Methodology for probability density function estimation

Metodología para la estimación de la función de densidad de probabilidad

Beatriz Poggi[1], Martel Tobío[2], Levy Abad[3]
[1]*Department of Mathematics, Universidad de Antioquia, Colombia*
[2]*Department of Mathematics, Universidad de Los Andes, Colombia*
[3]*Department of Mathematics, Universidad del Valle, Cali, Colombia*

Abstracto
Los datos numéricos se encuentran en muchas aplicaciones de análisis de datos. La mayoría de los datos numéricos provienen de mediciones y experimentos, resultando así del muestreo de una variable aleatoria, el concepto matemático que caracteriza los resultados numéricos de los experimentos.

Palabras clave: función de densidad de probabilidad, estimación, metodología, muestreo.

Abstract
Numerical data are found in many applications of data analysis. Most numerical data come from measurements and experiments, thus resulting from the sampling of a random variable, the mathematical concept that characterizes the numerical results of experiments.

Keywords: Probability density function, estimation, methodology, sampling.

1.Introducción

Los datos numéricos se encuentran en muchas aplicaciones de datos análisis. La mayoría de los datos numéricos provienen de mediciones y experimentos, resultando así del muestreo de un variable aleatoria, el concepto matemático que caracteriza los resultados numéricos de los experimentos[1]. A analizar datos, uno puede optar por manejar directamente el resultados de los experimentos. Por ejemplo, datos simples métodos de análisis como el PCA lineal (Principal Análisis de componentes)[2], el MLP no lineal (multicapa Perceptron), y muchos otros, trabajan directamente en el valores numéricos de muestras[3]. Mientras esta forma de trabajar puede revelar adecuado en muchas situaciones, otras requieren trabajar con la variable aleatoria subyacente en lugar de la muestra numérica[4].

Una variable aleatoria se caracteriza completamente por su Funciones de densidad de probabilidad (PDF), es decir, una función que representa la probabilidad de que ocurra un evento cuando el variable aleatoria es igual a (o está contenida en un intervalo alrededor) un valor específico[5]. Se pueden utilizar dos ejemplos para ilustran esta necesidad.

En el marco bayesiano, por ejemplo en bayesiano clasificación, las decisiones se toman de acuerdo con la regla, que implica directamente la evaluación del PDF[6]. En un problema de dos clases, por ejemplo, la decisión de elegir una clase u otra se toma de acuerdo con el PDF más grande evaluado en los datos para clasificar, después multiplicación por alguna previa. Como los PDF son desconocidos en situaciones prácticas (solo se conocen las muestras), trabajar en tal marco requiere estimar la PDF de las variables aleatorias[7].

Otro ejemplo es la detección de novedades. Nos deja imagina un sensor dando, en condiciones normales, salida valores distribuidos según algún PDF[8]. Si el sensor las características se modifican, por ejemplo por el envejecimiento, El PDF de las medidas variará, aunque el las condiciones no han cambiado[9]. Evaluando la diferencia entre PDF es, por tanto, una forma de detectar el envejecimiento, es decir, de evaluar alguna novedad en el proceso de medición[10].

2.Revisión de literature

Por lo tanto, estimar archivos PDF basados en una muestra es de importancia primordial en muchos contextos. Existen mucho de métodos destinados a estimar archivos PDF (ver por ejemplo Árbitro[11]. para una descripción general); mientras que todos ellos son aplicables en el caso univariado, algunos de ellos también se pueden aplicar al multivariado. Sin embargo, y a pesar de la vasta literatura sobre el tema, incluso en el caso univariado no hay consenso sobre qué método utilizar, ni sobre los pros y los contras de estos métodos. En este documento, restringirá el estudio a la estimación de PDF univariante[12].

El objetivo de este documento es brindar algunas ideas sobre los métodos que tradicionalmente utilizan los analistas de datos. Pros y contras a priori y experimentales se proporcionarán comparaciones. Todas las preguntas sobre el uso de los métodos para estimar PDF no será respondido, ni se cubrirán todos los métodos. Sin embargo, este documento brinda pautas al usuario tener que elegir entre la estimación estándar de PDF métodos, cuando se enfrenta a diversos requisitos en cuanto a prestaciones, velocidad y niveles de robustez. Las pautas deben entenderse aquí en un sentido Amplio[13]. No es el propósito de este estudio proporcionar reglas estrictas, ni siquiera para tratar de generalizar los resultados numéricos encontrados en ilustrando ejemplos de situaciones reales. Simplemente creen que esto es imposible, ya que la amplia variedad de situaciones encontradas con varios PDF obviamente pueden conducir a varias conclusions[14]. Este papel lleva otro punto de vista. Mirando los modelos y la estimación algoritmos, es posible tener conocimientos sobre sus comportamiento esperado, por ejemplo, lo que calidad de la estimación, la robustez a pequeños y muestras grandes, las dificultades encontradas debido a parámetros que deben ajustarse, la robustez para estos parámetros, etc. La mayoría de estos Las características no se describen de forma independiente y forma objetiva en la literatura existente, en particular cuando podrían considerarse inconvenientes más que ventajas[15]. La originalidad de este trabajo es, comparar modelos de estimación PDF ampliamente utilizados de la punto de vista de sus características esperadas, más bien que de los resultados numéricos que inevitablemente serían obtenido en ejemplos específicos sin convencer poder de generalización. No obstante, para validar las afirmaciones, estas últimas se ilustran en un PDF ejemplo elegido por su variedad de características; esta parte experimental del papel no debe tomarse como un prueba sino más bien como una ilustración de las afirmaciones de que forman el núcleo de este document[16].

