alto. Así se produce un fenómeno adverso de infravaloración de la investigación básica en fundamentos de la EIO.

Las comisiones que valoran los curricula de los científicos muestran una tendencia progresiva a dividir el conjunto de revistas de una sección en tercios o cuartos, y a identificar la calidad de una revista con su ubicación dentro de la partición (categorización ordinal). Puesto que las secciones ISI de EIO no están divididas en Fundamentos y Aplicaciones (al contrario que en matemáticas), la citada práctica penaliza fuertemente a los eio de orientación más teórica. Posiblemente, tales comisiones deberían hacer el esfuerzo de asignar las etiquetas "teórica", "aplicada" o "ambas" a las revistas ISI de EIO y aplicar el proceso de categorización en las dos subclases resultantes.

Estas dos diferencias con el resto de las áreas de Matemáticas hacen que, en el caso de la EIO, las actividades de desarrollo no sean suficientemente valoradas y que las actividades de investigación sean generalmente infravaloradas por las comisiones de evaluación de carácter multi-disciplinar. Este fenómeno afecta a procesos de evaluación para la concesión de proyectos competitivos, acciones especiales, premios de doctorado, menciones de calidad a programas de doctorado, tramos de investigación, becarios, etc. El problema está ahí y en ocasiones se agudiza al participar en las citadas comisiones colegas matemáticos que desconocen la idiosincrasia y la relevancia de la EIO. Me temo que aun debemos hacer grandes esfuerzos por dar a conocer las especificidades de la EIO.

1. ARTÍCULOS DE ESTADÍSTICA

LITERATURA Y ESTADÍSTICA: EL PROBLEMA DE LA AUTORÍA DE TIRANT LO BLANC

F. J. Girón, J. Ginebra y A. Riba

1 INTRODUCCIÓN AL PROBLEMA

Tirant lo Blanc es una obra principal de la literatura catalana y para muchos —véase Cervantes (1605), D. Alonso (1951) y M. Vargas Llosa (1991)—, la primera novela moderna en Europa. Escrita entre 1460 y 1465, no fue publicada hasta 1490 en Valencia por Nicolau Spindeler. Consta de 487 capítulos de longitudes muy desiguales y de un total de 418 000 palabras aproximadamente.

Existe un debate, que viene de muy antiguo, acerca de su autoría. En la edición original hay un prólogo debido a Joanot Martorell y un colofón escrito por el que se supone pudiera ser el segundo autor, Martí Joan de Galba. Tanto Martorell como Galba fallecieron antes de que se publicase la primera edición.

Los argumentos a favor de la autoría única se basan en la ***dedicatoria*** y el análisis literario de la obra (véase, p. ej., Givanel i Mas (1911), Vaeth (1918), Marinesco (1978), Martín de Riquer (1990), Hauf (1993), Chiner (1991, 93), Casanova (1994), Badia (1993)), mientras que los de la doble autoría (véase, p. ej., Martínez y Martínez (1916), Entwistle (1927), Moll (1933), Menéndez y Pelayo (1934), Martín de Riquer (1947), Alonso (1951), Coromines (1956), Nicolau d'Olwer (1961), Goerz (1967), Ferrando (1987, 89, 95), Bosh (1987), Rubiera (1990, 92), Wittlin (1990, 93), Hintz (1992)) se basan en el colofón y en el estudio estilístico del lenguaje. La mayoría cree que Galba fue algo más que simplemente un editor de la novela. Capdevila en el prólogo a su edición de 1924–29 resolvió, al parecer, el misterio de las cuatro partes del libro a las que se refiere el ***colofón***, a saber: las aventuras en Inglaterra, la conquista de Rodas, el período en Constantinopla y las guerras del norte de África. Hay diversas y muy dispares opiniones acerca de las partes que escribió cada uno.

