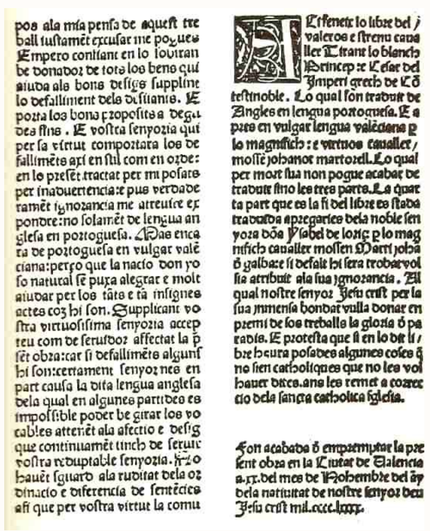
When did author B take over for author A? Confidence Distributions for Change-Points

Céline Cunen, Gudmund Hermansen, Nils Lid Hjort
Department of Mathematics, University of Oslo cmlcunen@math.uio.no

Introduction

The 15th century Catalan chivalry romance *Tirant lo Blanch* is an important work in medieval literature, and a perfect subject for a change-point analysis. The first author, Joanot Martorell, died before the completion of the novel and his friend and fellow knight, Marti Joan de Galba, took over the task. Many analyses, both statistical and literary, have attempted to identify exactly where the change-of-author takes place. Here, we propose our solution, using a new method which both estimates the change-point and provides the associated confidence curve.

Data


pos ala mia pensa de aquest tre
ball tustamẽt excusar me po
Empero confiant en lo sobiran
be donador de tots los bens qui
ajuda als bons desigs supplire
lo defalliment dels discrets. E
posta los bons propositos e deg
des fins. E vostra senyoria qui
per sa virtut comportara los de
fallimẽts axi en stil com en orde
en lo prefet tractat per mi posats
per inadvertencia: e pus verdade
ramẽt sgnorancia me atreuire ex
pondre: no solamẽt de lengua an
gleſa en portogueſa. Dias enca
ra de portogueſa en vulgar valẽ
ciana: perco que la nacio don yo
so natural ſe puxa alegrar e molt
ajudar per los tãt e tã insignes
actes cos hi son. Supplicant vo
ſtra virtuoſiſſima senyoria accep
teu com de ſeruidor affectat la p
fet obra: car ſi defallimẽts alguns
hi ſon: certament senyor nos en
part cause la dita lengua angleſa
dela qual en algunes partides es
impoſſible poder se girar los vo
cables atenẽt ala affectio e deſig
que continuamẽt tinch de ſeruir
voſtra reduptable senyoria. E
hauẽt ſguard ala rudicitat dela or
dinacio e diferencia de ſentẽcies
aſſi que per voſtra virtut la comu

L'effenit lo libre del ſ
valeros e ſtrenu caua
ller Tirant lo blanch
Prince: e Ceſar del
Imperi grech de Cõ
ſtentinoble. Lo qual ſon traduit de
Angles en lengua portogueſa. E a
pres en vulgar lengua valẽciana p
lo magnifich: e virtuos cauallier,
moſſẽ johanot martorell. Lo qual
per mort ſua non pogue acabar de
traduir ſino les trees parts. E a quar
ta part que es la ſi del libre es ſtada
trabuada apregaries dela noble ſen
yora bõa Viabel de lorig p lo mag
nifich cauallier moſſen Marti joha
ẽ galbace ſi defalt hi ſera trobat vol
ia attribuit ala ſua ignorancia. El
qual noſtre ſenyor Jeſu criſt per la
ſua immensa bondat vulla donar en
premi de los treballs la gloria ẽ pa
radis. E protesta que ſi en lo dit li
bre haura poſades algunes coſes ẽ
no ſien catholiques que no les vol
hauer dites. ans les remet a correc
cio dela ſancta catholica eſgleſia.

Fon acabada ẽ emprentar la pre
ſent obra en la Ciutat de Valencia
a.xx. del mes de Nohembre del any
dela natiuitat de noſtre ſenyor deu
Jeſu criſt mil.cccc.lxxx.

Our data is the text of an ancient Catalan book with 487 chapters and we search for the chapter where author B takes over from author A. From the text we can measure, for all chapters $i = 1, \ldots, 487$:

- y_i = average word length in chapter i, m_i = number of words in chapter i
- $\mathbf{z}_i$ = the vector of proportions of words of length 1 to 10 in chapter i.