Después de una introducción, la Sección describirá el los métodos más populares utilizados para estimar PDF, sin con el objetivo de ser exhaustivo: histogramas, Parzen ventanas, estimadores basados en cuantificación vectorial y Mezclas gaussianas finitas (FGM)[17]. La Sección 3 comentar estos métodos según criterios y restricciones dadas por aplicaciones prácticas: esperado error de estimación, varianza de este error, cálculo requisitos de esfuerzo y memoria, número de datos para la estimación, número de parámetros en el modelo, etc. Las características esperadas del modelo que se describen en esta sección forman el núcleo de este documento. A continuación, la Sección 4 presentará el experimento procedimiento. En la Sección 5, algunos experimentos seleccionados ilustrar las afirmaciones descritas en la Sección 3: los rendimientos esperados de los métodos se comparan con los resultados de los experimentos. Finalmente, la Sección 6 concluir el trabajo con algunas pautas para el uso de Métodos de estimación de PDF[18].

La estimación de la función de densidad de probabilidad (PDF) es una paso fundamental en las estadísticas como un PDF caracteriza completamente el "comportamiento" de una variable aleatoria. Eso proporciona una forma natural de investigar las propiedades de un determinado conjunto de datos, es decir, una realización de esta variable aleatoria, y llevar a cabo una minería de datos eficiente.

Cuando realizamos la estimación de densidad tres se pueden considerar alternativas. El primer acercamiento, conocido como estimación de densidad paramétrica, asume el los datos se extraen de un modelo de densidad específico. El modelo luego, los parámetros se ajustan a los datos. Desafortunadamente, un La elección a priori del modelo PDF en la práctica no adecuado ya que podría proporcionar una representación falsa de el verdadero PDF.

Una alternativa es construir PDF no paramétrico estimadores, como por ejemplo el histograma o el estimador de ventana de Parzen, para "dejar que los datos hablen por sí mismos". Un tercer enfoque consiste en utilizar semiparamétricos modelos 4. Como técnicas no paramétricas, no asuma la forma a priori del PDF para estimar. Sin embargo, a diferencia de los métodos no paramétricos, el La complejidad del modelo se fija de antemano, con el fin de evitar un aumento prohibitivo del número de parámetros con el tamaño del conjunto de datos, y para limitar el riesgo de sobreajuste. Los modelos de mezcla finita se utilizan comúnmente para sirven para este propósito.

En esta sección, recordamos brevemente el histograma y el estimador de ventana de Parzen, y muestre cómo el kernel El ancho se puede seleccionar a priori en el caso de Parzen. A continuación, presentamos una versión basada en cuantificación vectorial de Parzen, que nos permite reducir la complejidad de su modelo. Finalmente, presentamos modelos de mezcla finita gaussiana.

A continuación, nos referimos a X como una secuencia aleatoria continua. variable, pX (x) como su PDF y {} N xn n = 1 como muestra de X. Nada impide que el método se aplique, en teoría, a situaciones multidimensionales; en este caso X es un azar vector y xn son vectores. Sin embargo, veremos que varios métodos están limitados a pequeñas dimensiones en situaciones prácticas.

El histograma parece depender en gran medida de la elección de σ, ya que regula la suavidad del estimar. Además, la elección de las ubicaciones de las Los orígenes de los contenedores pueden influir en la calidad de la estimación como bien. Una forma de reducir esta dependencia es utilizar un histograma desplazado promediado2. Sin embargo, en este enfoque, uno debe elegir parámetros adicionales a priori, es decir, el número de turnos y el paso de turno.

3.Discusión

Al igual que los histogramas, el estimador de ventana de Parzen5 no asumir cualquier forma funcional del PDF desconocido, ya que permite que su forma se determine completamente a partir de los datos sin tener que elegir una ubicación de los centros.

Elegir a priori el ancho del núcleo σ definitivamente no es la mejor forma de utilizar las ventanas Parzen, como la óptima valor (es decir, el valor que minimiza la medida de disimilitud entre el PDF verdadero y su estimación) depende en gran medida del tipo de datos que estamos tratando con, su número y la cantidad de ruido que están corrompido por.

En es la desviación estándar empírica de los datos y R es el rango intercuartílico de la muestra. los motivación detrás de la introducción del intercuartil rango R es para reducir la sensibilidad de σSIL a valores atípicos, como R calcula el rango intercuartílico de la muestra a partir del Cuantil del 75% al 25%.

En teoría, la aplicación de la fórmula σopt conducir a un sobreajuste, ya que el valor óptimo sería calculado sobre los propios datos. En la práctica, sin embargo, el uso del estimador σSIL en lugar de σopt tiende a sobreestimar el ancho óptimo del grano, como ya mencionado por Silverman3; esto es cierto cuando se trata de la mayoría de los casos de distribuciones no gaussianas, en particular con poblaciones multimodales. Sobreajuste por lo tanto no ocurre en este caso. Sin embargo, como se discutirá en Secciones 3 y 4, solo se encontró la sobreestimación cierto para conjuntos de datos relativamente grandes.

