Definition 4 (Regularly varying function). A function $f : [x_0, \infty) \to (0, \infty)$, for $x_0 > 0$, is *regularly varying* with index $\alpha \in \mathbb{R}$ if, for all $\lambda > 0$,

$$\lim_{x\to\infty} \frac{f(\lambda x)}{f(x)} = \lambda^\alpha\,.$$

If $\alpha = 0$, f is said to be *slowly varying*.

In symmetric pdfs, the regular variation behavior is the same in both tails and requires $\alpha < -1$ since otherwise f would not be integrable (see the assumptions in Devroye and Györfi, 1985, Section 9.3, Theorem 2). An interesting consequence of the regularly varying definition is the universal representation $f(x) = x^\alpha L(x)$, where L is a slowly varying function (Seneta, 1976). The most paradigmatic regularly varying distributions belong to Student's t family (Maronna et al., 2019, Eq. 2.9), with pdf

$$f(x) \propto \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}\,, \tag{8}$$

parameterized by the *degrees of freedom* $\nu > 0$, corresponding to the regular variation index $\alpha = -(\nu + 1)$. The limit of (8) as $\nu \to \infty$ is the standard normal pdf, which is not regularly varying.

Theorem 5 below specifies conditions under which regularly varying pdfs, particularly Student's t pdfs, have a KME that is more efficient than the sample mean.

Theorem 5. *Assume that a symmetric and unimodal pdf f, with center θ, is such that $f_0(x) = f(\theta + x)$ is regularly varying with index $\alpha < -3$ and bounded slowly varying component L satisfying $\lim_{h\to\infty} L(h) = \ell > 0$. Suppose that a kernel K satisfies the hypotheses of Theorem 4, part 2, and the following two integral conditions: (i) $I_{K,\alpha} = \int_0^\infty x^\alpha [|K''(0)|^{-2} K'(x)^2 - x^2]\, dx$ is finite and negative, and (ii) $\int_0^\infty x^\alpha [1 + |K''(0)|^{-1} K''(x)]\, dx$ is finite. Then, $\lim_{h\to\infty} h^{-(\alpha+3)}(\sigma^2_{f,K,h} - \sigma^2) = 2\ell I_{K,\alpha}$. Consequently, for every sufficiently large h, we have $\sigma^2_{f,K,h} < \sigma^2$.*

Remark 1. In Theorem 5, if K has compact support $[-1, 1]$, we have:

(i) For the first integral condition, if $\bar{I}_{K,\alpha} = \int_0^1 x^\alpha [|K''(0)|^{-2} K'(x)^2 - x^2]\, dx$, then $I_{K,\alpha} = \bar{I}_{K,\alpha} + (\alpha + 3)^{-1}$. Consequently, it suffices to check that $\bar{I}_{K,\alpha}$ is finite and $\bar{I}_{K,\alpha} < -(\alpha + 3)^{-1}$. Since $\alpha < -3$ by hypothesis, the latter inequality is satisfied if $\bar{I}_{K,\alpha} < 0$.

(ii) For the second integral condition, since $\alpha < -3$ by hypothesis, it suffices to check that $\int_0^1 x^\alpha [1 + |K''(0)|^{-1} K''(x)]\, dx$ is finite.

Theorem 4 and Theorem 5 together allow searching for an h that minimizes the variance $\sigma^2_{f,K,h}$, following a similar scheme to that of Chacón et al. (2007).

Corollary 1. *Assume all the hypotheses in Theorem 4 and Theorem 5. Define $\mathcal{V} : (0, \infty) \to (0, \infty)$ given by $\mathcal{V}(h) = \sigma^2_{f,K,h}$. There exists $h^\star \in (0, \infty)$ such that $\mathcal{V}(h^\star) = \inf_{h\in(0,\infty)} \mathcal{V}(h) < \sigma^2$.*

Remark 2. Of course, if $\sigma^2 = \infty$, then any KME is more efficient than the sample mean, regardless of h.

Finding a kernel that satisfies the hypotheses of Theorem 5 is not trivial, especially one covering every $\alpha < -3$. The Epanechnikov kernel, defined as $K_{\mathrm{Ep}}(x) = 3(1 - x^2)/4 \cdot \mathbb{1}_{(-1,1)}(x)$ (see Eddy, 1980, p. 875), accomplishes the two integral conditions, but it is not differentiable. Though differentiable, other polynomial kernels, such as the *triweight* (or Tukey's *biweight* ψ, in the M-estimator parlance of Huber and Ronchetti, 2009, Eq. 4.92), only cover a limited range of α. In turn, the household Gaussian kernel is bell-shaped but only satisfies the first integral condition. We present now a parametric family of kernels compatible with the assumptions of Theorem 5 and whose parameter can be tuned to reach every α.

Theorem 6. *Define the family of bump-like functions B_β, with $\beta > 0$,*

$$B_\beta(x) = \begin{cases} e^{-1/(1-|x|^\beta)}, & \text{if } |x| < 1 \\ 0, & \text{otherwise} \end{cases}. \tag{9}$$

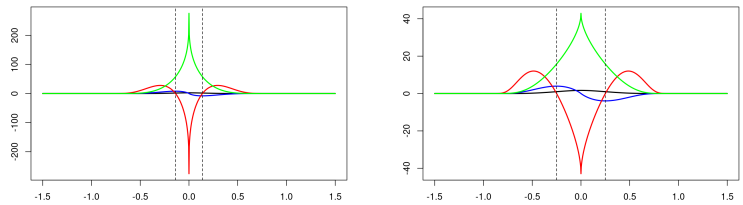
Then, consider the family of pdf kernels K_β whose derivative is given by

$$K'_\beta(x) \propto -x B_\beta(x). \tag{10}$$

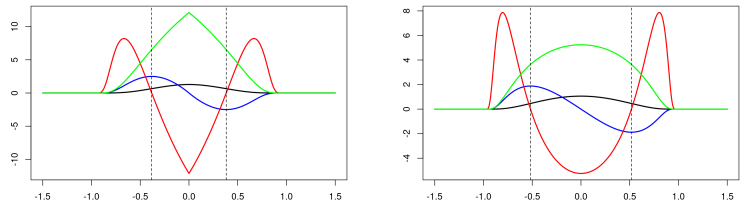
If $|\alpha + 1| < \beta$, then the kernel K_β satisfies the hypotheses of Corollary 1.

See in Fig. 3 several instances of K_β showcasing the flexibility of the new family. The inspiration for (10) comes from the redescending M-estimator perspective, as we derive ρ from ψ and not vice versa. Essentially, we multiply $-x$ by (9) to obtain a smooth version of K'_{Ep}, which, except for the points where K_{Ep} is not differentiable (± 1), corresponds to $\psi(x) \propto x \cdot \mathbb{1}_{(-1,1)}(x)$. The latter produces the most efficient redescending M-estimator for a pdf f satisfying $-f'_0(x)/f_0(x) \propto x$ among those ψ with support $[-1, 1]$ (Huber and Ronchetti, 2009, Eq. 4.88). One can easily see that such f is either a Gaussian or a truncated Gaussian.

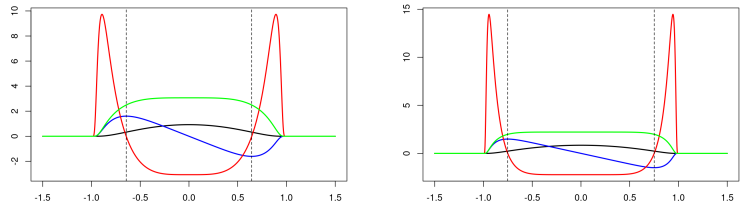
If we consider a Student's t pdf with $\nu > 2$ degrees of freedom, Theorem 6 says we should take $\beta > \nu$. Therefore, as $\nu \to \infty$, approaching Gaussianity, we must



(A) $\beta = 1/4$ (B) $\beta = 1/2$



(C) $\beta = 1$ (D) $\beta = 2$



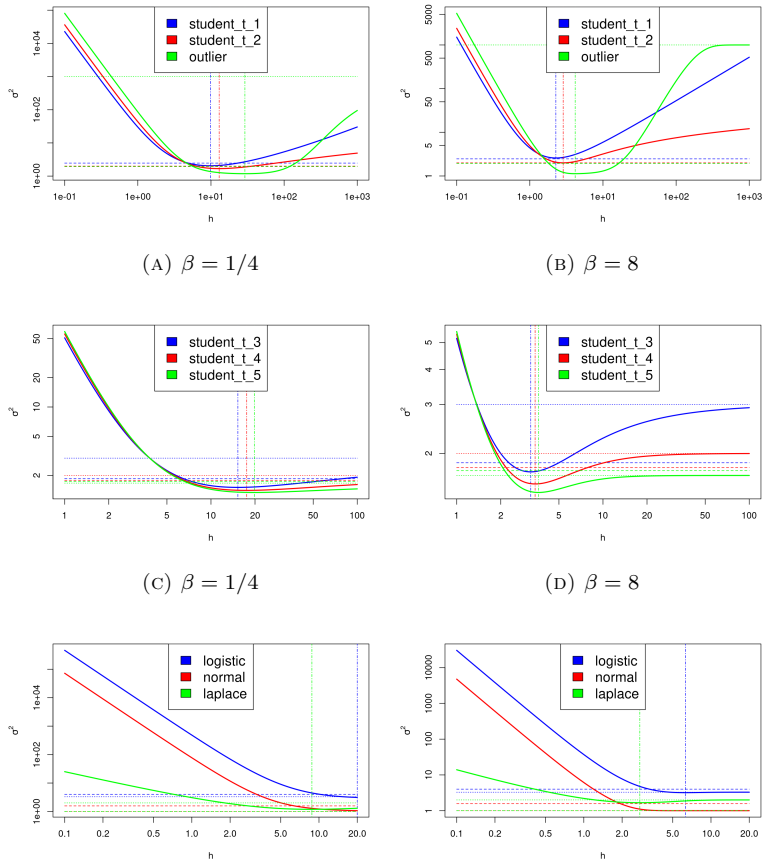
(E) $\beta = 4$ (F) $\beta = 8$

FIG. 3. Several instances of the parametric kernel family K_β defined in Theorem 6, along with some derived associated functions. The black, blue, and red lines are K_β, K'_β, and K''_β, respectively. The green line corresponds to the weight function (12), which is equal to the B_β (9) in the case of K_β. The inflection points of K_β appear as vertical dashed lines. As β grows, the inflection points diverge towards the edges of the support, while K'_β and B_β approach an oblique straight line and a rectangular function, respectively. Consequently, K''_β *redescends* increasingly more sharply. In turn, as β decreases to zero, the inflection points converge towards the center, making the slope of K'_β increasingly *steep* near zero. Indeed, for $\beta \leq 1$, we see that K''_β and B_β are not differentiable at zero.

also take $\beta \to \infty$, and we see that $\lim_{\beta\to\infty} K'_\beta(x) \propto -x \cdot \mathbb{1}_{(-1,1)}(x)$ for all $x \in \mathbb{R} \setminus \{\pm 1\}$. Consequently, $\lim_{\beta\to\infty} K_\beta = K_{\mathrm{Ep}}$ uniformly. In that sense, the kernel (10) asymptotically approaches the optimal behavior of the Epanechnikov kernel for Gaussian data. Interestingly, the KME $\hat{\mathsf{m}}_{n,h}$ with kernel K_{Ep} and bandwidth h is equivalent to the M-estimator with

$$\rho(x) = \begin{cases} x^2\,, & \text{if } |x| < h \\ h^2\,, & \text{otherwise} \end{cases}\,. \tag{11}$$

Therefore, in that case, $\hat{\mathsf{m}}_{n,h}$ has a straightforward interpretation as a sort of "trimmed mean", i.e., an average of those X_i satisfying $|X_i - \hat{\mathsf{m}}_{n,h}| < h$ (Huber, 1964, p. 79).



(A) $\beta = 1/4$ (B) $\beta = 8$

(C) $\beta = 1/4$ (D) $\beta = 8$

(E) $\beta = 1/4$ (F) $\beta = 8$

FIG. 4. Several curves of the function $h \mapsto \sigma^2_{f,K,h}$ for various combinations of pdf f and parameter β, taking $K \equiv K_\beta$. Both the horizontal and vertical axes are on a logarithmic scale. The subfigures from the left column correspond to $\beta = 1/4$, while those on the right consider $\beta = 8$. The pdfs are those in Fig. 2. The top row includes two instances from the Student's t, with $\nu \in \{1, 2\}$, and the *outlier* pdf. The mid row gathers the remaining representatives of the Student's t, with $\nu \in \{3, 4, 5\}$. The bottom row comprises the normal, logistic, and Laplace pdfs. All the subfigures have the same structure. The elements related to a given pdf appear in the same color whenever defined and finite. The solid curve is the main variance function. The values $\sigma^2 = n\mathrm{Var}(\hat{\mu}_n)$ and $[2f(M)]^{-2} \sim n\mathrm{Var}(\hat{M}_n)$ appear as the horizontal dotted and dashed lines, respectively. Finally, the optimal h value minimizing $\sigma^2_{f,K,h}$ is a vertical line.

Before moving on to the computational aspects of KMEs, the essential theoretical results from this section are graphically summed up in Fig. 4. There, we can see several plots of the variance function $h \mapsto \mathcal{V}(h)$ for various combinations of true pdf f and kernel K_β. First, the two parts of Theorem 4 are confirmed. Increasingly small values of h make the variance explode to infinity, while the variance of the KME approaches that of the true pdf as $h \to \infty$, including the case of the Student's t for $\nu \in \{1, 2\}$, where $\sigma^2 = \infty$. In the case of the Laplace pdf, not differentiable at θ, the explosion as $h \to 0^+$ turns out to be less dramatic.

Secondly, Fig. 4 depicts the existence of a bandwidth h^* minimizing the variance $\mathcal{V}(h)$, even for pdfs that are not regularly varying, falling out of the hypotheses of

Corollary 1. The only case in Fig. 4 for which the variance does not have a finite minimizer is for the normal distribution. Of course, this is because the sample mean $\hat{\mu}_n$ coincides with the maximum likelihood estimator (MLE) in this case, and, hence, it cannot be improved. On the other hand, in most of the remaining cases, the KME with bandwidth h^* not only improves over the sample mean, but it also presents lower variance than the sample median $\hat{M}_n$, which has $\mathrm{Var}(\hat{M}_n) \sim [4nf(\theta)^2]^{-1}$ (Lai et al., 1983, Eq. 14). The only exception is for the Laplace pdf, for which, again, the sample median is also the MLE of its location parameter (Maronna et al., 2019, p. 23). All in all, despite having some narrow margins for improvement (see, e.g., Fig. 4f), the KME should theoretically outperform the sample mean and median over these test-beds whenever possible.

Lastly, Fig. 4 exposes an intuitive yet not trivial fact from (7): the *shape* of the kernel K impacts the efficiency of the KME, in addition to the bandwidth. In some extreme cases, a poorly chosen kernel can make the KME severely underperform, having an optimal minimum variance that exceeds that of a much simpler method like the sample median. That is precisely the case in Fig. 4b for $\beta = 8$ and Student's t with $\nu = 1$. By contrast, in Fig. 4a, the kernel K_β with $\beta = 1/4$ outperforms the sample median for a range of values of h.

3.2. **Computation.** There is no closed-form antiderivative for (10). Nonetheless, we shall see that there is a straightforward iterative algorithm to calculate $\hat{\mathsf{m}}_{n,h}$ without evaluating K_β or even K'_β. Again, the kernel (10) has a convenient interpretation when the KME is regarded as an M-estimator. Adapting Maronna et al. (2019, Section 2.3.3) to our context, the weight function

$$W_h(x) = \begin{cases} -K'_h(x)/x\,, & \text{if } x \neq 0 \\ -K''_h(0)\,, & \text{if } x = 0 \end{cases}, \tag{12}$$

which is always positive, allows expressing the KME as the weighted mean

$$\hat{\mathsf{m}}_{n,h} = \frac{\sum_{i=1}^n W_h(X_i - \hat{\mathsf{m}}_{n,h})X_i}{\sum_{i=1}^n W_h(X_i - \hat{\mathsf{m}}_{n,h})}\,, \tag{13}$$

where the weights on the right-hand side also depend on $\hat{\mathsf{m}}_{n,h}$.