For example $z_1 = (0.08, 0.23, 0.17, 0.08, 0.13, 0.08, 0.06, 0.07, 0.04, 0.07)$, so 8% of the words in chapter 1 are one-letters words (length 1), and 7% of the words have more than 10 letters.

General method

We denote the unknown change-point by τ. To estimate τ and the associated uncertainty we use a method described in Cunen, Hermansen and Hjort (2016). In this case, we assume independent data y_i, following a parametric model with

$$Y_i \sim f(y, \theta_L) \quad \text{for } i \leq \tau$$
$$Y_i \sim f(y, \theta_R) \quad \text{for } i \geq \tau + 1$$

We can then form the profile log-likelihood function

$$\ell_{\text{prof}}(\tau) = \max\{\ell(\tau, \theta_L, \theta_R) \colon \text{all } \theta_L, \theta_R\}$$
$$= \sum_{i \leq \tau} \log f(y_i, \widehat{\theta}_L(\tau)) + \sum_{i \geq \tau+1} \log f(y_i, \widehat{\theta}_R(\tau)).$$

The maximiser $\widehat{\tau}$ of this function is a good estimate of the change-point. From this we form the deviance function

$$D(\tau, Y) = 2\{\ell_{\text{prof}}(\widehat{\tau}) - \ell_{\text{prof}}(\tau)\}.$$

In order to find the full confidence curve (essentially confidence intervals at different levels) we need the estimated distribution of $D(\tau, Y)$ for each position τ,

$$\text{cc}(\tau) = \Pr_{\tau, \widehat{\theta}_L, \widehat{\theta}_R}\{D(\tau, Y) < D(\tau, y_{\text{obs}})\}.$$

This can be computed via stochastic simulation: we generate a large number B of simulated datasets Y^* from $f(y, \widehat{\theta}_L)$ to the left of τ and $f(y, \widehat{\theta}_R)$ to the right, at each candidate value τ, and calculate

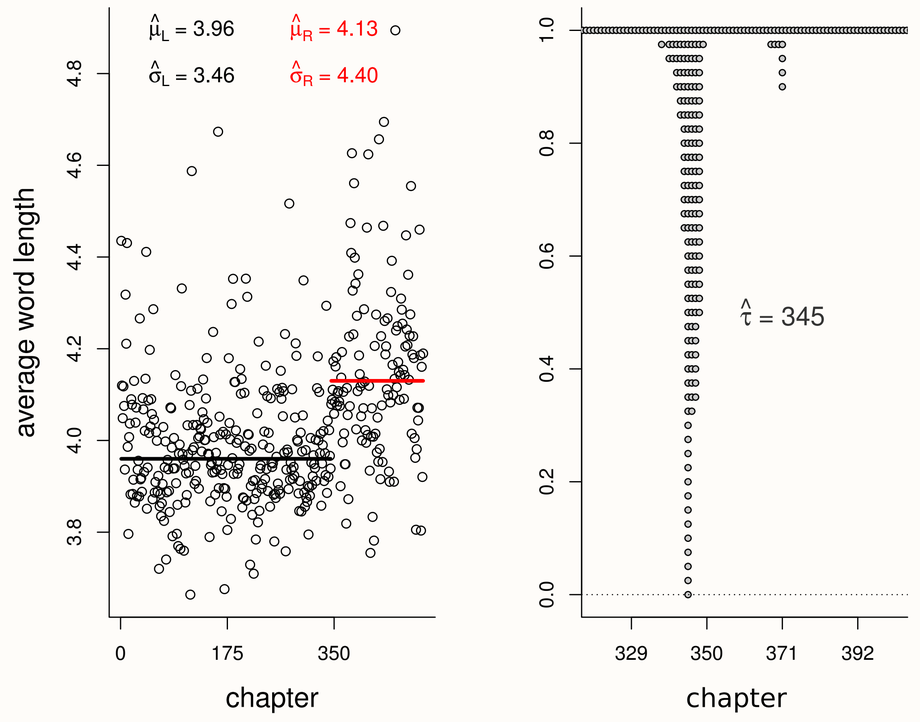
$$\text{cc}(\tau) = \frac{1}{B} \sum_{j=1}^{B} I\{D(\tau, Y_j^*) < D(\tau, y_{\text{obs}})\}.$$

Average word length

For the average word length in each chapter we assume the following model,

$$y_i \sim \text{N}(\mu_L, \sigma_L^2/m_i) \text{ for } i \leq \tau$$
$$y_i \sim \text{N}(\mu_R, \sigma_R^2/m_i) \text{ for } i \geq \tau + 1.$$

According to this model, our best guess is that the change of author takes place in chapter 345. Author B uses longer words on average, and is also more variable across chapters. (The mean number of words per chapter is $\bar{m} = 927$.)