Dedicatoria

Y para que en la presente obra ningún otro pueda ser increpado si algún error fuere encontrado, yo, Joanot Martorell, caballero, sólo yo quiero llevar la carga, y no otro conmigo; pues por mi sólo ha sido ventilada en servicio del muy ilustre Príncipe y señor rey expectante Don Fernando de Portugal la presente obra, y comenzada el dos de enero del año mil cuatrocientos sesenta.

Colofón

Aquí acaba el libro del virtuoso y valiente caballero Tirant lo Blanc, ..., que fue traducido del inglés a la lengua portuguesa, y después en vulgar lengua valenciana, por el magnífico y virtuoso caballero Mossèn Joanot Martorell el cual, a causa de su muerte, no pudo acabar de traducir más que tres partes. La cuarta parte, que es el final del libro, ha sido traducida, ..., por el magnífico caballero Mossèn Martí Joan de Galba; y si desfallecimiento fuera hallado, quiere sea atribuido a su ignorancia; ....

La ***Estilometría*** —es decir, el análisis estadístico de características cuantificables, no controlables de forma consciente y propias del autor y no del género, época o editor—, sería la herramienta adecuada para tratar el problema de la autoría de Tirant lo Blanc. En nuestro caso, se trataría de determinar si existe un estilo o más de un estilo y, en el caso en que se detecte más de un estilo, determinar la frontera (o fronteras) de estilo y qué es lo que caracteriza cada estilo; también saber si el cambio de estilo es progresivo o repentino y si éste se puede atribuir a la existencia de dos autores.

Lo que hace interesante y, a la vez, difícil el análisis de la autoría del Tirant lo Blanc comparado con otros problemas de autoría es que no tenemos textos de Martorell ni de Galba con los que comparar, por lo que hemos desarrollado técnicas estadísticas bayesianas novedosas para abordar el problema (véase Girón et al. 2005) en contraposición a los análisis basados en técnicas más informales del análisis de datos como el análisis de correspondencias o la regresión logística (véase, p. ej., Ginebra y Cabos (1998) y Riba y Ginebra (2000, 2005).

Para la obtención de los datos se ha utilizado la edición de Martín de Riquer de 1983 y se han excluído los títulos de los capítulos y las palabras en cursiva que suelen corresponder a citas en latín, con lo que queda un total de 398 242 palabras distribuídas en un total de 489 capítulos. En el análisis estadístico solamente se utilizan los 425 capítulos que tienen más de 200 palabras.

El criterio estilístico que hemos utilizado en esta nota se refiereala *longitud de palabra (número de letras)*, ya usado desde muy antiguo (Mendenhall, 1887) para discriminar entre obras de Shakespeare, Bacon y Marlowe. Mosteller y Wallace (1964, 1984) lo usaron en su famoso estudio de la autoría de los Papeles Federalistas. Los resultados del análisis usando otros criterios estilísticos pueden verse en las referencias.

L.P.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10+**	$\mathbf{N_i}$	$\mathbf{I_i}$
Cap.1	21	59	44	19	33	20	16	17	9	17	255	4.47
2	53	113	80	49	52	33	28	36	16	16	476	4.15
3	109	274	239	128	112	110	76	51	43	32	1174	4.06
4	69	150	126	71	60	71	47	32	23	21	670	4.14
...	...	...	...	...	...	...	...	...	...	...	...	...
484	59	67	68	37	26	32	15	14	17	6	341	3.82
485	96	174	106	57	77	86	42	54	24	25	741	4.18
486	45	88	91	46	40	28	13	30	11	10	402	3.94
487	48	49	62	53	41	36	21	9	16	13	348	4.2

Tabla 1. Longitud de palabra en número de letras por palabra: y_{ij} es el número de palabras de *j* letras en el capítulo *i*-ésimo.

Los datos, según este criterio, se categorizan en una tabla de contingencia de 425 filas ordenadas por 10 columnas de la que la Tabla 1 ofrece un extracto.

2 MODELOS DE PUNTO DE CAMBIO EN SUCESIONES DE DATOS MULTINOMIALES

El detectar un cambio de estilo a lo largo de la obra se puede abordar mediante un modelo estadístico estándar conocido como modelo de punto de cambio para las filas de una tabla de contingencia como es la distribución multinomial.