Equation (13) implies that $\hat{\mathsf{m}}_{n,h}$ is *one* of the fixed points of the function $\varpi_{n,h}$ given by

$$\varpi_{n,h}(x) = \frac{\sum_{i=1}^n W_h(X_i - x)X_i}{\sum_{i=1}^n W_h(X_i - x)}\,, \tag{14}$$

when the denominator in (14) does not vanish, and $\varpi_{n,h}(x) = x$ otherwise. This observation suggests an iterative procedure for computing $\hat{\mathsf{m}}_{n,h}$, known as *iterative reweighting* (IRW). Such an algorithm turns out to be numerically more stable than solving (4) or (5) through customary optimization or root-finding algorithms, respectively (Maronna et al., 2019, Section 2.10.5.1).

The subindex h on the left-hand side of (12) is a convenient notation to express the dependence on h. Despite W_h being similar to a kernel in the KDE sense, it is generally *not* a pdf. Also, if we denote W the weight function (12) corresponding to $h = 1$, we would have $W_h(x) \neq W(x/h)/h$ for all $h \neq 1$. To obtain pdf-like scaling behavior, we should take $x \mapsto h^2 W_h(x)$.

The following discussion of IRW remains valid for other kernels, but we shall focus on K_β. Considering $K \equiv K_\beta$ in (12), we get $W_h(x) \propto B_\beta(h^{-1}x)$, where the exact value of the proportionality constant does not matter, as it cancels out due to the normalizing denominator in (13). Taking $\beta \to \infty$, we get a *flat* weight function, i.e., $W_h(x) \propto 1$, if $|x| < h$, and zero elsewhere, retrieving the "trimmed mean" discussed above.

Let us assume that a choice of β and h has been made. The IRW procedure is implemented as follows. First, define the score function

$$\mathcal{Z}_\beta(x) = \begin{cases} -(1-|x|^\beta)^{-1}\,, & \text{if } |x|<1 \\ -\infty\,, & \text{otherwise} \end{cases}. \tag{15}$$

Starting from an initial guess $\hat{\mathsf{m}}_0$ at $\hat{\mathsf{m}}_{n,h}$, the transition from the k-th approximation $\hat{\mathsf{m}}_k$ to the $(k+1)$-th approximation $\hat{\mathsf{m}}_{k+1}$, for $k \geq 0$, is given by

$$\hat{\mathsf{m}}_{k+1} = \varpi_{n,h}(\hat{\mathsf{m}}_k) = \sum_{i=1}^n \mathbf{w}_i^{(k)} X_i\,, \tag{16}$$

where the k-th weight vector $\mathbf{w}^{(k)}$ depends on $\hat{\mathsf{m}}_k$ through

$$\mathbf{w}^{(k)} = \sigma\left[\mathcal{Z}_\beta\left(\frac{X_1-\hat{\mathsf{m}}_k}{h}\right),\ldots,\mathcal{Z}_\beta\left(\frac{X_n-\hat{\mathsf{m}}_k}{h}\right)\right],$$

and σ is the *softmax* function mapping $\mathbf{z}=(\mathbf{z}_1,\ldots,\mathbf{z}_n)$ to $\sigma(\mathbf{z}) = (\sigma(\mathbf{z})_1,\ldots,\sigma(\mathbf{z})_n)$, for $\sigma(\mathbf{z})_i = e^{\mathbf{z}_i}/\sum_{j=1}^n e^{\mathbf{z}_j}$, using the convention $e^{-\infty}=0$. More abstractly, for $k\geq 1$, if we define $\varpi_{n,h}^k = \varpi_{n,h}\circ\cdots\circ\varpi_{n,h}$ (k times), then $\hat{\mathsf{m}}_k = \varpi_{n,h}^k(\hat{\mathsf{m}}_0)$. Iterations stop on some convergence criterion, such as $|\hat{\mathsf{m}}_{k+1}-\hat{\mathsf{m}}_k|\leq \varepsilon h$, where $\varepsilon>0$ is a typically small tolerance parameter (Maronna et al., 2019, Section 2.8.1).

Following Maronna et al. (2019, Section 2.8.1), the convergence of IRW in our context is guaranteed because K_β is sufficiently smooth, and its corresponding W_h is bounded and monotonically decreasing as a function of $|x|$. If there is a unique maximizer $\hat{\mathsf{m}}_{n,h}$ without other local maxima, then $\lim_{k\to\infty}\hat{\mathsf{m}}_k = \hat{\mathsf{m}}_{n,h}$, regardless of the starting point $\hat{\mathsf{m}}_0$. Otherwise, $\hat{\mathsf{m}}_0$ must be close to $\hat{\mathsf{m}}_{n,h}$ to avoid "bad solutions". A standard choice to prevent the worst-case scenario is $\hat{\mathsf{m}}_0 = \hat{M}_n$, the sample median. The KME calculated this way also inherits the optimal breakdown point of the sample median (Huber and Ronchetti, 2009, p. 55). Nevertheless, at the expense of a computational cost increase, an even *safer*, minimal-risk choice is $\hat{\mathsf{m}}_0 = X_j$, where $j = \arg\max_{1\leq i\leq n}\hat{f}_{n,h}(X_i)$. However, the latter requires explicitly evaluating the KDE, which IRW was meant to avoid.

Remark 3. It is worth noting here the link between the IRW algorithm and modal clustering (Chacón, 2015). Indeed, the update scheme (16) is known in clustering as the *mean shift* algorithm (Fukunaga and Hostetler, 1975), which iteratively translates any initial point through the steepest density ascent path until it reaches a local maximum, and then clusters together all the points that converge to the same local density mode after such a translation.

3.3. **Parameter optimization.** As mentioned above, carefully choosing β and h is critical for obtaining an efficient KME $\hat{\mathsf{m}}_{n,h}$. We propose optimizing β and h based on the data. Let us denote $\sigma^2_{f,\beta,h}$ the asymptotic variance (7) considering a kernel $K\equiv K_\beta$. Then, the optimal shape and bandwidth parameters are those that minimize $\sigma^2_{f,\beta,h}$, i.e., $(\beta^*,h^*) = \arg\min_{\beta>0,h>0}\sigma^2_{f,\beta,h}$. Let us define a *standardized* version of the random variable $X\sim f$ centered on its actual center of symmetry θ as $Z_h = (X-\theta)/h$, where the bandwidth h plays the role of a scaling parameter. Also, denoting $\psi(x;\beta) = -xB_\beta(x)$, if we define the auxiliary functions $\Psi_1(x;\beta) = \psi(x;\beta)^2$ and $\Psi_2(x;\beta) = \partial\psi(x;\beta)/\partial x$, we have

$$\sigma^2_{f,\beta,h} = h^2\,\frac{\mathbb{E}[\Psi_1(Z_h;\beta)]}{\mathbb{E}[\Psi_2(Z_h;\beta)]^2}\,. \tag{17}$$

Using the symmetry of the auxiliary functions and the target pdf, and the compact support $[-1, 1]$ of the former, inherited from B_β, we get, for $\eta \in \{1, 2\}$,

$$\mathbb{E}[\Psi_\eta(Z_h; \beta)] = 2h \int_0^1 \Psi_\eta(x; \beta)\ f_0(hx)\ dx\,, \tag{18}$$

where we recall that $f_0(x) = f(\theta + x)$ is the centered version of f in Definition 2. Both integrals (18) can be accurately computed using standard numerical methods.

Minimizing (17), a two-dimensional optimization problem, is more complicated than the one-dimensional scenario depicted in Fig. 4, where β was fixed. We propose employing a gradient-free optimization algorithm such as Nelder and Mead (1965), which produces more than satisfactory results, as Section 5 will demonstrate. In this respect, we should emphasize that, rather than necessarily finding the global optimum, our goal is to *improve* on some default sensible parameter guesses $\beta = 1$ and $h = \mathrm{MADN}(X_1, \ldots, X_n)$, where the latter is the normalized *median absolute deviation about the median* defined in Maronna et al. (2019, p. 5), a robust scale estimator in the context of M-estimators.

Since f is a priori unknown in (17), we propose employing a *plug-in*-type estimator, replacing f_0 with a convenient estimate $\tilde{f}_0$ in (18). Specifically, following Meloche (1991), we can take $\tilde{f}_0(x) = \hat{f}^*_{n,g}(\hat{M}_n + x)$, where

$$\hat{f}^*_{n,g}(x) = \frac{\hat{f}_{n,g}(x) + \hat{f}_{n,g}(2\hat{M}_n - x)}{2}\,, \tag{19}$$

and $\hat{f}_{n,g}$ is the KDE in (1) but with a custom bandwidth g independent of the h in (17). Since $\tilde{f}_0$ will be evaluated many times when computing (18), we recommend employing, especially for large samples, a *binned* interpolated approximation (Wand and Jones, 1995, Appendix D.2) for $\hat{f}_{n,g}$ over $[\hat{M}_n - h_{\max}, \hat{M}_n + h_{\max}]$ using a fine grid, where $h_{\max} > 0$ is some reasonable upper bound for the optimal h.

The function (19) is a modified version of the household KDE that considers the symmetry of f about θ. Indeed, (19) is a symmetric pdf about the sample median $\hat{M}_n$, an estimator of θ. Alternatively, (19) can be seen as the KDE with an *augmented* sample $(X_1, \ldots, X_n, 2\hat{M}_n - X_1, \ldots, 2\hat{M}_n - X_n)$, establishing connections with similar procedures in robust statistics such as Mehrotra et al. (1991) (see also Chacón et al., 2009, Example 1). Results by Meloche (1991) show that, under mild assumptions, symmetrization of the KDE about a well-behaved estimator of θ, such as the sample median, halves the variance term in the *mean integrated square error* for pdf estimation.

Given the stringent assumptions of symmetry and unimodality, a basic and fast *rule-of-thumb* bandwidth selection criterion, such as that of Silverman (1986, p. 48), should provide a reasonable estimate for g. Such a procedure relies on relatively minor departures from normality, usually producing over-smoothed KDEs in general multimodality settings (Wand and Jones, 1995, Section 3.2.1). Nonetheless, over-smoothing should play to our advantage, producing tails in $\tilde{f}_0$ that are robust against outliers and enforce unimodality. Even though other more sophisticated bandwidth selectors could potentially yield better results at ensuring unimodality, Silverman's bandwidth copes well with this particular scenario while keeping the computational cost to a minimum.

3.4. **Illustrative example.** Let us illustrate our KME proposal through a synthetic data example. Consider the sample realization given by $(x_1, \ldots, x_7) = (-2, -1, 0, 1, 2, 10, 11)$, with $n = 7$. The sample mean and median realizations are $\hat{\mu}_n = 3$ and $\hat{M}_n = 1$, respectively. None of them makes a convincing candidate for the center of symmetry, as the pilot KDE $\hat{f}_{n,g}$ shown in red in Fig. 5 suggests a

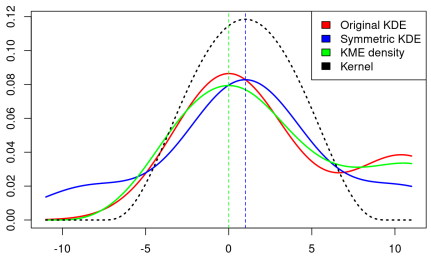


FIG. 5. Plot summary of all the relevant functions involved in IRW for the synthetic data example in Section 3.4. The pilot KDE $\hat{f}_{n,g}$ and its symmetrized version $\hat{f}^*_{n,g}$ about the median, represented as a blue vertical dashed line, are shown in red and blue, respectively. The black dotted curve is the rescaled kernel with optimal parameters (β^*, h^*). The subsequent KDE $\hat{f}_{n,h}$ underlying $\hat{\mathsf{m}}_{n,h}$ shows in green, while the green vertical dashed line corresponds to the IRW computation result.

bimodal structure with two subpopulations. The largest one, corresponding to the first five observations, reaches its maximum density near zero, which would be a perfect center of symmetry if we removed the last two observations (10 and 11) in the second cluster. Therefore, finding a good center of symmetry entails separating *bulk* data from *outlying* data. In other circumstances, if unimodality were not assumed, those outliers would be worth analyzing.

The symmetric KDE (19) is depicted in blue in Fig. 5. The median $\hat{M}_n$ about which the original KDE $\hat{f}_{n,g}$ is *mirrored* shows as the blue vertical dashed line. A slight shift in the maximum of $\hat{f}^*_{n,g}$ is the price for symmetric and thinner tails, more compatible with our assumptions. The optimal parameters for our KME are $(\beta^*, h^*) = (1.765101, 9.199545)$. Then, the corresponding *rescaled* kernel centered on the median, i.e., $x \mapsto K_\beta[h^{-1}(x - \hat{M}_n)]/h$, is displayed as the dotted black line in Fig. 5. Finally, the KDE underlying the KME, built from the previous kernel, is shown in solid green, while the KME is given by the green vertical dashed line. As we can see, the optimal KDE is slightly smoother than the original $\hat{f}_{n,g}$, but both ultimately reach their peak near zero.

Table 1 gathers the final and intermediate results from the IRW algorithm. Convergence was reached after eight iterations. From the first iteration, the observation x_7 has zero weight, while the weight of x_6 vanishes from the second one onwards. As the algorithm progresses, the weights stabilize symmetrically about $x_3 = 0$, which has the most mass. Ultimately, the value of $\hat{\mathsf{m}}_{n,h}$ is practically indistinguishable from zero.

4. CASE STUDY

The estimation of the center of symmetry naturally arises in physics. A measurement X can be described by the location model $X = \theta + \varepsilon$, where θ represents an unknown parameter of interest and ε is a random variable accounting for the measurement error (Maronna et al., 2019, Section 2.1). In this context, to dismiss the existence of systematic errors, the error variable ε is usually assumed to be symmetric about zero, physically meaning that overestimating and underestimating are equally likely (Taylor, 1997).

A classic example to illustrate the robust estimation of location is Simon Newcomb's experiment from 1882 for measuring the speed of light (Gelman et al., 2013,

k	$\mathbf{w}_1^{(k)}$	$\mathbf{w}_2^{(k)}$	$\mathbf{w}_3^{(k)}$	$\mathbf{w}_4^{(k)}$	$\mathbf{w}_5^{(k)}$	$\mathbf{w}_6^{(k)}$	$\mathbf{w}_7^{(k)}$	$\hat{\mathsf{m}}_{k+1}$
0	1.80E-01	1.96E-01	2.07E-01	2.11E-01	2.07E-01	2.09E-12	0	6.89E-02
1	1.92E-01	2.03E-01	2.07E-01	2.04E-01	1.94E-01	0	0	4.67E-03
2	1.93E-01	2.03E-01	2.07E-01	2.03E-01	1.93E-01	0	0	3.17E-04
3	1.93E-01	2.03E-01	2.07E-01	2.03E-01	1.93E-01	0	0	2.15E-05
4	1.93E-01	2.03E-01	2.07E-01	2.03E-01	1.93E-01	0	0	1.46E-06
5	1.93E-01	2.03E-01	2.07E-01	2.03E-01	1.93E-01	0	0	9.92E-08
6	1.93E-01	2.03E-01	2.07E-01	2.03E-01	1.93E-01	0	0	6.73E-09
7	1.93E-01	2.03E-01	2.07E-01	2.03E-01	1.93E-01	0	0	4.57E-10

TABLE 1. IRW weights and mode approximations from the update scheme (16) for the synthetic data example.

pp. 66–67; Maronna et al., 2019, Example 1.2; Stigler, 1977). Newcomb measured the time it takes light to cover a distance of 7,442 meters. The recorded unique values and their number of repetitions are collected in the first two columns of Table 2. See Gelman et al. (2013, Figure 3.1) for a sample histogram. A total of $n = 66$ measurements were taken, the lowest two of which (-44 and -2) are outliers. The sample mean, 26.2, is much more affected by the two outliers than the sample median, 27. Despite the latter being a more reasonable centrality measure, the value 28 is still the most repeated in the sample, i.e., the *discrete* mode. For that matter, if we removed the two low outliers, the value 28 would be at the same distance from the new minimum (i.e., 16) and the maximum (i.e., 40). We shall see that the KME provides an elegant solution to this problem closer to 28.

The optimal shape and bandwidth parameters for the KME using our kernel proposal (10) are $(\beta^*, h^*) = (97.03537, 21.23523)$. Then, the IRW algorithm yields $\hat{\mathsf{m}}_{n,h} = 27.75$. The corresponding IRW unitary and total weights at the last iteration for each sample value are shown in the third and fourth columns of Table 2. As we can see, the two outliers have no weight, while the rest have the same unit weight of $1/64$. Therefore, IRW computes the "trimmed mean" M-estimator corresponding to (11), equivalent to the KME with Epanechnikov kernel, employing a bandwidth $h = h^*$. As it turns out, considering the finite computer precision, the score function $\mathcal{Z}_\beta$ in (15) with β equal to the large optimal β^* above is numerically indistinguishable from the constant -1 over all the standardized random variables $(X_i - \hat{\mathsf{m}}_{n,h})/h$ underlying (13). Finally, the fifth column in Table 2 shows the value of the KDE behind the KME at each sample observation, reaching its maximum at 28, the discrete mode.

