CONFIDENCE DISTRIBUTIONS FOR CHANGE-POINTS AND REGIME SHIFTS

Céline Cunen, Gudmund Hermansen and Nils Lid Hjort

Department of Mathematics, University of Oslo

ABSTRACT. Suppose observations $y_1, \ldots, y_n$ stem from a parametric model $f(y, \theta)$, with the parameter taking one value θ_L for $y_1, \ldots, y_\tau$ and another value θ_R for $y_{\tau+1}, \ldots, y_n$. This article provides and examines two different general strategies for not merely estimating the break point τ but also to complement such an estimate with full confidence distributions, both for the change-point τ and for associated measures of differences between the two levels of θ. The first idea worked with involves testing homogeneity for the two segments to the left and the right of a candidate change-point value at a fine-tuned level of significance. Carrying out such a scheme requires having a goodness-of-fit test for constancy of the θ parameter over a segment of indices, and we also develop classes of such tests. These also have some independent interest. The second general method uses the log-likelihood function, profiled over the other parameters, and we show how this may lead to confidence inference for τ. Our methods are illustrated for four real data stories, with these meeting different types of challenges.

Key words: change-points, confidence distributions, homogeneity testing, log-likelihood profiling, monitoring bridges, regime shifts, Tirant lo Blanch

1. Introduction and summary

Many types of processes and natural phenomena experience change-points, sometimes via a jump in mean level and on other occasions via different and perhaps more subtle changes of behaviour. Such changes and discontinuities, when parameters of a model change from one state to another, are variously called break-points, tipping points, paradigm or regime shifts, structural changes or critical transitions, depending on the type or school of application. There is naturally a vast literature inside several areas of application, from engineering (see e.g. Frick et al. (2014)), economics and finance, to biology (e.g. Gould & Eldredge (1977)), meteorology, geology, climate, sociology and history (cf. Spengler (1918), Fukuyama (1992)). As Gladwell (2000) writes in *The Tipping Point*, "the tipping point is that magic moment when an idea, trend, or social behavior crosses a threshold, tips, and spreads like wildfire". There is similarly a large literature regarding aspects of estimation and assessment of change-points inside statistical methodology. The present paper is

Date: August 2017.

a contribution to the methodological side but also presents real data applications stories. Our methods aim at spotting change-points, but, importantly, along with a full assessment of uncertainty, in the form of confidence distributions.

A fruitful statistical framework is as follows. Suppose $y_1, \ldots, y_n$ are independent from a model with density say $f(y, \theta)$, with θ of dimension say p. Our theme is that of pinpointing and providing full inference for the break point τ, assumed to exist, where the θ associated with $y_1, \ldots, y_\tau$ is equal to one value, say θ_L, whereas the parameter vector behind $y_{\tau+1}, \ldots, y_n$, say θ_R, is different. For various applications it may be necessary to extend this framework to models with dependence, as for time series, and several of our methods work also for such cases. The statistical challenge is to estimate τ, along with measures of uncertainty. The traditional ways of reporting precision of parameter estimates are via standard errors (estimates of standard deviation) or say 95% confidence intervals. Our preferred format is that of a full confidence curve, say $\mathrm{cc}(\tau, y_{\mathrm{obs}})$, based on the observed dataset y_{obs}. Its interpretation is that, at the true change-point parameter τ, the set $R(\alpha) = \{\tau\colon \mathrm{cc}(\tau, Y) \leq \alpha\}$ ought to have probability approximately equal to α, with Y denoting a random dataset drawn from the model; see Schweder & Hjort (2016) for a full account of confidence distributions. In particular, confidence sets at any confidence level can be read off from the confidence curve.

The theory and applications of confidence distributions work out more easily for continuous parameters in smooth models, for several reasons. First, for a continuous parameter there is then a possibility of having exact or nearly exact confidence distributions, in the sense that $R(\alpha)$ given above has probability equal to or very close to α, for each confidence level α. This is not fully attainable for the present case of change-point parameters, as the natural statistics informative for τ, like a point estimator $\widehat{\tau}$, have discrete distributions. Secondly, various methods and results pertaining to continuous parameters of smooth models, related to exact or approximate distributions for such statistics, like large-sample normality or chi-squaredness of deviances, are not valid and have no clear parallels when it comes to inference for τ. Confidence distributions and confidence curves may nevertheless be fruitfully constructed for various situations with discrete parameters, as developed in Schweder & Hjort (2016, Ch. 3). This is also the line of development and investigation for the present paper.

In Sections 2 and 3 we propose two different general methods for obtaining such confidence curves for change-points. The first of these requires having a homogeneity test for each given segment of data points where the hypothesis of no change can be accurately examined. For this reason we develop classes of general goodness-of-fit tests for such homogeneity hypotheses in Section 4. Tests we develop there, based on

successive log-likelihood maxima, ought also to have independent interest. Questions regarding behaviour and performance of our different confidence distributions are then treated in Section 5.

The methods we develop in this paper are then shown at work for four different stories with real data, each involving separate challenges. In Section 6 a Poisson model is used to assess British mining disasters 1851–1962, with confidence inference for both the change-point and the relative change. In Section 7 we use different versions of our methodology to pinpoint precisely where the second author (Marti Joan de Galba) took over for the first author (Joanot Martorell), in what is arguably the world's first proper novel, *Tirant lo Blanch*, published in València 1490. Several scholars have previously worked with multinomial models for word sizes, but we demonstrate that the data are overdispersed, inviting the use of multinomial-Dirichlet type models. Then in Section 8 we consider a time series of the number of skiings days at a certain place near Oslo, from 1897 to 2014, and where the question is precisely when Nature started changing her ways. Finally in Section 9 we examine an important and long-running time series from fisheries sciences, consisting of the liver quality of skrei (the North-East Atlantic cod, *Gadus Morhua*), from 1859 to 2013, along with potentially influencing covariates. The aim is again to pinpoint where a moderately complex model for such data experiences a regime shift. We end our article by offering a list of concluding remarks in Section 10, some pointing to further relevant research for change-point inference.

For further pointers to the statistics literature, regarding both methods and applications, see e.g. Frigessi & Hjort (2002) for a general discussion of discontinuities in statistical models, in their introduction to a special issue of *Journal of Nonparametric Statistics* on such topics, and the edited volume Carlstein et al. (1994). Frick et al. (2014) develop methods for joint inference of multiple jumps in a certain class of models, and references in that paper give pointers to several other approaches to change-point analyses. For Bayesian approaches, consult Carlin et al. (1992), Fearnhead (2006), and also Section 5.3 below.

2. General method A: via tests of homogeneity

Though we shall work with models with dependence later in our paper, we assume in the present section that the Y_i are independent, with density $f(y,\theta_i)$ for observation i. Assume further that we for each given n have managed to construct a well-working goodness-of-fit test for the homogeneity hypothesis $H_{1,n}\colon \theta_1=\cdots=\theta_n$, say $Z_{1,n}$, with null distribution $G_{1,n}$. Testing $H_{1,n}$ at level 0.05, for example, is then carried out by rejecting if $Z_{1,n}>G_{1,n}^{-1}(0.95)$, etc. We shall come back to classes of such tests in Section 4.

Consider now the regime shift setup, where the θ_i are equal to a θ_L for $i = 1, \ldots, \tau$ but equal to a different θ_R for $i = \tau+1, \ldots, n$. To form a confidence set for τ, at confidence level α, we suggest forming

$$\begin{aligned} R(\alpha) &= \{\tau\colon H_{1,\tau} \text{ is accepted at level } \sqrt{\alpha}, H_{\tau+1,n} \text{ is accepted at level } \sqrt{\alpha}\} \\ &= \{\tau\colon Z_{1,\tau} \le G^{-1}_{1,\tau}(\sqrt{\alpha}), Z_{\tau+1,n} \le G^{-1}_{\tau+1,n}(\sqrt{\alpha})\} \end{aligned} \tag{2.1}$$

for each of a grid of α values. The probability that τ belongs to this random set, under the true τ, is then

$$P_\tau\{\tau \in R(\alpha)\} = \sqrt{\alpha}\sqrt{\alpha} = \alpha.$$

Note that $R(\alpha)$ consists of points, seen as candidate values for τ at level confidence α, not an interval, per se; also, it may not be connected, as seen in e.g. Figure 9.2.

For an easy illustration, suppose $Y_i \sim \mathrm{N}(\theta_i, 1)$. Here there is a simple test for homogeneity for any given segment of observations using $Q_{a+1,a+b} = \sum_{i=a+1}^{a+b}(Y_i - \bar{Y})^2$, with $\bar{Y} = \bar{Y}_{a+1,a+b}$ the average, and which has a simple χ^2_{b-1} null distribution (other tests will be considered in Section 5). Thus we may easily find the set

$$R(\alpha) = \{\tau\colon Q_{1,\tau} \le H^{-1}_{\tau-1}(\sqrt{\alpha}), Q_{\tau+1,n} \le H^{-1}_{n-\tau-1}(\sqrt{\alpha})\} \tag{2.2}$$

for each confidence level α, writing $H_\nu(\cdot)$ for the distribution function of a χ^2_ν. The $R(\alpha)$ sets can then be displayed for α values $0.01, 0.02, 0.04, \ldots, 0.96, 0.98, 0.99$, say. Simple simulations reveal that the $R(\alpha)$ sets for given confidence level α might not be connected, i.e. may consist of a union of different connected sets, and also that they may be empty for smaller levels. The sets $R(\alpha)$ can be used to define a confidence curve for our method A,

$$\mathrm{cc}_A(\tau, y) = \min\{\alpha\colon \tau \in R(\alpha)\}. \tag{2.3}$$

This curve fulfils the important property that it will be uniformly distributed at the true change-point value τ_0. This is easily established by realising that $\mathrm{cc}_A(\tau_0, Y) \le \alpha$ is equivalent to $\tau_0 \in R(\alpha)$. For some further properties of this confidence curve, see Section 5.

The $\sqrt{\alpha}\sqrt{\alpha} = \alpha$ idea works of course also with other combinations, like using $\alpha^{\tau/n}\alpha^{1-\tau/n}$ for the τ under scrutiny, which we find tends to work slightly better in terms of leading to somewhat slimmer confidence sets; see Section 5. For the illustration just considered, cf. (2.2), there are other homogeneity tests that may be used, in addition to the simple chi-squared method used there, and some alternatives are worked with in Section 5.

In more complex models the situation is less clear-cut than for the illustration around eq. (2.2), not due to any conceptual difficulties with method (2.1), but because we may not have a test of homogeneity with an exact null distribution fully free of parameters. As long as there is a decent test $Z_{1,n}$, for each stretch 1 to n, with

a null distribution exactly or approximately independent of any underlying parameters, we are very much in business, however. We discuss two classes of such tests in Section 4. It is also important to realise that the (2.1) method works in complicated and perhaps high-dimensional situations, as long as there is such a homogeneity test. A case in point is the nonparametric graph-based scan statistics method of Chen & Zhang (2015), which may be put to work as long as a similarity measure on the sample space can be given. In fact Chen & Zhang (2015) utilise an idea similar to our (2.1) above, but in the context of constructing a single confidence interval inside a special model framework only; our concern is that of a full confidence curve, and we emphasise the broad generality of the approach. One may also find traces of related ideas, such as for blocking parameters into groups and identifying splits, in Cox & Spjøtvoll (1982); Worsley (1986). We take time to mention that a Bonferroni version of the argument may be used in cases where data from the left and right segments are dependent, with $\frac12+\frac12\alpha$ replacing $\sqrt{\alpha}$ in (2.1, yielding an alternative set $R_b(\alpha)$; this secures a conservative $P_\tau\{\tau\in R_b(\alpha)\}\geq\alpha$. The difference is actually slight for confidence levels $\alpha>\frac12$ and very small for the higher levels.

3. General method B: profiled log-likelihood and deviance

Suppose in general terms that $Y_1,\ldots,Y_\tau$ come from $f(y,\theta_L)$ and $Y_{\tau+1},\ldots,Y_n$ stem from $f(y,\theta_R)$. This corresponds to a log-likelihood function of the form

$$\ell(\tau,\theta_L,\theta_R)=\sum_{i\leq\tau}\log f(y_i,\theta_L)+\sum_{i\geq\tau+1}\log f(y_i,\theta_R)=\ell_{1,\tau}(\theta_L)+\ell_{\tau+1,n}(\theta_R).$$

We shall see how profiled versions may lead to confidence distributions, for both the breakpoint position τ and for the degree of change, suitably measured.

3.1. **Confidence for the breakpoint.** From the function above we may compute the profile log-likelihood function

$$\begin{aligned}\ell_{\rm prof}(\tau)&=\max_{\theta_L,\theta_R}\ell(\tau,\theta_L,\theta_R)=\ell(\tau,\widehat\theta_L(\tau),\widehat\theta_R(\tau))\\&=\ell_{1,\tau}(\widehat\theta_L(\tau))+\ell_{\tau+1,n}(\widehat\theta_R(\tau)),\end{aligned}\tag{3.1}$$

involving the maximisers of $\ell(\tau,\theta_L,\theta_R)$ over θ_L and θ_R for given τ. The maximiser of $\ell_{\rm prof}$ is the maximum likelihood (ML) estimator $\widehat\tau$, yielding also the ML estimators $\widehat\theta_L=\widehat\theta_L(\widehat\tau)$ to the left, $\widehat\theta_R=\widehat\theta_R(\widehat\tau)$ to the right. From the profile we form and display the deviance function

$$D(\tau,Y)=2\{\ell_{\rm prof}(\widehat\tau)-\ell_{\rm prof}(\tau)\}.\tag{3.2}$$

To construct a confidence curve for τ based on the deviance, consider the estimated distribution of $D(\tau,Y)$ at position τ,

$$K_\tau(x)=P_{\tau,\widehat\theta_L,\widehat\theta_R}\{D(\tau,Y)<x\}.$$

The Wilks theorem says that $K_\tau(x)$ is approximately the distribution function of a χ^2_1, in the case of parametric models smooth in its continuous parameters. There is no Wilks theorem in the present case of a discrete-valued parameter τ, however, so we typically need to resort to computing $K_\tau(x)$ by simulation. Also, $D(\tau, Y)$ has a discrete distribution, say with positive point probabilities $k_\tau(x)$ for certain x; in particular, there is a positive probability $k_\tau(0) = P_{\tau,\widehat{\theta}_L,\widehat{\theta}_R}\{\widehat{\tau} = \tau\}$ that the deviance is zero. Hence the probability transform $K_\tau(D(\tau, Y))$ does not have an exact uniform distribution.