Considere un criterio de error E bien definido y suponga podemos evaluarlo. Dejando variar el ancho del grano σ sobre un cierto rango, uno puede determinar fácilmente la ancho de grano óptimo σopt seleccionando σ que minimiza E. Sin embargo, tenga en cuenta que se debe tener cuidado para evitar sobreajuste al evaluar el rendimiento del modelo: parte de los datos, el conjunto de aprendizaje, debe utilizarse para construir la estimación de Parzen, mientras que un conjunto de validación debe dejarse de lado para evaluar su generalización rendimiento y seleccionando σopt. De hecho, al contrario de Solución enchufable de Silverman, sin estimador de suavizado entra en la selección del ancho óptimo en este caso. Además, deben utilizarse técnicas de remuestreo. en la práctica, como por ejemplo dejar uno fuera, doblar K validación cruzada, validación cruzada de Monte Carlo o Bootstrap para obtener estimaciones fiables. Para una detallada descripción general de estos métodos, nos referimos.

Considerando que se espera que este enfoque supere La regla de oro de Silverman, el precio a pagar es su complejidad computacional. Además, como no sabemos el verdadero PDF en la práctica, este enfoque es inútil para la mayoría de los criterios de error. Sin embargo, este método es muy conveniente en dos casos particulares:
• al considerar problemas con los juguetes (conociendo así el densidad aproximada) para evaluar objetivamente la calidad de los modelos o tener una referencia para comparar;
• al minimizar ISE, porque su dependencia en el PDF desconocido se puede eliminar como se discutió abajo.

Más recientemente, técnicas elaboradas, como la El complemento Sheather-Jones o el bootstrap suavizado, fueron propuesto para seleccionar σ automáticamente. Para revisión completa nos referimos el referencias en el mismo. Sin embargo, estos métodos no suelen ser utilizados por el médico debido a su técnica. complejidad, mientras que muestra poca mejora efectiva en aplicaciones reales. Por lo tanto, excluimos estos técnicas de nuestra discusión.

Considerando que la complejidad computacional de Parzen es relativamente pequeña, la complejidad de su modelo es proporcional a el número de muestras de datos. Esto puede conducir rápidamente a problemas de almacenamiento de memoria. Además, como el núcleo el ancho es común a todos los núcleos, puede ser localmente no coinciden. Esto, a su vez, puede provocar oscilaciones en el colas de distribución o en regiones de baja densidad de la entrada espacio.

Una forma sencilla de evitar estos problemas es realizar la cuantificación vectorial (VQ) de antemano. Una vez que se elige el número de prototipos K, el VQ El algoritmo calcula sus ubicaciones en el espacio de entrada mediante minimizar un error de cuantificación y se asocia a cada prototipo de una región de influencia, llamada región de Voronoi.

Una de las técnicas VQ más populares es la Aprendizaje competitivo (CL) 9-10. En el resto lo haremos limitarnos a CL ya que es suficientemente flexible, limitando al mismo tiempo el número de parámetros. Otro enfoques generalizados son K-medias, Neural Gas11 o Mapas autoorganizados 12. Esto último generalmente resulta en un mejor cuantificación de los datos, al precio de una mayor complejidad. Cuando la cuantificación vectorial las actuaciones son el objetivo a perseguir, tales Los métodos avanzados de VQ son ciertamente preferibles a CL. Sin embargo, en nuestro caso, VQ solo se utiliza como preprocesamiento para reducir la cantidad de datos y aproximadamente Coloque los centroides en ubicaciones apropiadas en los datos. espacio. Usar un método VQ u otro no resulta en diferencias significativas en la aproximación de PDF calidad; además, las variaciones debidas a la VQ inicializacin son ciertamente tan grandes como los debidos a la elección del algoritmo VQ. La influencia del VQ La inicialización se discute en la parte experimental de este papel.

Repita hasta convergencia. En las ecuaciones, á es la tasa de aprendizaje. Generalmente, á disminuye exponencialmente durante las iteraciones del algoritmo.

Al colocar un kernel gaussiano en cada prototipo ck, el PDF desconocido se puede aproximar en un Parzen-like conducta. Además, a cada kernel k, podemos asociar un ancho de kernel local σk, correspondiente al experimental desviación estándar de la correspondiente región de Voronoi. Esto permite tener en cuenta las variaciones de la PDF, estableciendo diferentes anchos de kernel para diferentes partes del espacio X o rango. Finalmente, para forzar un cierta cantidad de superposición entre los granos y aumentar así el rendimiento de generalización, un Se introduce el factor de escala de ancho común h. Un similar La idea se introdujo en para kernels supervisados (en Redes de funciones de base radial); la escala de ancho El factor permite controlar la cantidad de regularización, que se ajustará a un criterio de Dejar uno fuera.

Similar a Parzen, donde el ancho de kernel común debe optimizarse, en Parzen basado en VQ la escala de ancho El factor h también debería optimizarse. De nuevo, por fines de comparación, h se puede seleccionar mediante buscar minimizando un criterio de error bien definido E. En la práctica, el criterio de validación cruzada de dejar uno fuera también podría usarse. De hecho, siguiendo el mismo derivación como la que se utiliza para las ventanas Parzen, podemos reescriba ELOO usando la estimación de densidad basada en VQ.

Las distribuciones de mezcla finita pueden aproximarse a cualquier PDF continuo, siempre que el modelo tenga suficiente número de componentes y proporcionó los parámetros de el modelo se elige correctamente.
En es la probabilidad de x dada la distribución de componentes k y P (k) son la mezcla proporciones o priores.

El EM opera en dos etapas. Primero, en el paso E, el valor esperado de algunos datos "no observados" es calculado, utilizando las estimaciones de los parámetros actuales y la datos observados. Aquí los datos "no observados" son los datos etiquetas de las muestras. Corresponden a la número de identificación de la mezcla diferente componentes y especifique cuál generó cada dato.