We can better assess the situation by looking at Fig. 6. In comparison with the synthetic data example in Section 3.4 and Fig. 5, the isolated local modes of $\hat{f}_{n,g}$, corresponding to the two outliers, are so much less pronounced that they are almost entirely removed from $\hat{f}_{n,h}$. Therefore, the risk of getting trapped in any local maxima of the KDE is minimal. We also see that the original $\hat{f}_{n,g}$ is already mostly symmetric about the sample median, nearly coinciding with $\hat{f}^*_{n,g}$. Moreover, the optimal rescaled kernel K_β, closely resembling the Epanechnikov, supports the Gaussianity hypothesis in Gelman et al. (2013), except for the two outliers.

Detecting and completely discarding the two outliers is an interesting feat of our KME proposal. Choosing a trimming level $\alpha = 2/66$ in a (symmetrically) trimmed mean would also affect the upper tail observations 39 and 40, which, though extreme, are not outliers. As a result, the trimmed mean would be 27.37, which is lower than our KME. In any case, despite the numerous tools for estimating the center of symmetry, the actual value of the speed of light known today is 33,

Value	Count	Unit weight	Total weight	Density
-44	1	0	0	5.41522E-04
-2	1	0	0	8.30601E-04
16	2	1.56250E-02	3.12500E-02	2.22093E-02
19	1	1.56250E-02	1.56250E-02	2.66347E-02
20	1	1.56250E-02	1.56250E-02	2.79265E-02
21	2	1.56250E-02	3.12500E-02	2.90680E-02
22	2	1.56250E-02	3.12500E-02	3.00520E-02
23	3	1.56250E-02	4.68750E-02	3.08786E-02
24	5	1.56250E-02	7.81250E-02	3.15478E-02
25	5	1.56250E-02	7.81250E-02	3.20595E-02
26	5	1.56250E-02	7.81250E-02	3.24138E-02
27	6	1.56250E-02	9.37500E-02	3.26106E-02
28	7	1.56250E-02	1.09375E-01	3.26499E-02
29	5	1.56250E-02	7.81250E-02	3.25318E-02
30	3	1.56250E-02	4.68750E-02	3.22563E-02
31	2	1.56250E-02	3.12500E-02	3.18233E-02
32	5	1.56250E-02	7.81250E-02	3.12329E-02
33	2	1.56250E-02	3.12500E-02	3.04850E-02
34	1	1.56250E-02	1.56250E-02	2.95797E-02
36	4	1.56250E-02	6.25000E-02	2.72967E-02
37	1	1.56250E-02	1.56250E-02	2.59271E-02
39	1	1.56250E-02	1.56250E-02	2.29097E-02
40	1	1.56250E-02	1.56250E-02	2.11794E-02

TABLE 2. Newcomb's measurements of the speed of light as deviations from 24,800 nanoseconds. The first two columns represent the unique measured values and the number of times they occur, respectively. The third column shows the IRW unitary weights corresponding to the first column values. Then, the fourth column is the unit weight times the number of repetitions of each unique value. Finally, the last column gives the value of the KDE underlying the KME at each observation.

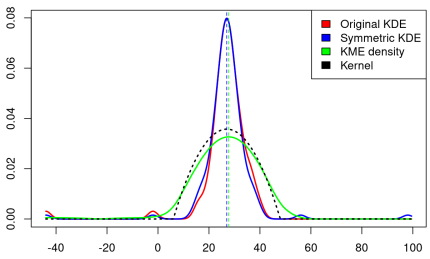


FIG. 6. Plot summary of all the relevant functions involved in IRW for the case study. The structure is the same as in Fig. 5.

far from any of the estimates we have seen. Therefore, as Gelman et al. remind us, data analysis can only be as good as the experiment that produces the data.

5. Simulation study

This section demonstrates the practical effectiveness of KMEs for estimating the center of symmetry, with particular attention paid to our proposal. The latter comprises the parameter optimization in Section 3.3 for the novel kernel family (10), followed by a run of the IRW algorithm in Section 3.2. We shall compare our KME with two classic M-estimators, the sample mean and median, and two redescending

M-estimators, Tukey's *biweight* and Andrew's *sine* (Huber and Ronchetti, 2009, p. 100). For completeness, we also include in the study two classic L-estimators, the trimmed and winsorized means (Maronna et al., 2019, Section 2.4).

The two considered redescending M-estimators are also KMEs. The kernel corresponding to Tukey's biweight is the *triweight* (Wand and Jones, 1995, p. 31) $K(x) \propto (1-x^2)^3 \cdot \mathbb{1}_{(-1,1)}(x)$, while that of Andrew's sine is the *raised cosine* $K(x) \propto [1+\cos(\pi x)] \cdot \mathbb{1}_{(-1,1)}(x)$. Both kernels and their respective associated functions are shown in Fig. 7. The scale parameter (bandwidth) h for these M-estimators is usually tuned by assuming data from a *contaminated* Gaussian distribution. This leads to bandwidths that are *prefixed* multiples of some robust scale estimate, contrary to the full data-driven optimization in Section 3.3. Hence, these methods are considered *non-adaptive* (Hogg, 1974). Specifically, letting $S = \mathrm{MADN}(X_1, \ldots, X_n)$ be as in Section 3.3, we consider $h = 6S$ for Tukey's biweight, and $h = 2.1\pi S$ for Andrew's sine. The factor S is included because of the recommendation in Maronna et al. (2019, Section 2.8.1) (see also Hogg, 1974), whereas the "magical" constants 6 and 2.1π, taken from Stigler (1977), were initially proposed by Tukey and Andrew themselves. The same IRW algorithm in Section 3.2 was used to compute both redescending M-estimators.

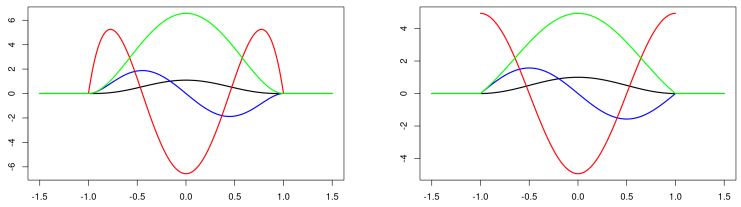
For the trimmed and winsorized means, users typically select one of the widespread values $\alpha \in \{0.1, 0.15, 0.25\}$ for the trimming level, as in Stigler (1977). Here, however, we shall attempt to optimize $\alpha \in [0, 1/2)$, allowing both L-estimators to range between a *mean-like* ($\alpha = 0$) or a *median-like* ($\alpha \lesssim 1/2$) behavior, depending on the data. To do so, we implemented the straightforward *bootstrap* variance-minimizing procedure by Mehrotra et al. (1991) that yielded good results for trimmed means over finite samples. Specifically, we employed their augmented sample strategy about the median (which they call Estimate 3), analogous to the one behind (19). These methods are considered *adaptive* by Hogg (1974), who praises the adaptive version of the trimmed mean for the symmetric case.

Simulated data will be drawn from each of the nine symmetric, unimodal pdfs about $\theta = 0$ shown in Fig. 2. These are the following:

- `normal`: Standard Gaussian pdf $f(x) = \phi(x) \propto e^{-x^2/2}$.
- `logistic`: Logistic pdf $f(x) = (e^{x/2} + e^{-x/2})^{-2}$.
- `laplace`: Laplace pdf $f(x) \propto e^{-|x|}$.
- `student_t_<ν>` ($\nu \in \{1, \ldots, 5\}$): Student's t pdf (8) with $\nu \in \{1, \ldots, 5\}$.
- `outlier`: Gaussian mixture pdf $f(x) = (9/10)\phi(x) + (1/10)\phi(x/100)/100$.

Mehrotra et al. (1991) previously used the normal, logistic, and Laplace (double exponential) distributions as test-beds. These authors also included Student's t with $\nu = 1$, the Cauchy distribution. In turn, the so-called *outlier* distribution is a variant of the classic homonym pdf in Marron and Wand (1992), where the original 10% component of the mixture was $\phi(x/10)/10$, i.e., a Gaussian with $\sigma = 10$ instead of our choice of $\sigma = 100$.

Compared to the normal distribution, the rest of the test-bed pdfs exhibit *heavy-tailedness*. The logistic, Laplace, and outlier distributions have a larger *kurtosis* (Maronna et al., 2019, p. 228) than the normal, meaning they have a moderate but more significant proportion of outliers. Then, all instances from Student's t family are regularly varying, with tails that decay according to a *power law*, which implies a slower rate than the exponential-like tails of the rest. In particular, when $\nu = 1$, the tails are so heavy that even the pdf expectation is undefined. The expectation does exist for $\nu = 2$, as well as the variance, but the latter is still infinite.



(A) Triweight kernel (B) Raised cosine kernel

FIG. 7. Kernels and associated functions corresponding to the redescending M-estimators Tukey's biweight (left) and Andrew's sine (right) used in the simulation study. The plot structure is the same as in Fig. 3, excluding the vertical lines that marked the inflection points. The left and right sets of functions are similar, except for the second derivative (red), which is discontinuous at $|x| = 1$ for the raised cosine.

Each of the nine test-bed pdfs will be paired with three sample sizes, $n = 100$ (small), $n = 1,000$ (medium-sized), and $n = 10,000$ (large), giving rise to twenty-seven sampling configurations. Then, $m = 1,000$ repetitions of each random experiment will be carried out to obtain significant results. For each i-th replication and each contending estimator $\mathtt{M}$, we shall compute an estimate $\hat{\theta}_i^{\mathtt{M}}$ of the center of symmetry $\theta = 0$. Let us assume that $\mathtt{M}_1$ is our KME proposal while $\mathtt{M}_2$ is some other method. Then, letting $\mathrm{MSE}(\mathtt{M}) = m^{-1}\sum_{i=1}^{m}(\hat{\theta}_i^{\mathtt{M}} - \theta)^2$ be the mean squared error (MSE) of the estimator $\mathtt{M}$, a relative efficiency measure comparing two methods $\mathtt{M}_1$ and $\mathtt{M}_2$ is given by $\mathrm{MSE}(\mathtt{M}_1)/\mathrm{MSE}(\mathtt{M}_2)$. Values less than one favor $\mathtt{M}_1$, while those greater than one give the advantage to $\mathtt{M}_2$. On the other hand, for evaluating on a *per-sample* basis, in addition to counting how many experiments satisfy $|\hat{\theta}_i^{\mathtt{M}_1} - \theta| < |\hat{\theta}_i^{\mathtt{M}_2} - \theta|$, a matched-samples Wilcoxon test can be performed to reject the *null* hypothesis that the sequence $(|\hat{\theta}_i^{\mathtt{M}_2} - \theta|)_{i=1}^m$ has an equal or lower *median* value than the sequence $(|\hat{\theta}_i^{\mathtt{M}_1} - \theta|)_{i=1}^m$. A low p-value will reject the null hypothesis in favor of the *alternative* that our KME has a lower median value.

The results from the simulation study are split across three tables with the same structure. Let us analyze each of them separately. First, Table 3 compares our KME with the two classic M-estimators: the sample mean and median. The overall results largely favor our proposal, especially with the three most heavy-tailed test-beds: `student_t_1`, `student_t_2`, and `outlier`. There are only two understandable exceptions, the `normal` and `laplace`, for which, as mentioned above, the sample mean and median are MLEs, respectively. Even so, the performance gap reduces for $n = 10,000$, especially with respect to the sample mean, which, though not significantly, is outperformed.

The results in Table 4 are the least advantageous for our proposal, which nonetheless prevails. The trimmed and winsorized means, especially the former, perform similarly to our KME in many settings. For most Student's t pdfs, namely $\nu \in \{2, 3, 4\}$, our KME is neither significantly better nor worse than the modified means. The same could be said about `logistic`. Our method only performs consistently better with `student_t_1`, `student_t_5`, and `outlier`. Surprisingly, the modified means do not behave like the sample mean with the test-bed `normal`, losing the advantage to our KME in small and medium-sized samples. In turn, the trimmed mean performs especially well with `laplace` and $n < 10,000$, contrary to the winsorized mean.

		Mean		Median	
Test-bed	n	MSE	Paired samples	MSE	Paired samples
`student_t_1`	100	6.43E-06	.925 (p < .001)	9.10E-01	.539 (p = .0016)
	1000	2.52E-06	.972 (p < .001)	7.98E-01	.581 (p < .001)
	10000	2.98E-08	.988 (p < .001)	8.52E-01	.545 (p < .001)
`student_t_2`	100	1.55E-01	.705 (p < .001)	8.99E-01	.565 (p < .001)
	1000	1.57E-01	.757 (p < .001)	8.50E-01	.547 (p < .001)
	10000	1.10E-01	.776 (p < .001)	8.48E-01	.568 (p < .001)
`student_t_3`	100	5.52E-01	.610 (p < .001)	8.69E-01	.564 (p < .001)
	1000	5.10E-01	.659 (p < .001)	8.28E-01	.563 (p < .001)
	10000	5.03E-01	.649 (p < .001)	8.31E-01	.567 (p < .001)
`student_t_4`	100	8.19E-01	.566 (p < .001)	8.39E-01	.576 (p < .001)
	1000	7.16E-01	.596 (p < .001)	8.06E-01	.551 (p < .001)
	10000	7.05E-01	.594 (p < .001)	8.15E-01	.554 (p < .001)
`student_t_5`	100	1.04E+00	.536 (p = .0012)	9.41E-01	.589 (p < .001)
	1000	7.68E-01	.594 (p < .001)	7.68E-01	.583 (p < .001)
	10000	7.82E-01	.585 (p < .001)	7.60E-01	.576 (p < .001)
`logistic`	100	1.00E+00	.532 (p = .24)	8.52E-01	.574 (p < .001)
	1000	9.21E-01	.555 (p < .001)	7.37E-01	.597 (p < .001)
	10000	9.27E-01	.534 (p = .0033)	7.13E-01	.560 (p < .001)
`outlier`	100	1.34E-03	.972 (p < .001)	6.44E-01	.641 (p < .001)
	1000	1.13E-03	.975 (p < .001)	5.60E-01	.640 (p < .001)
	10000	1.09E-03	.986 (p < .001)	5.68E-01	.647 (p < .001)
`normal`	100	1.67E+00	.450 (p = 1)	1.08E+00	.612 (p < .001)
	1000	1.01E+00	.487 (p = .85)	6.68E-01	.601 (p < .001)
	10000	9.99E-01	.514 (p = .06)	6.20E-01	.617 (p < .001)
`laplace`	100	6.42E-01	.618 (p < .001)	1.12E+00	.422 (p = 1)
	1000	5.50E-01	.635 (p < .001)	1.05E+00	.461 (p = .99)
	10000	5.17E-01	.632 (p < .001)	1.03E+00	.486 (p = .97)

TABLE 3. Part one of the simulation study results, comparing our KME with the sample mean and median. The columns MSE represent the ratio $\text{MSE}(\mathtt{M}_1)/\text{MSE}(\mathtt{M}_2)$, where $\mathtt{M}_1$ is our KME proposal, and $\mathtt{M}_2$ is either the sample mean or the sample median. The columns "Paired samples" represent a comparison on a *per-sample* basis. The fractional number is the proportion of the m replications of the random experiment in which $|\hat{\theta}_i^{\mathtt{M}_1} - \theta| < |\hat{\theta}_i^{\mathtt{M}_2} - \theta|$. The p-value in parentheses derives from a Wilcoxon paired-samples test with the alternative hypothesis that our KME has a lower *median* value of $|\hat{\theta}_i^{\mathtt{M}} - \theta|$.

Lastly, Table 5 compares the redescending M-estimators. Both Tukey's biweight and Andrew's sine have similar performances. The results show that our KME is more versatile, outperforming the other two in most settings, namely, in all the pdfs of Student's t family and the `laplace`. In particular, our KME outperforms the recommended method by Maronna et al. for the Cauchy distribution, which is the biweight (Maronna et al., 2019, p. 65). Notwithstanding, the redescending M-estimators demonstrate some proficiency in those scenarios that are more akin to robust statistics, where our KME proposal shows a similar performance. These cases include the `normal` and the similarly-shaped `logistic`, as well as the `outlier`, which is a *contaminated* normal.