We shall nevertheless work with the construction

$$\mathrm{cc}(\tau, y_{\mathrm{obs}}) = K_\tau(D(\tau, y_{\mathrm{obs}})) = P_{\tau,\widehat{\theta}_L,\widehat{\theta}_R}\{D(\tau, Y) < D(\tau, y_{\mathrm{obs}})\}. \tag{3.3}$$

The probability that $\mathrm{cc}(\tau, Y) \le \alpha$, under the true change-point parameter τ, is often well approximated with α, allowing the interpretation that confidence sets for τ can be read off from a plot of $\mathrm{cc}(\tau, y)$, which we call a confidence curve. The $\mathrm{cc}(\tau, y)$ of (3.3) is the acceptance probability for τ, or one minus the p-value for testing that value of τ, using the deviance based test which rejects for high values of $D(\tau, Y)$. We compute K_τ and hence $\mathrm{cc}(\tau, y)$ by simulation, i.e.

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

for a large enough number B of simulated copies of datasets Y^*. This needs to be carried out for each candidate value τ, with generated data Y_i^* from $f(y, \widehat{\theta}_L)$ to the left of τ and $f(y, \widehat{\theta}_R)$ to the right of τ. For a related idea see Section 10.1.

3.2. **The normal case.** An important special case of our general problem formulation is that of the normal with constant variance. Consider first the case where this variance is known, for convenience now taken to be one. With levels ξ_L and ξ_R to the left and to the right, the log-likelihood is

$$\ell(\tau, \xi_L, \xi_R) = -\tfrac{1}{2} \sum_{i \le \tau} (y_i - \xi_L)^2 - \tfrac{1}{2} \sum_{i \ge \tau+1} (y_i - \xi_R)^2,$$

leading to $\ell_{\mathrm{prof}}(\tau) = -\frac{1}{2}\{Q_L(\tau) + Q_R(\tau)\}$, with $Q_L(\tau) = \sum_{i \le \tau}\{y_i - \bar{y}_L(\tau)\}^2$ and $Q_R(\tau) = \sum_{i \ge \tau+1}\{y_i - \bar{y}_R(\tau)\}^2$, writing $\bar{y}_L(\tau)$ and $\bar{y}_R(\tau)$ for the averages to the left and the right of τ. The ML for τ is the value minimising the sum of these empirical variances to the left and the right, or, equivalently, maximising $\tau \bar{y}_L(\tau)^2 + (n - \tau)\bar{y}_R(\tau)^2$. Confidence statements for τ can then be reached via the recipe above, based on the deviance

$$D(\tau, y) = Q_L(\tau) + Q_R(\tau) - Q_L(\widehat{\tau}) - Q_R(\widehat{\tau}).$$

Next assume that the model takes $\mathrm{N}(\xi_L,\sigma_L^2)$ to the left and $\mathrm{N}(\xi_R,\sigma_R^2)$ to the right, with the four parameters being unknown, in addition to the breakpoint. Maximising the log-likelihood

$$\begin{aligned}\ell = &-\tau\log\sigma_L - \tfrac12(1/\sigma_L^2)[Q_L(\tau)+\tau\{\bar y_L(\tau)-\xi_L\}^2]\\ &-(n-\tau)\log\sigma_R - \tfrac12(1/\sigma_R^2)[Q_R(\tau)+(n-\tau)\{\bar y_R(\tau)-\xi_R\}^2]\end{aligned}$$

over first ξ_L,ξ_R and then σ_L,σ_R yields the profile log-likelihood function

$$\ell_{\rm prof}(\tau) = -\tau\log\widehat\sigma_L(\tau) - (n-\tau)\log\widehat\sigma_R(\tau),$$

where $\widehat\sigma_L(\tau)^2=(1/\tau)\sum_{i\le\tau}\{y_i-\bar y_L(\tau)\}^2$ and similarly with $\widehat\sigma_R(\tau)^2$. We see that the ML estimate of τ is the value that minimises $\widehat\sigma_L(\tau)^{\tau/n}\widehat\sigma_R(\tau)^{1-\tau/n}$. Also,

$$D(\tau,y) = 2\{\tau\log\widehat\sigma_L(\tau)+(n-\tau)\log\widehat\sigma_R(\tau)-\widehat\tau\log\widehat\sigma_L(\widehat\tau)-(n-\widehat\tau)\log\widehat\sigma_R(\widehat\tau)\},$$

and a confidence curve can be based on this, as per (3.3).

In this brief section on inference for the breakpoint in the normal model we finally include the important case where data follow $\mathrm{N}(\xi_L,\sigma^2)$ to the left and $\mathrm{N}(\xi_R,\sigma^2)$ to the right, i.e. with a common σ. This requires a modest extension of (3.1)–(3.2) to the case of common parameters being present on both sides of the breakpoint. The point is that recipe (3.3) for the confidence curve is still operable and valid. The log-likelihood function for this four-parameter model becomes

$$-n\log\sigma - \tfrac12(1/\sigma^2)[Q_L(\tau)+\tau\{\bar y_L(\tau)-\xi_L\}^2+Q_R(\tau)+(n-\tau)\{\bar y_R(\tau)-\xi_R\}^2],$$

which is easily maximised over (ξ_L,ξ_R,σ) for each fixed τ. We find

$$\widehat\sigma(\tau)^2 = n^{-1}\{Q_L(\tau)+Q_R(\tau)\},$$

and $\ell_{\rm prof}(\tau)=-n\log\widehat\sigma(\tau)$. The ML for τ is the $\widehat\tau$ making $\widehat\sigma(\tau)$ smallest. Also,

$$D(\tau,y) = n\log\frac{\widehat\sigma^2(\tau)}{\widehat\sigma^2(\widehat\tau)}.$$

3.3. **The multinormal case.** Assume the observations y_i are multivariate and normally distributed. Here we derive the required formulae for log-likelihood maxima and deviance functions, under two scenarios, corresponding to having the variance matrix constant or not, for the left and the right part of the data.

When $y_1,\ldots,y_n$ are i.i.d. $\mathrm{N}_p(\xi,\Sigma)$, the log-likelihood function is

$$\ell_n = -\tfrac12 n\log|\Sigma| - \tfrac12\sum_{i=1}^n(y_i-\xi)^{\rm t}\Sigma^{-1}(y_i-\xi) - \tfrac12 n\log(2\pi).$$

This is maximised by $\widehat\xi=\bar y$ and $\widehat\Sigma = n^{-1}\sum_{i=1}^n(y_i-\widehat\xi)(y_i-\widehat\xi)^{\rm t}$, see e.g. Mardia et al. (1979), with ensuing maximum $\ell_{n,\max}=-\frac12 n\log|\widehat\Sigma|-\frac12 np\{1+\log(2\pi)\}$. This leads

to a clear formula for the profile log-likelihood function for the model which takes $\mathrm{N}_p(\xi_L, \Sigma_L)$ to the left and $\mathrm{N}_p(\xi_R, \Sigma_R)$ to the right; indeed,

$$\ell_{\mathrm{prof}}(\tau) = -\tfrac{1}{2}\tau \log|\widehat{\Sigma}_L(\tau)| - \tfrac{1}{2}(n-\tau)\log|\widehat{\Sigma}_R(\tau)| \tag{3.4}$$

plus irrelevant constants. Here $\widehat{\Sigma}_L(\tau) = (1/\tau)\sum_{i\le\tau}(y_i - \bar{y}_L)(y_i - \bar{y}_L)^{\mathrm{t}}$, with $\bar{y}_L$ the average to the left, and similarly for $\widehat{\Sigma}_R(\tau)$.

Analogous calculations for the case of a common Σ across the range of data, but with different mean levels ξ_L and ξ_R, lead to

$$\ell_{\mathrm{prof}}(\tau) = -\tfrac{1}{2}n\log|\widehat{\Sigma}(\tau)|, \quad \text{with } \Sigma(\tau) = (\tau/n)\widehat{\Sigma}_L(\tau) + (1-\tau/n)\widehat{\Sigma}_R(\tau). \tag{3.5}$$

In particular, the ML estimator $\widehat{\tau}$ for this model is the value of τ minimising $\log|\widehat{\Sigma}(\tau)|$. This also yields the deviance formula

$$D(\tau, y) = n\{\log|\widehat{\Sigma}(\tau)| - \log|\widehat{\Sigma}(\widehat{\tau})|\},$$

with $\widehat{\tau}$ the ML estimator.

3.4. **Confidence for the degree of change.** In addition to spotting the real breakpoint τ_{true} itself there is often interest in the degree of change taking place, say via a suitable one-dimensional distance measure $\delta = \delta(\theta_L, \theta_R)$. The natural estimator is

$$\widehat{\delta} = \delta(\widehat{\theta}_L(\widehat{\tau}), \widehat{\theta}_R(\widehat{\tau})), \tag{3.6}$$

featuring the ML estimators of the left and right parameters, calculated at the ML position $\widehat{\tau}$. The distribution of $\delta(\widehat{\theta}_L(\tau), \widehat{\theta}_R(\tau))$, for a given τ, is typically close to a normal, but with a statistical bias of size $O(|\tau - \tau_{\mathrm{true}}|/\tau) + O(|\tau - \tau_{\mathrm{true}}|/(n-\tau))$, depending on how close τ is to the real value. The distribution of $\widehat{\delta}$ of (3.6) is a complex mixture of many such approximate normals, and with variable biases, depending also on how precise $\widehat{\tau}$ is for estimating τ_{true}.

In our investigations we have found it a sounder general approach to go for the profiled log-likelihood and deviance function, as for method B above, but now profiling for δ. The recipe is hence to compute

$$\ell_{\mathrm{prof}}(\delta) = \max\{\ell(\tau, \theta_L, \theta_R) \colon \delta(\theta_L, \theta_R) = \delta\},$$

then the deviance $D(\delta, y_{\mathrm{obs}}) = 2\{\ell_{\mathrm{prof}}(\widehat{\delta}) - \ell_{\mathrm{prof}}(\delta)\}$, followed by

$$\mathrm{cc}(\delta, y_{\mathrm{obs}}) = P_\delta\{D(\delta, Y) \le D(\delta, y_{\mathrm{obs}})\} = L_\delta(D(\delta, y_{\mathrm{obs}})). \tag{3.7}$$

There are two options here, for defining and then computing the probability distribution L_δ of $D(\delta, Y)$; these are related and often lead to very nearly the same results. The first is to fix τ at the ML position $\widehat{\tau}$ and compute L_δ under $(\widehat{\tau}, \widehat{\theta}_L(\delta), \widehat{\theta}_R(\delta))$, the position in the parameter space maximising $\ell(\tau, \theta_L, \theta_R)$ under the profiling constraint $\delta(\theta_L, \theta_R) = \delta$. The second is to follow the full profiling also for the τ part,

i.e. finding for given δ the point $(\widehat{\tau}(\delta), \widehat{\theta}_L(\delta), \widehat{\theta}_R(\delta))$ at which the log-likelihood is maximised, again under $\delta(\theta_L, \theta_R) = \delta$ but without fixing τ at the ML position. In both cases one computes $L_\delta(\cdot)$ and hence $\mathrm{cc}(\delta, y)$ of (3.7) by simulating datasets y_L^* and y_R^* to the left and the right from the estimated models and then computing the log-likelihood functions and hence $D(\delta, Y)$. This approach to reaching confidence inference for a degree of change parameter is illustrated for the ratio of Poisson rate parameters in Section 6 and for a ratio of standard deviances inside a broader model in Section 8.

4. Monitoring bridges and homogeneity tests

To apply the general method proposed in Section 2, particularly when encountering models outside the slim standard list where explicit tests might be available, one needs methods for testing distributional homogeneity of a sequence of observations, i.e. that $\theta_{a+1}, \ldots, \theta_{a+b}$ associated with observations $a+1, \ldots, a+b$ have remained unchanged. Here we describe some tests of this type. The monitoring bridges we construct based on log-likelihood maxima, along with associated goodness-of-fit tests, appear to be new and ought to have independent interest.

4.1. **Monitoring bridges.** Consider the sequence $y_1, \ldots, y_n$, with $y_i \sim f(y, \theta_i)$. Classes of such tests for constancy of the θ_i have been worked with in Hjort & Koning (2002). In particular, one may use the monitoring process

$$M_n(t) = n^{-1/2} \widehat{J}^{-1/2} \sum_{i \le nt} u(Y_i, \widehat{\theta}) \quad \text{for } 0 \le t \le 1, \tag{4.1}$$

in terms of the score function $u(y, \theta) = \partial \log f(y, \theta)/\partial\theta$ and the maximum likelihood estimator $\widehat{\theta}$ assuming $\theta_i = \theta$. Also, $\widehat{J}$ is the $p \times p$ estimated Fisher information matrix, with p the dimension of the model. We note that $M_n(\cdot)$ consists of p components, each starting and ending in zero; also, M_n is constant on each cell $[k/n, (k+1)/n)$, with $M_n(k/n) = n^{-1/2} \widehat{J}^{-1/2} \sum_{i \le k} u(Y_i, \widehat{\theta})$. Hjort and Koning prove that $M_n \to_d M$, the limit having p independent components $W_1^0, \ldots, W_p^0$, each a Brownian bridge. Hence plotting $M_{n,j}$ and checking various behavioural aspects, like their maxima or minima, leads to clear tests for homogeneity. These may be used in connection with the general construction of Section 2.