Se puede mencionar que la mutilación genital femenina podría verse como Máquinas de vectores de soporte específicas en el contexto de métodos del kernel 16. Sin embargo, en este caso, y al contrario a FGM, los anchos de grano se fijan de antemano y el los centros del kernel están restringidos a los puntos de datos. Incluso si un escasa solución puede obtenerse, tales restricciones en el los anchos y centros de los granos conducen a una cobertura ineficaz de la distribución efectiva, a diferencia de la MGF donde se optimizan los centros y anchos de kernel.

No se espera que todos los métodos de estimación de PDF funcionen de manera similar en todas las situaciones. Ciertamente hay pros y contras a todos los métodos, haciéndolos todos adecuados para algún tipo de aplicaciones. Lo siguiente de esta sección tiene como objetivo dar algunas ideas sobre el esperado desempeños de cada método, con respecto a varios criterios y restricciones impuestas por las aplicaciones. Más precisamente, las actuaciones esperadas a priori de Se investigan los métodos (junto con la varianza de las estimaciones). Como métodos de estimación de PDF a veces tienen requisitos computacionales que no son compatible con las limitaciones de la aplicación, El esfuerzo computacional y los requisitos de memoria serán discutido también. Finalmente, todas estas actuaciones serán relacionados con el número N de datos que son necesarios para obtener una aproximación "adecuada" del subyacente PDF y a la complejidad de los modelos.

En función de las características de los modelos, el Se pueden hacer los siguientes comentarios sobre los métodos. descrito.
• Los histogramas adolecen de varios inconvenientes. Primero, no son continuos por definición. Utilizando no continuo Las estimaciones de PDF pueden revelar problemas en algunos contextos, por ejemplo en binario

bayesiano clasificación, cuando la intersección entre el Se debe encontrar el PDF de las dos clases. Además, la elección de los contenedores (anchos y centros) puede ser demasiado difícil, especialmente en el caso de no liso PDF o de PDF con soporte ilimitado. Excepto su simplicidad computacional, los histogramas no tienen ventaja en comparación con los estimadores Parzen, mientras que tienen inconvenientes complementarios; Parzen estimadores pueden verse como un continuo (y derivable) extensión a histogramas. Finalmente, observemos que los histogramas no se extienden fácilmente a casos multivariados, ya que los contenedores tienden a estar vacíos rápidamente en promedio. Por todas estas razones, los histogramas no ser considerado en la parte experimental de este papel.

• Los estimadores de parzen requieren una elección correcta del parámetro de suavizado σ. Una pequeña σ resultará en sobreajuste: la estimación mostrará picos alrededor cada dato. Por otro lado, una σ grande resultará en un exceso de regularización: la estimación será más suave que el verdadero PDF. A menudo, Silvermann se aplica la regla de oro. De acuerdo con la literatura 2-3, esta regla sobreestima σ para grandes valores del número N de datos, y lo estima más o menos correctamente cuando N es pequeño (al menos para distribuciones unimodales). La sobreestimación principalmente resulta del carácter no gaussiano de el PDF.

• Como σ actúa como parámetro de suavizado en Parzen estimador, σ debe ser naturalmente grande para N pequeños y pequeño para grandes N.

• Excepto si se utiliza la aproximación Leave-One-Out detallado, una estimación de validación cruzada de un valor óptimo de σ es computacionalmente costoso si N es grande. Además, dicha validación cruzada es no es necesario si N es bajo, ya que las reglas del pulgar proporcionan valores adecuados.

• Los métodos basados en la cuantificación vectorial no funcionan bien. Tienden a sobreestimar las colas de los distribuciones. De hecho, en las colas, el número de datos disponible se reduce, lo que conduce a una propiedades de cuantificación. Además, la mayoría de los vectores Los métodos de cuantificación conducen a los llamados efecto "factor de aumento", que consiste en el hecho de que un PDF posterior a VQ refleja el original PDF elevado a una potencia inferior a uno (en otros palabras, la cuantificación vectorial en sí, antes de cualquier Estimación de PDF, sobreestima las colas y subestima los picos de una distribución).

• La desviación estándar de las estimaciones es una preocupación importante también. De hecho, la mayoría de los métodos requieren algo de inicialización, lo que lleva a diferentes estimaciones para diferentes valores iniciales de algunos parámetros. Tener grandes variaciones entre estimaciones, por supuesto, no es una buena propiedad de la método de estimación. Desviación estándar y media son útiles respectivamente para evaluar el real desempeño de un método y su robustez con respecto a la aleatoriedad de inicialización. Parzen los estimadores no sufren este inconveniente, ya que no hay inicialización. La mutilación genital femenina se basa en Potente algoritmo de optimización (EM), por lo tanto, generalmente ofrecen una baja desviación estándar de la estimar. La desviación estándar podría aumentar aunque: cuando un gran nmero K de gauss se utilizan componentes, EM puede producir diferentes soluciones aceptables correspondientes a diferentes configuraciones estables de las mezclas de componentes. Sin embargo, se pueden utilizar varias inicializaciones para reducir la varianza, si un mayor costo computacional es aceptado; múltiples inicializaciones se pueden utilizar en el estimador basado en cuantificación vectorial también para el mismo propósito.

