Wars and Whales: Extensions and Applications of Confidence Curves and Focused Model Selection

Céline Cunen

Dissertation presented for the degree of
Philosophiae Doctor (PhD)

Department of Mathematics
University of Oslo
August 2018

Preface

Four years ago, almost to the day, I started my PhD period at the University of Oslo. I was excited, but also slightly worried. With my background in applied statistics and biology from the Norwegian University of Life Science (NMBU), I felt a certain awe upon entering the Department of Mathematics. Looking back I can attest that my worries were not misguided, the PhD process has partly been difficult. I had much more to learn than I anticipated, and during the first two years I considered giving up multiple times. Luckily, I stayed, and I have been rewarded with a much deeper and broader understanding of the field of statistics than I started out with.

The original plans for this thesis were somewhat different from the final result. They were primarily concerned with ideas related to the last paper in this thesis, Paper IV. Eventually, other ideas emerged and caught our interest, sometimes originating from applied problems, for example leading to Papers II and III. However, in all these papers, similar methods came into use, and I therefore maintain that I did not stray too far from the intended path.

My main supervisor, Nils Lid Hjort, has played an important role in this thesis. He has been a great support throughout the process; helpful and always enthusiastic. Importantly, he has encouraged me to pursue various activities outside of 'regular' PhD work, for example the writing of blog posts, which has led to a number of popular science talks. Also, Nils is the leader of the FocuStat group, and I am very grateful for having been part of such an active and close-knit research group. My deep thanks to all FocuStat members for a large number of interesting discussions and pleasant company.

I am also very grateful to my other co-authors, Gudmund Horn Hermansen, Lars Walløe and Håvard Mokleiv Nygård. Lars, in particular, gave me the opportunity to present and defend my work at the Scientific Committee of the International Whaling Commission. Thanks to my co-supervisor Bo Lindqvist for pleasant conversations and meetings, most memorably at the ISI World Statistics Congress in Marrakech. I first learnt to love statistics at NMBU, for this I owe particular thanks to Trygve Almøy and Solve Sæbø. Further, I am very grateful to professor Sylvia Richardson at the MRC Biostatistical Unit in Cambridge, where I was a visitor for the autumn semester in 2017, and to all the PhD students at the BSU for making my stay so pleasant and memorable. Special thanks to all colleagues at UiO, particularly the PhD students.

Finally, I am grateful to friends and family for filling my life with joy and distractions. Particularly Jonas, for numerous discussions on statistical topics and great support, especially in the final weeks before submission; Sigrun, my oldest and best friend; my parents Josiane and Per Axel, for their help and encouragement; and my dear siblings, Christophe and Mélanie.

Céline Cunen, Oslo, August 2018

List of papers and manuscripts

Paper I

Cunen, C., Hermansen, G., and Hjort, N. L. (2018). Confidence distributions for change-points and regime shifts. *Journal of Statistical Planning and Inference* 195, 14–34.

Paper II

Cunen, C., Hjort, N. L., and Nygård, H. M. (2018). Statistical Sightings of Better Angels: Analysing the Distribution of Battle Deaths in Interstate Conflict over Time. *Invited to submit a revision to Journal of Peace Research*.

Paper III

Cunen, C., Walløe, L., and Hjort, N. L. (2018). Focused model selection for linear mixed models, with an application to whale ecology. *Invited to resubmit to Annals of Applied Statistics*.

Paper IV

Cunen, C., and Hjort, N (2018). Combining information across diverse sources: The II-CC-FF paradigm. *Submitted for publication*.

Contents

1 Introduction

As is apparent from the title, this thesis consists of several themes. Besides attempting to catch the reader's attention, the first part is a reference to *applications*. The second part of the title designates areas of methodological effort. This unification of applications and methods, or theory, if you want, is what characterises the field of statistics in general, and also this thesis in particular.

The applications appearing in this thesis come from a variety of scientific fields. One of the benefits of being a statistician is the opportunity to "play in everyone's backyard", as stated by Tukey. The "wars" designate Paper II, where we have involved ourselves in a current debate in the field of peace research. The "whales" appear in Paper III and refer to an ecological dataset concerning Antarctic Minke Whales. Other applications are treated in all four papers, as illustrations or examples. These involve for instance analyses of literary style in a medieval novel, measurements of a liver index in Arctic cod and meta-analyses of clinical trials.

The four papers in this thesis are connected via a shared attention to *modelling*. They also make use of similar statistical methods. Particularly, the framework of confidence distributions, or similarly *confidence curves*, plays a role in all four papers. Both of these concepts will be described in the next chapter. Some researchers see confidence distributions as an attempt at unification of the Bayesian and frequentist school of thinking (Schweder & Hjort, 2016; Xie & Singh, 2013), but I do not necessarily ascribe to this view. I see confidence distributions as an inherently frequentist framework, but which opens the door to Bayesian methods having good frequentist properties. Some differences between frequentist and Bayesian methods will be discussed in Section 5.1.

Another general theme is *statistical inference*, i.e. making statements with an attached uncertainty about unknown parameters. Some authors contrast inference with prediction, where one seeks to make statements about an unobserved variable, usually also with some attached uncertainty. Considerable efforts in modern statistics are devoted to prediction methods, but prediction will not play a direct role in this thesis.

Statistical inference is sometimes also contrasted with model selection, which concerns the broader question of selecting the most suitable statistical model among a set of different models. Model selection, particularly of the focused type, appears in this thesis, mostly in Paper III. The concept of the *focus* will be a regular feature throughout the papers. On one hand, the focus parameter simply designates the one-dimensional parameter of primary interest in both inference and model selection problems. More generally, the focus encapsulates the idea that statistical analyses are always undertaken with a specific goal in mind. The goal could be to decide between two competing scientific theories, or to estimate a certain quantity with precision. The goal should influence the modelling and the choice of statistical tools.

Inference in a special setting, *change-point* analysis, appears in three of the papers. Given a

sequence of ordered observations, change-point methods aim at identifying the position where the distribution of the data changes. Methods for change-point analysis are the main theme of Paper I. One of these methods is applied in Paper II, and also in one of the illustrations in Paper IV.

Before coming to the four papers that constitute this thesis, I will start by providing some explanations of general concepts, specifically confidence distributions in Chapter 2 and focused model selection in Chapter 3. These two methodological chapters are followed by brief summaries of the four papers in Chapter 4, with some emphasis on elements that will be relevant for the discussion in Chapter 5. This discussion chapter is meant to be different from the discussions already present within each paper, particularly concerning two aspects.

First, I have aimed at addressing more general questions than in the papers. In the first notes for the discussion chapter, these questions took the informal forms: why do people care about coverage properties? Is model selection really useful? Is statistics beneficial for science? These three questions have been slightly moderated and sharpened, as you will see, but still constitute the essence of the three sections making up the discussion chapter.

Secondly, I have tried to be more critical. It is an empirical fact that new methods in statistical papers usually appear in a favourable light. Similarly in applied papers, the authors usually give the impression that they have made an interesting, and even surprising, discovery. Sometimes these claims are warranted, often they are at least slightly exaggerated. Usually, this is not due to dishonesty on the parts of the authors (and I am not suggesting my co-authors and I have been dishonest), but could be due to various overlapping factors: (i) publishing practices encouraging scientists to make exaggerated claims about their results or their methods, for example in a methodological publication one would typically have to claim that the proposed method is somehow better than existing methods; (ii) psychological mechanisms which hinder us to discover flaws in our own contributions; (iii) the many levels of criticism possible. This last point concerns the field of statistics in particular: our purpose is to construct methods for answering real-world questions, but these are developed within a mathematical framework which seldom is entirely realistic. Thus, for each method we propose, there is a ladder of possible criticism, from "does it fulfil the advertised goals under the mathematical assumptions made" to "can it actually be useful in a realistic scenario". This ladder is the reason why most publications in statistics have a section which investigates performance under the assumptions (often by simulations) and a section where the method is applied to a real dataset. The full scope of potential criticism is seldom explored in any single paper, however; for obvious reasons, since this would make each paper impractically long and complex. I do not claim that I will manage to present all possible criticism of the papers in this thesis, but I have enjoyed the attempt.

2 Confidence distributions and confidence curves

Unlike the near-ubiquitous confidence *intervals*, confidence *distributions* are not a mainstream statistical concept; at least not yet. Since they play a role in all of the papers that constitute this thesis I will start off with a general presentation of confidence distributions (CDs) and related approaches. In Section 2.1, I briefly describe the general motivation behind CDs, as well as some historical roots. Then, I provide the definition of CDs and confidence curves in Section 2.2, and some methods for constructing CDs in Section 2.3. I will return to properties of CDs in the discussion, in Section 5.1.

2.1 Knights of the Holy Grail

Let us start, as we often do in statistics, with some data Y and a statistical model parametrised with an unknown θ. The parameter θ may be multi-dimensional, but suppose we are primarily interested in a scalar parameter γ, our focus parameter, which is a function of θ (or simply one of the elements in θ). The focus parameter γ could be a regression coefficient, some measure of the effect of a treatment in a clinical trial, or a change-point, to pick some examples which will be encountered in this thesis. The central theme of statistical inference is to produce data-dependent statements about γ, usually with an attached *uncertainty measurement*. These statements can take various forms depending on the particular application and also the statistical school we belong to. Typical examples are point-estimates for parameters of relevance and interest, confidence or credibility intervals, providing a set of parameter values agreeing sufficiently well with data via a model, and hypothesis tests. The statement can also take the form of a full probability distribution over the space of possible γ values, which subsumes all the types of statements mentioned above.

The most common example of such distributions over the parameter space is the Bayesian posterior distribution $p(\gamma \,|\, Y)$. I will not go through particulars of Bayesian inference, but simply remind readers that Bayes' formula provides a clear recipe for computation of such a posterior distribution. Bayes' formula requires that users specify a parametric model, and crucially, that they are willing and able to provide a prior distribution $p(\theta)$ reflecting their beliefs and their uncertainty concerning the values of the parameters, prior to observing the data. I will come back to differences between Bayesian and frequentist inference in Section 5.1, but for now it suffices to say that the subjective element inherent in the prior does not sit well with all statisticians – especially in situations where prior information is missing or difficult to translate into a clear distribution.

Obtaining posterior distributions without having to specify subjective priors is the Holy Grail of statistical theory, according to Efron (2010), or as Hjort & Schweder (2018) put it, obtaining

frequentist posterior distributions. Precisely what is meant by frequentist posterior distributions is not immediately clear, and might be understood differently by different authors. I will make a distinctions between two related goals:

- a graphical summary of the inference taking the form of a distribution over the parameter space;
- an epistemic interpretation of this distribution, i.e. obtain a valid frequentist post-data inference.