Para cada capítulo *i*, las filas y_i de las Tabla 1 siguen una distribución multinomial

$$y_i \mid N_i, \boldsymbol{\theta_i} \sim Mu_{l-1}(N_i, \boldsymbol{\theta_i})$$

donde N_i es el total de la fila i-ésima, $\boldsymbol{\theta}_i$ el vector de probabilidades de las 1 categorías de la fila i-ésima.

Una sucesión de variables multinomiales ordenadas $(\mathbf{y}_1,...,\mathbf{y}_n)$ presenta un cambio de modelo (change-point) en el punto *r* si

$$\mathbf{y}_i \mid N_i, \theta_i, r \sim \begin{cases} Mu_{l-1}(N_i, \boldsymbol{\theta}_a) & si \quad i \le r, \\ Mu_{l-1}(N_i, \boldsymbol{\theta}_d) & si \quad i > r, \end{cases}$$

con $\boldsymbol{\theta}_a \neq \boldsymbol{\theta}_d$.

El análisis bayesiano del modelo de un punto de cambio se basa en el cálculo de la distribución a posteriori conjunta de los tres parámetros de interés: el punto de cambio r y los parámetros de las distribuciones multinomiales antes del cambio $\boldsymbol{\theta}_a$ y después del cambio $\boldsymbol{\theta}_d$.

Como las dos teorías sobre la autoría están en marcado conflicto, para ser neutrales usamos como distribuciones a priori sobre *r*, $\boldsymbol{\theta}_a$, $\boldsymbol{\theta}_d$ distribuciones no informativas e independientes. Así, contrastar la hipótesis de que solamente hay un autor es equivalente a que no hay punto de cambio en la sucesión de capítulos; es decir, contrastar

$$H_0: r = n \quad \text{frente a} \quad H_1: r = n.$$

La evidencia a favor de la hipótesis de la autoría única se calcula a partir de la probabilidad a posteriori de H_0 para la Tabla 1, y resultó ser muy próxima a *0*.

Como las distribuciones a posteriori de $\boldsymbol{\theta}_a$ y $\boldsymbol{\theta}_d$ son complejas, su comparación se hace a través de las muestras simuladas lo que nos permite determinar las diferencias estilísticas que hay antes y después del cambio.

El análisis de la distribución a posteriori de r reveló la existencia de un cambio de estilo principal a partir del capítulo 372 y de un número indeterminado de pequeños cambios de estilo, probablemente fruto de intervenciones menores de cada autor en la parte del otro.

Se decidió, en vez de estudiar los posibles cambios de estilo múltiple, realizar un análisis bayesiano de conglomerados, lo que además nos proporcionó mayor información acerca de la autoría de cada uno de los capítulos.

3 ANÁLISIS BAYESIANO DE CONGLOMERADOS

En el anterior análisis de cambio de estilo, la sucesión de filas original se particiona en dos subsucesiones más homogéneas que la original, forzando a que se respete el orden original de las observaciones, mientras que el análisis de conglomerados particiona el conjunto de todos los datos en dos grupos más homogéneos que el total pero sin imponer ninguna restricción en el orden para formar los dos grupos.

El análisis de conglomerados bayesiano se basa en modelos de mixtura tal como se describe a continuación.

Cada una de lasfilas yi de la Tabla proviene de una distribución multinomial $Mu_{l-1}(N_i, \boldsymbol{\theta}_1)$ con probabilidad *p*, y con probabilidad *1- p* de una distribución multinomial $Mu_{l-1}(N_i, \boldsymbol{\theta}_2)$, es decir

$$\boldsymbol{y}_i \mid N_i, p, \boldsymbol{\theta_1}, \boldsymbol{\theta_2} \sim p\,Mu_{l-1}(N_i, \boldsymbol{\theta_1}) + (1-p)\,Mu_{l-1}(N_i, \boldsymbol{\theta_2}),$$

donde *p* representa la proporción de capítulos escritos por el primer autor.