• Más dramáticamente, la optimización FGM (EM algoritmo) puede colapsar si el número K de Las funciones gaussianas son grandes, o de manera equivalente, si el el número N de datos es pequeño 18. Aunque reciente técnicas como la FGM19 máxima penalizada o FGM20 variacional se puede utilizar para evitar esto problema, el colapso ocurre naturalmente cuando uno intenta incorporar demasiada estructura en el modelo de densidad en comparación con el número de disponibles datos. En esta situación, la estimación de Parzen es una alternativa interesante. También tiene la característica de tener un solo parámetro para optimizar (en comparación con varios parámetros para cada función gaussiana en un MGF), que revela una ventaja cuando la El número N de datos disponibles es pequeño (el La minimización de la probabilidad está atrapada en diferentes mínimos locales).

• Si este problema relacionado con un número bajo de datos N disponible no se encuentra, la MGF se espera que funcione bien, probablemente mejor que otros metodos. Si N es grande, uno tiene más conocimiento sobre el PDF subyacente; creciente el número K de componentes en la MGF es por tanto Se espera que aumente el rendimiento de la estimador.

Finalmente, sobre los requisitos de memoria, histogramas y los estimadores Parzen tienen la ventaja de que no requiere ningún aprendizaje. Sin embargo, si N es grande, ambos los requisitos de memoria y la evaluación de la La estimación para cualquier valor de x puede exceder cualquier nivel impuesto por el contexto de la aplicación. los La complejidad de la evaluación es una consecuencia directa de la fórmula de Parzen (es una suma de todas las N muestras). Esta es otro argumento a favor de la mutilación genital femenina.

Sección 3 propiedades detalladas y comportamiento esperado de los métodos de estimación de PDF que resultan directamente de los algoritmos. En el resto de este artículo, muestran principalmente resultados experimentales que sí confirman el comportamiento esperado. También detallaremos otras propiedades que resultan de

experimentos de simulación. Este último puede se considerarán válidos solo para el PDF utilizado y no Se espera que dibuje propiedades generales.

El tamaño N de la muestra utilizada para estimar un PDF será investigado. De hecho, como se introdujo, se espera que algunos métodos no funcionen igualmente cuando N es bajo y cuando N es grande; evaluando qué método utilizar en qué situación es uno de los objetivos de este estudio.

Los experimentos se realizan con un PDF generado. Por tanto, es posible comparar el estimación al PDF real. La comparación tiene que apoyarse en una medida de distancia (entre PDF) que debe ser elegido correctamente también.

El PDF de referencia ilustrado en la Figura 1 ha sido elegido para las simulaciones.
Ha sido construido para incluir una amplia variedad de comportamientos que influyen en el desempeño de los algoritmos: regiones planas y pendientes, afiladas y suaves picos, sesgos y una cola derecha suave y escalada lado izquierdo. Por construcciones, es obviamente diferenciable.

Más precisamente, el PDF de referencia es un mezcla de cuatro PDF gaussianos, es decir, N1 (x | 3.5,1.6), N2 (x | 7.5,2), N3 (x | 12,2), N4 (x | 16.5,1.5), y de uno Gamma PDF con parámetros iguales a 8 y 0,2. Las muestras con N realizaciones se han extraído de la PDF de referencia según un dibujo de Montecarlo esquema.

Se pueden utilizar muchas medidas de distancia para comparar PDF, al menos en el caso unidimensional.
Tenga en cuenta que el factor 1 / (b-a) en la definición de MSE es nada más que una constante de normalización que podría ser añadido a las otras dos medidas también.

Todas las distancias se han utilizado en experimentos condiciones. Desde un punto de vista cualitativo, todos lideran a conclusiones similares sobre el uso del PDF métodos de estimación. Por esta razón, Hellinger la distancia solo se utiliza en lo siguiente de este documento; el La distancia de Hellinger es simétrica (a diferencia del Kullback divergencia) y no demasiado sensible a grandes diferencias entre PDF limitado a pequeñas regiones del espacio (a diferencia del error cuadrático medio). Hay que insistir en el hecho de que la distancia Hellinger haya sido elegida para su adecuación al objetivo de este documento (comparación cualitativa experimental entre PDF métodos de estimación); otras medidas de distancia podrían Sin embargo, revelan más apropiado en contextos reales o situaciones más específicas; por ejemplo, si los picos altos superan un rango limitado en la diferencia entre el PDF tiene que evitarse, uno preferiría el criterio de MSE, debido al mayor exponente en el PDF.

Ventanas Parzen, Parzen basado en cuantificación vectorial ventanas, y se han utilizado mezclas gaussianas finitas para estimar el PDF de referencia detallado en la Sección 4. Por las razones discutidas en la Sección 3, no tomaremos en cuenta cuenta histogramas en los resultados experimentales. El último dos modelos usados para los experimentos incluyen definidos por el usuario parámetros (número de núcleos) que pueden tener influencia en la calidad de los resultados. Además, el el tamaño N de la muestra influye en gran medida en los rendimientos también. Los siguientes experimentos muestran la calidad del métodos de estimación con respecto a los parámetros y el tamaño de la muestra: los experimentos están destinados a ilustrar el análisis. A menos que de otra manera mencionado, los experimentos se han realizado con un tamaño N de la muestra variando según.

Las ventanas parzen son un método determinista una vez que el ancho del núcleo σ es fijo: los rendimientos son determinista también, y sin desviación estándar de la los resultados se pueden calcular. Para cada número N de datos (tamaño de la muestra) como se detalla anteriormente, σ se varió de 0,15 a 2 en pasos de 0,1. La muestra el resultados en términos de distancia Hellinger entre el verdadero PDF y su estimación, con respecto a σ y con N como parámetro.