The first goal is relatively easy to understand and not very hard to obtain. The second is more controversial and I will come back to this in more detail in the discussion.

The Holy Grail has been sought by many researchers throughout statistical history. Fisher proposed a framework to construct distributions over the parameter space called fiducial distributions, see for instance Fisher (1930, 1935). Fisher's fiducial distributions were on one hand a reaction against the ad hoc use of uniform priors to represent ignorance (Schweder & Hjort, 2016, Chapter 6), while on the other hand, a reaction against the orthodox frequentism of for example Neyman (Efron, 1998). Fisher's ideas on this topic created a large controversy. A number of statisticians pointed out various flaws in the fiducial construction, particularly concerning Fisher's insistence that fiducial distributions could be handled by ordinary probability calculus. For example, a joint fiducial distribution for a pair of parameters will not in general yield valid marginal fiducial densities by integration, see a detailed overview of the controversy in Schweder & Hjort (2016, Chapter 6). Following the controversy, fiducial distributions were largely abandoned for many years.

In the last twenty years, there has been a renewed interest in various approaches seeking the Holy Grail. In addition to confidence distributions, the interested reader may investigate Generalised Fiducial Distributions (Hannig, 2009; Hannig et al., 2016) and Inferential Models (Martin & Liu, 2013, 2015); see also the special issue *Inference with Confidence* in the *Journal of Statistical Planning and Inference* 2018. The sub-field of Objective Bayes aims at constructing posterior distributions without any subjective elements in the prior, and with good frequentist properties (Kass & Wasserman, 1996; Berger, 2006). Researchers involved in these topics and others interested in foundational issues in statistics meet at the yearly BFF conferences (for "Bayes, Fiducial, Frequentist" or, allegedly, Best Friends Forever).

2.2 Defining CDs

After the short historical and motivational section above, I turn to the formal definitions of confidence distributions and confidence curves. Again, θ represents the full parameter vector, and the focus parameter $\gamma = a(\theta)$ is a function $a(\cdot)$ of θ. Let θ_0 be the true parameter vector and $\gamma_0 = a(\theta_0)$. The definition used here is from Schweder & Hjort (2002, 2016) and is also used by Xie & Singh (2013) and in related papers.

Definition 2.2.1 (Confidence distribution)**.** *A function $C(\gamma, Y)$ of the scalar γ depending on the data Y is a confidence distribution for γ if*

1. *$C(\gamma, Y)$ is a cumulative distribution function on the parameter space of γ; and*
2. *whatever the value of θ_0, $C(\gamma_0, Y)$ has a uniform distribution $U(0, 1)$ at the true parameter value $\gamma_0 = a(\theta_0)$.*

The first part of the definition ensures that a CD actually is a probability distribution, but the second part is more fundamental. It ensures that tests and confidence intervals constructed based on the CD have the correct coverage, i.e. that a 95% interval obtained from a CD really covers the true unknown parameter 95% of the time if we have repeated observations of the same process. A CD is thus defined based on its performance in repeated applications – an inherently frequentist concept; more on this in Section 5.1.

The uniform distribution in the second requirement of Definition 2.2.1 will often be obtained asymptotically, i.e. when the sample size increases to infinity. Such CDs are sometimes called *approximate* CDs to distinguish them from *exact* CDs, where the uniform distribution is obtained even for finite n.

The following example is meant to illustrate the definition and the CD concept. Assume we have a sample of $n = 5$ observations $Y_1, \ldots, Y_5$ and assume that these observations are independently sampled from a normal distribution $N(\mu, \sigma^2)$ with unknown parameters. Let the focus parameter be the expectation μ. From introductory statistics courses we know that the statistic

$$T = \frac{\mu - \bar{Y}}{\widehat{\sigma}/\sqrt{n}} \tag{2.1}$$

with $\bar{Y} = \sum_{i=1}^{n} Y_i/n$ and $\widehat{\sigma}^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2/(n-1)$, has a Student's t-distribution with $n - 1$ degrees of freedom. This distribution is independent of the two model parameters. The statistic T is famously the basis for hypothesis test on the value of μ and also provides a valid confidence distribution for μ. With G_{n-1} being the cumulative distribution function of the Student's t-distribution, the function

$$C(\mu, Y) = G_{n-1}\left(\frac{\mu - \bar{Y}}{\widehat{\sigma}/\sqrt{n}}\right)$$

is a confidence distribution; it is a cumulative distribution function over the space of potential μ values and provides valid confidence intervals and tests for the focus parameter, in the sense of Definition 2.2.1. By the probability integral transform it is clear that $C(\mu, Y)$ will have a uniform distribution at the true μ value, and this CD is therefore exact.

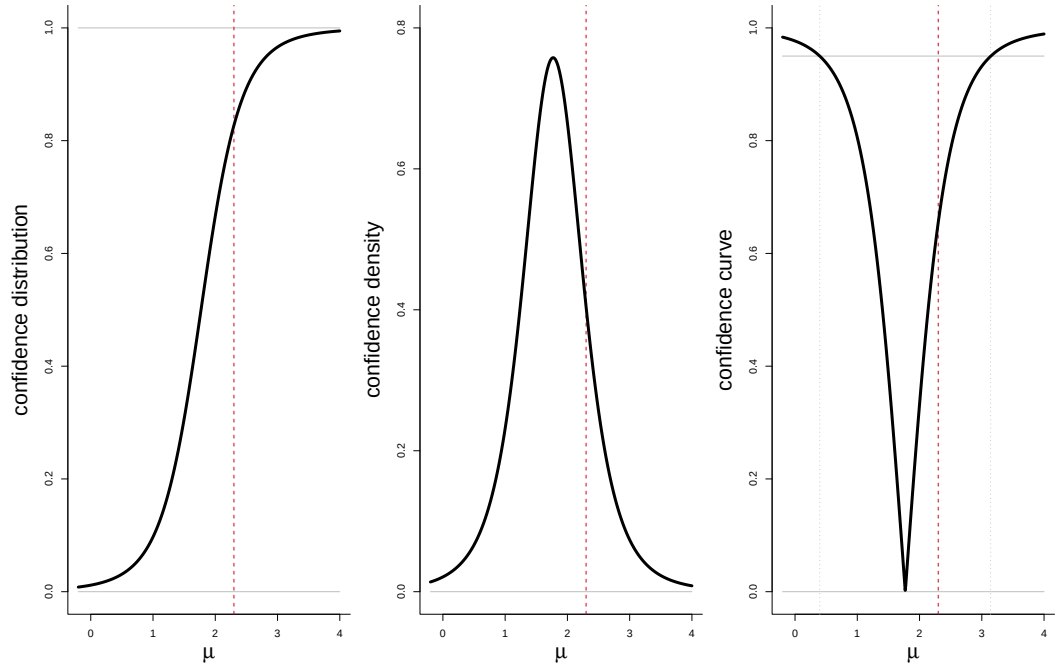

Figure 2.1: *Three representations of a confidence distribution, calculated based on a sample of 5 observations with* $\mu = 2.3$ *and* $\sigma = 1$*. Left panel: as a cumulative distribution function; middle panel: as a density; right panel: as a confidence curve. The vertical red line shows the true* μ *value. The median confidence estimate is* $\widehat{\mu} = 1.77$*, and the dotted vertical lines in the right panel indicate a 95% interval,* $[0.40, 3.14]$.

Figure 2.1 displays three representations of the same CD. We call these the confidence distribution $C(\mu, Y)$, the confidence density $c(\mu, Y) = \partial C(\mu, Y)/\partial\mu$, and the confidence curve $\text{cc}(\mu, Y) = |1 - 2C(\mu, Y)|$. Keep in mind that all of these three objects are random variables, or realisations of random variables. Often we write $C(\mu, Y)$ and $C(\mu, y)$ to make this clear. The true parameter, shown by the red vertical line, is fixed, while the CD will be different for each new sample. In fact by Definition 2.2.1, the value where $C(\mu, Y)$ crosses the red line is a random variable with a uniform distribution.

From a graphical point of view, the confidence density and curve are more informative summaries than $C(\mu, Y)$. The confidence density is similar to the form in which we are used to see Bayesian posteriors, but throughout this thesis the confidence curve will be the favoured CD representation. Point-estimates and confidence intervals can be obtained from all of the three representations, but we find they are most straightforwardly obtained through the confidence curve. The confidence curve points to a point-estimate for μ: the median confidence estimate where $C(\mu, y) = 0.5$. Equi-tailed confidence intervals of all levels can directly be read off from the right panel of Figure 2.1; for example to get a 95% interval, read off the two μ values where the horizontal 0.95 line crosses the confidence curve.

All confidence distributions can be transformed to confidence curves by $\text{cc}(\mu, Y) = |1 - 2C(\mu, Y)|$, and then we automatically have that

$$\text{cc}(\gamma_0, Y) \sim U(0, 1). \tag{2.2}$$

However, this property can be fulfilled for functions of γ which are not necessarily constructed based on a CD, or which might not even have any correspondence to a CD. In other words, there are functions satisfying the property in (2.2) and which we will call confidence curves, but which cannot be turned into valid cumulative distribution functions and therefore do not fulfil the first requirement in Definition 2.2.1. For example, any method capable of producing valid confidence sets at all levels gives rise to a confidence curve satisfying (2.2). Confidence sets can consist of disjoint intervals and a confidence curve based on such disjoint intervals will have multiple valleys, like the confidence curves for change-points in Papers I and II. If one attempts to transform such a multi-valley confidence curve to a CD, the resulting function will violate the monotonicity requirement of a cumulative distribution and thus fall outside Definition 2.2.1. The confidence curve can therefore be considered a broader, more general concept than the confidence distribution itself (Hjort & Schweder, 2018).

2.3 Constructing CDs

Definition 2.2.1 does not provide a method for constructing CDs, but simply defines the concept in terms of some desirable properties. In this section, I will describe some general methods used to construct CDs.

Fisher's fiducial distributions were based on pivots, like in the normal example in Section 2.2. A pivot $\mathrm{piv}(Y, \gamma)$ is a function of the data and the focus parameter which has a distribution function independent of the full parameter vector θ. The T statistic in (2.1) is a well-known example, but there are many others in various model classes. In order to construct an exact CD for γ we need a pivot $\mathrm{piv}(Y, \gamma)$ which is monotone in γ and where we can derive or simulate its cumulative distribution function $G(\cdot)$. If the pivot is increasing in γ,

$$C(\gamma, Y) = G(\mathrm{piv}(Y, \gamma)).$$

See more on pivots in Schweder & Hjort (2016, Chapter 3). If the pivot is not monotone, it may be used to construct a confidence curve directly. Their exactness makes pivots attractive, but they are not always easy to establish in applied settings. Usually, they must be worked out in a case by case manner, and cannot be automatised.

