Test de comparación de densidades con el paquete sm

XOEL MONTES VARELA, BORJA SOUTO PREGO, UXÍO MERINO CURRÁS*

Máster en Técnicas Estadísticas

j.montesv@udc.es, borja.souto@udc.es, uxio.merino@udc.es

Resumen

Este artículo investiga métodos no paramétricos para comparar funciones de densidad entre grupos, utilizando la función `sm.density.compare` del paquete `sm` en `R`. La metodología combina: (1) contrastes de hipótesis mediante permutaciones, (2) selección óptima de ventanas de suavizado, y (3) visualización con bandas de referencia. La metodología se aplica a datos biomédicos de supervivencia en tumores cerebrales, demostrando su capacidad para detectar diferencias en las funciones de densidad estimadas más allá de cambios en localización o escala. En el trabajo se comentan las propiedades teóricas del estimador tipo núcleo y se proporcionan directrices prácticas para la interpretación de resultados.

1. Introducción

La estimación no paramétrica de la densidad ofrece un enfoque flexible para comprender la distribución subyacente de los datos sin imponer supuestos paramétricos restrictivos. En muchas aplicaciones, los investigadores comparan cómo se distribuyen los datos entre diferentes grupos para identificar diferencias y determinar si las discrepancias son aleatorias o podrían deberse a patrones sistemáticos. El paquete `sm` de `R` (Bowman y Azzalini, 2024) proporciona herramientas completas para este propósito a través de su función `sm.density.compare`. Estos métodos son ampliamente aplicables en diversos ámbitos científicos donde resulta de interés comparar distribuciones, desde la ingeniería hasta las ciencias sociales.

Esta función se basa en el marco metodológico descrito en la Sección 6.2 de la obra de referencia de (Bowman y Azzalini, 1997). Combina técnicas de visualización (comparación gráfica de densidades), métricas cuantitativas (error cuadrático integrado, ISE) y métodos inferenciales no paramétricos (pruebas de permutación), ofreciendo así un análisis exhaustivo. Su principal ventaja radica en la capacidad de detectar diferencias complejas en la forma de las distribuciones, más allá de simples desplazamientos o cambios de escala, lo que resulta especialmente útil cuando las distribuciones difieren estructuralmente entre grupos.

En este artículo, revisamos y comentamos detalles de implementación en `R` R Core Team (2023), y demostramos su aplicación mediante un estudio con datos reales.

2. Comparación de densidades mediante estimación kernel

Esta sección describe el marco metodológico para comparar densidades mediante estimación kernel, centrándose en el contraste de hipótesis y la implementación computacional.

2.1. Contraste de igualdad de densidades

La función `sm.density.compare` implementa un contraste no paramétrico para evaluar la igualdad de densidades entre k grupos. Formalmente, el test plantea:

*Universidade da Coruña

$$H_0: \quad f_1(y) = f_2(y) = \cdots = f_k(y) \quad \forall y$$
$$H_1: \quad \exists\, i, j, y \mid f_i(y) \neq f_j(y)$$

Para realizar el contraste de hipótesis se utiliza el estadístico de error cuadrático integrado (ISE):

$$T = \begin{cases} \int \left(\hat{f}_1(y) - \hat{f}_2(y)\right)^2 dy & , k = 2 \\ \sum_{i=1}^{k} n_i \int \left(\hat{f}_i(y) - \hat{f}_{\text{global}}(y)\right)^2 dy & , k > 2 \end{cases} \tag{1}$$

Para el caso de más de dos grupos, $\hat{f}_{\text{global}}$ representa la densidad estimada combinando todos los grupos.

2.2. Implementación computacional

La implementación computacional del contraste de densidades en la función `sm.density.compare` sigue un proceso estructurado en cuatro componentes principales. En primer lugar, la selección del parámetro de suavizado se realiza calculando el ancho de banda óptimo (h_{opt}) para cada grupo mediante validación cruzada, seguido del cálculo de la media geométrica de estos valores ($h_{\text{común}} = \left(\prod_{i=1}^{k} h_i\right)^{1/k}$). Este procedimiento asegura que $\mathbb{E}[\hat{f}_i(y) - \hat{f}_j(y)] = 0$ bajo H_0, garantizando insesgadez cuando $f_i = f_j$.

La estimación kernel propiamente dicha se implementa evaluando las funciones de densidad en una rejilla de puntos equiespaciados, utilizando como función núcleo la densidad normal estándar $K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$. Para el contraste de hipótesis formal, la función emplea un test de permutación que redistribuye aleatoriamente las etiquetas de grupo entre las observaciones, generando así una distribución nula empírica del estadístico de prueba. El p-valor resultante se calcula como la proporción de permutaciones en las que el estadístico excede al valor observado en los datos originales ($\hat{p} = \frac{1}{B} \sum_{b=1}^{B} I(T_b \geq T_{\text{obs}})$).