Como era de esperar, el ancho de grano óptimo σ es pequeño para valores grandes de N y grandes para valores pequeños. De hecho el número de núcleos utilizados en las ventanas Parzen es igual a el número de datos; por lo tanto, la varianza del kernel debe aumentar para N pequeña para crear un PDF fluido estimar.
Se pueden extraer otras conclusiones. Primero, se ve que cualquiera que sea el número N de datos, la curva de error tiene un mínimo. La confirma que se produce un sobreajuste para σ demasiado pequeño y para σ demasiado grande.

Otra conclusión es que, para un número N de datos, el error de aproximación resultante (en σ óptimo) es siempre menor que el error resultante de menor N. Esto significa que aumentar el número de granos siempre es beneficioso, al menos cuando σ es aproximadamente optimizado. Sin embargo, lo mismo no es cierto para una σ fija: en el

centro y la derecha, se puede ver que aumentar el número de granos a fijo σ podría aumentar el error de aproximación.

Lo que es más sorprendente es la cantidad de dependencia observada al variar ó, para un fijo número N de datos. Teniendo en cuenta el logarítmico escala, se ve que con N grande, un error elección de ó puede conducir a errores que son dos órdenes de magnitud mayor que la óptima. Inesperadamente, el dependencia al ancho del kernel ó es mucho mayor para ¡Muestras grandes que pequeñas! Este resultado debe ser en paralelo con heurísticas ad-hoc que normalmente seleccionan ó de una manera no óptima, como el complemento de Silverman: en muchas situaciones, incluidas las no difíciles, sólo una búsqueda exhaustiva (o la correspondiente exclusión elección) puede conducir a una elección adecuada del kernel ancho ó.

Para evaluar tanto el complemento de Silverman como el procedimientos de validación cruzada de omisión para elegir σ, la Figura 3 muestra los rendimientos (nuevamente en términos de Distancia de Hellinger) versus el número N de datos con σ como parámetro. Los resultados del complemento Silvermann y el método de validación cruzada de dejar uno fuera son se muestra también.

La muestra que el complemento de Silverman es un método eficiente para establecer σ sólo cuando un pequeño número de se consideran los datos. Para muestras más grandes, Silverman Los resultados del complemento no son aceptables. El dejar uno fuera El método de validación cruzada da mejores resultados casi En todas partes; sin embargo, debido a su mayor carga computacional, debe usarse solo para grandes muestras.

Al contrario de las ventanas de Parzen donde solo el kernel El ancho σ debe optimizarse, la cuantificación vectorial El método basado en Parzen necesita optimizar dos parámetros: el ancho del kernel a través de la escala de ancho factor hy el número K de granos. Usando un doble método de búsqueda exhaustivo integrado en una validación cruzada procedimiento llevaría a inasequible tiempos de cálculo en la mayoría de las aplicaciones. Por tanto, es necesario desarrollar elecciones heurísticas para establecer al menos una de los dos parámetros. Para evaluar si el la dependencia a uno de los dos parámetros es lo suficientemente baja para desarrollar una heurística eficiente, se realizan experimentos variando el factor de escala de ancho h de 1 a 100 y el número K de granos de 5 a 300. Como método utiliza una inicialización aleatoria de los centros del kernel, cada El experimento se repite 20 veces y la media y se calculan las desviaciones estándar.

La muestra la media de la distancia Hellinger entre el PDF verdadero y su aproximación con respecto al factor de escala de ancho hy el número de granos K.Una dependencia relativamente baja del ancho Se observa el factor de escala h, pero solo para un gran número de granos; la dependencia es mucho mayor para bajas número de granos.

Si para cada número K de granos consideramos el valor de h que da la media mínima de Hellinger distancia, obtenemos los resultados que se muestran. La La desviación estándar experimental también se muestra, junto con el valor óptimo de h en cada experimento.

Las muestran que los resultados mejorados son obtenido al aumentar el número de granos; el el factor de escala de ancho óptimo h debería aumentar en paralela. Los mejores resultados se obtienen con la mayor número de granos usados en los experimentos; esto refleja el hecho de que, en general, partiendo del estándar Parzen estimador de ventanas a la cuantificación vectorial basada en Parzen one disminuye la calidad de la aproximación. Además, el uso de una gran cantidad de granos en el último método conduce a la inestabilidad, como lo confirma el gran desviación estándar observada.

La muestra un detalle de los experimentos realizado con un número fijo de núcleos: K = 300. Se Se observa que la desviación estándar es siempre grande en comparación con la mejora que se podría obtener optimizar el factor de escala de ancho h. Por un lado este es una buena noticia para una heurística esperada para establecer el valor de h, pero por otro lado este resultado muestra que el naturaleza estocástica de la cuantificación vectorial basada El método conduce a una falta de robustez.

En el PDF estimado con la mejor parámetros resultantes de los experimentos (K = 300 y h = 15) se muestra.

La inestabilidad (gran desviación estándar), los pobres resultados en comparación con las ventanas Parzen, y el excesivo tiempo de cálculo necesario para encontrar un valor óptimo de los dos parámetros (en un esquema de validación cruzada) son argumentos para usar las ventanas estándar de Parzen estimador en lugar del basado en cuantificación vectorial, excepto quizás en una situación donde el número N de datos es tan grande que usarlos todos sin preliminares la cuantificación sería poco práctica.