Thus, there is a need for a more general recipe. The following method is used to construct confidence curves directly. It is based on the log-likelihood profile and is used extensively, especially in Paper IV. Assume that our data Y come from a parametric model $f(y, \theta)$. As before, θ represents the full parameter vector and $\gamma = a(\theta)$ is the focus parameter. Sometimes it is convenient to re-parametrise θ so that $\theta = (\gamma, \lambda)$ where λ is a p-dimensional nuisance parameter. Let $\ell_n(\theta)$ be the log-likelihood function of θ based on n observations. The profile log-likelihood function for γ is then

$$\ell_{n,\mathrm{prof}}(\gamma) = \max\{\ell_n(\theta) \colon a(\theta) = \gamma\} = \ell_n(\gamma, \widehat{\lambda}(\gamma)). \tag{2.3}$$

The second equality shows that we obtain the profile by finding $\widehat{\lambda}(\gamma)$, the maximum likelihood estimator of λ, for each fixed γ value. Profiling is a convenient way of eliminating the nui-

sance parameters and obtaining a function of γ with similar properties to a real log-likelihood function, under some assumptions.

We define the profile deviance as

$$D_n(\gamma) = 2\{\ell_{n,\text{prof}}(\widehat{\gamma}) - \ell_{n,\text{prof}}(\gamma)\}, \tag{2.4}$$

with $\widehat{\gamma}$ the maximum likelihood estimate. Let θ_0 be the true parameter vector with $\gamma_0 = a(\theta_0)$, and assume that θ_0 is an inner point in the parameter space, and that $a(\cdot)$ is a suitably smooth function. Under the model and some regularity conditions described in Schweder & Hjort (2016, Chapter 2), we have that

$$D_n(\gamma_0) = 2\{\ell_{n,\text{prof}}(\widehat{\gamma}) - \ell_{n,\text{prof}}(\gamma_0)\} \to_d \chi^2_1. \tag{2.5}$$

Throughout the thesis, this theorem will be referred to as Wilks' theorem. From this result, it is clear that one may consider the profile deviance an approximate pivot (Reid, 2015), or perhaps more precisely, a large-sample pivot (Schweder & Hjort, 2016). Letting Γ_1 be the cumulative distribution function of a χ^2_1, it follows from (2.5) that

$$\text{cc}(\gamma, y) = \Gamma_1(D_n(\gamma)) \tag{2.6}$$

will tend to a uniform distribution at the true parameter value as the sample size increases, and $\text{cc}(\gamma, y)$ is therefore an approximate confidence curve. Informally, the approximation gets better when there is enough information to accurately estimate the nuisance parameters. When the log-likelihood profile is unimodal, this confidence curve will have a corresponding CD. The result in (2.5) also extends to k-dimensional focus parameters, then with the deviance function tending to a χ^2_k.

Consider again the example in Section 2.2 with n observations from $N(\mu, \sigma^2)$. There we used an exact pivot to construct a confidence distribution for μ, but let us now investigate the result if we had used the profile log-likelihood method introduced in this section instead. Profiling the normal log-likelihood with respect to μ yields

$$\ell_{n,\text{prof}}(\mu) = -\tfrac{1}{2} n \log \left\{ \frac{n-1}{n} \widehat{\sigma}^2 + (\bar{y} - \mu)^2 \right\},$$

with $\widehat{\sigma}^2 = \sum_{i=1}^5 (y_i - \bar{y})^2/(n-1)$. We compute the deviance and find $\text{cc}(\mu)$ using (2.6). The result is shown in Figure 2.2, with the exact confidence curve from Section 2.2 for reference. We see that the profile-based curve gives the same point-estimate as before, but is too narrow compared to the exact one. However, the profile-based confidence curve will quickly approach the exact one as n increases.

The profile method explained above is useful and adaptable to a wide range of situations, including situations where $\gamma = a(\theta)$ is a complicated function of the model parameters. The resulting confidence curve can be non-symmetric and with multiple valleys. In some settings, especially in the context of meta-analysis, one may need to put in additional efforts to ensure that the result is a valid confidence curve. These efforts might consist of modifications to the profile

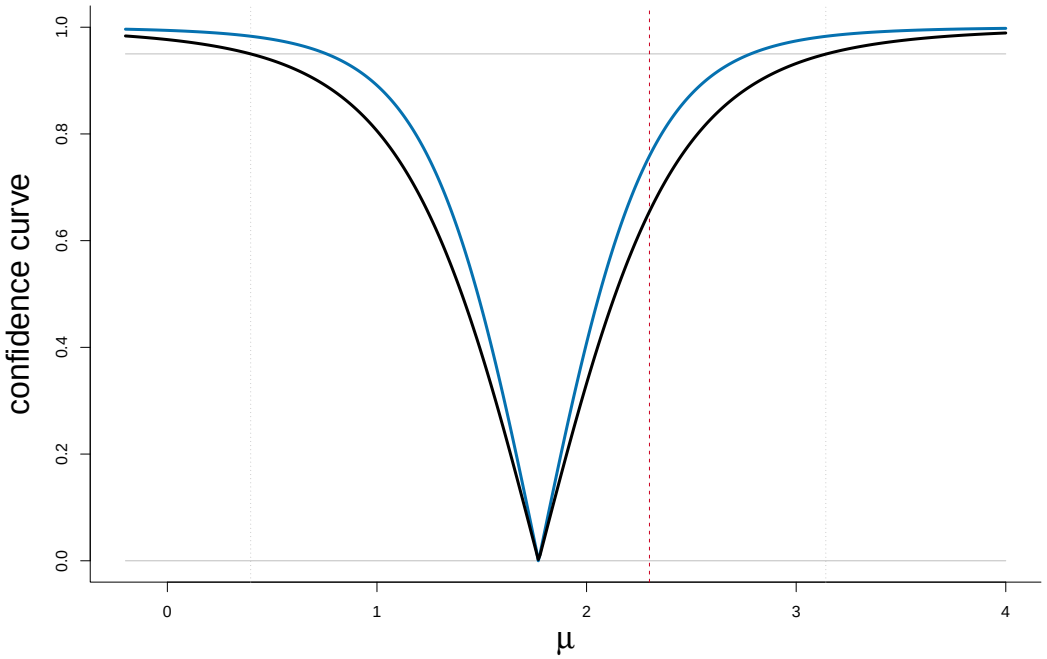

Figure 2.2: *Confidence curves based on a sample of $n = 5$ observations, with $\mu = 2.3$ and $\sigma = 1$. The exact curve in black, and the approximate curve based on the profile log-likelihood in blue. The vertical red line shows the true μ value.*

log-likelihood, which are briefly discussed in Paper IV; see also more general frameworks for improving the distributional approximation in Schweder & Hjort (2016, Chapter 7).

There are situations were the Wilks approximation might be unsuited, but the profile log-likelihood, and profile deviance, still may be fruitful starting points for building confidence curves. See for instance in Paper I where we build confidence curves for a change-point parameter by simulating the distribution of the profile deviance function for each potential change-point value.

There are other general methods for finding approximate pivots to use in the construction of confidence curves. Xie & Singh (2013) and co-authors often make use of the approximate normality of maximum likelihood estimators in regular models (which is closely related to the Wilks approximation above). Given the broadness of the CD definition any procedure which can produce confidence intervals having the correct coverage, either exactly or approximately, can be used to construct CDs, or at least confidence curves. Thus, any distribution over the parameter space having frequentist coverage properties might be called a CD; even if it is constructed by Bayesian or fiducial methods, for instance using the frameworks of generalised fiducial inference (Hannig, 2009).

Finally, how should one decide which CD to use when faced with multiple candidates in a particular application? Tools and concepts to compare CDs for the same parameter are treated in Schweder & Hjort (2016, Chapter 5) and also in Xie & Singh (2013). Informally, a CD is considered superior to another if it is more concentrated around the true value of the parameter. More formally, one can compare the risk functions of different CDs. In Schweder & Hjort (2016, Chapter 5) a theory for loss and risk functions for CDs is developed. Let $C(\gamma, Y)$ be a

confidence distribution and let $\gamma_{\rm cd}$ be a random draw from this distribution. The confidence risk function of $C(\gamma, Y)$, using the squared error loss, is defined as

$$R(C,\gamma) = \mathrm{E}_\gamma \left\{ \int (\gamma_{\rm cd} - \gamma)^2 \ \mathrm{d}C(\gamma_{\rm cd}, Y) \right\}. \tag{2.7}$$

The inner integral evaluates the spread among $\gamma_{\rm cd}$ draws. Then, the outer expectation is taken over the data Y under parameter value γ. The squared error loss may be replaced by other loss functions.

For certain parameters in continuous exponential models, there exists a CD which has a lower risk than all other CDs for any γ value and all convex and non-negative loss functions, see theorems in Schweder & Hjort (2016, Chapter 5). This uniformly optimal confidence distribution for the focus parameter γ exists when the log-likelihood can be written in the form

$$\ell(\gamma,\lambda) = \gamma A + \lambda_1 B_1 + \cdots + \lambda_p B_p - d(\gamma, \lambda_1, \ldots, \lambda_p) + h(y)$$

where A and $B = (B_1, \ldots, B_p)$ are statistics, with observed values $a_{\rm obs}$ and $b_{\rm obs}$, and $\lambda = (\lambda_1, \ldots, \lambda_p)$ denotes the p-dimensional nuisance parameter. Then, the optimal CD for γ is

$$C(\gamma, y) = P_\gamma\{A \geq a_{\rm obs} \,|\, B_1 = b_{1,\rm obs}, \ldots, B_p = b_{p,\rm obs}\}. \tag{2.8}$$

This type of construction is used in some of the examples in Paper IV.

3 Focused model selection

In many applications the analyst will have to choose between several reasonable statistical models. The models can for instance differ in the choice of probability distribution for the unexplained variation or in the set of covariates in a regression setting. Sometimes these models represent competing hypothesis about the phenomenon we are modelling, while at other times we would simply like to have a model which is appropriate for a certain task. In the latter case especially, we might want to use a *focused* model selection approach, which allows us to take the purpose of the data-analysis into account when choosing between models. Paper III concerns focused model selection explicitly, and the theme also appears in Paper II. I will return to model selection with FIC in the discussion, in Section 5.2.