Finalmente, cuando se comparan exactamente dos grupos, la función ofrece la posibilidad de construir bandas de referencia gráficas mediante una transformación estabilizadora de varianza. Esta aproximación se basa en la transformación raíz cuadrada de las densidades estimadas, cuya varianza puede aproximarse por $\frac{R(K)}{2nh_{\text{común}}}$, donde $R(K)$ representa la integral del cuadrado del núcleo. Las bandas de confianza al $95\,\%$ se construyen posteriormente transformando de vuelta a la escala original los intervalos derivados para las densidades transformadas. Este enfoque combinado proporciona tanto una evaluación cuantitativa como una representación visual intuitiva de las diferencias distribucionales entre grupos.

3. Análisis del estimador

Esta sección examina tanto los fundamentos teóricos como el comportamiento práctico del estimador implementado en `sm.density.compare`. En la primera parte, se revisan las propiedades estadísticas clave que garantizan la validez del método, incluyendo su sesgo, varianza y consistencia bajo condiciones generales. Posteriormente, se ilustra su aplicación mediante un estudio de caso con datos biomédicos, demostrando cómo las propiedades teóricas se traducen en resultados interpretables en contextos reales.

3.1. Propiedades teóricas

El estimador kernel para comparación de densidades implementado en `sm.density.compare` satisface propiedades teóricas fundamentales que garantizan su validez estadística. Bajo la hipótesis nula de igualdad de densidades, el uso de un ancho de banda común h asegura que el sesgo del estimador de diferencia sea nulo, cumpliéndose $\mathbb{E}[\hat{f}(y) - \hat{g}(y)] = \int K_h(y-z)(f(z) - g(z))dz = 0$ para todo y. Esta propiedad clave permite comparaciones no distorsionadas entre grupos.

La varianza del estimador sigue la expresión asintótica $\text{Var}[\hat{f}(y) - \hat{g}(y)] \approx \frac{R(K)}{nh}(f(y) + g(y))$,

donde $R(K) = \int K^2(u)du$, mostrando la dependencia del tamaño muestral n y el ancho de banda h. Esta estructura de varianza, combinada con las condiciones de consistencia ($h \to 0$ y $nh \to \infty$ cuando $n \to \infty$), garantiza que el estadístico de prueba sea consistente contra alternativas fijas.

Desde la perspectiva de optimalidad, el estimador minimiza el Error Cuadrático Medio Integrado (MISE) asintótico entre la clase de estimadores lineales cuando se emplean kernels simétricos no negativos. Esta propiedad óptima se mantiene bajo transformaciones monótonas de los datos, preservándose la relación $\hat{f}_{T(Y)}(y) - \hat{g}_{T(Y)}(y) \approx \hat{f}_Y(T^{-1}(y)) - \hat{g}_Y(T^{-1}(y))$, lo que permite extender las comparaciones a escalas transformadas sin perder validez inferencial.

El conjunto de estas propiedades teóricas (sesgo controlado bajo H_0, estructura de varianza conocida, consistencia, optimalidad e invarianza bajo transformaciones) hace del estimador una herramienta robusta para detectar diferencias distribucionales que van más allá de simples cambios de ubicación o escala.

3.2. Ilustración con datos reales

Para ilustrar las capacidades de la función `sm.density.compare`, se empleó un conjunto de datos reales sobre pacientes diagnosticados con tumores cerebrales, categorizados en tres tipos: meningioma, glioma de bajo grado (LG glioma) y glioma de alto grado (HG glioma). Estos datos están disponibles en el paquete `ISLR2` James et al. (2022) de `R`, concretamente en el conjunto de datos `BrainCancer`. El objetivo fue comparar las distribuciones del tiempo de supervivencia (en meses) entre los distintos tipos de tumor, a través de la estimación de sus funciones de densidad.

Dado que la variable de interés —tiempo de vida— está acotada inferiormente por cero (es decir, se encuentra en $\mathbb{R}^+$), y se observan valores cercanos a este límite, la estimación directa de la densidad desplaza parte de la masa de probabilidad hacia valores negativos que no son compatibles con el soporte real de la variable. Para mitigar este problema, se transformaron los datos aplicando el logaritmo antes de realizar la estimación. Esta transformación respeta el soporte positivo de la variable original.