One particular version of this strategy is to test homogeneity using

$$Z_n = \max_{j \le p} \|M_{n,j}\| = \max_{j \le p} \max_{k \le n} |M_{n,j}(k/n)|.$$

Under homogeneity, $Z_n \to_d Z = \max_{j\le p}\max_{0\le t\le 1}|W_j^0(t)|$. The distribution for a single of these maxima of a Brownian bridge can be expressed as

$$H(z) = P\{\max_{0\le t\le 1}|W^0(t)| \le z\} = 1 + 2\sum_{k=1}^{\infty}(-1)^k \exp(-2k^2z^2), \tag{4.2}$$

as proved in Billingsley (1968, Ch. 3). For the case the maximum over several asymptotically independent components, as with the construction (4.1), we have $P\{Z_n \le z\} \to H(z)^p$, which is easily computed. Other variations can of course be used here, like the sum of maxima rather than the maximum of maxima, or the sum of Cramér–von Mises type statistics $n^{-1}\sum_{k=1}^n\{M_{n,1}(k/n)^2 + \cdots + M_{n,p}(k/n)^2\}$. The latter tends in distribution to $\sum_{j=1}^p \int_0^1 W_j^0(t)^2\,\mathrm{d}t$, which can be computed and tabled via simulations, or via results obtained in Csörgő & Faraway (1996).

4.2. **New monitoring bridges for model homogeneity.** Here we are however eager to build a new type of test, using the succession of attained log-likelihood maxima. Assume homogeneity, i.e. that there is a common θ_0 underlying the observations. With $\ell_j = \ell_j(\theta)$ the log-likelihood function based on $y_1,\ldots,y_j$, we compute the maximum likelihood estimate $\widehat\theta_j$ and the associated log-likelihood maximum $\widehat\ell_j = \ell_j(\widehat\theta_j)$. Computing the maximum likelihood estimator takes at least p observations. Our monitoring bridges take the form

$$\widehat B_{n,j} = n^{-1/2}\{\widehat\ell_j - (j/n)\widehat\ell_n\}/\widehat\kappa \quad \text{for } j = p,\ldots,n. \tag{4.3}$$

Here $\widehat\kappa$ is a consistent estimator of the standard deviation κ of $\log f(Y,\theta_0)$, e.g.

$$\widehat\kappa^2 = \frac{1}{n}\sum_{i=1}^n\{\log f(y_i,\widehat\theta) - \widehat\xi\}^2,$$

where $\widehat\xi = \widehat\ell_n/n$ the estimate of $\xi = \mathrm{E}_{\theta_0}\,\log f(y,\theta_0) = \int f_{\theta_0}\log f_{\theta_0}\,\mathrm{d}y$.

We show below that the process with these $\widehat B_{n,j}$ values tends to a Brownian bridge, under the null hypothesis of homogeneity. More precisely, consider the piecewise constant process $\widehat B_n$ on $[0,1]$ with values $\widehat B_{n,j}$ on $[j/n,(j+1)/n)$ for $j \ge p$, and zero for $[0,p/n)$. The claim is that under unchanging model conditions,

$$\widehat B_n \to_d W^0 \quad \text{in } D[0,1], \tag{4.4}$$

the limit being a Brownian bridge (a zero-mean Gaussian process with covariance function $s(1-t)$ for $s \le t$). The convergence in distribution in question takes place in the space of all functions $x\colon [0,1]\to\mathbb{R}$, right continuous with limits from the left, equipped with the Skorohod topology; cf. Billingsley (1968). Plotting the $\widehat B_{n,j}$, therefore, gives a monitoring bridge which should behave like a Brownian bridge under homogeneity conditions. The weak convergence result (4.4) implies $h(\widehat B_n) \to_d h(W^0)$ for all continuous functionals, so that $\max_{p\le j\le n}|\widehat B_{n,j}| \to_d \max_{0\le t\le 1}|W^0(t)|$,

$(n-p+1)^{-1}\sum_{j=p}^n \widehat{B}_{n,j}^2 \to_d \int_0^1 W^0(t)^2\,\mathrm{d}t$, etc. Among the benefits of the new goodness-of-fit construction (4.3) is that a multidimensional parametric family is mapped directly into a one-dimensional monitoring bridge.

To prove (4.4), start out considering the partial-sum process

$$A_n(t) = n^{-1/2}\sum_{i\le[nt]}\{\log f(y_i,\theta_0)-\xi\}/\kappa = n^{-1/2}(\ell_j - j\xi)/\kappa \quad \text{for } 0\le t\le 1,$$

writing $\ell_j = \sum_{i\le j}\log f(y_i,\theta_0)$ and $j=[nt]$ (so that j/n tends to t). From Donsker's theorem, cf. Billingsley (1968, Ch. 3), $A_n\to_d A$, the Brownian motion process. It then follows that the process B_n, defined by $B_n(t)=A_n(t)-tA_n(1)$, converges in distribution to the process B, defined by $B(t)=A(t)-tA(1)$, and this limit is demonstrably a Brownian bridge process on $[0,1]$. Also,

$$B_n(j/n) = n^{-1/2}\{\ell_j-(j/n)\ell_n\}/\kappa,$$

i.e. the tying-down has caused ξ to not being present.

We may now prove that $\widehat{A}_n$ and $\widehat{B}_n$ have the same limits as A_n and B_n, where

$$\widehat{A}_n(t) = n^{-1/2}(\widehat{\ell}_j - j\xi)/\widehat{\kappa} \quad\text{and}\quad \widehat{B}_n(t) = n^{-1/2}\{\widehat{\ell}_j-(j/n)\widehat{\ell}_n\}/\widehat{\kappa}$$

for $t\in[j/n,(j+1)/n)$. To show this, note from a Taylor expansion argument that $\ell_j(\theta_0)=\ell_j(\widehat{\theta}_j)+\frac12(\theta_0-\widehat{\theta}_j)^{\mathrm{t}}\ell_j''(\widehat{\theta}_j)(\theta_0-\widehat{\theta}_j)+o_{\mathrm{pr}}(1)$, which leads to

$$\widehat{\ell}_j = \ell_j(\theta_0)+\tfrac12 W_j + o_{\mathrm{pr}}(1), \tag{4.5}$$

where $W_j = j(\widehat{\theta}_j-\theta_0)^{\mathrm{t}}\widehat{J}_j(\widehat{\theta}_j-\theta_0)+o_{\mathrm{pr}}(1)$, with $\widehat{J}_j=-(1/j)\ell_j''(\widehat{\theta}_j)$ being the normalised observed Fisher information after j data points. The W_j tends to a χ^2_p as j increases. Hence the differences $\max|\widehat{A}_n-A_n|$ and $\max|\widehat{B}_n-B_n|$ are both $O_{\mathrm{pr}}(p/\sqrt{n})$, which goes to zero in probability. This proves claim (4.4).

Note that $\widehat{\ell}_j$ of $\widehat{A}_n$ overshoots $\ell_j(\theta_0)$ of A_n, essentially with the amount $\frac12 W_j$, a random variable with distribution tending to a half a χ^2_p, with mean value $\frac12 p$. This suggests using the sample-size modification $n^{-1/2}(\widehat{\ell}_j-\frac12 p-j\xi)/\widehat{\kappa}$ for $\widehat{A}_n$, which with a bit of algebra leads to the modification

$$B^*_{n,j} = n^{-1/2}\{\widehat{\ell}_j-(j/n)\widehat{\ell}_n-\tfrac12 p(1-j/n)\}/\widehat{\kappa}$$

for $\widehat{B}_{n,j}$ of (4.3). This version is closer in distribution to that of a Brownian bridge for finite n. We also point out that the key result (4.4) continues to hold also in situations with short-range dependence, as for most time series models. This is essentially since the partial-sum process A_n above still tends to the Brownian motion, under weak assumptions of this type; see Billingsley (1968, Ch. 4).

5. Performance

In previous sections we have developed a general machinery for confidence inference for change-points. It is clear from these developments that there are several available methods, for a given dataset and a given vehicle model. In particular, for general method A there is a choice to be made for the homogeneity test. In the present section we consider performance issues for the resulting confidence distributions, also comparing method A with method B. The primary performance aspect is that the confidence distributions really come close to delivering adequate coverage, which in our change-point context means that the confidence curve construction $\mathrm{cc}(\tau, y)$ should have $G(\alpha) = P_\tau\{\mathrm{cc}(\tau, Y) \le \alpha\}$ close to α. For method B, the distribution of $U_\tau = \mathrm{cc}(\tau, Y)$ is never perfectly uniform, since it is discrete, though $G(\alpha)$ is often seen to be close to α with our constructions. For method A, however, the uniformity at the true change-point value τ is exact (as long as the homogeneity tests on each side are exact), as demonstrated in Section 2, and will be retained even if the change-point is very clear and only one τ value, the true one, "survives" at all levels. In that case the minimum of $\mathrm{cc}_A(\tau)$ will be uniformly distributed. This has consequences for the interpretation of the confidence curve defined by method A: while the τ minimising $\mathrm{cc}_A(\tau, y)$ may be considered an estimate of the change-point, the actual minimal value of $\mathrm{cc}_A(\tau, y)$ is of limited interest, and should not be interpreted as a measure of certainty of the change-point estimate.

A second performance aspect, which is a measure of certainty of the change-point estimate, is that a $\mathrm{cc}(\tau, y)$ should lead to 'thin' or narrow confidence sets $\{\tau\colon \mathrm{cc}(\tau, y) \le \alpha\}$, for most or all values of the confidence level α. We measure such thinness or slimness here by the number of τ belonging to the confidence set where $\mathrm{cc}(\tau, y) \le \alpha$, for a range of α levels (rather than the width or range of the set, as the sets may be non-connected); for simplicity we use the term 'size' below to indicate such numbers.

Schweder & Hjort (2016, Ch. 5) offer a broad discussion of performance and risk functions for confidence distributions, also identifying classes of situations where there is a unique optimal confidence procedure; see also the discussion on performance in Xie & Singh (2013). Such clear results seem out of reach when it comes to confidence for change-points, however. Below we report briefly on investigations into the mentioned performance aspects for our confidence methods.

5.1. **Method A with different tests.** Method A is a general method for constructing confidence sets for a change-point, but depends on having a well-working test of homogeneity for the segments $1, \ldots, \tau$ and $\tau + 1, \ldots, n$. It may also depend upon the choice of the test levels at work in (2.1); here we compare having the

fixed level, i.e. $\sqrt{\alpha}\sqrt{\alpha}$, with the alternative where it depends on the sizes of the segments, via $\alpha^{\tau/n}\alpha^{1-\tau/n}$. Considering the simple model with $y_i \sim \mathrm{N}(\xi_L, 1)$ to the left of τ and $y_i \sim \mathrm{N}(\xi_R, 1)$ to the right, and with ξ_L and ξ_R unknown, we have investigated three different tests and the two different versions of test levels via simulations. The first test is that used in connection with the (2.2) illustration, using $Q_{a+1,a+b} = \sum_{i=a+1}^{a+b}(Y_i - \bar{Y})^2$ with a χ^2_{b-1} null distribution. The second test uses the regression slope coefficient from a regression model; on the segment $1, \ldots, \tau$ we consider $\widehat{b} = \sum_{i=1}^{\tau}(i - \bar{i})y_i/M(\tau)$, where we know that $M(\tau)\widehat{b}^2 \sim \chi^2_1$, under the homogeneity hypothesis; here $M(\tau) = \sum_{i\le\tau}(i-\bar{i})^2$ and $\bar{i}$ is the average of $1, \ldots, \tau$. The test for the segment $\tau+1, \ldots, n$ is similar. The third test uses monitoring bridges from Hjort & Koning (2002), as presented in Section 4.1. For this model the monitoring processes become

$$M_L(t) = \frac{1}{\tau^{1/2}} \sum_{i\le[\tau t]} (Y_i - \widehat{\xi}_L) \text{ and } M_R(t) = \frac{1}{(n-\tau)^{1/2}} \sum_{\tau+1\le i\le\tau+1+(n-\tau-1)t} (Y_i - \widehat{\xi}_R)$$

to the left and to the right of τ, respectively. From these processes we use $V_L = \max_t |M_L(t)|$ and $V_R = \max_t |M_R(t)|$ as test statistics, with the theory from Hjort & Koning (2002) implying that these are asymptotically distributed as maxima of Brownian bridges; cf. (4.2).

method	50% coverage		50% size		90% coverage		90% size		95% coverage		95% size	
A-I	0.49	0.50	34.01	17.19	0.88	0.88	112.72	66.28	0.94	0.95	135.86	86.13
A-II	0.52	0.51	10.30	8.91	0.90	0.90	33.64	23.62	0.94	0.95	44.96	28.72
A-III	0.60	0.58	10.53	9.28	0.93	0.93	28.88	23.43	0.97	0.96	37.57	27.81
B	0.50	0.51	2.47	2.11	0.90	0.90	10.59	7.97	0.95	0.95	14.60	10.51

TABLE 5.1. Coverage and mean size of confidence sets produced with method A with three different tests (and test level depending on τ) and with method B, applied to the normal model with known variance, with $n = 200$. Test A-I is the simple test, A-II is the regression test and A-III is the Hjort–Koning test. The leftmost numbers in each column are results from datasets with $\tau = 25$, the rightmost numbers are results from datasets with $\tau = 100$. Each number is based on 10^3 simulated datasets.

The simulations were carried out by generating datasets of size $n = 200$. We examined different combinations of position of τ, confidence levels, and difference between the left and right levels. Here we briefly report on the cases where τ positions were set to $25, 50, 75, 100$ (cases $175, 150, 125$ are fully symmetric with $25, 50, 75$), and with $\xi_L = 2.2$ and $\xi_R = 3.3$, indicating a difference not easy to tell immediately from the data. In order to evaluate the six different combinations

of tests and test levels, the coverage and size (number of τ values, rather than the range from smallest to largest value) of the confidence sets of level 0.50, 0.90 and 0.95 were recorded. One method is considered superior (more powerful) than another if it produces slimmer confidence sets while keeping the correct coverage. Method A with tests 1 and 2 has the correct coverage probability, per construction, and this is reflected in the simulations (see Table 5.1). The third test (Hjort–Koning) is based on an asymptotic result and therefore does not have exactly the right coverage, however. The simulations reveal that the deviation is generally small, for example a 95% confidence set typically covers the true τ value 97% of the time. Tests 2 and 3 produce confidence sets of very similar size, but the first test is clearly less powerful than the two others. For example, while test 2 and 3 produce confidence sets with a mean size of 29 and 28 at the 95% level for $\tau = 100$, the first test has confidence sets of mean size 86. When it comes to the choice of test level, it is beneficial to let the level depend on τ, in the manner of $\alpha^{\tau/n}\alpha^{1-\tau/n}$, rather than using $\sqrt{\alpha}\sqrt{\alpha}$ in (2.1), but the differences between these two alternatives tends to be small; in Table 5.1 we therefore include only the first choice. For all methods the resulting confidence sets are smaller for τ values near the middle of the data (close to 100). The opposite effect is most obvious for the datasets with $\tau = 25$, where the confidence sets typically are close to 1.5 times larger than the confidence sets from data with $\tau = 100$. The results for datasets with τ equal to 50 and 75 are not shown here, but have been seen to be fairly close to the results for $\tau = 100$. For the good performance of method B see Section 5.2

We also investigated the behaviour of the different versions of method A when datasets without change-points were generated. In these cases, the method produces extremely wide confidence sets, generally spanning nearly the entire set of possible τ values, thus indicating, as they should, that the data are homogeneous on both sides of nearly all possible choices of τ.