3.1 The FIC idea

Many popular model selection methods consist in finding the model maximising some measure of fit to the data, with a penalty for complexity (or, equivalently, against overfitting). This description encompasses AIC, BIC, some forms of cross-validation and various penalised methods. All these methods are different in terms of motivation and theoretical properties, but share a similar logic.

Maximising the fit to the data, with a trade-off against complexity, is not always the primary concern of the user. Data-analysis is often conducted with a specific goal in mind; this could be obtaining precise estimates of a quantity of interest, or predicting the outcome for a patient with a certain set of covariates; or determining the probability of some variable exceeding a threshold. In this sense, there often is a natural focus parameter, pertaining to the purpose of the analysis. We will continue to denote the one-dimensional focus parameter by γ. In this chapter, the focus γ should have a similar interpretation across all the models under consideration.

Selecting a model which is maximally suited to a certain goal seems like a good principle. In the world of focused model selection, this principle translates to selecting a model which minimises the estimated risk associated with the focus parameter. As usual, the risk is the expected loss. The focused information criterion (FIC) is defined as the estimated risk, a number to be computed for each candidate model. In this thesis, the risk measure will be the mean squared error (MSE), but in parts of the FIC literature other risk measures are occasionally used. The MSE conveniently separates into a variance and squared-bias part,

$$\mathrm{mse}(\widehat{\gamma},\gamma) = \mathrm{E}\{(\widehat{\gamma}-\gamma)^2\} = \mathrm{Var}(\widehat{\gamma}) + \{\mathrm{E}(\widehat{\gamma})-\gamma\}^2.$$

The overall FIC idea consists of estimating the MSE in the different candidate models and rank

them accordingly. For a candidate model denoted by M, the FIC formula will be of the form

$$\mathrm{fic}_M = \widehat{\mathrm{mse}}(\widehat{\gamma}_M) = \widehat{\mathrm{var}}_M + \widehat{\mathrm{bsq}}_M, \tag{3.1}$$

where $\widehat{\gamma}_M$ is the estimated focus parameter using model M; $\widehat{\mathrm{bsq}}_M$ is an estimate of the squared bias and $\widehat{\mathrm{var}}_M$ an estimate of the variance associated with using $\widehat{\gamma}_M$.

Before coming to details about computation of FIC scores, we might have a look at the typical outcome of model selection with FIC. Figure 3.1 is a FIC plot; on the vertical axis we have the estimated focus parameter values and on the horizontal axis we can read off the square-root of the FIC scores. The square-root brings the FIC scores back to the scale of the focus parameter and thereby make them easier to interpret. Here, we have compared six models. We see that the model M_5 has the smallest FIC score and is therefore considered the best model, in the sense that it gives the most precise estimates of the focus parameter. The figure presented here is actually from the illustration in Paper III and the data will be described briefly in Section 4.3.

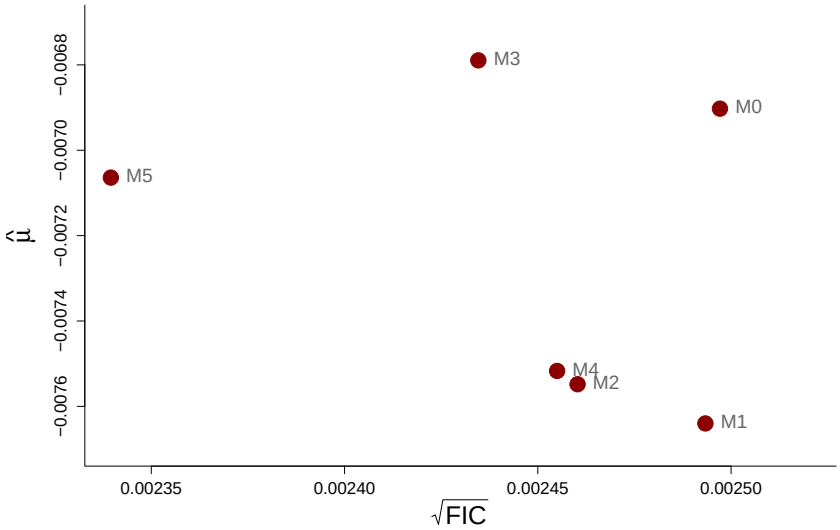

Figure 3.1: *A FIC plot belonging to the illustration in Paper III (see Section 4.3). The scale of measurements is tons. Here we have compared five candidate models against the wide model M_0. Actually, there is also a sixth candidate model in this illustration M_6 which is outside the range of the axes with an estimate of 0 and $\sqrt{\mathrm{fic}}$ score of 0.0064.*

One may identify three different FIC frameworks and these differ in their estimation schemes concerning the quantities bsq_M and var_M in (3.1). Each framework relies on obtaining good formulas for expectations and variances under model misspecification. Model misspecification is central to the FIC idea; because in order to be comparable, the MSE for each candidate model needs to be evaluated with respect to the same model, i.e. the assumed true, but unknown data-generating mechanism. In general situations the FIC formulas rely on large-sample approximations. I will describe the three FIC frameworks briefly in Section 3.3, but for concreteness I will first present a very simple example, where exact risk formulas can be obtained.

3.2 A simple illustration

Assume we are in a simple linear regression setting with n observations of the response variable and two covariates. The true data-generating mechanism is $Y_i = \alpha x_{1,i} + \beta x_{2,i} + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma^2)$. The idea underlying FIC, and which I hope to illustrate in this simple example, is that even though the two-covariate model M_0 is the true model in this case, a different model M_1 may provide more precise estimates of a parameter of interest, at least in parts of the parameter space.

We will assume that α is the focus parameter, and let the candidate model M_1 be $Y_i = \alpha x_{1,i} + \epsilon_i$ and $\epsilon_i \sim N(0, \sigma_1^2)$. Technically, α does not have the same interpretation in M_1 as in the true model M_0, but the smaller model can still serve as an approximation to the truth, producing estimates of α, which typically will be biased, but can have lower variance.

Let x_1 and x_2 be the column vectors of observed covariates and $X = [x_1, x_2]$ be the $n \times 2$ design matrix. The maximum likelihood estimator from the true model, $\widehat{\alpha}_0$, is the first element of $(X^{\mathrm{t}}X)^{-1}X^{\mathrm{t}}Y$, while $\widehat{\alpha}_1 = (x_1^{\mathrm{t}}x_1)^{-1}x_1^{\mathrm{t}}Y$ is the maximum likelihood estimator from the smaller model. The simplicity of the linear regression model allows us to work out exact risk formulas for the focus parameter *under the true model*. Let $\theta = (\alpha, \beta, \sigma)$ be the true parameters in the large model M_0. Again we use the mean squared error,

$$\begin{aligned} \mathrm{mse}(\widehat{\alpha}_0, \theta) &= \{\mathrm{E}_0(\widehat{\alpha}_0) - \alpha\}^2 + \mathrm{Var}_0(\widehat{\alpha}_0) = 0 + \sigma^2(x_1^{\mathrm{t}}x_1)^{-1}(1-\rho_x^2)^{-1} \\ \mathrm{mse}(\widehat{\alpha}_1, \theta) &= \{\mathrm{E}_0(\widehat{\alpha}_1) - \alpha\}^2 + \mathrm{Var}_0(\widehat{\alpha}_1) = \{(x_1^{\mathrm{t}}x_1)^{-1}x_1^{\mathrm{t}}x_2\beta\}^2 + \sigma^2(x_1^{\mathrm{t}}x_1)^{-1} \end{aligned} \tag{3.2}$$

where $\rho_x = (x_1^{\mathrm{t}}x_2)/\sqrt{(x_1^{\mathrm{t}}x_1)(x_2^{\mathrm{t}}x_2)}$, which corresponds to the correlation between the covariates when these are centred around 0. The subscript 0 in E_0 and Var_0 is meant to emphasise that the expectations and variances are taken with respect to the true model M_0. The true model has zero bias, as expected.

With the risk formulas from (3.2) we can investigate which model gives the most precise α estimates in different parts of the θ and covariate space. Here the risk formulas only depend on σ, β and the covariates x_1 and x_2, particularly through ρ_x. If $\beta = 0$, M_1 will have lower risk than M_0 as long as ρ_x is non-zero. With $\rho_x = 0$, it is clear that the two models will give the same variance, and they will actually have exactly equal risk too. To see this, note that the bias of M_1 also depends on ρ_x, $(x_1^{\mathrm{t}}x_1)^{-1}x_1^{\mathrm{t}}x_2\beta = (x_1^{\mathrm{t}}x_1)^{-1}\rho_x\sqrt{(x_1^{\mathrm{t}}x_1)(x_2^{\mathrm{t}}x_2)}\beta$. In Figure 3.2, I display the relative risk $\mathrm{mse}(\widehat{\alpha}_0)/\mathrm{mse}(\widehat{\alpha}_1)$ for $n = 20$, $\sigma = 2$ and a range of β and ρ_x values. For small values of β, the smaller model M_1 can outperform the true model, especially when the correlation between the covariates is large.