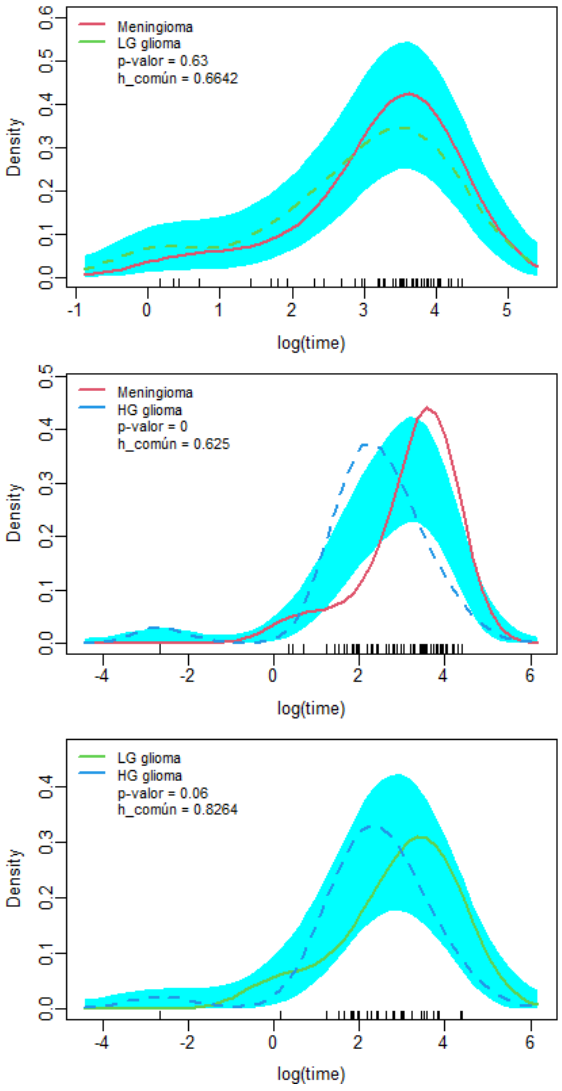

Figura 1: *Comparación de densidades del tiempo de supervivencia entre distintos tipos de tumores cerebrales mediante `sm.density.compare`. Se muestran las bandas de referencia, el p-valor del test de permutación y el ancho de banda común utilizado para el suavizado.*

La Figura 1 muestra tres comparaciones pareadas entre los tipos de diagnóstico. En cada panel, la función `sm.density.compare` estima

las densidades mediante curvas suaves, e incluye bandas de referencia al $95\,\%$, que ofrecen una guía visual sobre la significancia de las diferencias locales entre las curvas.

El gráfico muestra el p-valor del test de igualdad de densidades y el ancho de banda común usado para el suavizado, combinando evidencia gráfica y formal para facilitar tanto el análisis exploratorio como la inferencia estadística. Los resultados sugieren diferencias significativas en los tiempos de supervivencia entre meningiomas y gliomas de alto grado (HG glioma), demostrando la versatilidad de `sm.density.compare` para detectar diferencias distribucionales que van más allá de simples cambios en las medias, particularmente relevante en contextos biomédicos.

4. Conclusiones

El presente trabajo demuestra la utilidad de los métodos no paramétricos para comparar distribuciones de densidad, particularmente cuando las diferencias entre grupos trascienden variaciones en media o varianza. La función `sm.density.compare` emerge como una herramienta integral que combina: (1) visualización mediante estimación kernel, (2) contrastes de hipótesis robustos vía permutaciones, y (3) bandas de referencia para evaluación gráfica de significancia.

Tres aspectos metodológicos clave sustentan su validez: primero, el uso de un ancho de banda común que elimina el sesgo bajo H_0; segundo, la optimalidad del estimador en términos del MISE; y tercero, la adaptabilidad a diferentes estructuras de datos. La aplicación a supervivencia en tumores cerebrales evidenció su capacidad para detectar diferencias distribucionales que métodos tradicionales pasarían por alto, destacando su relevancia en investigación biomédica.

Estos resultados enfatizan el valor del suavizado kernel cuando se prioriza flexibilidad y robustez frente a supuestos paramétricos. Como recomendación práctica, se subraya la necesidad de seleccionar cuidadosamente el parámetro de suavizado y complementar siempre la inspección visual con tests formales, práctica que la función estudiada implementa eficientemente.

Referencias

Bowman, A.W. y Azzalini, A. (2024). *sm: nonparametric smoothing methods (version 2.2-6.0)*. R package manual. University of Glasgow, UK y Università di Padova, Italia. Disponible en: `http://www.stats.gla.ac.uk/~adrian/sm/`.

Bowman, A.W. y Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations*. Oxford University Press, Oxford.

R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

James, G., Witten, D., Hastie, T., Tibshirani, R. y Narasimhan, B. (2022). *ISLR2: Introduction to Statistical Learning, Second Edition (versión 1.3-2)*. R package. Disponible en: `https://CRAN.R-project.org/package=ISLR2`.