El modelo de mixtura (M) presenta un problema de identificabilidad que se resuelve imponiendo la restricción $p \geq .5$ compatible con las dos hipótesis de autoría.

La asignación de los capítulos a cada uno de los autores, que es problema central del análisis de conglomerados, no se deduce directamente del modelo de mixtura (M). Sin embargo, la posibilidad de asignación de cada capítulo a cada uno de los dos autores se consigue introduciendo variables latentes dicotómicas z_i, $i = 1,...,425$, definidas por

$$z_i = \begin{cases} \mathbf{1} \text{ si } y_i \text{ es del primer autor,} \\ \\ \mathbf{0} \text{ si } y_i \text{ es del segundo autor.} \end{cases}$$

La introducción de las variables latentes z_i permite simplificar no solo el modelo de mixtura (M), sino también el cálculo de la distribución a posteriori de los parámetros p, $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ y la asignación de los capítulos $z = (z_1,...,z_n)$ mediante la aplicación del algoritmo de muestreo de Gibbs cuando la distribución a priori es conjugada respecto de la verosimilitud, como ocurre en nuestro caso al utilizar distribuciones no informativas.

La probabilidad a posteriori de pertenencia del capítulo *i*-ésimo al primer autor es precisamente la esperanza a posteriori $E(z_i \,|\, y_1,...,y_n)$, que se obtiene fácilmente como subproducto de aplicar el algoritmo de Gibbs. Las inferencias sobre la proporción de capítulos del primer autor *p* se recogen en la siguiente gráfica que refuerza la hipótesis de que aproximadamente unas tres cuartas partes del manuscrito se deben al primer autor.

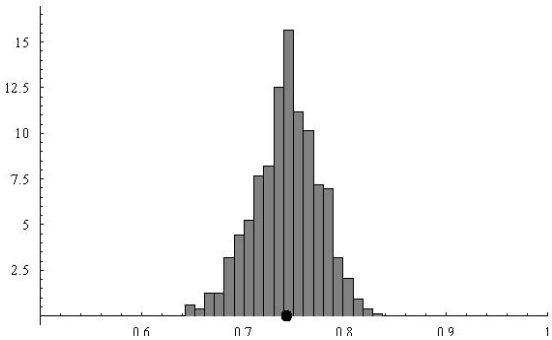

Figura 1. Histograma de una muestra simulada de la distribución a posteriori de p —proporción de capítulos del primer autor—, para las filas de la Tabla 1.

La evidencia a favor de la hipótesis de la autoría única (hipótesis H_0: $p = 1$) frente a la doble autoría (hipótesis H_1: $p < 1$) se calcula, en este caso, aplicando métodos bayesianos de selección de modelos. El resultado es que la $Pr(H_0 | \text{ datos}) \approx 0$.

Las muestras de las distribuciones a posteriori de los parámetros θ1y θ2, obtenidas de la muestra producida por el algoritmo de Gibbs, son muy similares a las resultantes del análisis del punto de cambio de los parámetros $\boldsymbol{\theta}_a$ y $\boldsymbol{\theta}_b$, lo que refuerza la consistencia de los dos procedimientos.

4 CONCLUSIONES

La longitud de palabra, usada como criterio estilístico, detecta un cambio de autor en la transición del capítulo 371 a l 372, en línea con lo dicho en el colofón, de modo que hay evidencia estadística muy fuerte a favor de la doble autoría, aunque no descartamos otras explicaciones alternativas.

El análisis de conglomerados bayesiano muestra además que hay consistencia entre los resultados de cambio en la autoría detectados por el análisis de un cambio de estilo al pasar del capítulo 371 al 372 y el análisis de conglomerados, en el sentido de que minimiza el número de capítulos mal clasificados por el punto de cambio.

Al parecer, hay intervenciones menores —retoques en algunos capítulos— de ambos autores en las partes respectivas atribuidas al otro autor. Serían los capítulos mal clasificados por el cambio de estilo.