The relative risk surface in Figure 3.2 illustrates the general motivation behind model selection with FIC: finding a candidate model which performs better than the true model given a certain purpose. Here the purpose was to estimate α. In practice, the risk functions are unavailable, since they depend on the unknown parameters. When faced with a real dataset, one needs to estimate the risk functions in a specific position of the θ and covariate space by plugging-in the maximum likelihood estimates of the necessary parameters into (3.2). These estimated mean squared errors are the FIC scores of the two models. FIC aims at determining which model has

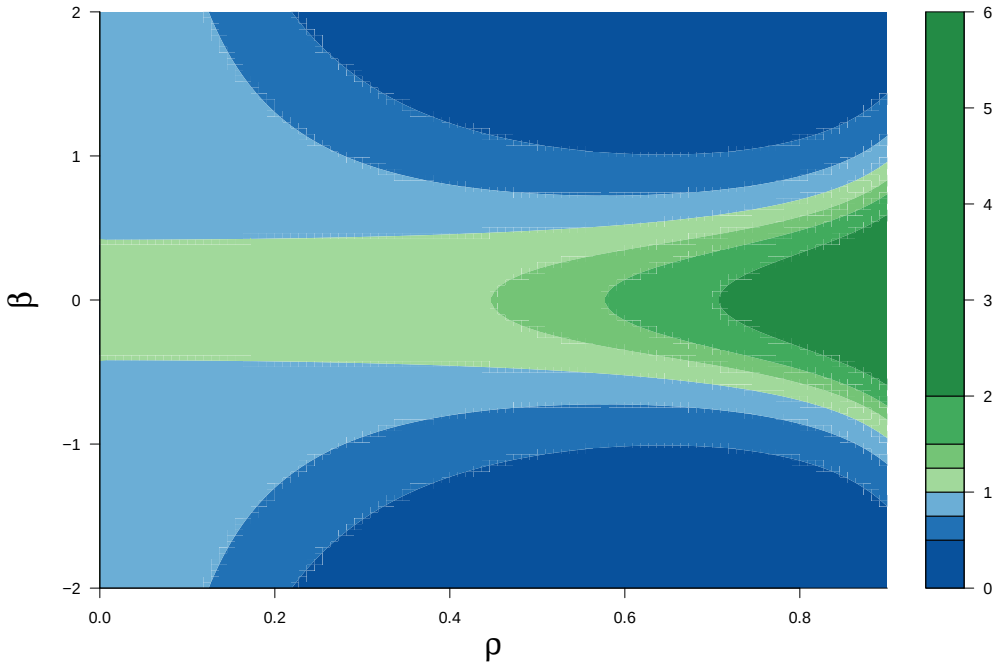

Figure 3.2: *The relative risk* $\mathrm{mse}(\widehat{\alpha}_0)/\mathrm{mse}(\widehat{\alpha}_1)$ *for* $n = 20$ *and different values of* β *and* ρ_x*, the correlation between the covariates. Blue colours indicate that* M_0 *gives the most precise estimate, while green colours indicate that* M_1 *is better. Note that when* $\rho_x = 0$ *they have exactly equal risk.*

the smallest risk in estimating α for a specific dataset.

In simple cases like this one, the risk formulas and their estimates are immediately available from basic statistical knowledge. In general situations, exact formulas are not available and one needs to settle for approximations of the true risk functions, usually obtained through large-sample considerations. The three FIC frameworks described in the next section offer ways to obtain estimated risks in fairly general settings.

3.3 Three FIC frameworks

Three different FIC frameworks may be identified. I will call them FIC type I, type II and type III as there are no canonical names in the literature. Potential names could be "local misspecification"-FIC, "fixed, non-parametric truth"-FIC and "fixed, parametric truth"-FIC. As these names indicate, the frameworks differ in the assumptions made in order to obtain risk approximations and risk estimates, which lead to different FIC formulas. None of the frameworks will be described in detail, and readers actually interested in computing FIC scores should follow the various references. The point here is simply to sketch out the general ideas, and some of the differences between them. For concreteness, consider some candidate model M with an estimator for the focus parameter $\widehat{\gamma}_M$.

The FIC type I framework introduced in Claeskens & Hjort (2003), and developed in Claeskens & Hjort (2008), requires all candidate models to lie within a single family of density functions f, nested between the smallest, narrow, model and the largest, wide, model. The narrow model is parametrised by θ, the wide model has an additional parameter vector ψ. The density functions for some data y are then $f(y, \theta, \psi)$ for the wide model, and $f(y, \theta) = f(y, \theta, \psi_0)$ for the narrow,

where the additional parameter ψ takes a known value $\psi = \psi_0$ in the small model. The focus parameter is a function of both θ and ψ, $\gamma = \gamma(\theta, \psi)$.

As mentioned in Section 3.1, the computation of FIC formulas requires assumptions about the true data-generating mechanism. For FIC type I, the true data-generating density is assumed to be a function of the sample size n,

$$f(y, \theta_{\text{true}}, \psi_0 + \delta/\sqrt{n})$$

with some fixed, but arbitrary θ_{true}. This corresponds to a so-called local misspecification framework. The parameter δ determines the distance from the narrow model. This machinery is not meant to be entirely realistic, but to serve as a useful tool for obtaining FIC formulas. Particularly, it ensures that the squared bias and variance parts are on the same scale, so that neither dominates over the other in the limit.

Inside the local misspecification framework, Claeskens & Hjort (2003) work out the asymptotic distribution $\sqrt{n}(\widehat{\gamma}_M - \gamma_{\text{true}}) \to_d N(b_M, \nu_M)$ with formulas for b_M and ν_M. Here, the true focus parameter is $\gamma_{\text{true}} = \gamma(\theta_{\text{true}}, \psi_0 + \delta/\sqrt{n})$. The formulas for b_M and ν_M lead to estimators for the mean squared error components in (3.1). FIC type I has been adapted to a range of models, notably for variable selection in various regression models, and been used in several spheres of applications.

Recently, a different FIC framework has been proposed; giving a different set of formulas for the MSE components. In Jullum & Hjort (2017) the formulas are derived under the assumption of a fixed true model, with a fixed unknown focus parameter γ_{true}. No parametric assumptions are made about the true data-generating mechanism, and the variances and biases are computed with respect to this unknown distribution. Informally, one could say that the infinitely dimensional non-parametric model is assumed to be the true model. The framework thus requires the existence of a natural non-parametric estimator $\widehat{\gamma}_{\text{np}}$ of the focus, which will be competing against parametric candidate models with estimator $\widehat{\gamma}_M$. In this framework, the candidate models can be from completely different parametric families and do not need to be nested within any wide model. The FIC formulas are obtained from the joint asymptotic distribution of $\widehat{\gamma}_M$ and $\widehat{\gamma}_{\text{np}}$, which the authors derive in a quite general setting,

$$\begin{pmatrix} \sqrt{n}(\widehat{\gamma}_{\text{np}} - \gamma_{\text{true}}) \\ \sqrt{n}(\widehat{\gamma}_M - \gamma_{\text{LF}}) \end{pmatrix} \xrightarrow{d} N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \nu_{\text{np}} & \nu_c \\ \nu_c & \nu_M \end{pmatrix} \right). \tag{3.3}$$

Note that $\widehat{\gamma}_{\text{np}}$ converges to the true focus parameter, but $\widehat{\gamma}_M$ converges to γ_{LF}, the least false parameter, which minimises the Kullback-Leibler divergence from the true model to M. The difference $\gamma_{\text{LF}} - \gamma_{\text{true}}$ represents the bias. Jullum & Hjort (2017) provide estimators for ν_{np}, ν_M and ν_c and FIC formulas follow from these estimators. FIC type II is used in Paper II.

FIC type II differs from FIC type I in making no assumptions about the true data-generating mechanism and in allowing non-nested candidate models. Also, FIC type II allows non-parametric alternatives to be compared with parametric ones. The framework has been adapted to survival analysis (Jullum & Hjort, 2018) and certain types of time-series modelling (Hermansen, Hjort & Jullum, 2015), but is difficult to use in a regression setting when one wishes to select

between models with different sets of covariates (Jullum, 2015).

The third framework could be considered a variant of FIC type II. Instead of assuming a non-parametric truth, we assume that the true data-generating mechanism is a fixed parametric model. This model will be called the *wide model*, because it typically corresponds to the largest candidate model. The obvious benefit of this approach is that it is easily adapted to regression models, unlike FIC type II. Although FIC type III has the potential to be used quite generally, it has not been investigated in full generality yet. In Paper III, we work on FIC type III within the class of linear mixed effect (LME) models. Given that a user has several candidate models within the LME class, and is willing to assume that one of these encompasses the true data-generating mechanism, the FIC approach in Paper III offers a way to select the candidate models giving the most precise estimates of the focus parameter. The FIC formulas are obtained from a joint asymptotic distribution similar to (3.3), but with $\widehat{\gamma}_{\text{np}}$ replaced by $\widehat{\gamma}_{\text{wide}}$ and under the assumption that the wide LME model is true. Naturally, FIC type III could be extended to other model classes as well, and we point to some potential extensions of this type at the end of Paper III.

Unlike FIC type I, FIC type II and III assume a fixed true model, in the sense that the true data-generating mechanism is not a function of n. This entails that in these frameworks the bias part of the FIC formula will dominate the variance part when the sample size is large. In other words, the variances will disappear with n, but not the biases. Therefore, in the limit, models which estimate the focus without bias are assured to win. In practice, this means that FIC type II and III tend to select their assumed true model when n is large, meaning the non-parametric model or the wide model respectively.

4 Summary of papers

4.1 Paper I

Cunen, C., Hermansen, G., and Hjort, N. L. (2018). Confidence distributions for change-points and regime shifts. *Journal of Statistical Planning and Inference* 195, 14–34.

The paper presents, discusses and illustrates two general methods for inference about a single change-point parameter. Assume we have a sequence of observations $y_1, \ldots, y_n$ with the observations $y_1, \ldots, y_\tau$ coming from one distribution, and $y_{\tau+1}, \ldots, y_n$ coming from another. Change-point methodology concerns inference for the position where the distribution changes, τ. In addition to providing a point-estimate for τ, the methods in Paper I also assess the uncertainty in the estimation, which is presented in the form of a confidence curve for τ. The methods are fairly general and can discover changes in any aspect of the distribution, not just changes in for instance the mean level. They can also be adapted to regression situations. The two methods are quite different, both conceptually and practically. The first one, method A, only requires a valid test of homogeneity, which is applied to the sequence of observations on each side of all potential change-points. The choice of test is crucial, and may originate from a parametric model or be fully non-parametric. If a parametric model is assumed, general tests of homogeneity may be derived from certain monitoring bridges. These have also independent interest and are described in the paper.

The second method, method B, is a variant of the general method for constructing confidence curves in (2.6) and requires the user to assume a parametric model, say $f(y_i, \theta)$. In a change-point setting, we have θ_L for $i \leq \tau$ and θ_R for $i \geq \tau + 1$ and the full log-likelihood is $\ell_n(\tau, \theta_L, \theta_R) = \sum_{i \leq \tau} \log f(y_i, \theta_L) + \sum_{i \geq \tau+1} \log f(y_i, \theta_R)$. As in Section 2.3, we will need the log-likelihood profile for τ,

$$\ell_{\rm prof}(\tau) = \max\{\ell_n(\tau, \theta_L, \theta_R) \colon \theta_L, \theta_R\} = \ell_n(\tau, \widehat{\theta}_L(\tau), \widehat{\theta}_R(\tau)),$$

where $\widehat{\theta}_L(\tau)$ and $\widehat{\theta}_R(\tau)$ are the maximum likelihood estimates of θ_L and θ_R for each fixed τ value. We compute the deviance $D_n(\tau, Y)$ as in (2.4), but since τ is a discrete-valued parameter we cannot rely on the Wilks theorem from (2.5). We obtain the cumulative distribution function of $D_n(\tau, Y)$, K_τ, at each position τ by simulations. We then construct the confidence curve by

$$\mathrm{cc}(\tau, y_{\rm obs}) = K_\tau(D_n(\tau, y_{\rm obs})) = P_{\tau, \widehat{\theta}_L, \widehat{\theta}_R}\{D_n(\tau, Y) < D_n(\tau, y_{\rm obs})\}$$

with the maximum likelihood estimates $\widehat{\theta}_L = \widehat{\theta}_L(\widehat{\tau})$ and $\widehat{\theta}_R = \widehat{\theta}_R(\widehat{\tau})$. In practice, the confidence curve is computed by generating a large number of datasets Y, at each position τ, from the assumed model with $\widehat{\theta}_L$ and $\widehat{\theta}_R$. This yields an approximate confidence curve. The non-exactness comes from using the maximum likelihood estimates of the model-parameters, $\widehat{\theta}_L$ and $\widehat{\theta}_R$, in

the simulations instead of the true values. Investigations reveal that the approximation is good as long as there is a reasonable number of observations on each side of the true change-point.