Further examined were the two different tests for method A for the model where the variance is unknown (and potentially different on the two segments); $y_i \sim \mathrm{N}(\xi_L, \sigma_L^2)$ to the left of τ and $y_i \sim \mathrm{N}(\xi_R, \sigma_R^2)$ to the right. The first test is an extension of the regression based test above. Writing down the required formulae for the full segment $1, \ldots, n$ (and then applying these for the left and right segments later on), we have $\widehat{b} = \sum(i - \bar{i})y_i/M$ with $M = \sum_{i=1}^n (i - \bar{i})^2$, and employ $t = M^{1/2}\widehat{b}/\widehat{\sigma}$, where $\widehat{\sigma}^2 = \sum_{i=1}^n \{y_i - \bar{y} - \widehat{b}(i - \bar{i})\}^2/(n-2)$. Here $\bar{i}$ is the average of indexes employed. Under homogeneity, $t \sim t_{n-2}$. The second test is an application of the monitoring bridges from Hjort & Koning (2002). This time we have two unknown parameters and thus the monitoring process is two-dimensional. Following

the recipe for these monitoring processes, we have to the left of τ

$$M_L(t) = \tau^{-1/2} \sum_{i \le [\tau t]} \begin{pmatrix} Z_i \\ (Z_i^2 - 1)/\sqrt{2} \end{pmatrix} \qquad \text{for } 0 \le t \le 1,$$

with $Z_i = (Y_i - \widehat{\xi}_L)/\widehat{\sigma}_L$, and as the test statistic we use

$$V_L = \max\{\max_{t \le 1} |M_{L,1}(t)|, \max_{t \le 1} |M_{L,2}(t)|\},$$

the maximum of the absolute maxima of the two bridges. The asymptotic distribution of V_L (under homogeneity) can be easily computed via $H(z)^2$ with $H(z)$ from (4.2). We construct a similar test statistic for the segment to the right of τ.

Again we generated datasets of size $n = 200$ with four different τ values (25, 50, 75, 100), and again we recorded the coverage and size (number of τ values) of the confidence sets of level 0.50, 0.90 and 0.95. We studied the two tests for two different settings, one where the change-point is a change in the mean, with $\xi_L = 2.2$, $\xi_R = 3.3$ and $\sigma_L = \sigma_R = 1$, and the other where it is a change in the variance level, with $\xi_L = \xi_R = 2.2$, $\sigma_L = 1$ and $\sigma_R = 2$.

method	50% coverage		50% size		90% coverage		90% size		95% coverage		95% size	
A-I	0.49	0.52	6.32	6.14	0.92	0.90	28.44	20.53	0.96	0.96	58.03	34.03
A-II	0.62	0.60	14.88	11.92	0.94	0.94	43.03	28.65	0.97	0.98	40.97	26.40
B	0.49	0.51	2.73	2.19	0.86	0.89	12.72	8.57	0.92	0.95	18.00	11.36
A-I	0.50	0.52	99.27	99.08	0.92	0.90	177.26	176.86	0.96	0.95	186.76	185.93
A-II	0.63	0.64	35.24	16.59	0.94	0.94	130.31	37.57	0.97	0.98	155.42	43.82
B	0.46	0.49	4.79	3.24	0.85	0.90	21.59	12.28	0.89	0.96	30.95	16.28

TABLE 5.2. Coverage and mean size of confidence sets produced with method A with two different tests (and test level depending on τ) and with method B, applied to the normal model with unknown variance. The first three rows concern the case where the change-point is a change in the mean, and the second three concern the case where the change-point is a change in the variance. Test A-I is the regression test and test A-II is the Hjort–Koning test. The leftmost numbers in each column are results from datasets with $\tau = 25$, the rightmost numbers are results from datasets with $\tau = 100$. Each number is based on 10^3 simulated datasets.

For datasets with a change in the mean, the regression-based test was slightly more advantageous than the Hjort–Koning test, having the correct coverage and narrower confidence sets (see Table 5.2). However, the regression-based test is only constructed to discover changes in the mean levels and the Hjort–Koning test is therefore a more flexible test, able to discover change-points also when the change only affects the variance (see Table 5.2, lower part). For both settings, the resulting

confidence sets were much larger for datasets with $\tau = 25$. This was most apparent for the Hjort–Koning test in the second setting (change in the variance), where the size of the confidence sets increased from around 44 on the 95% level (for τ equal to 50, 75 or 100) to 155 for $\tau = 25$.

5.2. **Method A versus Method B.** The two methods proposed in this article have similar aims, but different points of departure and different performances. While method B assumes a model where there is a change-point (exactly one) on the whole data segment, method A considers possible τ values as points where the data on each side of τ are deemed homogeneous. The performance of method A is thus mostly dependent on the power of the chosen test in discovering lack of homogeneity. We included method B in the three simulation studies described above, and they reveal that method B produces clearly smaller confidence sets compared to the different versions of method A we have included. However, method B has a tendency to produce confidence sets with slightly lower coverage than the specified level (and the confidence sets should therefore be larger). The coverage problem is more apparent when the change-point is far from the centre of the data, especially for the more complex model with unknown variance (see Table 5.2). Method B seems nonetheless to outperform A in these simulations. We still consider method A to be fruitful, with a higher degree of flexibility considering the choice of test and more applicable to complicated high-dimensional or even nonparametric situations.

5.3. **The Bayesian approach.** Bayesian solutions to the change-point problem are not hard to put up, but they require of course a prior to be set up for $(\tau, \theta_L, \theta_R)$, sometimes with ad hoc constructions. This leads to a posterior distribution for τ. Suppose in particular that τ is given the prior $\pi_0(\tau)$, independently of priors π_L and π_R for θ_L and θ_R. This leads to the posterior distribution

$$\pi(\tau \,|\, \text{data}) \propto \pi_0(\tau)\lambda_L(\tau)\lambda_R(\tau),$$

expressed via the marginal left and right likelihoods

$$\lambda_L(\tau) = \int L_L(\theta_L)\pi_L(\theta_L)\,\mathrm{d}\theta_L \quad \text{and} \quad \lambda_R(\tau) = \int L_R(\theta_R)\pi_R(\theta_R)\,\mathrm{d}\theta_R.$$

These can be computed explicitly in a few models, and lead to a clear Bayesian posterior for τ. Via numerical integration methods or MCMC one may also compute such a $\pi(\tau \,|\, \text{data})$ in a range of other situations, even without clear formulae for the marginal likelihoods; see Carlin et al. (1992) and Fearnhead (2006).

Useful approximations emerge via the following Taylor expansion arguments, which we first put up for the case of n observations from the same model with a

prior $\pi(\cdot)$ for the same θ parameter, of dimension p:

$$\begin{aligned}\lambda &= \int \exp\{\ell_n(\theta)\}\pi(\theta)\,\mathrm{d}\theta \\ &\doteq \int \exp\{\ell_{\max} - \tfrac{1}{2}(\theta-\widehat{\theta})^{\mathrm{t}} n\widehat{J}(\theta-\widehat{\theta})\}\pi(\theta)\,\mathrm{d}\theta \\ &\doteq \exp(\ell_{\max})|n\widehat{J}|^{-1/2}\pi(\widehat{\theta})(2\pi)^{p/2}.\end{aligned}$$

Here $\widehat{\theta}$ is the ML estimator and $\widehat{J} = -n^{-1}\partial^2\ell(\widehat{\theta})/\partial\theta\partial\theta^{\mathrm{t}}$ the normalised Hessian matrix, converging with increasing n to a certain matrix. This leads to $\log\lambda = \ell_{\max} - \frac{1}{2}p\log n + O_{\mathrm{pr}}(p)$, akin to the approximation leading to the Bayesian information criterion BIC (see Claeskens & Hjort (2008, Ch. 4)).

Now going back to the change-point analysis, and keeping the leading terms only, we are led to the approximation

$$\begin{aligned}\pi(\tau\,|\,\mathrm{data}) &\propto \pi_0(\tau)\exp\big[\ell_{\mathrm{prof}}(\tau) - \tfrac{1}{2}p\log\{\tau(n-\tau)\}\big] \\ &= \pi_0(\tau)\exp\{\ell_{\mathrm{prof}}(\tau)\}\{\tau(n-\tau)\}^{-p/2}.\end{aligned} \tag{5.1}$$

This assumes that the left and right priors for θ_L and θ_R are not overly different. At any rate, the sizes of the leading terms of $\log\pi(\tau\,|\,\mathrm{data})$ are $O_{\mathrm{pr}}(n)$ and $O(p\log\tau + p\log(n-\tau))$, with remainder terms of size $O_{\mathrm{pr}}(p)$. The approximation is useful for the computational side of things, as it bypasses the need for high-dimensional integration or for MCMC setups, but also for shedding light on the behaviour of the posterior distribution and for how it differs from the frequentist approaches we are developing and advocating in the present paper. We learn e.g. that the Bayesian posterior has a tendency to push τ towards the extreme ends. The (5.1) formula is incidentally exactly correct for the case of the model $\mathrm{N}_p(\xi_L, I_p)$ to the left and $\mathrm{N}_p(\xi_R, I_p)$ to the right, and with flat priors for ξ_L and ξ_R.

For quantities associated with smooth parametric models one is used to the phenomenon described via so-called Bernshteĭn–von Mises theorems, that Bayesian and frequentist inference tend to agree well, and with the prior in question being reasonably quickly washed out by the data provided the parameter dimension being low; see e.g. the discussion in Hjort et al. (2010, Introduction). This is different here, however, in view of (5.1) and its consequent

$$\log\pi(\tau\,|\,\mathrm{data}) = \log\pi_0(\tau) + \ell_{\mathrm{prof}}(\tau) - \tfrac{1}{2}p\{\log\tau + \log(n-\tau)\} + O_{\mathrm{pr}}(1),$$

which shows both that there is a certain bias inherent in the Bayes construction and that this bias is more slowly disappearing with increasing sample size than in regular parametric models. Simple simulation exercises reveal that the distribution of $U_\tau = \sum_{\tau_B<\tau}\pi(\tau_B\,|\,Y) + \frac{1}{2}\pi(\tau\,|\,Y)$, the half-correction version of the cumulative posterior distribution for the Bayesian parameter τ_B, computed under the true τ, is often far from the uniform, even for moderately large n. From such investigations,

along with those reported on earlier in this section, it is apparent that confidence distribution Method B based on the deviance and its distribution does a much better job than the Bayesian apparatus when it comes to delivering confidence intervals with correct coverage. The reason the Bayes method does badly in this regard is partly that there is an inherited bias of size $O(p \log n + p \log(n-\tau))$ in the log-posterior, but even more so that the distribution

$$\pi(\tau \,|\, y) \propto \exp\{\ell_{\rm prof}(\tau)\}$$

also often delivers inaccurate confidence, even for moderately large n in simple models. This is in contrast to how things pan out for parameters of smooth regular models, where such a recipe typically leads to accurate coverage with increasing sample size, as per the Bernshteĭn–von Mises theorems (here in the form of the Laplace type inverse probability, Bayes with a flat prior). In this connection see also Fraser (2011), who argues that Bayes is sometimes only 'quick and dirty confidence', and Efron (2015), who is concerned with frequentist accuracy of Bayes solutions in a general perspective. We also mention that the ML estimator $\widehat{\tau}$ typically does better than the Bayes estimator $\widehat{\tau}_B$ maximising the posterior distribution (i.e. the Bayes solution under a 0-1 loss function), as judged by e.g. mean absolute deviation, as seen via simulation experiments.

6. Application 1: British mining accidents

As a first, simple illustration, we apply method B of Section 3 to a dataset from the change-point literature, the number of British coal-mining disasters from 1851 to 1962; see Jarrett (1979) for relevant background and for certain corrections that were made to earlier accounts. With y_i the number of mining disasters in year i, we take these to be independent and Poisson distributed with parameter θ_L for $i \leq \tau$ and θ_R for $i \geq \tau+1$. This is the model used for these data by Carlin et al. (1992), for a Bayesian analysis, where clear posterior distributions are found for the parameters based on their given prior for $(\tau, \theta_L, \theta_R)$. They also provided a posterior density for the relative change parameter $\rho = \theta_L/\theta_R$. In this case, our methods give very similar results to the above-mentioned Bayesian analysis; with our confidence distributions matching their posteriors, but without priors.

In order to compute the confidence curve for the breakpoint, we need the deviance function and the profile log-likelihood function, here taking the form

$$\ell_{\rm prof}(\tau) = \tau \bar{y}_L(\tau)\{\log \bar{y}_L(\tau) - 1\} + (n-\tau)\bar{y}_R(\tau)\{\log \bar{y}_R(\tau) - 1\}.$$

The ML estimates are $\widehat{\tau} = 41$ (corresponding to year 1891), $\widehat{\theta}_L = 3.098$ and $\widehat{\theta}_R = 0.901$. From the estimates of θ_L and θ_R we simulated datasets under each possible change-point value (that is, for all years between 1851 and 1961), calculated the

deviance functions and computed the confidence curve from recipe (3.3); this yields the left panel of Figure 6.1. The curve agrees with the posterior distribution for the change-point given in Carlin et al. (1992), where the posterior mode also agrees with the ML estimate.