Apart from the simulation study results, *debugging* the outputs of the parameter optimization process of Section 3.3 is very instructive. The underlying figures can be checked in Table 6, but we shall focus our comments on Fig. 8. There, we can see the weight function (12) for $K \equiv K_\beta$, and $(\beta, h) = (\beta^*, h^*)$. The optimal parameters (β^*, h^*) are obtained in Fig. 8d considering the true centered pdf f_0.

		Trimmed mean		Winsorized mean	
Test-bed	n	MSE	Paired samples	MSE	Paired samples
`student_t_1`	100	9.03E-01	.536 (p = .001)	8.55E-01	.550 (p < .001)
	1000	8.65E-01	.546 (p < .001)	8.45E-01	.553 (p < .001)
	10000	9.02E-01	.518 (p = .01)	8.44E-01	.548 (p < .001)
`student_t_2`	100	9.88E-01	.537 (p = .02)	9.74E-01	.518 (p = .04)
	1000	9.72E-01	.512 (p = .43)	9.44E-01	.516 (p = .08)
	10000	9.61E-01	.497 (p = .04)	9.24E-01	.549 (p < .001)
`student_t_3`	100	1.03E+00	.497 (p = .65)	1.04E+00	.483 (p = .92)
	1000	1.01E+00	.515 (p = .14)	9.86E-01	.508 (p = .15)
	10000	9.81E-01	.524 (p = .08)	9.56E-01	.535 (p = .0093)
`student_t_4`	100	9.93E-01	.493 (p = .78)	1.01E+00	.495 (p = .83)
	1000	9.80E-01	.499 (p = .41)	9.78E-01	.507 (p = .35)
	10000	9.94E-01	.484 (p = .6)	9.72E-01	.510 (p = .07)
`student_t_5`	100	1.13E+00	.520 (p = .05)	1.19E+00	.502 (p = .53)
	1000	9.57E-01	.525 (p = .01)	9.46E-01	.554 (p < .001)
	10000	9.67E-01	.544 (p < .001)	9.35E-01	.545 (p < .001)
`logistic`	100	1.01E+00	.491 (p = .77)	1.04E+00	.485 (p = .95)
	1000	9.86E-01	.517 (p = .16)	9.54E-01	.518 (p = .02)
	10000	1.00E+00	.498 (p = .5)	9.69E-01	.505 (p = .29)
`outlier`	100	7.77E-01	.597 (p < .001)	8.05E-01	.601 (p < .001)
	1000	7.33E-01	.599 (p < .001)	7.05E-01	.607 (p < .001)
	10000	7.52E-01	.591 (p < .001)	7.35E-01	.596 (p < .001)
`normal`	100	1.43E+00	.516 (p = .0024)	1.45E+00	.559 (p < .001)
	1000	9.75E-01	.538 (p = .0023)	9.19E-01	.534 (p < .001)
	10000	9.99E-01	.508 (p = .7)	9.48E-01	.511 (p = .01)
`laplace`	100	1.11E+00	.438 (p = 1)	9.94E-01	.511 (p = .27)
	1000	1.01E+00	.455 (p = .98)	8.71E-01	.589 (p < .001)
	10000	1.01E+00	.503 (p = .52)	8.91E-01	.558 (p < .001)

TABLE 4. Part two of the simulation study results, comparing our KME with the trimmed and winsorized means, with the same structure as Table 3.

The "optimal parameters" in the remaining subfigures are an average of the m optimal parameter vectors obtained considering an n-size sample estimate $\tilde{f}_0$.

The first aspect worth noticing in Fig. 8 is the natural progression of the weight function estimates as n grows, from Fig. 8a to Fig. 8c, towards the true optimal weight functions in Fig. 8d. In Fig. 8a, our KME produces nearly *flat* weight functions in most sampling configurations, effectively behaving like a sample mean. Only the three most heavy-tailed test-beds (`student_t_1`, `student_t_2`, and `outlier`) make the estimates slightly *bend*. The situation starts changing in Fig. 8b and finally almost settles in Fig. 8c. The second remarkable point about Fig. 8 is that, except for the `normal` and the `outlier`, all test-beds *prefer* a rapid weight decay ($\beta < 1$), being the most extreme case that of the `laplace`.

Last but not least, the optimal values β in Table 6 for the Student's t test-beds show that the condition imposed by Theorem 6 on β is far stronger than necessary in practice. Even though β grows with ν, we do not have $\beta > \nu$. Indeed, Theorem 6 establishes a very stringent hypothesis on the kernel so that the sample mean is outperformed regardless of the bandwidth, providing that h is sufficiently large. In our optimization scenario, however, we neither rely on a fixed kernel nor need an unbounded subset of bandwidths to outperform the sample mean.

		Tukey's biweight		Andrew's sine	
Test-bed	n	MSE	Paired samples	MSE	Paired samples
student_t_1	100	6.90E-01	.614 (p < .001)	5.91E-01	.626 (p < .001)
	1000	6.12E-01	.640 (p < .001)	5.23E-01	.648 (p < .001)
	10000	6.32E-01	.626 (p < .001)	5.48E-01	.636 (p < .001)
student_t_2	100	9.28E-01	.544 (p = .0019)	8.56E-01	.543 (p < .001)
	1000	8.85E-01	.556 (p < .001)	8.26E-01	.581 (p < .001)
	10000	8.41E-01	.580 (p < .001)	7.85E-01	.585 (p < .001)
student_t_3	100	9.88E-01	.518 (p = .13)	9.44E-01	.551 (p < .001)
	1000	9.26E-01	.557 (p < .001)	8.89E-01	.555 (p < .001)
	10000	9.07E-01	.576 (p < .001)	8.67E-01	.583 (p < .001)
student_t_4	100	1.02E+00	.469 (p = .98)	1.00E+00	.499 (p = .73)
	1000	9.63E-01	.551 (p < .001)	9.27E-01	.552 (p < .001)
	10000	9.54E-01	.530 (p < .001)	9.23E-01	.534 (p < .001)
student_t_5	100	1.22E+00	.471 (p = .9)	1.20E+00	.502 (p = .57)
	1000	9.87E-01	.517 (p = .06)	9.58E-01	.546 (p = .0025)
	10000	9.78E-01	.534 (p = .0089)	9.54E-01	.538 (p = .0028)
logistic	100	1.05E+00	.469 (p = .99)	1.05E+00	.464 (p = .98)
	1000	9.93E-01	.506 (p = .1)	9.85E-01	.536 (p = .02)
	10000	9.93E-01	.507 (p = .2)	9.87E-01	.508 (p = .21)
outlier	100	1.10E+00	.530 (p = .03)	1.11E+00	.514 (p = .2)
	1000	9.92E-01	.525 (p = .06)	9.84E-01	.530 (p = .01)
	10000	1.00E+00	.492 (p = .59)	1.01E+00	.489 (p = .54)
normal	100	1.64E+00	.520 (p = .18)	1.66E+00	.507 (p = .55)
	1000	9.90E-01	.512 (p = .05)	1.00E+00	.503 (p = .32)
	10000	9.92E-01	.492 (p = .36)	9.99E-01	.486 (p = .56)
laplace	100	8.42E-01	.553 (p < .001)	8.10E-01	.576 (p < .001)
	1000	7.30E-01	.613 (p < .001)	6.93E-01	.613 (p < .001)
	10000	7.01E-01	.589 (p < .001)	6.67E-01	.593 (p < .001)

TABLE 5. Part three of the simulation study results, comparing our KME with the redescending M-estimators, with the same structure as Table 3.

6. Discussion

This paper provides theoretical and practical results about estimating the center of symmetry from a modal point of view, covering a clear *gap* between the nonparametrics and robust statistics communities. Specifically, we adopt a purely nonparametric approach to a classic theme in robust statistics such as location estimation, assuming only symmetry and unimodality. Along the way, interesting connections with KDEs are enabled.

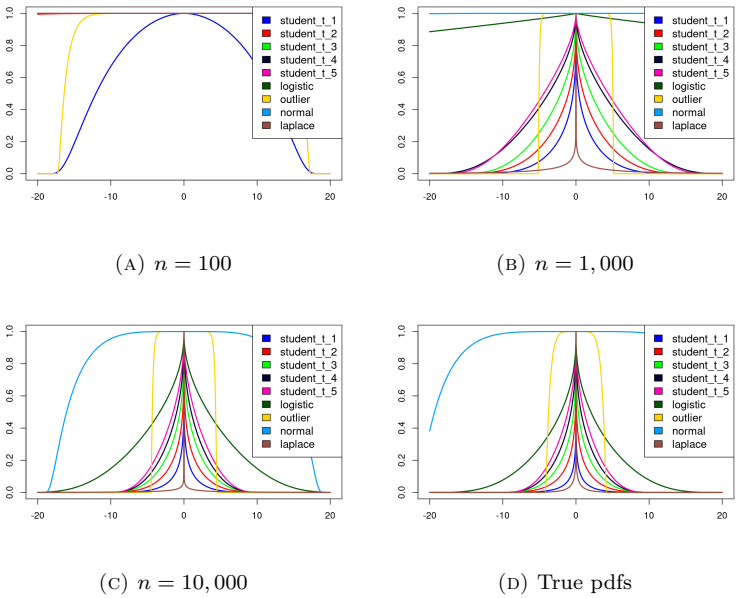
We have studied the efficiency of KMEs in terms of the bandwidth h, formalizing results without any parametric assumption, contrary to robust statistics. On the one hand, a bandwidth too small dangerously increases the variance, supporting the *empirical* caveats of Huber and Ronchetti (2009) against wrong "scaling". Conversely, increasingly large bandwidths make the KME behave like the sample mean. However, it turns out that the KME can asymptotically outperform the sample mean with heavy-tailed, non-Gaussian data. Namely, considering regularly varying pdfs, some kernels K allow finding an optimal bandwidth that minimizes the variance of the asymptotic distribution beyond that of the sample mean M-estimator. In that sense, we also provide a novel parametric family of kernels K_β whose parameter β, connected to the regular variation index α, can be tuned to achieve that goal.

Test-bed	Optimal β	Optimal h	n	Mean β	Mean h
student_t_1	9.69E-02	3.04E+01	100	1.93E+00	1.91E+01
			1000	2.19E-01	2.25E+01
			10000	1.27E-01	2.47E+01
student_t_2	1.57E-01	2.16E+01	100	5.30E+00	5.43E+01
			1000	2.90E-01	2.23E+01
			10000	1.87E-01	2.01E+01
student_t_3	2.23E-01	1.73E+01	100	8.23E+00	1.25E+02
			1000	3.89E-01	2.07E+01
			10000	2.63E-01	1.64E+01
student_t_4	2.91E-01	1.49E+01	100	8.86E+00	1.86E+02
			1000	4.74E-01	2.47E+01
			10000	3.34E-01	1.44E+01
student_t_5	3.60E-01	1.33E+01	100	7.97E+00	2.51E+02
			1000	5.55E-01	2.20E+01
			10000	4.05E-01	1.33E+01
logistic	3.54E-01	2.54E+01	100	9.14E+00	5.57E+02
			1000	8.53E-01	2.70E+02
			10000	4.27E-01	2.69E+01
outlier	6.34E+00	4.31E+00	100	1.39E+01	1.75E+01
			1000	3.45E+01	5.19E+00
			10000	1.88E+01	4.55E+00
normal	5.34E+00	2.29E+01	100	5.62E+00	2.13E+02
			1000	5.45E+00	6.00E+01
			10000	4.72E+00	1.94E+01
laplace	7.07E-02	3.53E+01	100	3.58E+00	3.31E+02
			1000	5.50E-02	3.44E+02
			10000	3.33E-02	9.99E+02

TABLE 6. Optimal parameters (β, h) for each sampling configuration of test-bed and sample size n. The "Optimal" columns represent the optimal values of β and h using the true f_0 in the optimization process. The "Mean" columns are averages of the optimal values of β and h obtained for each random sample using the estimate $\tilde{f}_0$ in the optimization.

The theoretical efficiency calculations in Fig. 4 demonstrate that, contrary to the widespread *belief* shared by Huber and Ronchetti (2009, p. 99), the shape of the kernel K_β, i.e., correctly choosing β, might be critical with specific data, to the point of rendering useless any optimization of h if the kernel is not suitable. These considerations lead to a natural two-step KME procedure consisting of an IRW run preceded by a double joint optimization of the shape parameter β and the bandwidth h, based on a *plug-in* estimate of the KME variance. Technical details on obtaining an estimate $\tilde{f}_0$ and performing the two-dimensional optimization were also provided.

In addition to the above theoretical guarantees, the simulation study shed some light on the actual performance of our proposal compared to the sample mean and other more competitive methods. The results are very favorable for our proposal. The sample mean and median were easily outperformed whenever possible. Then, two more sophisticated methods, the trimmed and winsorized means, though having very few weak points, were outperformed more often than not by our KME, especially with heavy-tailed pdfs. Finally, the two redescending M-estimators, also KMEs, were not nearly as versatile as our proposal, showing that tuning both the scale and shape parameters makes the difference.



(A) $n = 100$ (B) $n = 1,000$

(C) $n = 10,000$ (D) True pdfs

FIG. 8. IRW weight functions (12). For the bottom right subfigure, the underlying parameters (β, h) are the optimal ones (β^*, h^*), considering the true centered pdf f_0. For the remaining subfigures, we computed an average of the optimal parameters for each n-size sample, considering the estimate $\tilde{f}_0$.

The simulation and case studies revealed the compelling flexibility of the kernel family K_β. For the more Gaussian-like pdfs and data, values $\beta > 1$ were chosen, producing smooth, nearly flat weight functions. By contrast, for most test-beds, including all the regularly varying pdfs, values $\beta < 1$ were *preferred*, producing a *steep* slope near zero for the kernel derivative that resembles the ψ of the sample median. The latter shape, pervasive in our results, is *rare* among usual redescending ψ functions such as those in Fig. 7, designed with the typical robust statistics assumption of Gaussianity in mind. However, in a fully nonparametric context, we see that those prefixed designs need not be optimal anymore. All in all, the novel kernel family allows the KME to display a *mean-like* or a *median-like* behavior as needed. These pieces of evidence deserve further investigation, reopening the quest for flexible redescending M-estimators.

Our results by no means invalidate those of robust statistics. The latter framework is theoretically appealing and well-founded for many applications, where a default parametric model, typically Gaussian, can be assumed. However, the robust statistics community has virtually ignored the relationship between redescending M-estimators and KMEs. On the other hand, the nonparametrics community has bypassed the symmetric, unimodal case so far, which is an interesting situation where asymptotically small bandwidths are paradoxically counterproductive. By focusing on efficiency, beyond the robustness gained by employing a kernel with compact support (always welcomed), our research demonstrates that KMEs provide an effective alternative as estimators of the center of symmetry.

ACKNOWLEDGEMENTS

The research of the first author has been supported by the MICINN grant PID2021-124051NB-I00. The second author would like to thank Professor Amparo

Baíllo Moreno for her advice as a doctoral counselor at the Autonomous University of Madrid.

Appendix A. Proofs

This appendix provides proof of the results in Section 3.1.

Proof of Theorem 1. Let us first check the symmetry of $\bar{f}_h$ about θ. Using the symmetry of f and K, we can easily see that if $X \sim f$,

$$\begin{aligned}\bar{f}_h(\theta + x) &= \mathbb{E}[K_h(\theta + x - X)] = \mathbb{E}[K_h(x + X - \theta)] \\ &= \mathbb{E}[K_h(\theta - x - X)] = \bar{f}_h(\theta - x)\,.\end{aligned}$$

Now, to see the unimodality of $\bar{f}_h$, we will check that $x \mapsto \bar{f}_h(\theta + x)$ is a strictly decreasing function of $|x|$ over its support by calculating its derivative. Without loss of generality, we will assume that $x \geq 0$ since the symmetry proved above will automatically cover the case of a negative argument. Therefore, straightforward calculations yield

$$\begin{aligned}\bar{f}_h(\theta + x) &= \int_{-\infty}^{\infty} K_h(\theta + x - y) f(y)\; dy = \int_{-\infty}^{\infty} K_h(x - y) f(\theta + y)\; dy \\ &= \int_{-\infty}^{\infty} K_h(x - y) f_0(y)\; dy\,, \end{aligned} \tag{20}$$

and, subsequently, since K_h' being bounded allows differentiation under the integral,

$$\begin{aligned}\frac{d}{dx}\bar{f}_h(\theta + x) &= \int_{-\infty}^{\infty} K_h'(x - y) f_0(y)\; dy = \int_{-\infty}^{\infty} f_0(x - y) K_h'(y)\; dy \\ &= \int_{0}^{\infty} [f_0(x - y) - f_0(x + y)] K_h'(y)\; dy\,. \end{aligned} \tag{21}$$

If f_0 and K_h have support $[-a, a]$ and $[-b, b]$, respectively, for $a, b \in (0, \infty]$, one can easily see from (20) that the support of $x \mapsto \bar{f}_h(\theta + x)$ is $[-a - b, a + b]$. Hence, we shall check that (21) is zero for $x = 0$ and negative for $x \in (0, a + b)$.