Como Parzen basado en cuantificación vectorial, gaussiano finito Las mezclas (MGF) se basan en dos parámetros: el número K de kernels y el número de iteraciones. Todos los demás Los parámetros, incluidos los anchos del kernel, se fijan mediante aprendizaje a través del algoritmo EM. Sin embargo, como el Este último no tiene un criterio de parada definido, el número de iteraciones de los pasos E y M generalmente se fija en avance, constituyendo por tanto un complemento parámetro.

Los experimentos con 15000 datos fueron realizados por variando el número de iteraciones y el número de granos. La figura 8 muestra el valor medio de Hellinger distancia entre el PDF verdadero y su aproximación, versus estos dos parámetros.

No es sorprendente observar que, en la mayoría de los casos, un un mayor número de iteraciones conduce a mejores resultados; Sin embargo, esto se obtiene al precio de un aumento costo de computación.

En el siguiente experimento, el número de iteraciones se fija en 200 para observar el dependencia de otros parámetros sin interferencias del número de iteraciones. Dependiendo de número de datos y número de núcleos, convergencia Se observaron problemas durante la ejecución del EM algoritmo. Estos problemas se deben al colapso de algunos granos, como se menciona en la literatura18. De hecho si el número de datos es bajo en comparación con el número de kernels, puede suceder que solo un dato esté asociado a un núcleo; la desviación estándar de este último, estimado en el paso M del algoritmo EM, entonces se convierte en cero, lo que conduce a problemas numéricos obvios. La muestra las configuraciones (es decir, el número de datos y número de núcleos) donde el algoritmo EM converge siempre a una solución, donde nunca lo hace y donde a veces converge, entre las 20 corridas realizadas en cada configuración.

La muestra claramente una ejecución sin problemas del EM algoritmo cuando el número de datos es grande o equivalentemente cuando el número de granos es bajo, una falta de convergencia en el caso contrario, y un intermedio situación para el número medio de núcleos y datos.

La muestra la distancia Hellinger entre verdadero PDF y su aproximación con respecto a la número de datos, siendo el número de núcleos un parámetro.

Es interesante observar el bajo nivel desviación obtenida con el algoritmo EM, en el situaciones donde converge. Aunque el estándar La desviación podría aumentar en situaciones específicas, esto resulta generalmente de una configuración particular del número de datos y de núcleos, en algún lugar del región intermedia donde el estándar La desviación se puede calcular pero incluye resultados con inestabilidad numérica.

La comparación de los métodos enfatiza su complementariedades. Primero, como ya se mencionó en el sección anterior, Parzen basado en cuantificación vectorial no aporta ventajas en comparación con el Parzen estándar ventanas; requieren la estimación de un parámetro suplementario (que conduce a un aumento tiempos de cálculo en un esquema de validación cruzada), su el rendimiento es menor, y finalmente su naturaleza estocástica (debido a la inicialización de la cuantificación del vector) conduce a una fuerte falta de robustez. Comparado con el estándar Parzen Windows, Parzen basado en cuantificación vectorial Recomendar solo cuando una gran cantidad de datos es disponible; la cuantificación vectorial puede verse entonces como una preprocesamiento destinado a reducir el tamaño de la muestra, por razones de costos computacionales. Si el tamaño de la muestra es realmente grande, entonces el paso de cuantificación vectorial será inofensivo con respecto a la información contenida en el datos y beneficioso en el nivel de costos computacionales.

4.Conclusión

Excepto en esta situación, dos métodos comparten el liderazgo: mezclas gaussianas finitas (MGF) y parzen ventanas. Para un número reducido de datos, la experiencia de la MGF dificultades numéricas debido al colapso del grano; Por tanto, se prefieren las ventanas Parzen. El ancho del núcleo debe fijarse adecuadamente: a pesar de la sensibilidad del resultados al ancho del grano es menor con una muestra pequeña que con uno más grande, queda que el observado Los resultados varían en un orden de magnitud si el kernel el ancho se elige adecuadamente. La regla de Silverman se puede utilizar en este caso, ya que proporciona resultados con un número reducido de muestras. Para más seguridad, el procedimiento de validación cruzada de dejar uno fuera debería ser usado; proporciona al menos resultados de la misma calidad, y supera en gran medida la regla de oro de Silverman cuando el número de datos es mayor.

Para grandes conjuntos de datos, la mutilación genital femenina no sufre colapsando problemas más. Ellos no son demasiado sensible a la elección de sus parámetros (número de iteraciones y número de núcleos). Además, en este caso de que las ventanas Parzen tuvieran que integrarse en un dejar un esquema de validación cruzada para

configurar su kernel parámetro de ancho; Por tanto, la MGF ofrece una buena alternativa, con una menor complejidad computacional. Esta ventaja, junto con una baja varianza de los resultados y rendimientos comparables a las ventanas Parzen, hace que la mutilación genital femenina sea más adecuada cuando se disponible. Estos resultados se resumen.

Este documento compara la estimación de PDF ampliamente utilizada modelos desde el punto de vista de sus esperados características, en lugar de resultados numéricos que inevitablemente se obtendría en ejemplos específicos sin poder de generalización convincente. Sin embargo, para validar las afirmaciones, estas últimas son ilustrado en un ejemplo de PDF elegido por su variedad de caracteristicas. Otras simulaciones realizadas en otros Los PDF muestran resultados cualitativos similares.