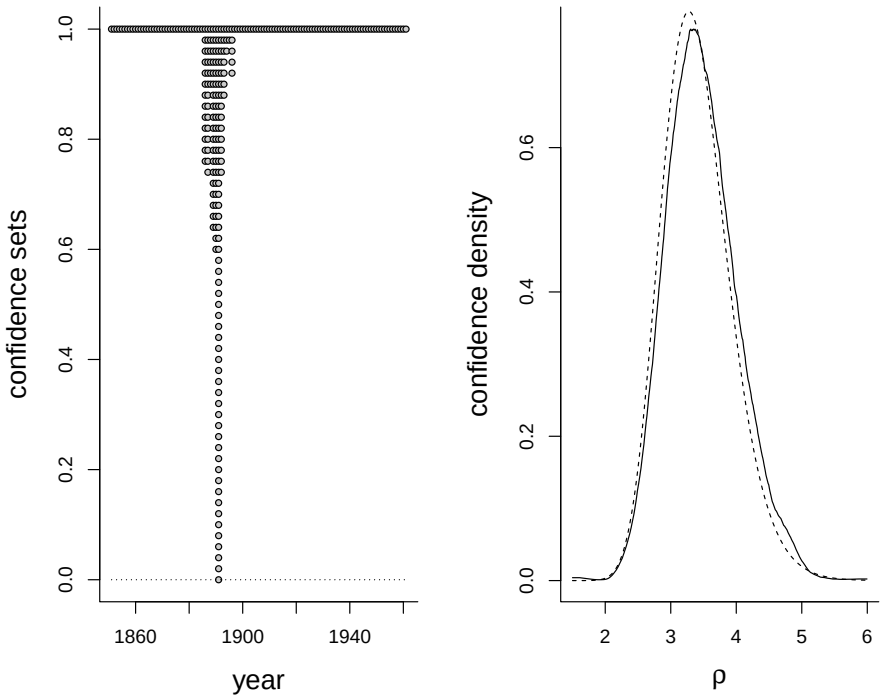

FIGURE 6.1. Left panel: Confidence curve for the change-point τ, using the deviance based method B. Right panel: Confidence density for the degree of change $\rho = \theta_L/\theta_R$, via method B (full line), and the Bayesian method (dashed line).

In order to analyse the degree of change ρ (the ratio between the rates of disasters, for the past state of affairs and for the present), we reparametrise the model as $y_i \sim \text{Pois}(\rho\theta)$ for $i \le \tau$ and $y_i \sim \text{Pois}(\theta)$ for $i \ge \tau + 1$. The log-likelihood function is

$$\ell(\tau, \rho\theta, \theta) = \tau\{-\rho\theta + \bar{y}_L \log(\rho\theta)\} + (n-\tau)(-\theta + \bar{y}_R \log\theta),$$

which we then maximise over θ and τ to reach the profile log-likelihood function

$$\begin{aligned}\ell_{\text{prof}}(\rho) = \widehat{\tau}(\rho)\big[-\rho\widehat{\theta}(\rho) + \bar{y}_L(\widehat{\tau}(\rho))\log\{\rho\widehat{\theta}(\rho)\}\big] \\ +\{n-\widehat{\tau}(\rho)\}\{-\widehat{\theta}(\rho) + \bar{y}_R(\widehat{\tau}(\rho))\log\widehat{\theta}(\rho)\},\end{aligned}$$

with $\widehat{\theta}(\rho,\tau) = \{\tau\bar{y}_L + (n-\tau)\bar{y}_R\}/\{\tau\rho + n - \tau\} = n\bar{y}/\{\tau\rho + n - \tau\}$ and $\widehat{\tau}$ obtained by maximising over all possible τ values. The ML estimate for the degree of change was 3.437, and the confidence curve was obtained by (3.3) by simulating datasets from

a grid of ρ values, using the overall ML estimate for τ along with $\widehat{\theta}_L(\rho)$ and $\widehat{\theta}_R(\rho)$, following the recipe of Section 3.4. The confidence curve $\mathrm{cc}(\rho, y)$ can be converted to a cumulative confidence distribution $C(\rho, y)$, via $\mathrm{cc}(\rho, y) = |1 - 2\,C(\rho, y)|$, which then via numerical derivation yields a confidence density, say $c(\rho, y)$, displayed in the right panel of Figure 6.1. This may now be compared to the posterior density for ρ arrived at with any reasonable start prior for $(\tau, \theta_L, \theta_R)$, e.g. from MCMC methods presented in Carlin et al. (1992). Our prior-free method gives results very similar to those of the Bayesian machinery, with the almost noninformative priors used by Carlin et al. (1992). The right panel of Figure 6.1 displays two very similar curves; the confidence density and the Bayesian posterior calculated using a flat prior for τ and independent almost noninformative Gamma priors with parameters $(\frac{1}{2}, \frac{1}{2})$ for the two levels.

The simulations required for constructing the confidence curve with method B can be time-consuming, but here we may resort to an approximate solution based on the Wilks theorem. If we fix τ at the ML value $\widehat{\tau}$, and proceed with deviance calculus profiling over (θ_L, θ_R) subject to $\theta_L/\theta_R = \rho$, then the $D(\rho, Y)$ is very closely approximated with a χ^2_1, leading to a confidence curve for ρ via $\mathrm{cc}(\rho, y_{\mathrm{obs}}) = \Gamma_1(D(\rho, y_{\mathrm{obs}}))$, where Γ_1 is the cumulative distribution function of a χ^2_1 distribution. The resulting confidence curve is indistinguishable from the one computed with simulations and displayed in the right panel of Figure 6.1, demonstrating that τ is sufficiently well estimated in this case.

7. Application 2: Tirant lo Blanch

Our next change-point challenge concerns the Catalan novel *Tirant lo Blanch*. This chivalry romance, written in the 1460s, can be considered the world's first novel, and was incidentally much admired by Cervantes (who wrote the more famous *Don Quixote* about 150 years later). Most scholars agree that the novel had two authors; the first author Joanot Martorell died before the completion of the novel, and Marti Joan de Galba claimed to have finished it. Hence there is a change-point problem, where we should hunt for the chapter number where the change from the first to the second author takes place. Earlier statistical analyses include Girón et al. (2005), Riba & Ginebra (2005), Koziol (2014) and Chen & Zhang (2015). Most researchers favouring the change-of-author hypothesis believe that the change takes place towards the end of the 487 chapter long book, more accurately between chapter 350 and 400 (Chen & Zhang, 2015).

Different aspects of the chapters and the writing may be considered for statistical measurements and then collected from the text. Analysing a quarrel between Nobel Prize winners, Hjort (2007) used statistical modelling of sentence lengths to

discriminate between two literary corpora, for example, and in Section 10.2 we are indeed using such information to assist us in pinpointing the author change-point. Presently we are concentrating on the word lengths in each chapter, and we have only considered the 425 chapters with more than 200 words. From these we collect vectors y_i of dimension 10, displaying the number of words of length 1, 2, 3 and so on, up to the number of words equal to or longer than 10 letters. The aforementioned authors have used the same dataset, and all, except for Chen & Zhang (2015), model the 425 word count vectors as multinomially distributed. Chen & Zhang (2015) propose a graph based, nonparametric change-point method. Girón et al. (2005) adopt a Bayesian framework and provide a posterior distribution for the change-point τ. Similar models as in Girón et al. (2005) are assumed in Riba & Ginebra (2005), but in a frequentist framework and without providing any uncertainty around the change-point estimates. Koziol (2014) approaches the change-point problem with Lancaster partitions of chi-squared tests of homogeneity.

Initial goodness-of-fit checks demonstrate that the word lengths in the different chapters of the book have heterogeneous distributions; in particular, the pure multinomial model favoured by several previous scholars, with fixed probabilities of word lengths from chapter to chapter inside a segment, does not fit well/ allows for too little variability between chapters. We therefore investigated three other models: an overdispersed multinomial, that is the Dirichlet-multinomial distribution, and two different multinormal models. The first one allows the change-point to affect both the mean vectors and the covariance matrices, while the second assumes that the authors differ in the mean vector only. To judge between candidate models we have computed values of the Akaike information criterion, cf. Claeskens & Hjort (2008, Ch. 3), defined as $\text{AIC} = 2\,\ell_{\max} - 2\,\text{dim}$, with dim the number of parameters estimated in the model and $\ell_{\max}$ the associated maximum of the log-likelihood function (see Table 7.1). These AIC values give a clear indication that the multinormal model has a better fit to the data, and we thus used the multinormal for the construction of the confidence curve for the change-point.

model	$\ell_{\max}$	dim	AIC
multinomial	$-14,449$	19	$-28,936$
Dirichlet-multinomial	$-13,870$	21	$-27,780$
multinormal 1	$-13,722$	109	$-27,660$
multinormal 2	$-13,771$	64	$-27,671$

TABLE 7.1. Number of parameters and AIC values for different models: multinormal 1 is the model assuming that the two authors both have different mean vectors and different covariance matrices, while multinormal 2 assumes that the authors differ only in the mean vector.

The multinormal model assumes that the observed proportions $z_i = y_i/m_i$ in each chapter follow a multinormal distribution, with precision related to the sample size. Disregarding element no. 10, since the proportions sum to one for each chapter, the model used is

$$z_i \sim \mathrm{N}_9(\xi_L, \Sigma_L/m_i) \text{ for } i \le \tau \quad \text{and} \quad z_i \sim \mathrm{N}_9(\xi_R, \Sigma_R/m_i) \text{ for } i \ge \tau + 1.$$

Here ξ_L and ξ_R are the mean vectors of these distributions of proportions, and Σ_L and Σ_R appropriate 9×9 covariance matrices. The confidence curve was obtained by method B (Section 3). First we find the profile log-likelihood function, which in generalisation of the result (3.5) to the present case with variance matrices Σ/m_i becomes

$$\ell_{\mathrm{prof}}(\tau) = -\tfrac{1}{2}\tau \log|\widehat{\Sigma}_L(\tau)| - \tfrac{1}{2}(n-\tau)\log|\widehat{\Sigma}_R(\tau)|,$$

now with

$$\widehat{\Sigma}_L(\tau) = \frac{1}{\tau}\sum_{i\le\tau} m_i\{z_i - \widehat{\xi}_L(\tau)\}\{z_i - \widehat{\xi}_L(\tau)\}^{\mathrm{t}},$$

$$\widehat{\Sigma}_R(\tau) = \frac{1}{n-\tau}\sum_{i\ge\tau+1} m_i\{z_i - \widehat{\xi}_R(\tau)\}\{z_i - \widehat{\xi}_R(\tau)\}^{\mathrm{t}},$$

where $\widehat{\xi}_L(\tau) = \sum_{i\le\tau} m_i z_i / \sum_{i\le\tau} m_i$ and similarly for $\widehat{\xi}_R(\tau)$. There is a consequent formula for the deviance function for τ. The ML estimate for the change-point was found to be $\widehat{\tau} = 320$. In the original numbering of the chapters this corresponds to chapter 371, which is the same point estimate as with the ordinary multinomial model in Riba & Ginebra (2005), and also the mode of the change-point posterior distribution in Girón et al. (2005). The multinormal model gave the following ML estimates for the mean of the world length proportions (the covariances matrices are not given):

	1	2	3	4	5	6	7	8	9	10
left	0.106	0.222	0.209	0.103	0.105	0.104	0.053	0.045	0.029	0.024
right	0.114	0.209	0.190	0.098	0.103	0.105	0.058	0.050	0.038	0.035

TABLE 7.2. Estimated proportions of words of different lengths, before and after estimated change-point $\widehat{\tau} = 320$, via the multinormal model.

By simulating the distribution of $D(\tau, Y)$ we obtain the confidence curve for τ, shown in Figure 7.1. Interestingly, the curve indicates some confidence in $\tau = 295$, which corresponds to chapter 345, which is in accordance with the Bayesian posterior distribution in Girón et al. (2005) and with some of the analyses based on summary measures presented on the next page.

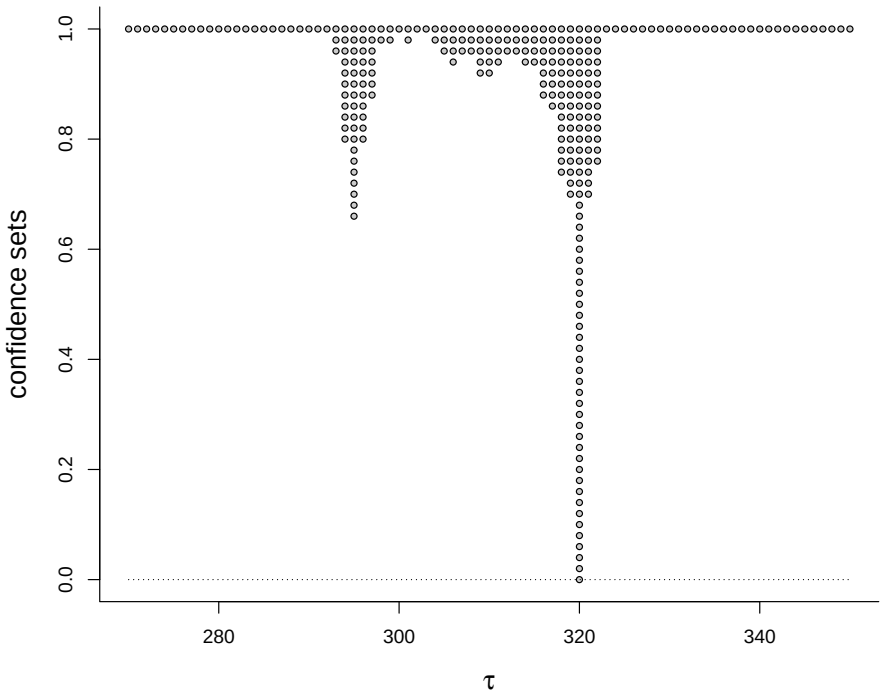

FIGURE 7.1. Confidence curve for the change-point τ, using method B based on the multinormal model.