Paper I also provides a method for constructing confidence curves for the degree of change. The function measuring the degree of change must be chosen by the user, and could for instance be the ratio $\gamma = \theta_R/\theta_L$ in the case of one-dimensional model-parameters. The confidence curve for the degree of change takes into account the uncertainty in the change-point estimate, and is constructed by a similar simulation-based method as the one for τ described above.

Method B explicitly assumes that there is a change-point in the sequence of observation. Thus, the method will provide an estimate $\widehat{\tau}$ and a confidence curve $\mathrm{cc}(\tau, y)$ even when there really is no change in the sequence. This issue is only briefly touched in Paper I. A general advice is to use this change-point method only when one has good reasons to believe that there is exactly one change-point, either from prior knowledge or by performing some initial test of homogeneity on the full sequence of observations. Another recommendation is to always compute a confidence curve for the degree of change in addition to $\mathrm{cc}(\tau, y)$. If the confidence curve for the degree of change encompasses values indicating no-change (typically 1 or 0, depending on the choice of function) at most levels, then there is little reasons to believe that the point-estimate indicated by $\mathrm{cc}(\tau, y)$ represents an actual change. Also, our investigations reveal that when the change is small or non-existent, $\mathrm{cc}(\tau, y)$ will typically have very wide confidence intervals at many levels. Intuitively, the level where $\mathrm{cc}(\tau, y)$ covers the entire sequence of observations could be taken as an implicit test of homogeneity. If that level is close to 1, it seems unlikely that the sequence is homogeneous. These last comments are only based on empirical investigations, and should be investigated in a more systematic way in future publications.

The main results in this paper are the two methods themselves, which have potential in real applications, as we demonstrate. The paper differs from parts of the change-point literature by its emphasis on uncertainty considerations. The performance of the methods is briefly studied by way of simulations, some extensions to multiple change-point situations are briefly discussed and four applications are provided. Some of these have significant interest in their own right. One of the applications concerns the medieval chivalry novel *Tirant lo Blanch*, and provides a perhaps rare example of a dataset where we can be certain that there is a change. The novel had two authors; the main author Joanot Martorell died before its completion, and his friend and fellow knight, Martí Joan de Galba finished it. Different aspects of literary style may be collected from the text, for instance concerning the typical word lengths used in each of the 487 chapters of the book. The sequence of word lengths was modelled by a suitable statistical model and we constructed the confidence curve for the change-point. The curve indicated, with surprising precision, that the likely change-point was either around chapter 345 or around chapter 371.

4.2 Paper II

Cunen, C., Hjort, N. L., and Nygård, H. M. (2018). Statistical Sightings of Better Angels: Analysing the Distribution of Battle Deaths in Interstate Conflict over Time. ***Invited to submit a revision to Journal of Peace Research*****.**

Paper II addresses a much-debated question in the field of peace and conflict research: is there statistical evidence for *the long peace* – the period of reduced interstate violence in the time after the second world war? Specifically, we examine the sequence of interstate wars from 1823 till today, focusing on the *war sizes*, the number of battle deaths in each war. Different models for the war size distribution are investigated, with the aim of discovering a change-point in the sequence of war sizes. Methodologically, this paper mainly rests on Paper I, and we make use of method B described in Section 4.1. The paper also introduces new fat-tailed models to the peace research community along with various statistical methods, for instance model selection with FIC and using confidence curves as an inferential summary. We also investigate the use of covariates, specifically democracy scores, in the change-point models for the wars.

The paper represents a contribution in an active debate in the peace research community, and could be understood as a response to the conclusions in Clauset (2017, 2018). There, the author concluded that the period of reduced violence after the second world war could not be considered "statistically significant", i.e. sufficiently unusual compared to the inherent variability in the data. In Paper II, we arrived at slightly different conclusions. When modelling the full battle death distribution, we found a confidence curve pointing to 1950, i.e. the Korean war, as the point-estimate for the year of change. This confidence curve displayed considerable uncertainty, but the accompanying confidence for the degree of change indicated a significant change. The degree of change was defined as the ratio $\phi_{0.75,L}/\phi_{0.75,R}$, where $\phi_{0.75,L}$ denotes the 0.75 quantile of the battle death distribution before the change-point, and $\phi_{0.75,R}$ denotes the same quantity after the change-point.

At any rate, we acknowledge that there remained appreciable uncertainty in our conclusions, but see our contribution as a form of hypothesis generation: instead of focusing on 1945 as the presumed year of change, our change-point methods indicated that the change could have taken place later, specifically in 1950, or even in 1965 (see more details in the paper). These tentative results could warrant further investigation. I will come back to some discussion on this application in Section 5.3.1.

4.3 Paper III

Cunen, C., Walløe, L., and Hjort, N. L. (2018). Focused model selection for linear mixed models, with an application to whale ecology. ***Invited to resubmit to Annals of Applied Statistics*****.**

This paper extends the FIC methodology to the class of linear mixed effect (LME) models using FIC type III. The aim is to choose between different LME models with a specific focus parameter in mind. LME models are common in many fields, for instance ecology, and are particularly useful when the observed data form natural groups and observations within the same group may be correlated. These groups can for instance refer to observations from the same time-unit or from the same spatial location. LMEs are extensions of ordinary linear regression models and take the form

$$y_i \sim N_{m_i}\left(X_i\beta, \sigma^2(I + Z_i D Z_i^{\mathrm{t}})\right) \quad \text{for } i = 1, \ldots, n,$$

with y_i an $m_i \times 1$ vector of responses for the ith group, X_i a known $m_i \times p$ design matrix of covariates, Z_i a known $m_i \times k$ design matrix for the random effects and $\sigma^2 D$ the $k \times k$ covariance matrix for the random effects.

As articulated above, FIC type III requires the user to specify a wide model. In this setting, this could be the largest possible (in terms of parameters) LME model considered plausible with respect to the system under study. The candidate models do not need to be simplifications of the wide model, but can for instance use a different set of covariates in their model matrices, $X_{M,i}$ and $Z_{M,i}$. Here M identifies the candidate model matrices. The focus parameter can be a function of the fixed effect parameters β, or of the variance components (σ, D), or both.

In order to obtain FIC formulas we derived the asymptotic distribution of estimators from LME models under the assumption that the true model is a different LME model. In the appendix of Paper III, we provided explicit formulas for the asymptotic variance and covariance matrices. These formulas were used in the computation of the FIC scores. The performance of the FIC method was evaluated in a simulation study, and some emphasis was placed on an illustration of the method on a real dataset.

The illustration concerns a long-standing debate in the Scientific Committee of the International Whaling Commission (IWC) regarding energy storage in Antarctic Minke whales. The dataset originates from the Japanese Whale Research Program under Special Permit in the Antarctic and consists of various measurements on more then 4000 adult Minke whales. The main research question was whether the body condition of the whales decreased during the 18 years of study (1988 to 2005). Some scientists in the IWC hold that there has been a documented decline in body condition, while others do not agree with this conclusion.

Measurements of energy storage, i.e. blubber thickness, girth and fat weight, were taken as proxies for the body condition of each whale, and used as the response variables in the analyses. Several potential covariates were also recorded, including the year of capture, the date within each year, the sex, the body length and the spatial location. The natural focus parameter was the yearly change in body condition, i.e. the effect of year. In most models, the yearly change

in body condition corresponds to the regression coefficient of the linear effect of year. In Paper III an analysis of this dataset was provided, with fat weight in tons as the response variable. A wide model, denoted by M_0, was assumed and compared against six candidate models. One of the models, M_6, did not contain any effect of year at all. The FIC results were presented in the form of the FIC plot in Figure 3.2. According to FIC, model M_5 provided the most precise estimates of the effect of year. A confidence curve for the focus parameter from the selected model was given. These results are discussed in Section 5.2.4.

In Paper III, the FIC methodology was the principal contribution and the Minke whale question served as an illustration. The same authors have conducted more extensive analyses of the Minke whale data, and these were presented in a report to the Scientific Committee of the IWC in 2017 (Cunen, Walløe & Hjort, 2017). The analysis of the Minke whale dataset has been discussed in the IWC for more than ten years, see various references in Paper III. Some of the central questions in the discussion concerned model selection and the Minke whale data was therefore the motivation behind the derivation of the FIC in Paper III. We will come back to this application in Section 5.3.2.

4.4 Paper IV

Cunen, C., and Hjort, N (2018). Combining information across diverse sources: The II-CC-FF paradigm. *Submitted for publication*.

This paper presents a general framework for combination of information across different, independent sources. The framework could be considered as a generalisation of many existing meta-analysis methods, but it also extends to situations outside the traditional meta-analysis setting and invites the analyst to consider combination problems more broadly.

The framework consists of three general steps, where details may change depending on the model, the data and the purpose of the analysis. We have k independent sources which inform on some parameters $\psi_1, \ldots, \psi_k$. For the sake of this summary, I will let these ψ_js be one-dimensional. Assume we have an overall scalar parameter of main interest, ϕ, which is related to $\psi_1, \ldots, \psi_k$, either via a deterministic function $\phi = \phi(\psi_1, \ldots, \psi_k)$ or through some type of random effect distribution. The first step is the Independent Inspection (II) where the analyst examines the data, y_j, in each source separately and produces a confidence curve $\mathrm{cc}_j(\psi_j, y_j)$ for ψ_j. In what we may call the standard setting, this will involve assuming a parametric model, putting up the log-likelihood function and profile out the parameters that are considered source-specific nuisance parameters. In that way, one obtains a profile log-likelihood $\ell_{\mathrm{prof},j}(\psi_j)$ which ideally contains all the information relating to ϕ from source j. In some situations, the analyst will only have access to non-sufficient summary statistics from each source, for example a point-estimate and a 95% confidence interval. We discuss methods to construct $\mathrm{cc}_j(\psi_j, y_j)$ in that case also.