When $x = 0$, since f_0 is symmetric about zero, the derivative is zero, and thus the maximum is reached. For $x > 0$, we will split the proof into two separate cases: $a = \infty$, on the one hand, and $a < \infty$, on the other.

First, let us assume that $a = \infty$. If $x > 0$, then x and y in (21) are positive, and we have $|x - y| < x + y$. Hence, $f_0(x - y) > f_0(x + y)$ in (21) for all $y > 0$ since both arguments lie in the support of f_0. Finally, the net sign of (21) is negative because K_h' is negative over $(0, b)$.

Secondly, let us assume that $a < \infty$. The conclusion easily follows if $x < a$, similarly to the case $a = \infty$ above, so we shall suppose $x \geq a$. In such case, we have $f_0(x + y) = 0$ in (21) for all $y > 0$, and

$$\begin{aligned}\frac{d}{dx}\bar{f}_h(\theta + x) &= \int_{0}^{\infty} f_0(x - y) K_h'(y)\; dy = \int_{-\infty}^{x} K_h'(x - y) f_0(y)\; dy \\ &= \int_{\max\{x-b,-a\}}^{a} K_h'(x - y) f_0(y)\; dy\,,\end{aligned}$$

where the last equality is obtained by intersecting the supports of $y \mapsto K_h'(x-y)$ and f_0 with $(-\infty, x)$. Finally, the result follows after observing that $y \mapsto K_h'(x-y) f_0(y)$ is negative over $(\max\{x - b, -a\}, a)$, while the latter interval is not empty because $x < a + b$. ∎

Proof of Theorem 2. Let $X \equiv_d Y$ denote equality in distribution between two random variables X and Y. We have to verify that $\hat{\mathsf{m}}_{n,h} - \theta \equiv_d \theta - \hat{\mathsf{m}}_{n,h}$, which is equivalent to $\hat{\mathsf{m}}_{n,h} \equiv_d 2\theta - \hat{\mathsf{m}}_{n,h}$. Note that the right-hand side of the last equality

in distribution is the KME reflected about θ, which is equal to the KME corresponding to the KDE based on the reflected sample $(2\theta - X_1, \ldots, 2\theta - X_n)$. Since (i) the latter i.i.d. sample follows the same distribution as the original one, (ii) the kernel K is symmetric, and (iii) ties among candidate modes are broken uniformly at random in a symmetric way, we necessarily have that both KMEs have the same distribution. Finally, the unbiasedness of $\hat{\mathsf{m}}_{n,h}$ is derived by taking expectations at both sides of the equality in distribution condition of symmetry. ■

Proof of Theorem 3. Most of the proof is reduced to applying Maronna et al. (2019, Theorem 10.7). Indeed, (7) corresponds to Equation (2.24) in that same reference, with $\psi(x) = -K_h'(x)$. There only remains to check that the numerator and denominator in (7) are finite, and that the latter is non-null. Since K_h' and K_h'' are bounded, the numerator and denominator are finite. Also due to the boundedness of K_h'', differentiation can go under the integral sign in the denominator as in Maronna et al. (2019, Theorem 10.7).

Finally, let us see that the denominator is not zero. Let a be the positive inflection point of K_h. Note that

$$\begin{aligned}\int_0^a |K_h''(x)|\, f_0(x)\, dx > f_0(a) \int_0^a |K_h''(x)|\, dx = f_0(a) \int_a^\infty |K_h''(x)|\, dx \\ \geq \int_a^\infty |K_h''(x)|\, f_0(x)\, dx\,,\end{aligned}$$

where we have used that f_0 is decreasing for positive x and that, since K_h is bell-shaped, $\int_0^a |K_h''(x)|\, dx = \int_a^\infty |K_h''(x)|\, dx > 0$. Therefore, we finally have

$$\begin{aligned}\int_{-\infty}^\infty K_h''(\theta - x) f(x)\, dx &= \int_{-\infty}^\infty K_h''(x) f_0(x)\, dx = 2\int_0^\infty K_h''(x) f_0(x)\, dx \\ &= 2\left(\int_a^\infty |K_h''(x)|\, f_0(x)\, dx - \int_0^a |K_h''(x)|\, f_0(x)\, dx\right) < 0\,.\end{aligned}$$

■

Proof of Theorem 4. To prove part 1, we note that, given the smoothness of f, and the fact that K' and f' are bounded, we have $(K_h'' * f)(\theta) = (K_h * f'')(\theta)$ (Wand and Jones, 1995, Exercise 2.25). Then, $\lim_{h\to 0^+}(K_h * f'')(\theta) = f''(\theta)$ since K_h is an approximation to the identity for the convolution operation (Parzen, 1962, Theorem 1A), given the remaining assumptions on f and K. Therefore, using that $f(\theta)$ and $f''(\theta)$ are finite, and f is bounded and continuous, calculations yield

$$\begin{aligned}\lim_{h\to 0^+} \sigma^2_{f,K,h} &= |f''(\theta)|^{-2} \lim_{h\to 0^+} [(K_h')^2 * f](\theta) \\ &= |f''(\theta)|^{-2} \lim_{h\to 0^+} \int_{-\infty}^\infty f(\theta - x) K_h'(x)^2\, dx \\ &= |f''(\theta)|^{-2} \lim_{h\to 0^+} h^{-3} \int_{-\infty}^\infty f(\theta - hx) K'(x)^2\, dx \\ &= |f''(\theta)|^{-2} f(\theta) \mathcal{R}(K') \lim_{h\to 0^+} h^{-3} = \infty\,,\end{aligned}$$

regardless of $f''(\theta)$ being zero or not. Finally, if $f''(\theta) < 0$, nearly identical steps show that $\lim_{h\to 0^+} h^3 \sigma^2_{f,K,h}/\sigma^2_{f,K} = 1$.

To prove part 2, first note that

$$hK'(h^{-1}x) = h[K'(h^{-1}x) - K'(0)] = h\int_0^{x/h} K''(y)\, dy = \int_0^x K''(h^{-1}y)\, dy\,.$$

From last equation, since K'' is bounded, we get $\lim_{h\to\infty} hK'(h^{-1}x) = xK''(0)$. Similarly, there exists some $C > 0$ such that $|hK'(h^{-1}x)| \leq C|x|$ for all $x \in \mathbb{R}$ and

all $h > 0$. Then, calculations using the dominated convergence theorem yield

$$\begin{aligned}\lim_{h\to\infty} \sigma^2_{f,K,h} &= \lim_{h\to\infty} \frac{h^{-6}\int_{-\infty}^{\infty} f(\theta-x)[hK'(h^{-1}x)]^2\,dx}{h^{-6}\left(\int_{-\infty}^{\infty} f(\theta-x)K''(h^{-1}x)\,dx\right)^2}\\ &= \frac{K''(0)^2\int_{-\infty}^{\infty} x^2 f(\theta-x)\,dx}{\left(K''(0)\int_{-\infty}^{\infty} f(\theta-x)\,dx\right)^2}\\ &= \frac{\int_{-\infty}^{\infty}(x-\theta)^2 f(x)\,dx}{\left(\int_{-\infty}^{\infty} f(x)\,dx\right)^2} = \sigma^2\,.\end{aligned}$$

■

Proof of Theorem 5. Without loss of generality, let us assume that $|K''(0)| = 1$. Otherwise, simply consider the normalized kernel $x \mapsto K(x)/|K''(0)|$. Defining

$$\tilde{\sigma}^2_{f,K,h} = h^2\int_{-\infty}^{\infty} f(\theta-x)K'(h^{-1}x)^2\,dx\,,$$

we shall prove that, on the one hand, $\lim_{h\to\infty} h^{-(\alpha+3)}(\tilde{\sigma}^2_{f,K,h} - \sigma^2) = 2\ell I_{K,\alpha}$, and, on the other hand, $\lim_{h\to\infty} h^{-(\alpha+3)}(\sigma^2_{f,K,h} - \tilde{\sigma}^2_{f,K,h}) = 0$.

To check the first limit, we note that

$$\begin{aligned}\tilde{\sigma}^2_{f,K,h} - \sigma^2 &= h^2\int_{-\infty}^{\infty} f(\theta-x)[K'(h^{-1}x)^2 - (h^{-1}x)^2]\,dx\\ &= 2h^3\int_0^{\infty} f(\theta+hx)[K'(x)^2 - x^2]\,dx\\ &= 2h^3 f(\theta+h)\int_0^{\infty}\frac{f(\theta+hx)}{f(\theta+h)}[K'(x)^2 - x^2]\,dx\\ &= 2h^{\alpha+3}L(h)\int_0^{\infty} x^\alpha\frac{L(hx)}{L(h)}[K'(x)^2 - x^2]\,dx\\ &\sim 2h^{\alpha+3}L(h)\int_0^{\infty} x^\alpha[K'(x)^2 - x^2]\,dx\,,\end{aligned} \tag{22}$$

where the boundedness of L is used to bring the limit inside the integral.

To check the second limit, let us first define

$$R_{f,K,h} = \int_{-\infty}^{\infty} f(\theta-x)K''(h^{-1}x)\,dx\,.$$

Then, we note that

$$\sigma^2_{f,K,h} - \tilde{\sigma}^2_{f,K,h} = \sigma^2_{f,K,h}(1+R_{f,K,h})(1-R_{f,K,h}) \sim 2\sigma^2(1+R_{f,K,h})\,.$$

Finally, similarly to (22),

$$\begin{aligned}1 + R_{f,K,h} &= \int_{-\infty}^{\infty} f(\theta-x)[1+K''(h^{-1}x)]\,dx\\ &= 2h\int_0^{\infty} f(\theta+hx)[1+K''(x)]\,dx\\ &= 2hf(\theta+h)\int_0^{\infty}\frac{f(\theta+hx)}{f(\theta+h)}[1+K''(x)]\,dx\\ &\sim 2h^{\alpha+1}L(h)\int_0^{\infty} x^\alpha[1+K''(x)]\,dx\,,\end{aligned}$$

and the result follows after $\lim_{h\to\infty} h^{-(\alpha+3)}h^{\alpha+1} = \lim_{h\to\infty} h^{-2} = 0$. ■

Proof of Corollary 1. It is a direct consequence of Theorem 4 and Theorem 5, with analogous reasoning to the proof of Chacón et al. (2007, Theorem 1). The result follows from the following facts: (i) $\mathcal{V}$ is continuous, (ii) $\lim_{h\to 0^+} \mathcal{V}(h) = \infty$, (iii) $\lim_{h\to\infty} \mathcal{V}(h) = \sigma^2$, and (iv) $\mathcal{V}(h) < \sigma^2$ for all sufficiently large h. ■

The proof of Theorem 6 relies on the following lemma.

Lemma 1. *Let $\varphi : [0,1] \to \mathbb{R}$ be strictly increasing, continuous, and such that $\varphi(0) = 0$ and $\varphi(1) = \infty$. Given $a < 0$, we have $\int_0^1 x^a \left|\exp[-\varphi(x)] - 1\right| \, dx < \infty$ whenever $x \mapsto x^a\varphi(x)$ is integrable near zero.*

Proof. Using a Taylor expansion, for every $x \in (0,1)$, we have

$$\exp[-\varphi(x)] - 1 = -\varphi(x) + \frac{e^{\xi_x}}{2}\varphi(x)^2 \,,$$

for some $\xi_x \in (-\varphi(x), 0)$. Let $x_0 \in (0,1)$ be the unique point such that $\varphi(x_0) = 1$. Then, for every $x \in (0, x_0)$,

$$\left|\exp[-\varphi(x)] - 1\right| \le \varphi(x) + \frac{1}{2}\varphi(x)^2 \le \frac{3}{2}\varphi(x) \,.$$

Consequently, given $a < 0$, it follows that

$$\int_0^1 x^a \left|\exp[-\varphi(x)] - 1\right| \, dx \le \frac{3}{2}\int_0^{x_0} x^a\varphi(x) \, dx + \int_{x_0}^1 x^a \left|\exp[-\varphi(x)] - 1\right| \, dx \,,$$

and the integrability of the left-hand side holds if $x^a\varphi(x)$ is integrable near zero. ■

Proof of Theorem 6. Let us first check the hypotheses of Theorem 4. The second derivative of the kernel is

$$K''_\beta(x) \propto \left[\frac{\beta\,|x|^\beta}{(1-|x|^\beta)^2} - 1\right] B_\beta(x) \,. \tag{23}$$

From (23), it is not difficult to see that K_β is bell-shaped, having as positive inflection point $a = [(1+\beta/2) - \sqrt{(1+\beta/2)^2 - 1}]^{1/\beta} \in (0,1)$. Also, $K''_\beta(0) \propto -1/e < 0$. The remaining assumptions in Theorem 4, part 2, follow from K_β having compact support.

Now, let us check the two integral conditions in Theorem 5. Remark 1 applies since K_β has compact support $[-1,1]$. Therefore, for the first integral condition, we have

$$x^\alpha \left[\frac{K'_\beta(x)^2}{K''_\beta(0)^2} - x^2\right] = x^{\alpha+2}\left[\exp\left(\frac{-2x^\beta}{1-x^\beta}\right) - 1\right] \,, \tag{24}$$

which is non-positive, and for the second one,

$$x^\alpha \left[1 + \frac{K''_\beta(x)}{|K''_\beta(0)|}\right] = \frac{\beta x^{\alpha+\beta}}{(1-x^\beta)^2}\exp\left(\frac{-x^\beta}{1-x^\beta}\right) - x^\alpha\left[\exp\left(\frac{-x^\beta}{1-x^\beta}\right) - 1\right] \,. \tag{25}$$

We shall see that (24) and (25) are integrable over $[0,1]$. In both cases, we can use Lemma 1. For (24), integrability holds if $|\alpha+3| < \beta$; for (25), if $|\alpha+1| < \beta$. Finally, since $|\alpha+3| < |\alpha+1|$, the result follows. ■

References

Chacón, J. E. (2015). A population background for nonparametric density-based clustering. *Statistical Science* **30**, 518–32.

Chacón, J. E. (2020). The modal age of statistics. *International Statistical Review* **88**, 122–41.

Chacón, J. E. and Duong, T. (2018). *Multivariate kernel smoothing and its applications*. Boca Raton, FL: Chapman and Hall/CRC.

Chacón, J. E., Montanero, J., Nogales, A. G., and Pérez, P. (2007). On the existence and limit behavior of the optimal bandwidth in kernel density estimation. *Statistica Sinica* **17**, 289–300.

Chacón, J. E., Montanero, J., Nogales, A. G., and Pérez, P. (2009). Partial sufficiency and density estimation. *Journal of Nonparametric Statistics* **21**, 969–75.

Chernoff, H. (1964). Estimation of the mode. *Annals of the Institute of Statistical Mathematics* **16**, 31–41.

Devroye, L. and Györfi, L. (1985). *Nonparametric density estimation: the L1 view*. Wiley Interscience Series in Discrete Mathematics. New York, NY: Wiley.

Eddy, W. F. (1980). Optimum kernel estimators of the mode. *The Annals of Statistics* **8**, 870–82.

Eddy, W. F. (1982). The asymptotic distributions of kernel estimators of the mode. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **59**, 279–90.

Fukunaga, K. and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory* **21**, 32–40.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Grenander, U. (1965). Some direct estimates of the mode. *The Annals of Mathematical Statistics* **36**, 131–8.

Grund, B. and Hall, P. (1995). On the minimisation of L^p error in mode estimation. *The Annals of Statistics* **23**, 2264–84.

Hogg, R. V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association* **69**, 909–23.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35**, 73–101.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust statistics*. Hoboken, NJ: Wiley.

Lai, T. L., Robbins, H., and Yu, K. F. (1983). Adaptive choice of mean or median in estimating the center of a symmetric distribution. *Proceedings of the National Academy of Sciences* **80**, 5803–6.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*. Second edition. New York, NY: Springer-Verlag.

Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2019). *Robust statistics: theory and methods (with R)*. Hoboken, NJ: Wiley.

Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *The Annals of Statistics* **20**, 712–36.