In addition to modelling the whole vector of proportions we can look at different summary measures for each chapter, for example the average word length per chapter. This was also used in Riba & Ginebra (2005), where they assumed a normal model for the average word length per chapter and for the value taken by the first principal component from correspondence analysis, yielding point estimates corresponding to chapters 345 and 371, respectively. We also consider the average word length, and in addition the standard deviation of word lengths, the proportions of words of length 3 or less, and the proportions of words of length at least 8 letters, in each chapter. The two last summary statistics are motivated by the fact that the change in author seems to be mostly reflected in the proportions of short and long words, cf. Table 7.2. Each of these summary measures, say w_i, can be modelled as

$$w_i \sim \mathrm{N}(\theta_L, \sigma_L^2/m_i) \text{ for } i \le \tau, \quad \text{and} \quad w_i \sim \mathrm{N}(\theta_R, \sigma_R^2/m_i) \text{ for } i \ge \tau + 1,$$

and confidence curves can easily be constructed by method B; see Figure 7.2.

Three of the summary measures give the most confidence to $\widehat{\tau} = 295$, corresponding to chapter 345, but all of these curves also place some confidence on the change-point taking place in chapter 371. When looking at the proportions of short words (of length 1 or 2 or 3 letters) in each chapter, the most confidence is however

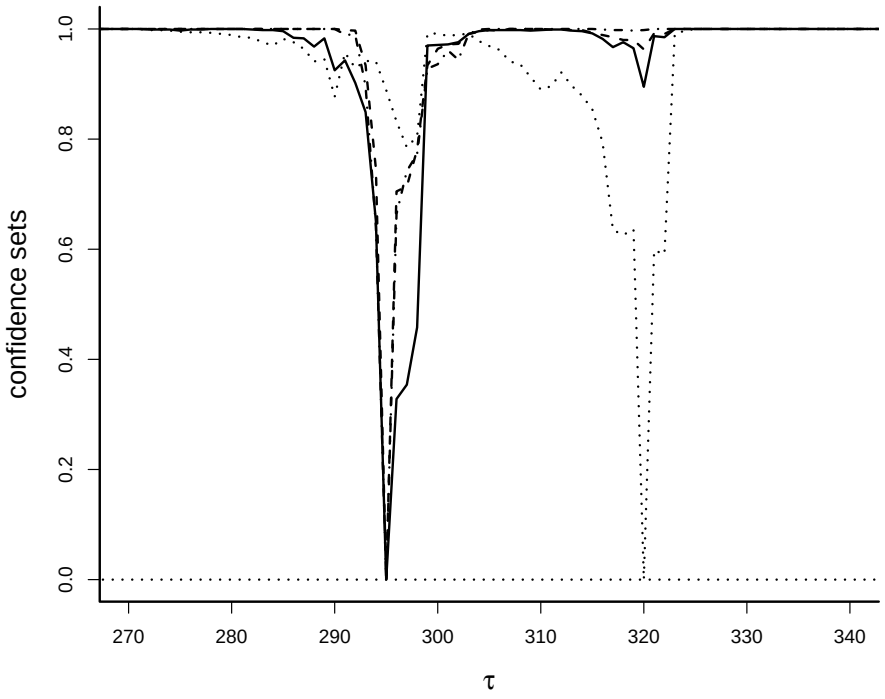

FIGURE 7.2. Confidence curves for the change-point τ, using method B: Full line, based on the average word length per chapter; dashed line, based on the standard deviation in word lengths; dotted line, based on the proportions of words of length 3 letters or less; and dot-dashed line, based on the proportions of words of length 8 or more.

placed on $\widehat{\tau} = 320$, corresponding to chapter 371, consistent with the multinormal analysis above. All our analyses for *Tirant lo Blanch* indicate that the change of authors takes place towards the end of the book, with the most confidence placed on the chapters 345 and 371. These results are consistent with previous statistical analyses of the work (Riba & Ginebra, 2005; Girón et al., 2005; Koziol, 2014; Chen & Zhang, 2015), with aspects of literary analyses (Rosenthal, 1984, preface), and with the assertion made by the second author himself, who, in the afterword of the book, writes that he completed the final quarter of the book.

8. Application 3: skiings days at Bjørnholt

The number of skiing days in a winter season is defined as the number of days with at least 25 cm snow.[1] In Figure 8.1 the number of such days at the particular location of Bjørnholt in Oslo's skiing and recreation area Nordmarka are plotted

[1]The definition and term 'skiing day' was introduced by the Norwegian meteorologist Gustav Bjørbæk as the least amount of snow needed to avoid injury in case of a fall.

for the winter seasons 1896-97 to 2014-15. The expected number of skiing days and the future prospects for snowy winters are of especially great interest to the skiing enthusiasts. Moreover, these numbers are good indicators of how cold winters are and provide indications of the general trend of temperature over a given period of time. This suggests that joint analysis of such skiing days time series form yet another potential source for studying climate change.

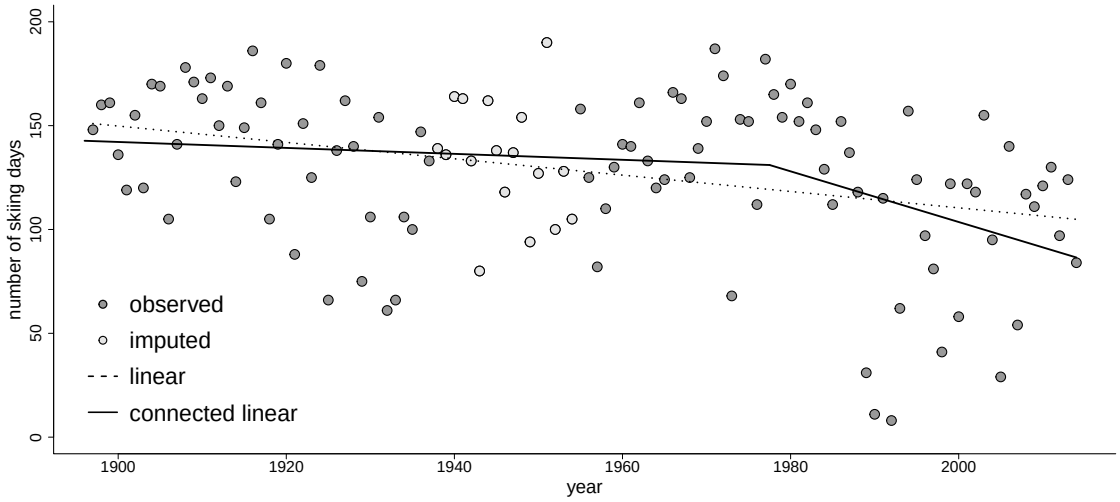

FIGURE 8.1. The number of skiing days for the winter seasons 1896-97 to 2013-14 at Bjørnholt. The imputed data points are meteorologists' reconstructions making use of nearby locations. The global linear trend (dashed line) decreases with an estimated slope of about -0.40. Some time after 1960 there appears to be a structural change in the series. The estimated year for the change-point of the connected linear model (full line) is at 1977, where the relative slope changes from a modest -0.14 to a dramatic -1.22.

Letting Y_t be the number of skiing days for year t, we consider change-point models of the type

$$Y_t = m(\beta_L, t) + \varepsilon_t \text{ for } t \le \tau \quad \text{and} \quad Y_t = m(\beta_R, t) + \varepsilon_t \text{ for } t \ge \tau + 1, \tag{8.1}$$

with $m(\beta, t)$ being suitable trend functions (here taken constant or linear), and where $\{\varepsilon_t\}$ is an autoregressive time series model of order one, i.e. an AR(1). The latter is defined via the representation $\varepsilon_t = \rho\varepsilon_{t-1} + \sigma\delta_t$, with the δ_t being independent and standard normal. Some analysis suggests an AR(1) captures the essential dependency structure here, with higher order autoregressions leading to overfitting. For our analyses we do use the full data sequence, but to avoid instability in the estimated model at the edges we only consider change-point candidates $\tau \in \{1907, \ldots, 2004\}$.

We will actually go through and briefly compare four different specialisations of the model above. The three first take an unchanged AR(1) process for the ε_t

but three different trend functions, each with a change-point: (i) constant, where $m(\beta_L, t) = \beta_L$ and $m(\beta_R, t) = \beta_R$; (ii) linear, using disconnected linear regression models of time, one to the left and one to the right of τ; and (iii) connected linear, meaning two separate linear models with the additional restriction of continuity, i.e. using trend $a_L + b_L t$ to the left and $a_R + b_R t$ to the right, but with $a_L + b_L(\tau + \frac{1}{2}) = a_R + b_R(\tau + \frac{1}{2})$. Model (iv) uses a common linear $a + bt$ trend across the 118 winters but allows the σ associated with the ε_t of (8.1) to jump from some σ_L to some σ_R. The number of unknown parameters for these four models, including the change-point τ itself, are 5, 7, 6, 5, respectively.

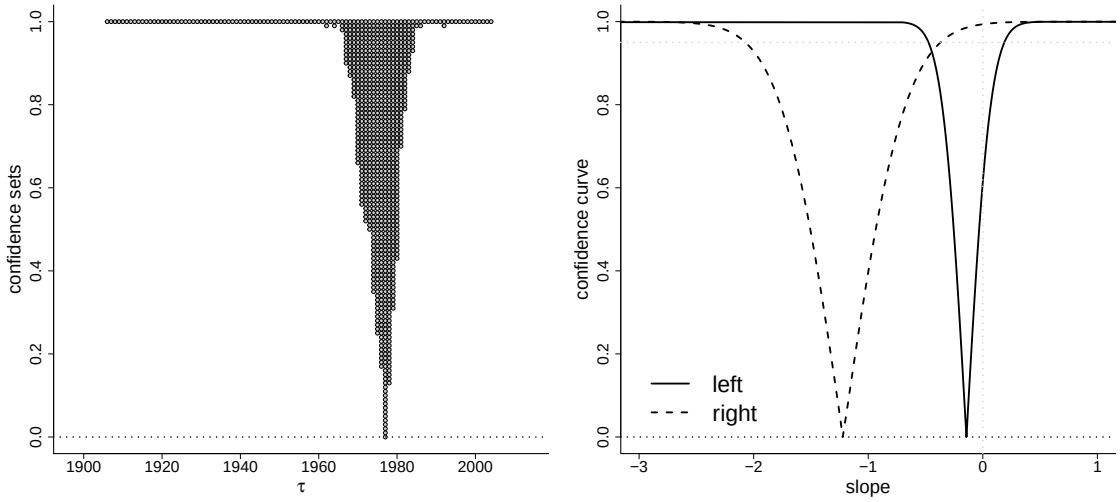

FIGURE 8.2. The confidence sets for the change-point τ (left) and the confidence curves for the two slopes (right) in the connected linear model. The estimated change-point for the connected linear model (with its six parameters) is 1977, where the mean slope changes from -0.14 to -1.22.

Our intention here is not to go into a detailed analysis of the underlying meteorological phenomena. Instead we aim at demonstrating the usefulness of our general method B of Section 3 for reaching confidence distributions, within the framework of change-point models with dependent errors. The main focus is on τ, but we also take an interest in the effect a change-point has on the estimated slopes. Method B of Section 3 yields confidence inference for τ and for degree of change parameters, within each of the four models just described.

The data in Figure 8.1 appear to indicate either a strong decreasing trend, or a change in the structure of the underlying model, perhaps some time after 1960. Our model (i), which has a change in a constant mean, finds via ML that the most likely year for a change is 1988, with $\widehat{\delta} = \widehat{\beta}_L - \widehat{\beta}_R = 45.65$. In words, everything is stable until 1988, then there is a massive drop, from 138.00 to 92.35, in the

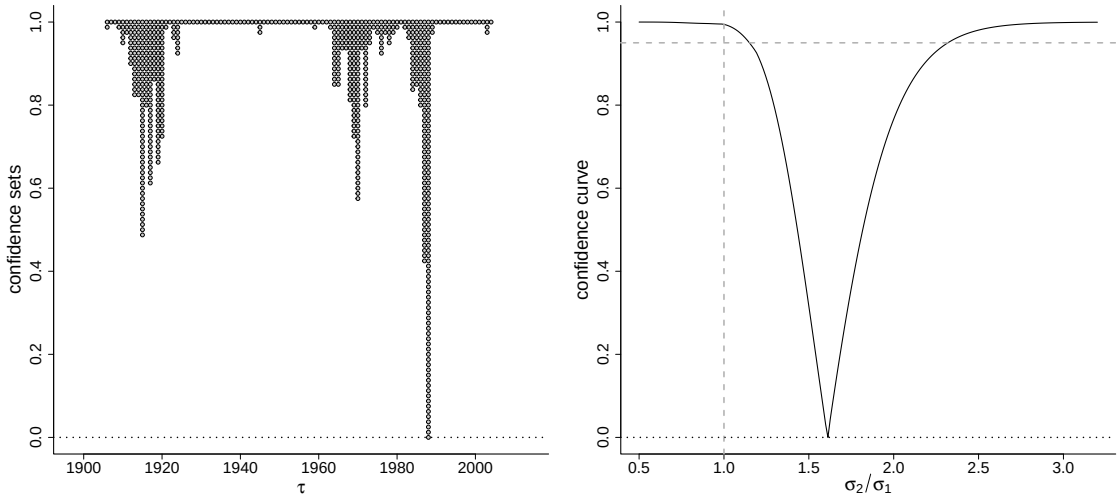

FIGURE 8.3. The confidence curve (left) suggests three possible locations for a change in σ, viz. 1915, 1970 and 1988, with the latter given most confidence. The confidence curve for σ_R/σ_L (right) indicates that the σ of the AR(1) part of (8.1) has increased, around 1988, with a factor of about 1.61.

expected number of skiing days per winter. Then consider model (ii), with two separate linear trends. The mean slope for the first part is approximately zero, with $\widehat{\beta}_L = (139.84, -0.04)$ given as (intercept, slope). Then there is a sudden drop and change in the expectation at the break-point $\widehat{\tau} = 1988$, now with a steady increase thereafter, with $\widehat{\beta}_R = (-131.57, 2.12)$, almost returning to the pre 1988 change-point level with an expected 118 and 120 skiing days for the 2013-14 and 2014-15 winter seasons. The abrupt changes found when analysing these two models do not match prior meteorological conceptions well, and indicate overfitting. For these reasons we prefer models (iii) and (iv). Figure 8.2 pertains to change-point analysis within model (iii), with connected linear trends, displaying a confidence curve for τ, with point estimate 1977, and confidence curves for the (negative) slope parameters for the trend before and after the break point.