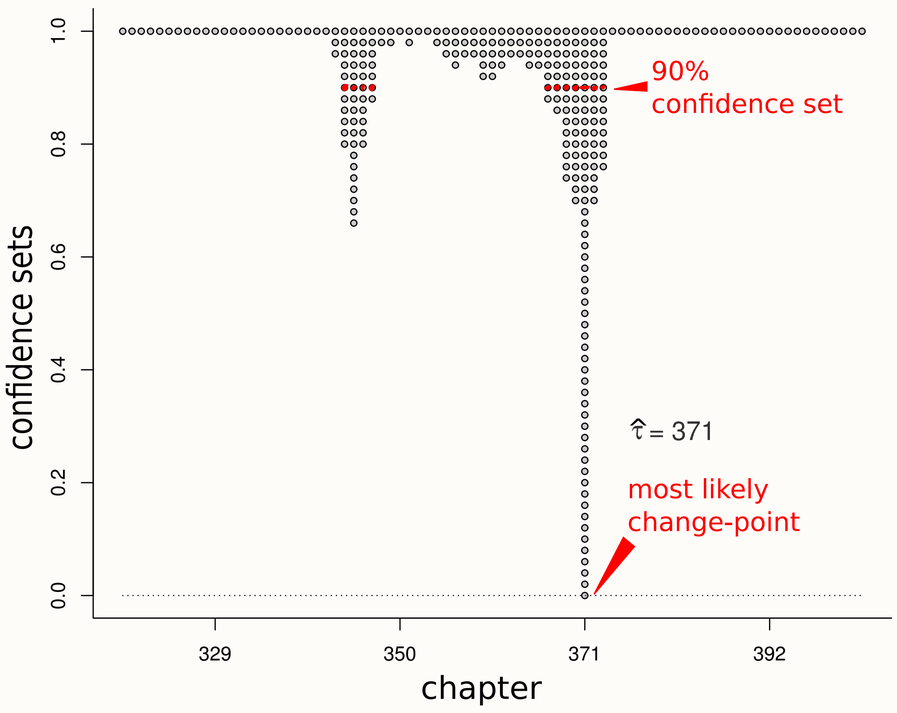


Proportion of words of different length

For the vector of proportions of words of length 1 to 9 (disregarding element no. 10, since the proportions sum to one) we assume the following multinormal distribution,

$$\boldsymbol{z}_i \sim \text{N}_9(\boldsymbol{\xi}_L, \Sigma_L/m_i) \text{ for } i \leq \tau$$
$$\boldsymbol{z}_i \sim \text{N}_9(\boldsymbol{\xi}_R, \Sigma_R/m_i) \text{ for } i \geq \tau + 1.$$

The estimated mean vectors indicate that author B differs from author A in using more one-letter words and also using more long words (with more than 7 letters). The change-of-author point-estimate according to this model is chapter 371, but the model also places some confidence in the change taking place earlier, in chapter 345.



Conclusion and references

Using word lengths as a measure of literary style, we have discovered two clear candidates for the chapter where Marti Joan de Galba took over from Joanot Martorell. Both our analyses indicate that the change in *Tirant lo Blanch* takes place in the last quarter of the book. This is consistent with previous studies, both statistical and literary.

Chen, H. & Zhang, N. (2015). Graph-based change-point detection. *Annals of Statistics* 43, 139–176.

Cunen, C., Hermansen, G. and Hjort, N.L. (2016). Confidence distributions for change-points and regime shifts. *To appear in Journal of Statistical Planning and Inference.*

Riba, A. & Ginebra, J. (2005). Change-point estimation in a multinomial sequence and homogeneity of literary style. *Journal of Applied Statistics* 32, 61–74.

focustat
FOCUS DRIVEN STATISTICAL INFERENCE WITH COMPLEX DATA