Mehrotra, K., Jackson, P., and Schick, A. (1991). On choosing an optimally trimmed mean. *Communications in Statistics - Simulation and Computation* **20**, 73–80.

Meloche, J. (1991). Estimation of a symmetric density. *Canadian Journal of Statistics* **19**, 151–64.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal* **7**, 308–13.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics* **33**, 1065–76.

Purkayastha, S. (1998). Simple proofs of two results on convolutions of unimodal distributions. *Statistics & Probability Letters* **39**, 97–100.

Romano, J. P. (1986). "On bootstrapping the joint distribution of the location and size of the mode". PhD thesis. University of California, Berkeley.

Romano, J. P. (1988a). Bootstrapping the mode. *Annals of the Institute of Statistical Mathematics* **40**, 565–86.

Romano, J. P. (1988b). On weak convergence and optimality of kernel density estimates of the mode. *The Annals of Statistics* **16**, 629–47.

Seneta, E. (1976). *Regularly Varying Functions*. Heidelberg: Springer-Verlag.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall Monographs on Statistics & Applied Probability. London: Chapman and Hall.

Stigler, S. M. (1977). Do robust estimators work with real data? *The Annals of Statistics* **5**, 1055–98.

Taylor, J. R. (1997). *An Introduction to error analysis: the study of uncertainties in physical measurements*. Sausalito, CA: University Science Books.

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*. London: Chapman and Hall.

Wang, T. (2024). Nonlinear kernel mode-based regression for dependent data. *Journal of Time Series Analysis* **45**, 189–213.


UAM
Universidad Autónoma
de Madrid


UÁM
Universidad Autónoma
de Madrid

https://doi.org/10.1007/s10463-025-00942-z


UAM
Universidad Autónoma
de Madrid


UAM
Universidad Autónoma
de Madrid

Conclusiones

Esta tesis por compendio de publicaciones, enmarcada en la línea de investigación en Estadística y Probabilidad, presenta tres trabajos independientes articulados sobre ciertas características de las funciones de densidad. La materia troncal a lo largo de toda la obra es la estadística no paramétrica, aderezada con elementos de datos funcionales y métodos geométricos, según el capítulo, y de estadística computacional, de forma transversal. Todos son temas propios del proyecto de investigación *Statistical techniques in high-dimensional spaces* de la UAM (antes, *Infinite-dimensional statistics: mathematical models and computation*), del cual son miembros el director y el autor de esta tesis.

Los tres trabajos poseen un marcado carácter metodológico. Sin perjuicio de la debida fundamentación y justificación teórica, se ha prestado especial atención a la resolución de problemas prácticos y reales en análisis de datos, primando los aspectos computacionales y aplicados. Una tarea recurrente en el planteamiento de la investigación ha sido el análisis exploratorio de datos. Este término vagamente aglutina a una gran variedad de procedimientos preliminares para comprender la estructura de los datos, antes de acometer tareas más conocidas como la clasificación, la regresión o el análisis de conglomerados (*clustering*, en inglés). En este contexto, la función de densidad resulta una abstracción de gran utilidad.

En el Capítulo 1, la tarea de exploración busca comprender *qué* hay de relevante en el conjunto de datos, resumiendo este a través de subconjuntos de $\mathbb{R}^d$ en los que la hipersuperficie densidad satisface ciertas propiedades geométrico-diferenciales. El objetivo de resumir prosigue en el Capítulo 2, esta vez calibrando *cuántas* subpoblaciones hay en los datos, característica de complejidad identificada con el número de modas de la densidad: un número entero positivo. La síntesis alcanza su punto álgido en el Capítulo 3, donde se trata de averiguar cuál es el valor *central* $\theta \in \mathbb{R}$ de un experimento aleatorio, caracterizado como el máximo global de la densidad, bajo determinadas hipótesis.

El concepto de valor atípico (*outlier*, en inglés), esquivo en estadística, surge de manera natural, de forma más o menos explícita, en cada uno de los trabajos. Al invertir la foto resumen de unos datos, el negativo lo conforman precisamente los valores que no debían aparecer, siendo protagonistas inesperados. En el Capítulo 1, son atípicos aquellos valores no incluidos en ninguna región del *bump*, pudiendo albergar algún significado para el analista. Luego, en el Capítulo 2, los atípicos están detrás de uno de los males que aquejan al estimador núcleo de la densidad: la formación de modas espurias. Finalmente, en el Capítulo 3, los atípicos son valores que descartar al estimar θ, en aras de la eficiencia y la robustez.

Las características de la densidad estudiadas en esta tesis comparten un origen geométrico y topológico local. Este difiere de las motivaciones clásicas en estadística, centradas en el contenido probabilístico global. Por ejemplo, los *bumps* de curvatura en el Capítulo 1 son capaces de captar sutilezas que las regiones de alta densidad de Hyndman (1996) pasan por alto. Igualmente, como observó Donoho (1988), dos densidades próximas, atendiendo a un criterio de distancia global, pueden tener números de modas totalmente dispares. Por último, la moda destaca por ser más robusta ante valores extremos que los conceptos clásicos de media y mediana, como se comprobó en el Capítulo 3. Todo ello hace que esta tesis se encuentre alineada con las tendencias actuales en investigación estadística.

El estimador núcleo de la densidad (Parzen, 1962) demuestra su potencial a lo largo de toda la tesis. Su flexibilidad resulta clave para dar una respuesta satisfactoria a una pregunta

difusa como la de *bump hunting* en el Capítulo 1. En cambio, en el Capítulo 2, dicha flexibilidad se matiza con cierta estructura ante un problema difícil, pero que requiere una solución exacta. Por otra parte, la universalidad del estimador núcleo se ve reafirmada en el Capítulo 3, donde aparece en un contexto no habitual, mientras que la utilidad de suavizar la densidad también se ve reforzada con los Capítulos 1 y 3. En definitiva, la tesis contribuye a ampliar el abanico de aplicaciones del estimador núcleo de la densidad.

Asimismo, esta tesis introduce varias innovaciones en el campo del estimador núcleo. En el Capítulo 1 se emplea el estimador núcleo de las derivadas de orden dos de la densidad, siguiendo la estela de Casa (2019), quien ya abordó las de primer orden. A su vez, el Capítulo 2 retoma, desde un punto de vista bayesiano y con ayuda de los *splines*, la relación entre el estimador núcleo y los métodos basados en verosimilitud, dos corrientes tradicionalmente antagónicas para las que Bolón (2024, Sección 2.3) también ha planteado recientemente una solución de compromiso. Finalmente, el Capítulo 3 incluye como novedad la optimización de la *forma* del núcleo, aspecto que en su contexto habitual suele mantenerse fijo.

El espíritu de las "matemáticas experimentales" (Berrendero, 2015), basadas en la computación y la simulación (véase también el comentario de E. Parzen en Good y Gaskins, 1980), se halla presente en la concepción y el desarrollo de la tesis. Además de haber realizado simulaciones para evaluar métodos sobre muestras finitas, algunos de los conceptos recogidos por Berrendero (2015) han resultado cruciales. A saber, el *bootstrap* posibilita la inferencia en el Capítulo 1 —no sin cierto trabajo teórico—, mientras que la simulación MCMC permite calcular probabilidades *a posteriori* en el Capítulo 2. Ambos son ejemplos de cómo la aleatoriedad puede ayudar, paradójicamente, a calibrar la incertidumbre (Berrendero, 2015).

Esta tesis ilustra claramente la gran diversidad de facetas que hacen valiosa la investigación. En términos de originalidad, el Capítulo 1 supone un paso hacia delante, pero bastante lógico y natural, mientras que el Capítulo 2 da un salto más disruptivo. Por el contrario, el avance en el Capítulo 3 consiste en rellenar un hueco en la literatura. Del mismo modo, la mejora obtenida con los resultados ha sido dispar. Si bien en el Capítulo 3 se obtuvo una ventaja clara respecto a los métodos competidores, en el Capítulo 2 se dio por bueno el empate, ponderando algunos aspectos no funcionales de BTS. Entretanto, la validación en el Capítulo 1 fue puramente cualitativa, dada la naturaleza visual del método.

En esta tesis también se dan la mano ramas y perspectivas diversas de las matemáticas y la estadística. Al igual que los geómetras han realizado recientemente incursiones en la estadística, a través del análisis de datos topológico (Marron y Dryden, 2021), el Capítulo 1 posee una fuerte componente geométrica. Por su parte, en el Capítulo 2 se combinan dos mundos históricamente distantes: el del estimador núcleo (Parzen, 1962) y el de los *splines* (Wahba, 1990). El objetivo de tender puentes entre dos comunidades investigadoras vecinas se hace explícito en el Capítulo 3, donde se reúne a la estadística no paramétrica y a la estadística robusta en torno a un problema de interés para ambas, posibilitando la colaboración.

Aunque los trabajos han sido publicados y presentados oralmente, cabe mencionar algunas líneas de investigación futura. La inferencia en el Capítulo 1 resultó conservadora, algo que podría corregirse siguiendo enfoques más sofisticados como los de Chen y col. (2017, Sección 4.1) y Mammen y Polonik (2013). Respecto a BTS, en el Capítulo 2, se abren dos vías para extender el método a datos circulares (Ameijeiras-Alonso, 2017; Bolón, 2024), un área activa de investigación en estadística no paramétrica, y a datos bivariantes (Minnotte y Scott, 1993, Sección 5), empleando los recientes avances de Hron y col. (2022). Por último, el Capítulo 3 podría continuarse con las sugerencias de Hino (2025) y Pardo-Fernández (2025).

Considerando el auge de la IA, cabe esperar que el análisis de datos que ahora realiza un humano en "simbiosis" con la máquina pase a ser, más tarde o más temprano, responsabilidad de esta última en solitario (Good, 1966). Aunque ya existen arquitecturas profundas para procesar datos de tipo *conjunto*, tienen un alcance limitado y requieren un gran número de patrones (Zaheer y col., 2017). Así, por tanto, en este camino hacia la "máquina ultrainteligente", esta tesis proporciona estrategias para que la IA asimile la información de manera inteligente.


UAM
Universidad Autónoma
de Madrid


UAM
Universidad Autónoma
de Madrid

Conclusions

This thesis by publication, framed within the research line of Statistics and Probability, presents three independent works focusing on certain features of density functions. The core subject throughout the work is nonparametric statistics, enriched with elements of functional data and geometric methods, depending on the chapter, and computational statistics in a transversal manner. These topics are part of UAM's research project *Statistical techniques in high-dimensional spaces* (formerly, *Infinite-dimensional statistics: mathematical models and computation*), of which the supervisor and the author of this thesis are members.

The three works have a strong methodological focus. Without disregarding the necessary theoretical foundation and justification, special attention has been given to solving practical and real-world problems in data analysis, prioritizing computational and applied aspects. A recurring task in the research approach has been exploratory data analysis. This term loosely encompasses various preliminary procedures to understand the data structure before undertaking more well-known tasks such as classification, regression, or cluster analysis. In this context, the density function is a very useful abstraction.

In Chapter 1, the exploration task aims to understand *what* is relevant in the dataset, summarizing it through subsets of $\mathbb{R}^d$ where the density hypersurface satisfies certain geometric-differential properties. The goal of summarization continues in Chapter 2, this time focusing on determining *how many* subpopulations exist in the data, a complexity feature identified with the number of density modes: a positive integer. The synthesis reaches its peak in Chapter 3, where the objective is to discover the *central* value $\theta \in \mathbb{R}$ of a random experiment, characterized as the global maximum of the density under specific hypotheses.

The concept of an outlier, elusive in statistics, naturally arises, more or less explicitly, in each of the works. By inverting the summary snapshot of a dataset, the negative is precisely composed of the values that were not meant to appear, becoming unexpected protagonists. In Chapter 1, outliers are values not included in any bump region, potentially holding some meaning for the analyst. Then, in Chapter 2, outliers are behind one of the issues affecting the kernel density estimator: the formation of spurious modes. Finally, in Chapter 3, outliers are values to be discarded when estimating θ for the purposes of efficiency and robustness.

The density features studied in this thesis share a local geometric and topological origin. This differs from the classic motivations in statistics, centered on global probabilistic content. For instance, the curvature bumps in Chapter 1 can capture subtleties that the highest density regions of Hyndman (1996) overlook. Similarly, as Donoho (1988) observed, two densities that are close according to a global distance criterion can have entirely different numbers of modes. Lastly, the mode stands out for being more robust to extreme values than the classic concepts of mean and median, as demonstrated in Chapter 3. All of this makes this thesis aligned with current trends in statistical research.

The kernel density estimator (Parzen, 1962) demonstrates its potential throughout the thesis. Its flexibility proves key to providing a satisfactory answer to a diffuse question like bump hunting in Chapter 1. Conversely, in Chapter 2, this flexibility is tempered with some structure to address a challenging problem that demands an exact solution. Moreover, the universality of the kernel estimator is reaffirmed in Chapter 3, where it appears in an unusual context, while the utility of smoothing the density is reinforced in Chapters 1 and 3. In summary, the thesis contributes to broadening the range of applications of the kernel density estimator.

This thesis also introduces several innovations in the field of kernel density estimation. In Chapter 1, the kernel estimator for second-order derivatives of the density is employed, following in the footsteps of Casa (2019), who previously addressed first-order derivatives. In turn, Chapter 2 revisits, from a Bayesian perspective and with the aid of splines, the relationship between the kernel estimator and likelihood-based methods, two traditionally antagonistic approaches for which Bolón (2024, Section 2.3) has also recently proposed a compromise solution. Finally, Chapter 3 includes, as a novelty, the optimization of the *shape* of the kernel, an aspect that is typically kept fixed in its usual context.

The spirit of "experimental mathematics" (Berrendero, 2015), grounded in computation and simulation (see also E. Parzen's comment in Good and Gaskins, 1980), is present throughout the conception and development of this thesis. In addition to conducting simulations to evaluate methods on finite samples, several concepts outlined by Berrendero (2015) have proven crucial. Specifically, the bootstrap enables inference in Chapter 1—not without some theoretical groundwork—while MCMC simulation allows calculating posterior probabilities in Chapter 2. Both are examples of how randomness can paradoxically help to quantify uncertainty (Berrendero, 2015).

This thesis clearly illustrates the wide range of facets that make research valuable. Regarding originality, Chapter 1 represents a step forward, albeit fairly logical and natural, while Chapter 2 takes a more disruptive leap. By contrast, the advancement in Chapter 3 lies in filling a gap in the literature. Similarly, the improvement achieved with the results has been varied. While Chapter 3 demonstrated a clear advantage over competing methods, Chapter 2 settled for a tie, considering some non-functional aspects of BTS. Meanwhile, the validation in Chapter 1 was purely qualitative, given the visual nature of the method.

This thesis also brings together diverse branches and perspectives of mathematics and statistics. Just as geometers have recently ventured into statistics through topological data analysis (Marron and Dryden, 2021), Chapter 1 has a strong geometric component. Meanwhile, Chapter 2 combines two historically distant worlds: the kernel estimator (Parzen, 1962) and splines (Wahba, 1990). The aim of building bridges between neighboring research communities is made explicit in Chapter 3, where nonparametric statistics and robust statistics are brought together around a problem of mutual interest, enabling collaboration.

Although the works have been published and presented orally, it is worth mentioning some lines of future research. The inference in Chapter 1 turned out to be conservative, something that could be addressed by adopting more sophisticated approaches such as those of Chen et al. (2017, Section 4.1) and Mammen and Polonik (2013). Regarding BTS in Chapter 2, two paths emerge to extend the method to circular data (Ameijeiras-Alonso, 2017; Bolón, 2024), an active area of research in nonparametric statistics, and to bivariate data (Minnotte and Scott, 1993, Section 5), using recent advances by Hron et al. (2022). Finally, Chapter 3 could be continued with the suggestions by Hino (2025) and Pardo-Fernández (2025).

Considering the rise of AI, it is expected that data analysis, currently performed by humans in "symbiosis" with machines, will, sooner or later, become the sole responsibility of the latter (Good, 1966). While deep architectures for processing *set-type* data already exist, they have a limited scope and require a large number of samples (Zaheer et al., 2017). Thus, along this path toward the "ultraintelligent machine," this thesis provides strategies for AI to assimilate information intelligently.


UAM
Universidad Autónoma
de Madrid


UAM
Universidad Autónoma
de Madrid

Referencias

ALBERT, J., BENNETT, J. y COCHRAN, J. J. (2005). *Anthology of statistics in sports*. ASA-SIAM Series on Statistics and Applied Probability. Society for Industrial and Applied Mathematics (SIAM). DOI: 10.1137/1.9780898718386.