At the outset it is by no means obvious that the heterogeneity seen in the data (interpreted in a wide sense) is a result of a change in mean structure. The apparent change of behaviour could potentially be caused by a sudden change in either dependence (i.e. the ρ parameter), the variability (i.e. the σ), or both. Investigations via method B do not provide any evidence of a change in the correlation structure. There is however some evidence that the standard deviation σ is not constant across years, see Figure 8.3. This model (iv) suggests that there is a change in σ around $\widehat{\tau} = 1988$. The estimated parameters are $\widehat{\beta} = (147.5, -0.27)$, $\widehat{\rho} = 0.31$, $(\widehat{\sigma}_L, \widehat{\sigma}_R) = (30.52, 49.15)$.

9. Application 4: the Hjort time series 1859-2012

As an illustration of our general method A of Section 2, we apply the new monitoring bridge plots from Section 4.2 to first test for full homogeneity of a long and prominent time series from fisheries sciences, and then to look for a regime shift. The time series in question is the Hjort liver quality index time series for the skrei, the Northeast Arctic cod. In marine biology this hepatosomatic index (HSI) is used as a measure or indicator for the 'quality of fish' in a certain population, and then typically studied as a time series; see Figure 9.1 (left panel). The index (in so-called bulk form) may be represented as

$$\text{HSI} = 100 \times \frac{\text{total amount of liver}}{\text{total amount of fish}} = 100 \times \frac{\sum x_i}{\sum y_i}, \tag{9.1}$$

where (x_i, y_i) represent the weight of the liver and the total weight of fish number i in one or several catches of fishes; in the Lofoten fishery tens of millions of fish are landed each year. The study of the liver quality index for the skrei goes back to Hjort (1914), where such measurements for the time period 1880–1912 were recorded and analysed, as part of his seminal work on the population dynamics underlying the fluctuations of the great fisheries. The series has since then been extended both forwards and backwards in time, to 1859–2012, yielding one of the longest time series of marine science; see Kjesbu et al. (2014) and Hermansen et al. (2016).

The underlying dynamics and evolution of such series are of great importance in marine biology. Studies of how the HSI evolves over time and interacts with and are influenced by associated factors include Kjesbu et al. (2014); Vasilakopoulos & Marshall (2015); Hermansen et al. (2016). Here we focus on a subset of this long time series, namely the years 1921–2012, where also the detailed monthly average temperatures for Kola are available, see Boitsov et al. (2012). From these monthly averages the average winter temperature can be constructed, averaging the monthly means from the start of October (previous year) to start of March (current year). Letting Y_i be HSI for year i, consider the model where

$$Y_i = \beta_0 + \beta_{\text{kola}} x_{i-1} + \varepsilon_i, \tag{9.2}$$

with $i = 1, \ldots, 90$ representing the years 1922–2012, and where $\{\varepsilon_i\}$ is an autoregressive process of order one and x_{i-1} is the winter Kola temperature for the previous year (checks suggest that there is no real model fit improvement using a higher order autoregressive model). Several tests indicate that last year's winter average temperature carries more relevant information for the present value of the HSI, than does the same year's winter temperature; this also matches biological arguments, see Hermansen et al. (2016) for additional discussion. Model (9.2) is quite simple and is

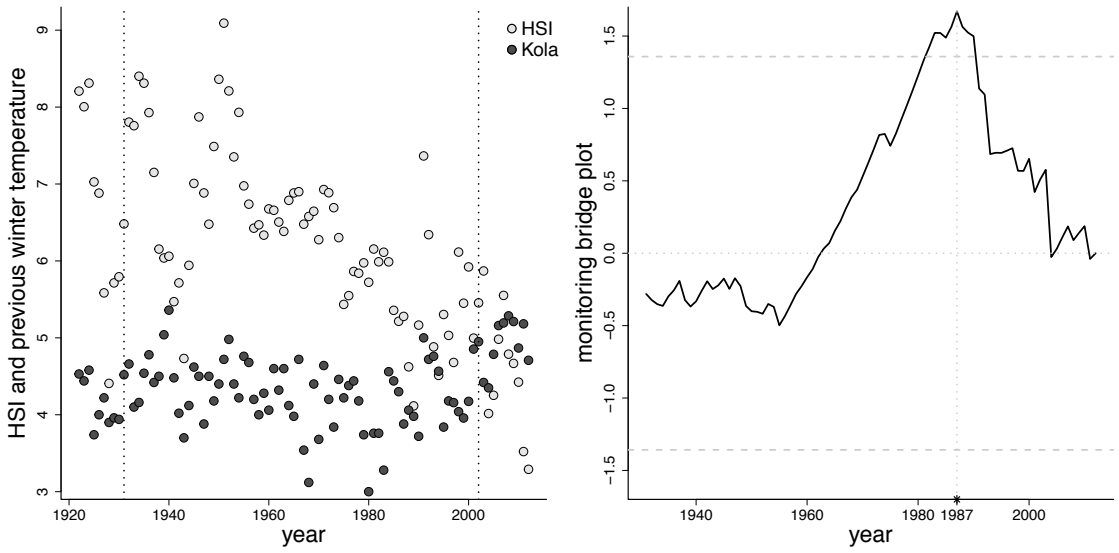

FIGURE 9.1. Left panel: The Hjort liver quality index (HSI) series for 1921–2012 (grey), along with average Kola winter temperature (black, in degrees Celsius). The vertical lines indicate the range (1927–2006) where we are searching for a potential change; the boundary points are excluded due to instability of the methods at the edges. Right panel: The monitoring bridge plot, reaching a maximum value of 1.67, with a corresponding p-value less than 0.01, suggesting a structural change in the model around 1987.

not meant to fully represent all the complex processes in the ocean influencing the HSI index. The goal here is to illustrate our regime shift assessment methodology.

We use the theory of Section 4.2 to compute the monitoring bridge plot for the HSI model (9.2), see Figure 9.1 (right panel). It indicates that the model is not sufficient for describing the underlying mechanism generating the full time series. The shape of the plot also suggests the existence of a regime shift. We shall search for such a change-point, here using the general method A of Section 2 to construct confidence sets for the location of this potential change. In short, the strategy is to test for homogeneity to the left and to the right of each candidate point τ, using our bridge plots. We do utilise the full data sequence in our analysis, but exclude the first and last ten years from the list of candidate values for τ, which we hence take as $1932, \ldots, 2002$. The resulting confidence curves are presented in Figure 9.2.

Our monitoring bridge tools are constructed to test the suitability of a model. A structural break should therefore be interpreted as indicating that the underlying model changes from one regime to another. Other terms used in marine science and biology include 'state shift' and 'critical transition'. A regime shift is characterised by "relatively rapid change (occurring within a year or two) from one decadal-scale

period of a persistent state (regime) to another decadal-scale period of a persistent state (regime)"; see King (2005) and also Brander (2010); Vasilakopoulos & Marshall (2015). Also, note that the underlying framework for our general method A assumes that the observations to the left and the right are independent of each other, as per (2.1); within a segment, however, the observations may be strongly dependent without violating the underlying assumptions of the method. For the time series framework there is not strict independence between goodness-of-fit statistics computed to the left and the right of a given τ; the dependency is however not strong here (the first order autoregressive model seems to capture most of structure), and such a mild deviation from the underlying assumptions does not invalidate the results using these versions of method A. A conservative Bonferroni correction, as spelled out at the end of Section 2, yields a fairly similar confidence curve, for all confidence levels above 0.60.

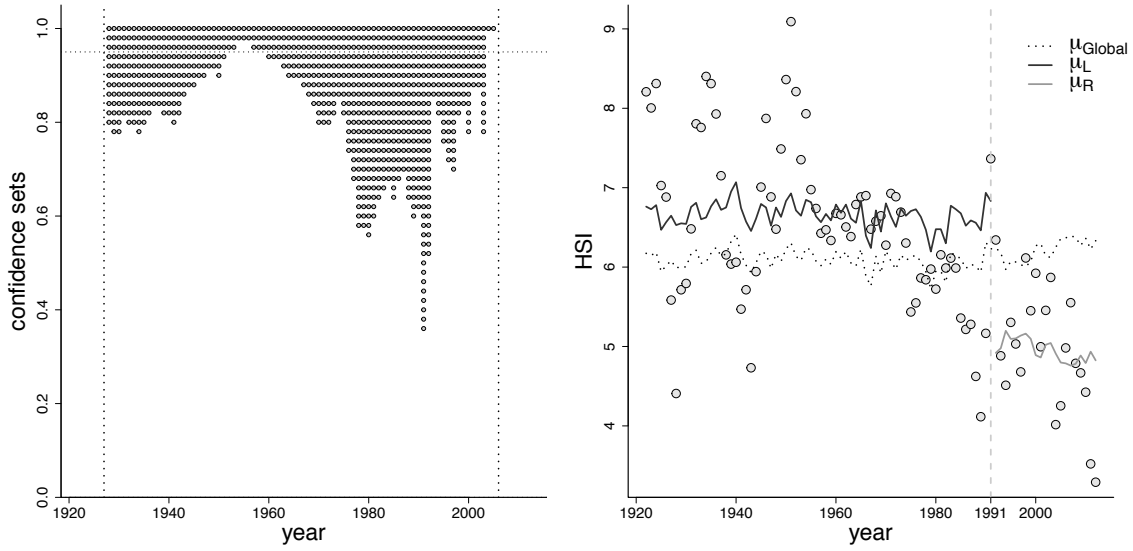

FIGURE 9.2. Left panel: The confidence curve for a regime shift τ obtained via method A, with absolute maxima of monitoring log-likelihood bridges, as in Section 4.2. The curve indicates two plausible regions for τ; one right before 1980 (which may be related to a decrease in variance, as suggested in Figure 9.1), and the second around 1991 (perhaps a change in the relationship between the HSI and the Kola winter temperature). Right panel: Estimates of HSI using the previous year's winter temperature, via (9.2), before and after the estimated regime shift.

We point out that a similar study, also involving the Kola winter temperature, is given in Hermansen et al. (2016), and that an investigation of structural breaks for this series is also conducted by Kjesbu et al. (2014); these studies tentatively identify a potential departure in the pattern connecting Kola temperature and HSI

model	$\widehat{\beta}_0$	$\widehat{\beta}_{\rm kola}$	$\widehat{\rho}$	$\widehat{\sigma}$
global	4.85 (0.85)	0.29 (0.16)	0.86 (0.06)	0.68 (0.05)
left	5.09 (0.77)	0.37 (0.16)	0.78 (0.08)	0.62 (0.05)
right	6.36 (2.26)	−0.30 (0.48)	0.39 (0.27)	0.71 (0.11)

TABLE 9.1. Estimated parameters (with standard errors in parentheses) for the model defined in (9.2), using the complete series observations, i.e. no change points (global), and the for the two sets 1922–1991 (left) and 1992–2012 (right) corresponding to the estimated change point between 1991 and 1992. The two most striking changes are the reversed influence of Kola temperature and change in correlation after 1991.

in beginning of the 1980s. Also, Vasilakopoulos & Marshall (2015) identified a regime shift having taken place in 1981 using principal component analyses on 13 North-East Atlantic cod population descriptors (including HSI) and 5 so-called stressors (also including Kola temperature). According to these authors, the shift in the early 1980s was largely driven by the combined effect of low temperature, high mortality rate and low stock size. Our methodology is capable not only of estimating the location of a potential change-point, but also to supplement such estimates with a measure of uncertainty using confidence sets; such questions are not touched upon in these other studies.

10. Concluding remarks

Below we offer a few concluding remarks, some pointing to further relevant research questions.

10.1. **Approximations and related approaches.** For our general method B we have relied on straight simulations to compute the required probabilities and confidence curves, as with (3.3). This brute force method works well, but approximations to the distributions of both the ML estimator $\widehat{\tau}$ and the deviance statistic $D(\tau, Y)$ can be worked with too; these are by necessity more complicated than the usual results concerning limiting normality and chi-squared-ness of deviances valid for continuous parameters of smooth parametric models. Such results have however the potential to both speed up calculations of confidence curves and to yield additional insights, also when it comes to comparing performances of different strategies. Methods initially worked with in Hinkley (1970), Hinkley & Hinkley (1970) and later on by Cobb (1978), Worsley (1986) and other authors are relevant here, and have the potential for being developed and finessed further. These lead in particular to certain approximations for the case where both τ and $n - \tau$ are large. Such envisioned results ought also to shed more light on questions of performance and for theoretical comparison of different confidence curve constructions.

Notably, Siegmund (1988) discusses the performance of several methods in a single change-point setting. He starts by comparing five different methods for the simple situation where the change-point τ is the only unknown parameter, i.e. when θ_L and θ_R are known. He also presents a method for the more general (and interesting) case, where θ_L and θ_R are unknown. The method produces exact confidence sets and is related to our method B. It can be re-written as a confidence curve and in our notation as

$$\mathrm{cc}(\tau, y_{\mathrm{obs}}) = P_\tau\{D(\tau, Y) < D(\tau, y_{\mathrm{obs}}) \,|\, \widehat{\theta}_L(\tau), \widehat{\theta}_R(\tau)\}. \tag{10.1}$$

The method is restricted to models within the exponential family, where we have sufficient statistics for the θ parameters and where one thus obtains a probability only dependent on τ by conditioning on the ML estimates $\widehat{\theta}_L(\tau)$ and $\widehat{\theta}_R(\tau)$ for each τ value. In practice this requires the user to simulate copies Y^* of the dataset from the conditional distribution of $Y \,|\, (\widehat{\theta}_L(\tau), \widehat{\theta}_R(\tau))$, as opposed to our method B where data are generated from $f(y, \widehat{\theta}_L)$ and $f(y, \widehat{\theta}_R)$, with the ML estimators $\widehat{\theta}_L = \widehat{\theta}_L(\widehat{\tau})$ and $\widehat{\theta}_R = \widehat{\theta}_R(\widehat{\tau})$. We have not yet undertaken a thorough comparison between our method B and Siegmund's method, but our initial investigations suggest that the two methods give very similar confidence curves in many cases. However, when either τ or $n - \tau$ are small, confidence sets from Siegmund's method appear to obtain more correct coverage than method B. This is not surprising as our method relies on estimating the θ parameters sufficiently well. Contrary to our method B, Siegmund's method is restricted to the class of exponential family models and is also more difficult to use in some cases, as generating datasets from $Y \,|\, (\widehat{\theta}_L(\tau), \widehat{\theta}_R(\tau))$ can be complicated. Siegmund (1988) also provides approximations to the conditional probability in (10.1). Again these rely on both τ and $n - \tau$ being large, and remain yet to be compared with our methods.