The next step is the Confidence Conversion (CC), which can be very simple or quite difficult, depending on what methods were used for the construction of $\mathrm{cc}_j(\psi_j, y_j)$ in the first step. The goal here is to convert $\mathrm{cc}_j(\psi_j, y_j)$ to a log-likelihood function $\ell_{c,j}(\psi_j)$. If the profile log-likelihood function was constructed in the first step we simply have $\ell_{c,j}(\psi_j) = \ell_{\mathrm{prof},j}(\psi_j)$, and

no more efforts are required. However, confidence curves can be constructed without making use of likelihood-based methods in the II-step, and then conversion is necessary. We discuss two method aimed at that type of conversion, one which is approximate, but widely applicable.

In the last step, we have to distinguish between situations with fixed and random effects. With random effects, $\psi_1, \ldots, \psi_k$ are assumed to come from some common distribution, for instance we may have that $\psi_j \sim N(\phi, \kappa^2)$, with the overall mean parameter ϕ being the parameter of primary interest. The case of random effects will involve the computation of an integral. The case of fixed effects is easier, then we have that $\phi = \phi(\psi_1, \ldots, \psi_k)$ with some function $\phi(\cdot)$. The log-likelihood contributions from the independent sources are simply summed and the final confidence curve $\mathrm{cc}^*(\phi, \mathrm{data})$ for ϕ is obtained through another round of profiling and typically the use of Wilks' theorem (2.5).

The II-CC-FF procedure can be considered successful if the final confidence curve $\mathrm{cc}^*(\phi, \mathrm{data})$ is valid, i.e. has the correct coverage properties, and benefits from the information in each of the k sources. Paper IV provides guidelines and admonitions to achieve this goal. Particularly, the user should be careful about so-called Neyman-Scott problems, situations where the profile log-likelihood can produce misleading results and where the approximation in (2.6) will not work. The paper also provides a number of illustrations, which illuminate different challenges, solutions and possibilities of the II-CC-FF framework. For instance, one illustration concerns an unusual meta-analysis case where some of the studies have reported continuous summary statistics, while others have reported binary outcomes. Estimating the abundance of humpback whales is treated in another illustration, in order to demonstrate the construction of confidence curves based on simply a point-estimate and a reported confidence interval. The final illustration touches the application in Paper II, and combines the change-point inference from the battle death data with change-point inference from an Ngram.

5 Discussion

A good discussion is meant to offer critical perspectives on the methods and results presented; possible weaknesses in the approaches, room for improvements and further work. Issues particular to the four papers comprising this thesis are treated in the different papers, and I have therefore chosen to restrict this discussion to three general questions. These apply to more than one paper, and also to works outside this thesis. Occasionally I will bring up specific topics from the papers. The first section concerns inference with CDs and coverage properties. The second examines somes aspects of model selection with FIC. The last section deals with a more philosophical issue, the role of statistics in scientific applications.

5.1 Inference with CDs

In this section, I will take a closer look at certain frequentist properties of CDs and their purported influence on the interpretation of CDs as epistemic probabilities. In this context, I will briefly treat some of the differences between Bayesian and frequentist methods and their interpretation, but without any aspirations of doing complete justice to that debate.

In Cox (1958) statistical inference is defined as statements made from observations *with measured uncertainty*. The effort to make a rigorous measurement of uncertainty is one of the prime characteristics of statistical inference in Cox's opinion. As a frequentist, Cox considers it natural to evaluate the uncertainty with respect "to the sample space of observations that might have been obtained", but he later stresses that the sample space should be defined to consist "of observations similar to the observed set, in all respects which do not give a basis for discrimination between the possible values of the unknown parameter of interest". The question on how the uncertainty in statistical statements should be evaluated will be the focus of this section.

First I need to clarify some of the terminology I will use.

- When I refer to a frequentist interpretation of probability I mean the interpretation of probability as the long-run frequency of an event in a large number of trials. This is sometimes contrasted with a subjective interpretation of probability, where probability is a quantification of our uncertainty about the world.
- Frequentist properties are properties of statistical methods which rely on a frequentist interpretation of probability.
- A Bayesian method or procedure is a statistical method whose inferential statements come from a posterior distribution, which depends on the specification of a prior distribution and the use of Bayes' formula. Frequentist methods are "non-Bayesian" methods. Both Bayesian and frequentist method can have good frequentist properties.
- When I write about Bayesians or Bayesian statisticians I refer to proponents of Bayesian

methods (at least part time); these may or may not have a subjective interpretation of probability (also dependent on context).

- Finally, I will discuss epistemic or post-data inferential statements. These are probability statements made about some dataset *after* the data has been observed.

5.1.1 Global coverage

Confidence distributions are defined by their frequentist coverage properties, i.e. the probability statements are tied to the number of times a confidence interval covers the parameter in repeated application of the same procedure. In Section 2.2 we saw that the CD label can be given to any distribution over the parameter space producing confidence intervals which obtain their assigned coverage probabilities over repeated experiments. The concept of coverage has a fairly long statistical history, dating back to Neyman (1934), and it is now almost universally accepted that coverage constitutes an attractive, and many would say necessary, property for a statistical procedure. Bayarri & Berger (2004) for instance claim that "essentially everyone" should agree with what they call the frequentist principle,

> In repeated practical use of a statistical procedure, the long-run average actual accuracy should not be less than (and ideally should equal) the long-run reported accuracy.

Some Bayesians are not convinced and do not require their Bayesian procedures to have exact frequentist coverage (Robert, 2011). Nonetheless, a large portion of statistical publications, both Bayesian and frequentist, devote considerable energy to demonstrations of the coverage properties of their methods, via theorems, or via simulation studies.

Coverage properties are appealing because they are seen as guarantees of average performance over repeated use of the same method, and even over whole scientific communities. If all scientists did their analyses in a correct way, the rate of errors among scientific findings would be controlled. This statement is of course only correct if all modelling assumptions made are valid, but I will bypass that complication here. Coverage properties are tied to the notion of calibration and a method with the correct coverage properties might be called a well-calibrated method (Gustafson & Greenland, 2009).

The coverage property discussed here is sometimes referred to as unconditional (Goutis & Casella, 1995) or *global* (Fraser, 2004). I will use the term global coverage in the following. A CD $C(\gamma, Y)$ has global coverage if

$$P(C(\gamma_0, Y) \leq \alpha) = \alpha, \quad \text{for each } \alpha,$$

and for all possible values of the true parameter γ_0. This is just another way of stating that $C(\gamma_0, Y)$ has a uniform distribution at the true parameter value, as in Definition 2.2.1. This property is global in the sense that it holds with respect to the full, unconditional distribution of Y.

In regular, finite-dimensional models, there is often agreement between Bayesian and frequentist methods when the sample size is large, see for instance Robins & Wasserman (2000). Still,

lack of coverage guarantees lies at the heart of frequentist criticism of Bayesian methods. Fraser (2011) points out that the agreement is only exact (in the small sample sense) for models that are linear in their parameters. When the model is nonlinear, the resulting inference can be misleading from a frequentist point of view and Fraser (2011) therefore asks whether the Bayesian posterior is just "quick and dirty confidence". This statement has been repeated by several proponents of confidence distributions and related methods (Schweder & Hjort, 2016; Xie & Singh, 2013; Martin & Liu, 2015). Bayesian methods can have particularly bad coverage properties for parameters lying near a finite boundary, even for large n (Fraser, 2011; Bayarri & Berger, 2004), and in high, or infinite, dimensional models (Robins & Wasserman, 2000; Hjort et al., 2010). Wasserman (2011) states that Bayesians "have an obligation to study the frequentist properties" of their methods. Many Bayesians agree and acknowledge that global coverage is important (Casella, 1992; Gelman & Hennig, 2017; Bayarri & Berger, 2004).

5.1.2 Conditional coverage

Bayesians accepting the importance of global coverage properties are often quick to point out that they also want good *conditional* coverage properties (Gelman & Hennig, 2017; Casella, 1992; Bayarri & Berger, 2004). Commonly, conditional properties are tied to the non-existence of *relevant* subsets of the sample space. A CD $C(\gamma, Y)$ admits a negatively biased relevant subset A if

$$P(C(\gamma_0, Y) \leq \alpha \,|\, Y \in A) \leq \alpha - \epsilon \tag{5.1}$$

for all possible values of the true parameter γ_0 and some positive number ϵ (Buehler, 1959). Similarly, a positively biased relevant subset A is defined as $P(C(\gamma_0, Y) \leq \alpha \,|\, Y \in A) \geq \alpha + \epsilon$ for all γ_0. The number $\pm\epsilon$ defines the bias in the conditional coverage $P(C(\gamma_0, Y) \leq \alpha \,|\, Y \in A)$ compared with the unconditional coverage level α. Buehler (1959) also introduces the related notion of *semi-relevant* subsets, where $\epsilon = 0$ and with strict inequalities holding for at least some γ_0. Generally, the existence of negatively biased subsets is considered more problematic than positively biased ones, since a CD admitting negatively biased subsets will have intervals with a realised coverage lower than their specified levels, and will then give a too optimistic outlook on the precision of the focus parameter.

The definition in (5.1) implies that the distribution of $C(\gamma_0, Y)$ is no longer uniform when evaluated with respect to the conditional distribution of $(Y \,|\, Y \in A)$. Some authors consider a confidence procedure to have good conditional coverage properties if it admits no relevant subsets, particularly not negatively biased ones. How common are relevant subsets? Note that although many subsets of the sample space may alter the conditional coverage of a procedure, they will not fulfil the definition of relevant subsets because they do not induce a bias of the same sign for all values of γ_0. For example, consider a single observation $X \sim N(\mu, 1)$ with $\mu = 2.2$ and the classical 95% confidence interval for μ, $[X - 1.96, X + 1.96]$. It is clear that $P(X - 1.96 < \mu < X + 1.96 \,|\, X \leq 1)$ will be far below 0.95, but the set $X \leq 1$ is not a relevant subset since the conditional coverage probability can be both larger or smaller than 0.95 depending on the true value of μ. In fact, confidence intervals for the parameter in simple location models admit no relevant subsets, and no semi-relevant subsets either (Buehler, 1959).