AMEIJEIRAS-ALONSO, J. (2017). Assessing simplifying hypotheses in density estimation. Tesis doctoral. Universidade de Santiago de Compostela. Handle: 10347/16416.

AMEIJEIRAS-ALONSO, J., CRUJEIRAS, R. M. y RODRÍGUEZ-CASAL, A. (2019). Mode testing, critical bandwidth and excess mass. *TEST* **28**, 900-19. DOI: 10.1007/s11749-018-0611-5.

BANKS, D. L. (1996). A conversation with I. J. Good. *Statistical Science* **11**, 1-19. DOI: 10.1214/ss/1032209661.

BERRENDERO, J. R. (2015). Simulación e inferencia estadística. *La Gaceta de la RSME* **18**, 45-65. URL: https://gaceta.rsme.es/abrir.php?id=1263 (visitado 18-08-2025).

BERRENDERO, J. R., COÍN, A. y CUEVAS, A. (2025). A Bayesian approach to functional regression: theory and computation. *Bayesian Analysis*. arXiv: 2312.14086.

BOLÓN, D. (2024). Object oriented inference methods. Tesis doctoral. Universidade de Santiago de Compostela. Handle: 10347/34217.

CASA, A. (2019). Climbing modes and exploring mixtures: a journey in density-based clustering. Tesis doctoral. Università degli studi di Padova. Handle: 11577/3422342.

CHACÓN, J. E. (2020). The modal age of statistics. *International Statistical Review* **88**, 122-41. DOI: 10.1111/insr.12340.

CHACÓN, J. E. y DUONG, T. (2018). *Multivariate kernel smoothing and its applications*. Chapman and Hall/CRC. DOI: 10.1201/9780429485572.

CHACÓN, J. E. y FERNÁNDEZ SERRANO, J. (2023). Bump hunting through density curvature features. *TEST* **32**, 1251-75. DOI: 10.1007/s11749-023-00872-z. arXiv: 2208.00174.

CHACÓN, J. E. y FERNÁNDEZ SERRANO, J. (2024). Bayesian taut splines for estimating the number of modes. *Computational Statistics & Data Analysis* **196**, 107961. DOI: 10.1016/j.csda.2024.107961. arXiv: 2307.05825.

CHACÓN, J. E. y FERNÁNDEZ SERRANO, J. (2025a). Mode-based estimation of the center of symmetry. *Annals of the Institute of Statistical Mathematics* **77**, 685-717. DOI: 10.1007/s10463-025-00942-z. arXiv: 2406.08241.

CHACÓN, J. E. y FERNÁNDEZ SERRANO, J. (2025b). Rejoinder to the discussion of "Mode-based estimation of the center of symmetry". *Annals of the Institute of Statistical Mathematics* **77**, 727-30. DOI: 10.1007/s10463-025-00945-w. arXiv: 2508.18909.

CHEN, Y.-C. (2022). Solution manifold and its statistical applications. *Electronic Journal of Statistics* **16**, 408-50. DOI: 10.1214/21-EJS1962.

CHEN, Y.-C., GENOVESE, C. R. y WASSERMAN, L. (2017). Density level sets: asymptotics, inference, and visualization. *Journal of the American Statistical Association* **112**, 1684-96. DOI: 10.1080/01621459.2016.1228536.

CHERNOFF, H. (1964). Estimation of the mode. *Annals of the Institute of Statistical Mathematics* **16**, 31-41. DOI: 10.1007/bf02868560.

DAVIES, L. y KOVAC, A. (2004). Densities, spectral densities and modality. *The Annals of Statistics* **32**, 1093-136. DOI: 10.1214/009053604000000364.

DEVROYE, L. y GYÖRFI, L. (1985). *Nonparametric density estimation: the L1 view*. Wiley Interscience Series in Discrete Mathematics. Wiley. URL: https://luc.devroye.org/L1bookBW.pdf (visitado 18-08-2025).

DONOHO, D. L. (1988). One-sided inference about functionals of a density. *The Annals of Statistics* **16**, 1390-420. DOI: 10.1214/aos/1176351045.

DUONG, T., COWLING, A., KOCH, I. y WAND, M. P. (2008). Feature significance for multivariate kernel density estimation. *Computational Statistics & Data Analysis* **52**, 4225-42. DOI: 10.1016/j.csda.2008.02.035.

FERNÁNDEZ SERRANO, J. (2021). Semiparametric bivariate extreme-value copulas. arXiv: 2109.11307.

FRIEDMAN, J. H. y FISHER, N. I. (1999). Bump hunting in high-dimensional data. *Statistics and Computing* **9**, 123-43. DOI: 10.1023/a:1008894516817.

GENOVESE, C., PERONE-PACIFICO, M., VERDINELLI, I. y WASSERMAN, L. (2016). Non-parametric inference for density modes. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **78**, 99-126. DOI: 10.1111/rssb.12111.

GODTLIEBSEN, F., MARRON, J. S. y CHAUDHURI, P. (2002). Significance in scale space for bivariate density estimation. *Journal of Computational and Graphical Statistics* **11**, 1-21. DOI: 10.1198/106186002317375596.

GOOD, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B (Methodological)* **14**, 107-14. DOI: 10.1111/j.2517-6161.1952.tb00104.x.

GOOD, I. J. (1966). Speculations concerning the first ultraintelligent machine. *Advances in Computers Volume 6*. Elsevier, 31-88. DOI: 10.1016/s0065-2458(08)60418-0.

GOOD, I. J. y GASKINS, R. A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association* **75**, 42-56. DOI: 10.1080/01621459.1980.10477419.

HINO, H. (2025). Discussion of "Mode-based estimation of the center of symmetry". *Annals of the Institute of Statistical Mathematics* **77**, 719-21. DOI: 10.1007/s10463-025-00943-y.

HRON, K., MENAFOGLIO, A., TEMPL, M., HRŮZOVÁ, K. y FILZMOSER, P. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics & Data Analysis* **94**, 330-50. DOI: 10.1016/j.csda.2015.07.007.

HRON, K., MACHALOVÁ, J. y MENAFOGLIO, A. (2022). Bivariate densities in Bayes spaces: orthogonal decomposition and spline representation. *Statistical Papers* **64**, 1629-67. DOI: 10.1007/s00362-022-01359-z.

HUBER, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35**, 73-101. DOI: 10.1214/aoms/1177703732.

HUBER, P. J. y RONCHETTI, E. M. (2009). *Robust statistics*. Wiley. DOI: 10.1002/9780470434697.

HYNDMAN, R. J. (1996). Computing and graphing highest density regions. *The American Statistician* **50**, 120-6. DOI: 10.2307/2684423.

MACHALOVÁ, J., TALSKÁ, R., HRON, K. y GÁBA, A. (2020). Compositional splines for representation of density functions. *Computational Statistics* **36**, 1031-64. DOI: 10.1007/s00180-020-01042-7.

MAMMEN, E. y POLONIK, W. (2013). Confidence regions for level sets. *Journal of Multivariate Analysis* **122**, 202-14. DOI: 10.1016/j.jmva.2013.07.017.

MARONNA, R. A., MARTIN, R. D., YOHAI, V. J. y SALIBIÁN-BARRERA, M. (2019). *Robust statistics: theory and methods (with R)*. Wiley. DOI: 10.1002/9781119214656.

MARRON, J. S. y DRYDEN, I. L. (2021). *Object oriented data analysis*. Chapman and Hall/CRC. DOI: 10.1201/9781351189675.

MINNOTTE, M. C. y SCOTT, D. W. (1993). The mode tree: a tool for visualization of nonparametric density features. *Journal of Computational and Graphical Statistics* **2**, 51-68. DOI: 10.2307/1390955.

PARDO-FERNÁNDEZ, J. C. (2025). Discussion of "Mode-based estimation of the center of symmetry". *Annals of the Institute of Statistical Mathematics* **77**, 723-5. DOI: 10.1007/s10463-025-00944-x.

PARZEN, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics* **33**, 1065-76. DOI: 10.1214/aoms/1177704472.

SENETA, E. (1976). *Regularly varying functions*. Springer Berlin Heidelberg. DOI: 10.1007/bfb0079658.

SEVERINI, T. A. (2020). *Analytic methods in sports. Using mathematics and statistics to understand data from baseball, football, basketball, and other sports*. Chapman and Hall/CRC. DOI: 10.1201/9780367252090.

STIGLER, S. M. (1977). Do robust estimators work with real data? *The Annals of Statistics* **5**, 1055-98. DOI: 10.1214/aos/1176343997.

TSYBAKOV, A. B. (2009). *Introduction to nonparametric estimation*. Springer New York. DOI: 10.1007/b13794.

WAHBA, G. (1990). *Spline models for observational data*. Society for Industrial and Applied Mathematics (SIAM). DOI: 10.1137/1.9781611970128.

WAND, M. P. y JONES, M. C. (1995). *Kernel smoothing*. Chapman and Hall. DOI: 10.1201/b14876.

WASSERMAN, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology* **44**, 92-107. DOI: 10.1006/jmps.1999.1278.

ZAHEER, M., KOTTUR, S., RAVANBHAKHSH, S., PÓCZOS, B., SALAKHUTDINOV, R. y SMOLA, A. J. (2017). Deep sets. *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 3394-404. arXiv: 1703.06114.

ZHANG, Y. y CHEN, Y.-C. (2025). Mode and ridge estimation in Euclidean and directional product spaces: a mean shift approach. *Journal of Computational and Graphical Statistics*, 1-20. DOI: 10.1080/10618600.2025.2505734.


UAM
Universidad Autónoma
de Madrid

Apéndice

Rejoinder to the discussion of "Mode-based estimation of the center of symmetry"

DECLASSIFIED

Annals of the Institute of Statistical Mathematics


UÁM
Universidad Autónoma
de Madrid

REJOINDER TO THE DISCUSSION OF "MODE-BASED ESTIMATION OF THE CENTER OF SYMMETRY"

JOSÉ E. CHACÓN[†] AND JAVIER FERNÁNDEZ SERRANO[‡]

ABSTRACT. Rejoinder to the discussion by Hino (2025) and Pardo-Fernández (2025) of Chacón and Fernández Serrano (2025), a special paper (with discussion) published in *Annals of the Institute of Statistical Mathematics*.

We want to express our sincere gratitude to Professor Hino and Professor Pardo-Fernández for their kind words and for providing such insightful and stimulating discussions. We are also extremely honored to have the opportunity to expand on some of their comments in this rejoinder. Hopefully, this series of papers will foster new developments in statistics from a modal perspective.

1. Professor Hino's comments

Professor Hino provides an excellent summary of our core contributions. His discussion rightly emphasizes the connection between the kernel mode estimator (KME) and robust M-estimators, as well as the novel finding that both kernel shape and bandwidth significantly impact the performance of the KME in the symmetric, unimodal setting.

In addition, Professor Hino conveniently points out alternative mode estimation techniques, such as k-NN-based methods and the half-sample mode. These are indeed important tools in the broader landscape of mode estimation. However, our paper focuses on kernel-based methods because they enjoy a unique property with symmetric data: producing unbiased estimates. The latter allows employing non-vanishing bandwidths designed to minimize the asymptotic variance. In this respect, it would be interesting to investigate if other mode estimators behave similarly.

On the other hand, we share Professor Hino's enthusiasm about the wide range of applications of the mode (Chacón, 2020). His works on modal linear regression (Sando et al., 2019) and modal principal component analysis (Sando and Hino, 2020) are great examples of how the mode can secure robustness in classic problems.

2. Professor Pardo-Fernández's comments

Professor Pardo-Fernández raises three specific and relevant potential extensions of our work, dealing with censored/truncated data, multivariate settings, and testing for symmetry. We briefly comment on each of them in the following.

2.1. **Censoring and truncation.** Censoring and truncation are pertinent subjects related to our research. As Professor Pardo-Fernández notes, estimators in these settings, like the Kaplan-Meier or Lynden-Bell estimators, involve random

[†]Departamento de Matemáticas, Universidad de Extremadura, Badajoz, Spain.
[‡]Departamento de Matemáticas, Universidad Autónoma de Madrid, Madrid, Spain.
E-mail addresses: [†]jechacon@unex.es ✉, [‡]javier.fernandezs01@estudiante.uam.es.
2020 *Mathematics Subject Classification.* 62G05 (Primary), 62G07, 62G35.
[†]https://orcid.org/0000-0002-3675-1960 .
[‡]https://orcid.org/0000-0001-5270-9941 .

weights W_i that depend on the full sample, unlike the uniform $1/n$ weights in the standard i.i.d. case. Extending our KME framework would require adapting the kernel density estimator definition (as in Equation (2) from the discussion) using these weighted distribution function estimators.

Certainly, KMEs have been studied under censoring and truncation (see Guessoum et al., 2018, and references therein). However, to our knowledge, the consequences of the additional symmetry assumption in these settings have *not* been investigated so far. In that context, the primary challenges would lie, on one hand, in deriving the asymptotic variance of the KME, and, on the other hand, in designing a suitable version of the iterative reweighting algorithm. For the former, the random weights W_i would imply a considerably more complex scenario due to their intricate dependence structure; for the latter, the additional layer of W_is might affect the stability, convergence properties, and interpretation of the algorithm.

Nevertheless, this is a valuable direction, as robust estimation under such data limitations is crucial in many fields (e.g., survival analysis or econometrics). Accordingly, this problem needs—and deserves—to be examined in greater depth in future publications.

2.2. **The multivariate case.** As Professor Pardo-Fernández suggests, extending the theory to the multivariate setting is a natural and important next step. Nonetheless, this problem appears challenging since our results rely on two essential concepts, symmetry and unimodality, for which several possible generalizations exist.

A multivariate density function $f\colon \mathbb{R}^d \to \mathbb{R}$ can be symmetric in many ways. Considering the origin as the center of symmetry, numerous definitions exist, including central symmetry, axial symmetry, rotational symmetry, radial symmetry, or elliptical symmetry, to name only a few. All these notions can be shifted to an arbitrary center of symmetry $\boldsymbol{\theta} \in \mathbb{R}^d$ through a simple translation. See Serfling (2006) for an interesting survey on multivariate symmetry.

Regarding unimodality, Dharmadhikari and Joag-Dev (1988) study various nonequivalent approaches in the multivariate context: star unimodality, block unimodality, linear unimodality or (central) convex unimodality, among others. They also show that the class of log-concave densities, which has recently been the focus of major research interest (for a review, see Samworth, 2018), can play an important role within these unimodal distributions.

Therefore, developing the corresponding multivariate theory necessarily includes, as a first task, identifying the appropriate class of symmetric and unimodal densities for which the convolution with a kernel leaves the mode invariant. That is the key feature that allows, in a second stage, focusing on deriving the asymptotic variance of the KME. Finally, obtaining a flexible parametric family of multivariate kernels (akin to our K_β), to fully leverage the efficiency and robustness of this nonparametric approach, surely represents a nontrivial task.

2.3. **Symmetry testing.** Using our KME to enhance the capabilities of existing symmetry tests is a very nice suggestion by Professor Pardo-Fernández. Since our optimized KME can offer higher efficiency in estimating the center of symmetry, incorporating it into a symmetry test statistic could lead to a significant power increase in some cases.

The trimmed mean was selected by Milošević and Obradović (2018) because of the flexibility of its trimming level $\alpha \in (0, 1/2)$, having the sample mean and the sample median as limiting cases as $\alpha \to 0$ and $\alpha \to 1/2$, respectively. Similarly,

in vanilla KME, the role of α is played by the bandwidth h. Furthermore, the resemblance between the KME and the trimmed mean when using the Epanechnikov kernel was noted in our work.

However, more importantly, the KME based on the parametric family K_β supplies additional flexibility with the shape parameter. Taking $\beta \to \infty$ allows retrieving the Epanechnikov kernel, whereas making $\beta \to 0$ produces median-like behavior. Moreover, interestingly, our variance-minimizing process favored small intermediate values $\beta < 1$ under heavier tails such as those of the Cauchy and the logistic (two of the three test-beds proposed in the study by Milošević and Obradović, 2018). In such scenarios, symmetry testing based on our full KME proposal could be more efficient than based on the trimmed mean.

Acknowledgements

The research of the first author has been supported by the MICINN grant PID2021-124051NB-I00, while both authors have been supported by the MICINN grant PID2023-148081NB-I00. We want to thank everyone involved in publishing this series of articles. Special thanks to the two discussants, Professor Hino and Professor Pardo-Fernández, and to the chief editor, Professor Ninomiya, for his decision to turn our original manuscript into a special paper.