10.2. **Combination of information.** There are sometimes several sources of information about a given change-point. In our analysis of *Tirant lo Blanch* in Section 7, for example, we investigated how the change of authors is reflected in aspects of the distribution of word lengths per chapter, such as the the mean word length chapter by chapter. There it is also worthwhile examining the sentence length distribution, through the chapters, to see if a change of author style can be detected there. Via a suitable R script operating on an electronic version of the Catalan 1490 manuscript we have indeed gotten hold of the string of the manuscript's 17593 sentence lengths. The mean sentence lengths, chapter by chapter, can be modelled as normally distributed on both sides of τ with (possibly) different mean and variance parameters and with the variance depending on m_i', the number of sentences in chapter i. The

mean word length and mean sentence length can be analysed separately by the methods developed in this paper, and as they can be considered independent sources of information, their inference on τ may be combined. One potential strategy is to use ideas related to combination of p-value functions in Liu et al. (2014), for a particular case involving discrete distributions, but there are better methods, as shown in Cunen & Hjort (2015), Cunen & Hjort (2016). The parallel for the present case is to stay with the log-likelihood profiles, naturally extending method B. Let $\ell_{\rm prof,1}(\tau)$ and $\ell_{\rm prof,2}$ be the profiled log-likelihoods function from information sources 1 and 2. These can be summed to $\ell_{\rm prof,comb}(\tau) = \ell_{\rm prof,1}(\tau) + \ell_{\rm prof,2}(\tau)$, from which we can find the combined maximum likelihood estimator $\widehat{\tau}$ and construct the combined deviance function, say $D_{\rm comb}(\tau, Y) = 2\{\ell_{\rm prof,comb}(\widehat{\tau}) - \ell_{\rm prof,comb}(\tau)\}$. Then we can simulate its distribution at each position τ by generating a large number of datasets Y_1^* and Y_2^* based on the first and second data source respectively. The result of the combination varies from case to case; if the two sources produce the same τ estimate then the combined confidence curve will also point to the same number, but with slimmer/smaller confidence sets, reflecting the increase in information. If the two sources have different estimates of τ, the combined confidence curve may give an estimate between the two sources (a compromise), but it may also favour the estimate from one source over the other. This is exactly what we observe with *Tirant lo Blanch*; in Figure 10.1 we see that the sentence length data indicate a much earlier change of author than the word length dataset (see also Section 7). The combined confidence curve is quite similar to the one from the word length data; the log-likelihoods and deviances thus appear to judge this source more informative than the sentence length data.

10.3. **More than one change-point.** The focus of our paper has been that of inference for a single change-point in a sequence of observations, under the operating assumption that precisely one such change-point exists. Sometimes there are strong a priori reasons for this, as with our application story of Section 7. In other cases it is useful to precede a change-point analysis with a test for full homogeneity; only when the data sequence fails such a test is it meaningful to go hunting for change-points. In various applications there may also be more than one breakpoint present. Some of our methods may be extended to cover such cases too, calling also for additional tools, such as model selection mechanisms to decide on the 'right' number of parameter discontinuities.

Our methods can be extended to the case of multiple change-points, but both become more complicated. In some cases we may perform a test, or have some a priori reasons to expect a specific number of change-points, for example two, say τ_1 and τ_2 (and assuming $\tau_1 < \tau_2$). Our method A then corresponds to identifying

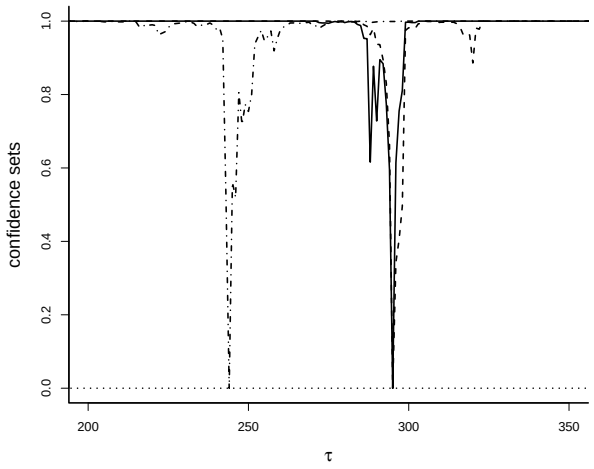

FIGURE 10.1. Confidence curves for the change-of-author-point τ, the dot-dashed line is the curve based on the mean sentence length in each chapter, the dashed line is the curve based on the mean word length in each chapter, and the full line is the combined confidence curve.

confidence regions, at level α, in the following way (see corresponding formula (2.1))

$$\begin{aligned} R(\alpha) &= \{\tau_1, \tau_2 \colon H_{1,\tau_1} \text{ accepted at level } \alpha^{1/3}, H_{\tau_1+1,\tau_2} \text{ accepted at level } \alpha^{1/3}, \\ &\qquad H_{\tau_2+1,n} \text{ accepted at level } \alpha^{1/3}\} \\ &= \{\tau_1, \tau_2 \colon Z_{1,\tau_1} \le G^{-1}_{1,\tau_1}(\alpha^{1/3}), Z_{\tau_1+1,\tau_2} \le G^{-1}_{\tau_1+1,\tau_2}(\alpha^{1/3}), Z_{\tau_2+1,n} \le G^{-1}_{\tau_2+1,n}(\alpha^{1/3})\}. \end{aligned}$$

This will produce joint confidence regions for τ_1 and τ_2. For method B, however, it is more natural to consider confidence curves for each of the change-points separately. With two change-points the likelihood takes the form

$$\ell(\tau_1, \tau_2, \theta_L, \theta_M, \theta_R) = \sum_{i=1}^{\tau_1} \log f(y_i, \theta_L) + \sum_{i=\tau_1+1}^{\tau_2} \log f(y_i, \theta_M) + \sum_{i=\tau_2+1}^{n} \log f(y_i, \theta_R),$$

where θ_M is the model parameter between the two change-points. In order to construct a confidence curve for one of the two change-points, say τ_1, we need (as before) the profile log-likelihood function,

$$\ell_{\text{prof}}(\tau_1) = \max_{\tau_2, \theta_L, \theta_M, \theta_R} \ell(\tau_1, \tau_2, \theta_L, \theta_M, \theta_R), \tag{10.2}$$

requiring $\ell(\tau_1, \tau_2, \theta_L, \theta_M, \theta_R)$ to be maximised over θ_L, θ_M, and θ_R as before, but also over all possible values of τ_2. The confidence curve is constructed in a similar way as before, but in this case the distribution of the deviance will be depending on τ_2 and the success of our simulation recipe will then depend on how well we estimate τ_2 from the data. Neither of these two suggestions has been tried out in detail. Further work in these directions could possibly follow ideas from Schweder (1976), Yao &

Au (1989), Bai & Perron (1998) and Braun et al. (2000). However, these articles do not treat change-points very generally. Yao & Au (1989) and Braun et al. (2000) consider segmentation problems, while Schweder (1976) and Bai & Perron (1998) work in a regression setting.

References

Bai, J. & Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* **66**, 47–78.

Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley.

Boitsov, V. D., Karsakov, A. L. & Trofimov, A. G. (2012). Atlantic water temperature and climate in the Barents Sea, 2000–2009. *ICES Journal of Marine Science* **69**, 833–840.

Brander, K. M. (2010). Impacts of climate change on fisheries. *Journal of Marine Systems* **79**, 389–402.

Braun, J. V., Braun, R. & Müller, H.-G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika* **87**, 301–314.

Carlin, B., Gelfand, E. & Smith, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Applied Statistics* **41**, 389–406.

Carlstein, E., Müller, H. & Siegmund, D. (1994). *Change-Point Problems*. New York: Institute of Mathematical Statistics.

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

Claeskens, G. & Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.

Cobb, G. W. (1978). The problem of the Nile: Conditional solution to a change-point problem. *Biometrika* **65**, 243–251.

Cox, D. R. & Spjøtvoll, E. (1982). On partitioning means into groups. *Scandinavian Journal of Statistics* **9**, 147–152.

Csörgő, S. & Faraway, J. (1996). The exact and asymptotic distributions of Cramér–von Mises statistics. *Journal of the Royal Statistical Society Series B* **58**, 221–234.

Cunen, C. & Hjort, N. L. (2015). Optimal inference via confidence distributions for two-by-two tables modelled as Poisson pairs: Fixed and random effects. In *Proceedings ofthe 60th World Statistics Congress*, F. Samaniego, ed. Rio de Janeiro: International Statistical Institute, pp. 3581–3586.

Cunen, C. & Hjort, N. L. (2016). Combining information across diverse sources: The ii-cc-ff paradigm. In *JSM Proceedings*. Alexandria, VA: American Statistical Association.

Efron, B. (2015). Frequentist accuracy of Bayes estimators. *Journal of the Royal Statistical Society Series B* **77**, 617–646.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple change-point problems. *Statistics and Computing* **16**, 203–213.

Fraser, D. A. S. (2011). Is Bayes posterior just quick and dirty confidence? [with discussion and a rejoinder]. *Statistial Science* **26**, 249–316.

Frick, S., Munk, A. & Sieling, H. (2014). Multiscale change-point inference [with discussion contributions]. *Journal of the Royal Statistical Society Series B* **76**, 495–580.

Frigessi, A. & Hjort, N. L. (2002). Statistical models and methods for discontinuous phenomena. *Journal of Nonparametric Statistics* **14**, 1–6.

Fukuyama, F. (1992). *The End of History and the Last Man*. Simon and Schuster.

Girón, J., Ginebra, J. & Riba, A. (2005). Bayesian analysis of a multinomial sequence and homogeneity of literary style. *American Statistician* **59**, 19–30.

Gladwell, M. (2000). *The Tipping Point: How Little Things Can Make a Big Difference*. New York: Little Brown.

Gould, S. J. & Eldredge, N. (1977). Punctuated equilibria: the tempo and mode of evolution reconsidered. *Paleobiology* **3**, 115–151.

Hermansen, G. H., Hjort, N. L. & Kjesbu, O. S. (2016). Recent advances in statistical methodology applied to the Hjort liver index time series (1859-2012) and associated influential factors. *Canadian Journal of Fisheries and Aquatic Sciences* **73**, 279–295.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika* **57**, 1–17.

Hinkley, D. V. & Hinkley, E. A. (1970). Inference about the change-point in a sequence of binomial random variables. *Biometrika* **57**, 477–488.

Hjort, J. (1914). *Fluctuations in the Great Fisheries of the Northern Europe Viewed in the Light of Biological Research*. Copenhagen: Rapports et Procès-Verbaux des Réunions du Conseil International pour l'Exploration de la Mer.

Hjort, N. L. (2007). And Quiet Does Not Flow the Don: Statistical analysis of a quarrel between Nobel laureates. In *Concilience*, W. Østreng, ed. Oslo: Centre for Advanced Research, pp. 134–140.

Hjort, N. L., Holmes, C., Müller, P. & Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge University Press.

Hjort, N. L. & Koning, A. (2002). Tests for constancy of model parameters over time. *Journal of Nonparametric Statistics* **14**, 113–132.

Jarrett, R. G. (1979). A note on the intervals between coal-mining disasters. *Biometrika* **66**, 191–193.

King, J. R. (2005). Report of the study group on fisheries and ecosystem responses to recent regime shifts. Tech. Rep. 28, North Pacific Marine Science Organization.

Kjesbu, O. S., Opdal, A. F., Korsbrekke, K., Devine, J. A. & Skjæraasen, J. E. (2014). Making use of Johan Hjort's 'unknown' legacy: reconstruction of a 150-year coastal time series on Northeast Arctic cod *(Gadus Morhua)* liver data reveals long-term trends in energy allocation patterns. *ICES Journal of Marine Science* **71**, 2053–2063.

Koziol, J. A. (2014). A note on change-point estimation in a multinomial sequence. *Enliven: Biostatistics and Metrics* **1**, 1–4.

Liu, D., Liu, R. & Xie, M. (2014). Exact meta-analysis approach for discrete data and its application to 2×2 tables with rare events. *Journal of the American Statistical Association* **109**, 1450–1465.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979). *Multivariate Analysis*. New York: Academic Press.

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

Rosenthal, D. H. (1984). Tirant lo Blanc: Foreword to the new translation.

Schweder, T. (1976). Some "optimal" methods to detect structural shift or outliers in regression. *Journal of the American Statistical Association* **71**, 491–501.

Schweder, T. & Hjort, N. L. (2016). *Confidence, Likelihood, Probability*. Cambridge: Cambridge University Press.

Siegmund, D. (1988). Confidence sets in change-point problems. *International Statistical Review/Revue Internationale de Statistique* **56**, 31–48.

Spengler, O. (1918). *Der Untergang des Abendlandes*. Wien: Braumüller.

Vasilakopoulos, P. & Marshall, C. T. (2015). Resilience and tipping points of an exploited fish population over six decades. *Global Change Biology* **21**, 1834–1847.

Worsley, K. J. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika* **73**, 91–104.

Xie, M.-g. & Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review. *International Statistical Review* **81**, 3–39.

Yao, Y.-C. & Au, S. (1989). Least-squares estimation of a step function. *Sankhyā: The Indian Journal of Statistics, Series A* , 370–381.