Se demuestra que las ventanas Parzen son las mejores estimador cuando el número de datos disponibles es bajo; bajo significa alrededor de 100-200 datos en una dimensión ejemplo mostrado en este documento, pero este número, por supuesto varía para diferentes PDF y aumenta para dimensiones superiores problemas. El parámetro de ancho del kernel en Las ventanas de parzen pueden configurarse según la regla de pulso de Silverman si el número de datos es definitivamente bajo; sin embargo el procedimiento de validación cruzada de dejar uno fuera debe ser se utiliza cuando no se sabe si la muestra es pequeña suficiente, a pesar del aumento del costo computacional.

Para conjuntos de datos más grandes, la mutilación genital femenina ofrece una interesante alternativa: su complejidad computacional se vuelve más pequeño que el de Parzen incrustado en una validación cruzada esquema, y muestran comparables rendimientos y baja varianza.

Finalmente, Parzen basado en cuantificación vectorial no agregar cualquier ventaja, excepto cuando la muestra es realmente Se prefieren las ventanas grandes y Parzen a pesar de la conclusión anterior; en este caso, la cuantificación vectorial puede considerarse como un preprocesamiento destinado a reducir el tamaño de la muestra, sin reducir el contenido información.

Referencias

[1] Neves, L.C., De Souza, J.B., De Sousa Vidal, C.M., Dias, A.N. “Nonlinear regression analysis applied to data from wastewater treatment of pulp and paper industry through ultrafiltration and microfiltration”, (2017) O Papel, 78 (6), pp. 88-93.

[2] Mariusz, T., Katarzyna, T. “Computer recognition of data structures using cluster analysis and the theory of mathematical records”, (2017) Journal of Konbin, 41 (1), pp. 5-20.

[3] Freitas, F., de Souza, F.N., Costa, A.P., Mendes, S. “The User Manual of qualitative data analysis software: The perceptions of webQDA users”, (2016) RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao, (19), pp. 107-117.

[4] Fontalvo, T.J., De La Hoz, E.J., Olivos, S. “Methodology of data envelopment analysis (DEA) - GLMNEt for assessment and forecasting of financial efficiency in a free trade zone - Colombia”, (2019) Informacion Tecnologica, 30 (5), pp. 263-270.

[5] De La Hoz, E., Fontalvo, T., López, L. “Data Envelopment Analysis and Multivariate Calculus to Assess, Classify and Predict the Productive Efficiency and Innovation of Companies in the Chemical sector”, (2019) Informacion Tecnologica, 30 (5), pp. 213-220.

[6] Buele, J., Espinoza, J., Fiallos, D., Mónica, R.R., Pilco-Salazar, A.-M., Reyes, J., Franklin, S.L. “Application of cobit model for data management (DS11): The case of a clinical analysis center”, (2019) RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao, (E17), pp. 1010-1021.

[7] Fontalvo, T., De La Hoz, E., De La Hoz, E. “Data envelopment analysis method and neural networks in the evaluation and prediction of the technical efficiency of small exporting companies”, (2018) Informacion Tecnologica, 29 (6), pp. 267-276.

[8] Bolivar, T.N., Rivas Almonte, F.U., Toro Flores, Y.A. “Security analysis in network traffic using data mining”, (2019) RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao, 2019 (E21), pp. 314-326.

[9] Recalde, H., Egas, P.F.B., Saltos, M.F.G., Toasa, R. “Temporal analysis and forecast of the ICT use, employing teaching evaluation data of a higher education institution”, (2019) RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao, 2019 (E22), pp. 425-436.

[10] Catricheo, H.R., Godino, J.D., Cezón, P.A. “Developing statistical knowledge in prospective primary school teachers through a data analysis project: Possibilities and limitations”, (2018) Educacion Matematica, 30 (3), pp. 83-100.

[11] Telles, W.R., Rodrigues, P.P.G.W., Silva Neto, A.J. “MOHID platform automatic calibration employing a stochastic optimization method and real data from an extreme climate event in Nova Friburgo-RJ: Part

2–sensitivity analysis and hydrological parameters estimation", (2017) Revista Internacional de Metodos Numericos para Calculo y Diseno en Ingenieria, 33 (3-4), pp. 204-211.
[12] Jiménez, M.Á.S. "Determinants of the competitiveness of the spanish sector of industrial computing, 1995-2015. A panel data analysis", (2018) RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao, 2018 (27), pp. 52-66.
[13] Cristovão, H.M., Fernandes, J.H.C. "Information retrieval in linked data: A model based on concept maps and complex networks analysis", (2018) Transinformacao, 30 (2), pp. 193-207.
[14] Fontalvo Herrera, T.J. "Efficiency of entities providers of health (EPH) in Colombia through data envelopment analysis", (2017) Ingeniare, 25 (4), pp. 681-692.
[15] Suescún, O.B., León, O.P., Britto Agudelo, R., Jaimes, W.A. "A proposed methodology for the selection of the inmotic distribution centers configuration using data envelopment analysis", (2016) Ingeniare, 24 (3), pp. 480-492.
[16] Caballero, R.E., Cadavid, D.V.L., Agudelo Toloza, J.M. "Efficiency of public educational institutions of the district of santa marta (Colombia) through "data envelopment analysis"", (2015) Ingeniare, 23 (4), pp. 579-593.
[17] De Queiroz Filho, A.P. "The definitions of slums and favelas and its implication on population data: A content analysis approach", (2015) Urbe, 7 (3), pp. 340-353.
[18] Altamiranda, J., Aguilar, J., Hernandez, L. "Pattern recognition system based on data mining for analysis of chemical substances in brain", (2015) Computacion y Sistemas, 19 (1), pp. 89-107.