References

Chacón, J. E. (2020). The modal age of statistics. *International Statistical Review* **88**, 122–41.

Chacón, J. E. and Fernández Serrano, J. (2025). Mode-based estimation of the center of symmetry. *Annals of the Institute of Statistical Mathematics*.

Dharmadhikari, S. W. and Joag-Dev, K. (1988). *Unimodality, convexity, and applications*. Boston, MA: Academic Press, Inc.

Guessoum, Z., Mansouri, M.-A., and Ould-Saïd, E. (2018). Asymptotic properties of the kernel mode estimator under twice censorship model. *Communications in Statistics - Theory and Methods* **47**, 2195–212.

Hino, H. (2025). Discussion of "Mode-based estimation of the center of symmetry". *Annals of the Institute of Statistical Mathematics*.

Milošević, B. and Obradović, M. (2018). Comparison of efficiencies of some symmetry tests around an unknown centre. *Statistics* **53**, 43–57.

Pardo-Fernández, J. C. (2025). Discussion of "Mode-based estimation of the center of symmetry". *Annals of the Institute of Statistical Mathematics*.

Samworth, R. J. (2018). Recent progress in log-concave density estimation. *Statistical Science* **33**, 493–509.

Sando, K., Akaho, S., Murata, N., and Hino, H. (2019). Information geometry of modal linear regression. *Information Geometry* **2**, 43–75.

Sando, K. and Hino, H. (2020). Modal principal component analysis. *Neural Computation* **32**, 1901–35.

Serfling, R. J. (2006). "Multivariate symmetry and asymmetry". *Encyclopedia of Statistical Sciences, Second Edition*. Vol. 8. Hoboken, NJ: Wiley, 5338–45.


UAM
Universidad Autónoma
de Madrid

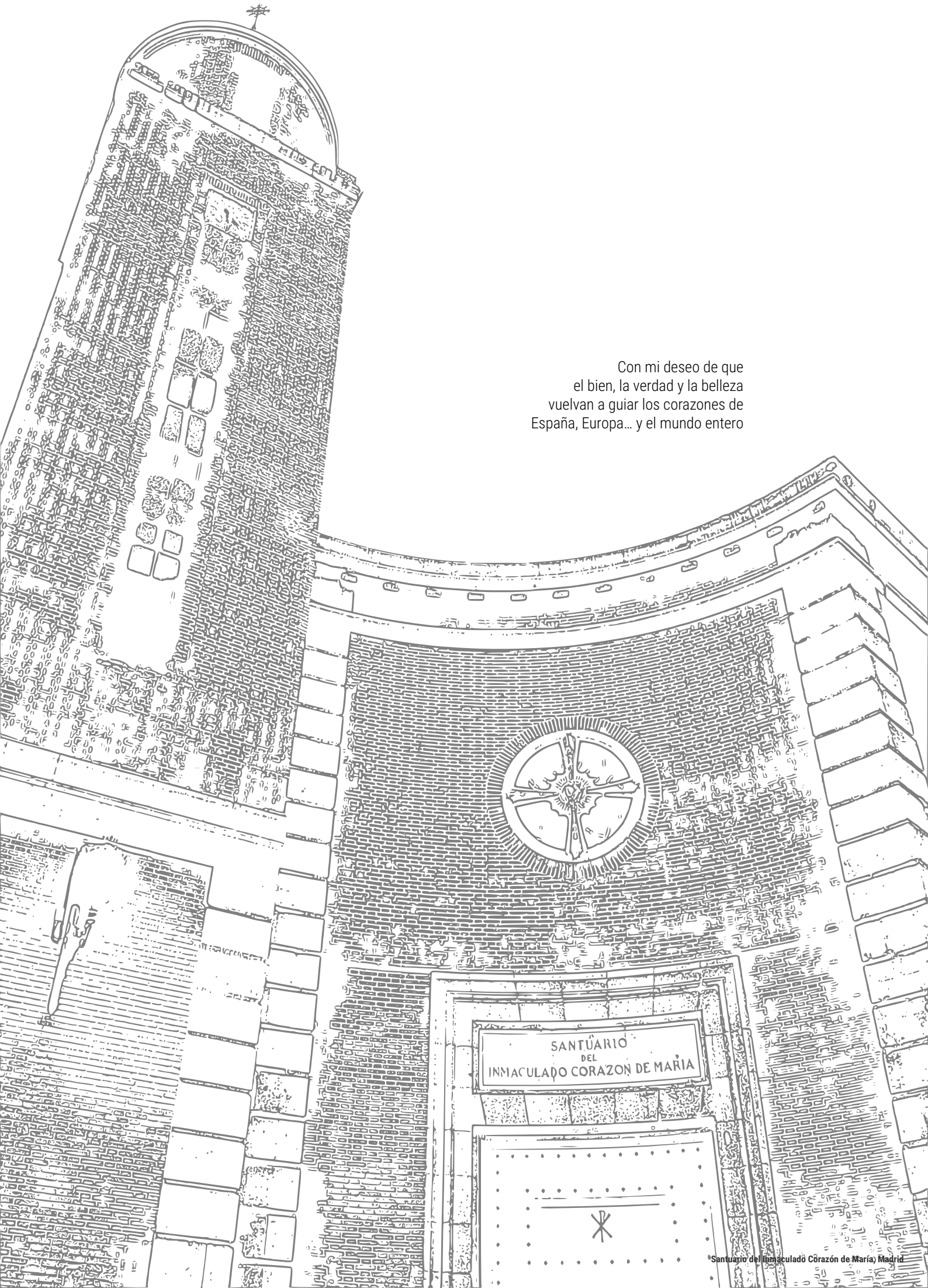

UÁM
Universidad Autónoma
de Madrid

https://doi.org/10.1007/s10463-025-00945-w


UAM
Universidad Autónoma
de Madrid


UAM
Universidad Autónoma
de Madrid

Con mi deseo de que
el bien, la verdad y la belleza
vuelvan a guiar los corazones de
España, Europa… y el mundo entero



Santuario del Inmaculado Corazón de María, Madrid


UAM
Universidad Autónoma
de Madrid

"¿Qué más debo hacer?"

Esta tesis se ha compuesto mediante el sistema LuaLaTeX y las clases AMS.
Los paquetes `pdfpages` y `newpax` han posibilitado incluir documentos PDF externos.
Este documento PDF se ha generado con la distribución TeX Live en entorno Linux Ubuntu.
Este documento PDF cumple con el estándar PDF/A-2b, según veraPDF y Adobe Acrobat.
Las fuentes tipográficas, colores y logos empleados son los oficiales de la UAM.*

*Con la excepción del precioso escudo no oficial de la UAM en este colofón, que es obra de `Asqueladd` y está disponible, bajo una licencia Creative Commons "Atribución-CompartirIgual 3.0 No portada", en `https://commons.wikimedia.org/wiki/File:Escudo_de_la_Universidad_Autónoma_de_Madrid.svg`.


UAM
Universidad Autónoma
de Madrid

Impreso y encuadernado en España

Dios guarde al lector muchos años


UAM
Universidad Autónoma
de Madrid

Javier Fernández Serrano es licenciado en Matemáticas e ingeniero en Informática por la Universidad Autónoma de Madrid (UAM). Posee másteres por la UAM en Matemáticas y Aplicaciones (especialidad en aplicaciones) y en Investigación e Innovación en TIC. Antes de acceder al Doctorado en Matemáticas, trabajó cinco años en una empresa aseguradora, ocupando distintos puestos en el ámbito de los datos y la tecnología. Al margen de la tesis, destaca entre sus intereses científicos la teoría de cópulas. En la actualidad, es miembro del proyecto de investigación *Statistical techniques in high-dimensional spaces*.

Publicaciones

José E. Chacón y Javier Fernández Serrano. Bump hunting through density curvature features. *TEST* (2023).

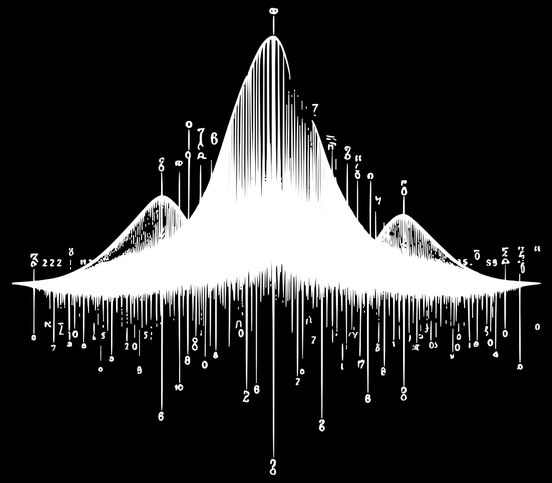
José E. Chacón y Javier Fernández Serrano. Bayesian taut splines for estimating the number of modes. *Computational Statistics & Data Analysis* (2024).

José E. Chacón y Javier Fernández Serrano. Mode-based estimation of the center of symmetry. *Annals of the Institute of Statistical Mathematics* (2025).

José E. Chacón y Javier Fernández Serrano. Rejoinder to the discussion of "Mode-based estimation of the center of symmetry". *Annals of the Institute of Statistical Mathematics* (2025).

UAM Universidad Autónoma de Madrid

La presente tesis doctoral, dirigida por José E. Chacón y organizada como compendio de publicaciones, aborda el estudio de diversas propiedades clave de las funciones de densidad de probabilidad, vistas como un medio para caracterizar y entender la estructura poblacional subyacente a conjuntos de datos. Basándose en resultados teóricos y prácticos, y prestando especial atención a los aspectos computacionales y aplicados, el autor presenta sendas propuestas metodológicas para estimar las regiones de curvatura, el número de modas y el centro de simetría. Entre las muchas contribuciones de esta investigación en estadística, además de técnicas de efectividad contrastada, el analista encontrará poderosas e innovadoras herramientas para la calibración de la incertidumbre y la visualización.



UAM Universidad Autónoma de Madrid

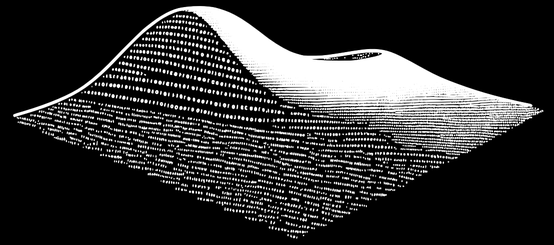
UAM Universidad Autónoma de Madrid

TESIS DOCTORAL

Javier Fernández Serrano

UAM Universidad Autónoma de Madrid

ESTIMACIÓN DE CARACTERÍSTICAS DIFERENCIALES Y TOPOLÓGICAS DE LAS FUNCIONES DE DENSIDAD DE PROBABILIDAD



Javier Fernández Serrano

El análisis de datos desempeña un papel crucial y transversal, potenciando cualquier rama del saber o actividad humana. A través de tres trabajos independientes, se presentan sendas propuestas metodológicas relativas a problemas clásicos en estadística de gran relevancia para el análisis de datos. El fundamento común a todos ellos es el estudio de la estructura poblacional subyacente a los datos, sustanciada en ciertas características de naturaleza diferencial y topológica de la función de densidad de probabilidad.

El primer trabajo aborda diversas propiedades de curvatura de la densidad, inéditas hasta ahora en estadística, con el fin de detectar subconjuntos significativos del espacio muestral. La investigación demuestra el buen comportamiento asintótico, tanto en consistencia como en inferencia, de los estimadores tipo núcleo de tales regiones de curvatura. Estas, asimismo, resultan de gran utilidad en aplicaciones del ámbito deportivo como herramienta de visualización para análisis exploratorio de datos multivariantes.

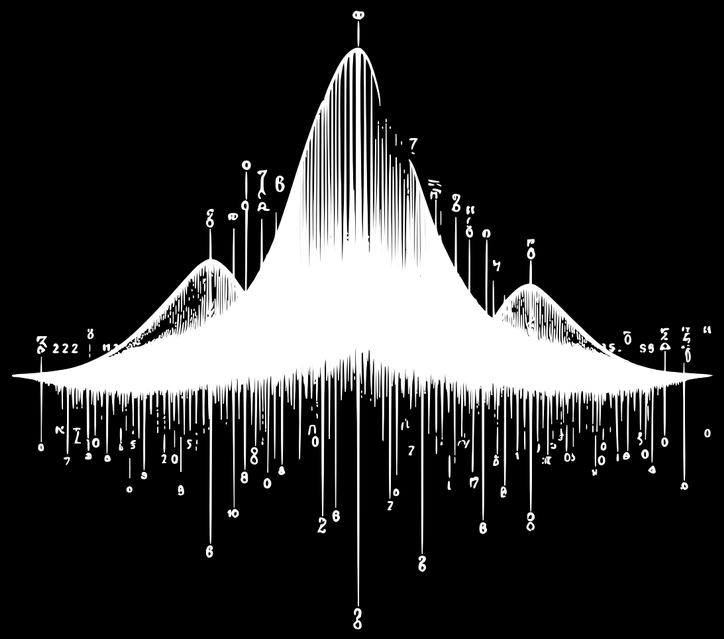
El segundo trabajo afronta el reto de estimar el número de modas de una densidad univariante, propiedad relacionada con la cantidad de subpoblaciones y, por tanto, con la complejidad de los datos. Marcando como meta la efectividad, se propone un nuevo método de estimación bayesiano que aúna múltiples perspectivas sobre la multimodalidad. El estudio de simulación llevado a cabo sitúa al nuevo método entre los más efectivos, a la vez que revela el mejorable rendimiento de algunas alternativas comúnmente empleadas.

El último trabajo revisa la estimación del centro de simetría de una densidad univariante, simétrica y unimodal, esto es, el valor poblacional más prominente. Los resultados asintóticos obtenidos, sobre el impacto del ancho de banda y la forma del núcleo en la eficiencia de la estimación, contribuyen a reducir la brecha entre las comunidades de estadística no paramétrica y estadística robusta. La propuesta subsiguiente de estimador adaptativo se contrasta con éxito en un estudio de simulación, destacando en escenarios con colas pesadas.

En conjunto, los tres trabajos ponen de relieve el potencial de estas características para extraer conocimiento de los datos, como evidencian los múltiples y variados casos de uso aportados. Los métodos propuestos suponen avances e innovaciones importantes respecto al estado del arte, tanto en el plano teórico como computacional, con especial énfasis en su uso práctico. En particular, se ahonda en la búsqueda de métodos flexibles en un contexto no paramétrico, ampliando el abanico de aplicaciones del estimador núcleo de la densidad.

UAM Universidad Autónoma de Madrid

La presente tesis doctoral, dirigida por José E. Chacón y organizada como compendio de publicaciones, aborda el estudio de diversas propiedades clave de las funciones de densidad de probabilidad, vistas como un medio para caracterizar y entender la estructura poblacional subyacente a conjuntos de datos. Basándose en resultados teóricos y prácticos, y prestando especial atención a los aspectos computacionales y aplicados, el autor presenta sendas propuestas metodológicas para estimar las regiones de curvatura, el número de modas y el centro de simetría. Entre las muchas contribuciones de esta investigación en estadística, además de técnicas de efectividad contrastada, el analista encontrará poderosas e innovadoras herramientas para la calibración de la incertidumbre y la visualización.



UAM Universidad Autónoma de Madrid

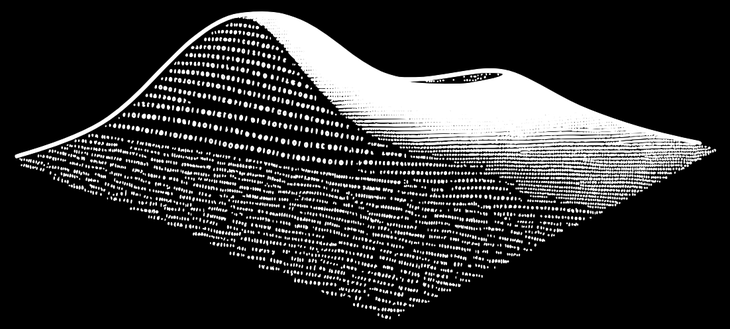
UAM Universidad Autónoma de Madrid

TESIS DOCTORAL

Javier Fernández Serrano

UAM
Universidad Autónoma
de Madrid

ESTIMACIÓN DE CARACTERÍSTICAS DIFERENCIALES Y TOPOLÓGICAS DE LAS FUNCIONES DE DENSIDAD DE PROBABILIDAD



Javier Fernández Serrano

f

π

θ

Universidad Autónoma
de Madrid

f

P

π

θ

Universidad Autónoma
de Madrid