Los gráficos de cajas, que no se incluyen en esta nota, son una herramienta muy útil a la hora de entender cuáles son las características que cambian en la frontera de estilo y detectar éstas.

Si comparamos el análisis Bayesiano del problema con los basados en el análisis de datos, resulta que aquel —el bayesiano— no solamente es más informativo sino que también es más sencillo.

5 REFERENCIAS

Ginebra, J. i Cabos, S. (1998). Anàlisi Estadística de l'estil Literari: Aproximació a l'autoria del Tirant lo Blanc. *Afers*, **29**, 185–206.

Riba, A. i Ginebra, J. (2000). Riquesa de Vocabulari i Homogeneitat d'estil en el Tirant lo Blanc. *Revista de Catalunya*, **13**, 99–118.

Riba, A. and Ginebra, J. (2005). Change-Point Estimation in a Multinomial Sequence and Homogeneity of Literary Style. *Journal of Applied Statistics*, (to appear).

Girón, J., Ginebra, J. and Riba, A. (2005). Bayesian Analysis of a Multinomial Sequence and Homogeneity of Literary Style. *The American Statistician*, **59**, **nº 1**, 1–12.

ESTRUCTURA Y ANÁLISIS DE MICROARRAYS

[1]Rivas-López, M.J., [1]Sánchez-Santos, J.M. y [2]De las Rivas, J.
[1] Dpto. Estadística. Universidad de Salamanca
[2] Centro de Investigación del Cáncer. Universidad de Salamanca

A menudo hemos escuchado que los genes, o sus mutaciones, pueden ser muy influyentes a la hora de desarrollar una determinada enfermedad. El DNA existente en el núcleo de las células contiene las instrucciones para la fabricación de proteínas. Un gen es un segmento de DNA que contiene la secuencia codificante específica para construir una proteína concreta. Cuando un gen está activo en una célula, decimos que ese gen está "expresado" en ella. En una célula hay una gran cantidad de genes diferentes activos, que dan lugar a todas las proteínas que funcionan en ese tipo celular. Así por ejemplo, las células humanas epiteliales son diferentes a las células humanas musculares porque a nivel biomolecular los genes expresados y las proteínas presentes en ellas son distintos.

Como la actividad y expresión de los genes da información sobre los procesos biológicos y el comportamiento celular, tanto en estados normales como patológicos, su medición es importante para el avance científico. Hasta hace poco tiempo en los laboratorios de biología y genética molecular se podía medir la actividad de cada gen por separado, pero actualmente con las modernas técnicas de genómica puede medirse simultáneamente la actividad de decenas de miles de genes usando una herramienta nueva conocida como microarray de oligonucleótidos de DNA (Harrington et al. 2000). Los microarrays son dispositivos construidos usando microtécnicas de alta precisión capaces de sintetizar en superficies muy pequeñas (del orden de uno o varios cm^2) miles de copias de moléculas. En el caso de los microarrays para medir la expresión génica lo que se sintetizan son oligonucleótidos de DNA; es decir, fragmentos cortos de DNA. La información obtenida de los microarrays de expresión génica ayuda a contestar importantes preguntas biológicas y biomédicas como son:

- ¿Existen diferencias de actividad génica entre personas sanas y personas con una determinada enfermedad?
- ¿Existen subgrupos genéticos de personas con una determinada enfermedad que responden positivamente a un tratamiento específico?

DISEÑO BASE DE UN MICROARRAY

El DNA es ácido desoxirribonucléico, una biomolécula formada por dos hebras ó cadenas cuyos eslabones son nucleótidos, cada uno incluyendo molecularmente un azúcar (la desoxirribosa), un fosfato y una base nitrogenada. Las bases nitrogenadas son de 4 tipos: adenina (A), citosina (C), guanina (G) y tiamina (T), de modo que cualquier DNA puede ser identificado específicamente por la secuencia lineal de las bases de sus nucleótidos, "secuencia genética", por ejemplo: ATTGCGCATA. De este modo, en