Still, relevant subsets are not unusual. The most famous example is the familiar Student's t interval for the expectation from a normal model with unknown variance. In that case the statistic $|\bar{X}|/S$, with $\bar{X}$ the sample mean and S the sample standard deviation, defines a relevant subset of the form $|\bar{X}|/S < k$, see for instance Goutis & Casella (1995). Other examples are treated in for example Cox (1958); Efron & Hinkley (1978); Fraser (2004).

Bayesian procedures using proper priors admit no relevant subsets (Casella, 1992). Bayesian posterior distributions are guaranteed to have good conditional coverage properties, since they always condition on the full dataset (Robins & Wasserman, 2000). At least this statement holds as long as the prior is "true", i.e. θ is really generated from the prior distribution, or if one manages to specify a suitable objective prior (Robinson, 1979; Goutis & Casella, 1995). Frequentist procedures have no such general guarantees, and the existence of relevant subsets needs to be investigated separately for all combinations of procedures and models. Some relevant subsets are tied to ancillary statistics, statistics whose distributions do not depend on the focus parameter, while others are not, like in the Student's t example mentioned above. The search for relevant subsets generated many efforts in the sixties and seventies, see Robinson (1979), and several references in Casella (1992). Note that when a relevant subset is discovered, it remains for the statistician to decide to what extent the subset is considered relevant in the non-technical sense, i.e. meaningful for the particular analysis we are faced with (Efron, 1998; Lehmann & Romano, 2006).

Despite the Bayesian embrace, it is important to keep in mind that conditional coverage as it is stated here is an inherently frequentist concept (Bayarri & Berger, 2004), which hinges on a frequentist interpretation of probability. In order to highlight the distinction between global and conditional coverage I will present a simple example in the next section. Variants of this example have been treated many places in the literature (Morey et al., 2016; Fraser, 2004; Goutis & Casella, 1995; Berger & Wolpert, 1988). The point of the example is to show that the global, or average, performance of a confidence procedure can hide poor performance in important subsets of the sample space (Cox, 1958; Fraser, 2004; Liu & Meng, 2016).

5.1.3 A simple illustration

Assume we have two independent observations Y_1 and Y_2 from $U(\theta - 5, \theta + 5)$ and we are interested in constructing a confidence curve for θ. I will present two different alternatives.

The first procedure considers the quantity $\bar{Y} - \theta$. This quantity has a triangular distribution on $(-5, 5)$, with cumulative distribution function denoted by G. The triangular distribution is independent of θ, hence $\bar{Y} - \theta$ is a pivot, and

$$C_1(\theta \,|\, \bar{Y}) = 1 - G(\bar{Y} - \theta) = G(\theta - \bar{Y})$$

is a valid confidence distribution for θ. The second procedure is the Bayesian posterior using a flat prior for θ. The Bayesian posterior is a uniform distribution on $(\bar{Y} - \{5 - \frac{1}{2}D\}, \bar{Y} + \{5 - \frac{1}{2}D\})$, with $D = |Y_1 - Y_2|$ and with cumulative distribution function F. The posterior can be

presented as a confidence distribution,

$$C_2(\theta \,|\, \bar{Y}, D) = F(\theta \,|\, \bar{Y}, D).$$

Separate investigations confirm that this is a valid confidence distribution for θ. This CD can also be obtained from fiducial arguments, see Welch (1939), or by deriving the conditional distribution of $(\bar{Y} \,|\, R)$ with $R = (Y_1 - Y_2)/2$ (Fraser, 2004). Intuitively, the resulting $a\%$ confidence intervals correspond to taking the central $a\%$ of the flat likelihood for θ, see further explanations in Morey et al. (2016).

We present both CDs in the form of confidence curves. In Figure 5.1 we observe the behaviour of these two confidence curves in two different situations. In the left panel, we have two observations with a small distance between them, and in that case both confidence curves display considerable uncertainty. Here the Bayesian curve is a bit wider than the curve based on the simple pivot $\bar{Y} - \theta$. In the right panel, however, we have two distant observations, and the Bayesian confidence curve is much narrower than in the left panel, while the red curve is just as wide.

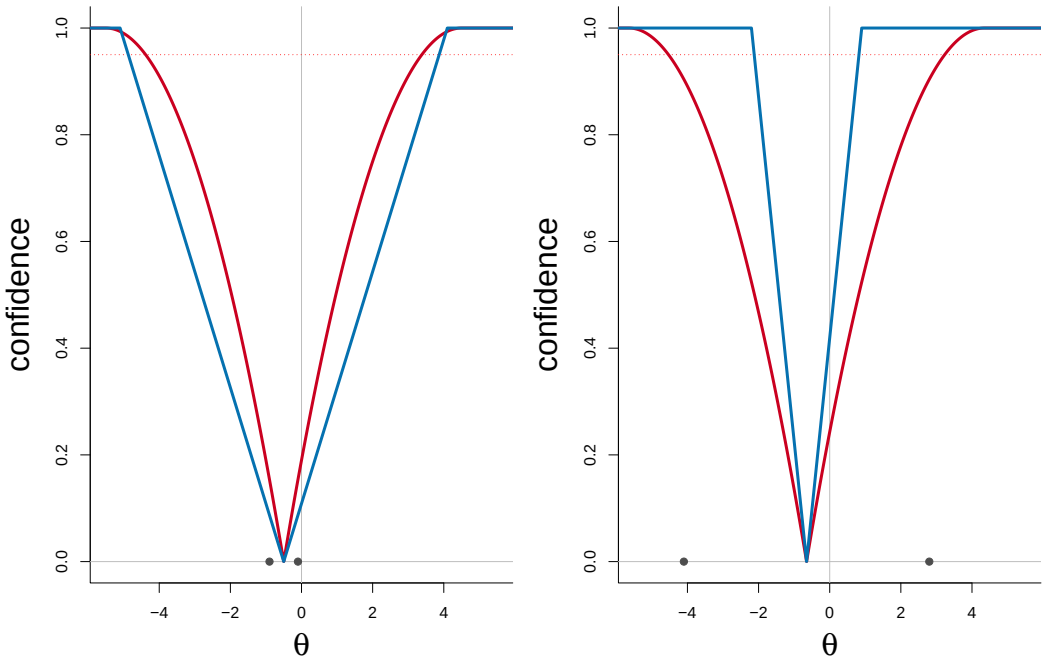

Figure 5.1: *In red the confidence curve based on the $\bar{Y} - \theta$ pivot, in blue the fiducial/Bayesian one. The two observations are displayed as black points at the bottom of each panel. Left panel: two close observations; right panel: two distant observations.*

Let us consider the right panel a bit further; where we observed $y_1 = -4.1$ and $y_2 = 2.8$. A bit of reasoning tells us that the true value of θ must be within ± 5 of each of these observations, therefore we can deduce that a 100% interval must be $[-2.2, 0.9]$ and this is exactly the 100% interval given by the blue curve. The red curve, however, only assigns about 0.50 confidence to the same interval, which is clearly not a sensible confidence assignment.

Both curves are valid CDs with global coverage properties, but only the blue curve has good conditional coverage properties. For instance if we condition on the distance between the

observation being less than five, it is easy to establish (by simulations or calculations) that $P(C_2(\theta_0) \leq \alpha \,|\, D < 5) = \alpha$, but $P(C_1(\theta_0) \leq \alpha \,|\, D < 5) < \alpha$, indicating that the red curve will be too narrow when the distance between the observations is small. Indeed, $D \leq k$ defines a negatively biased relevant subset for C_1, by (5.1), see Goutis & Casella (1995) for details. Similarly, $D \geq k$ defines a positively biased relevant subset for C_1, meaning that the red curve will be unnecessarily wide when the distance between the points is large, just as we see in the right panel of Figure 5.1.

This example is in some sense artificial. After some considerations, all reasonable statisticians will agree that the information in D, the distance between the observations, is somehow relevant to θ, and that the confidence curve procedure C_1 is worse than C_2 because C_1 fails to benefit from the information in D. More precisely, D is an ancillary statistic to θ, and there is an established literature treating the merits of conditioning on ancillary statistics (Fraser, 2004). Unfortunately, in more realistic problems, ancillary statistics may not exist, or there may be multiple ancillaries, and then it is not clear which characteristics of the data one should condition on (Goutis & Casella, 1995).

Moreover, this example seems problematic for the CD framework. Both C_1 and C_2 are valid CDs, but one seems clearly superior to the other. Tools and concepts to compare CDs for the same parameter were briefly treated in Section 2.3. Informally, a CD is considered superior to another if it is more concentrated around the true value of the parameter. In this case, both curves are equally concentrated around the truth when we average over the distribution of the data. To appreciate this, one can derive the confidence risk functions for both CDs, see Section 2.3. With squared error loss, both risk functions are constant with respect to θ and equal, $R(C_1, \theta) = R(C_2, \theta) = 100/12$. In order to differentiate between C_1 and C_2, we may have to explore the framework of conditional risk functions, see Berger (1985).

Finally, many authors would argue that only one of the two CDs considered here, C_2, can be said to offer valid post-data inference (Morey et al., 2016; Goutis & Casella, 1995), a concept that I will discuss presently.

5.1.4 Post-data inference

Good conditional properties are sought in order to avoid counter-intuitive results like in the example above. More generally, these properties are tied to the ability to make post-data probability statements. Let us consider a 95% confidence interval for some parameter γ, for instance $[2.1, 3.4]$, constructed with a frequentist procedure. According to the classical frequentist tradition, the probability 0.95 is *only* meaningful pre-data. Before the data have been collected and the interval computed we can say that the procedure will produce an interval covering the true value with probability 0.95. After the interval has been computed, the classical frequentist knows nothing of the probability that $[2.1, 3.4]$ covers the truth; it either does or does not. A post-data statement on the other hand would be "$[2.1, 3.4]$ covers the true parameter value with 0.95 probability", or (possibly even stronger) "the true parameter value lies within $[2.1, 3.4]$ with 0.95 probability". These kind of statements have traditionally been reserved to Bayesians.

Most introductory courses in statistics teach classical frequentists methods and often insist that

the all inferential statements should be pre-data statements. Despite the insistence on this interpretation there is widespread confusion concerning the correct interpretation of for instance confidence intervals, even among those assigning themselves to the frequentist school (Morey et al., 2016). The pre-data interpretation is perceived as unintuitive, and has led some frequentists to argue that we should allow post-data, or epistemic interpretations of probability statements from frequentist methods (Schweder, 2018). In some sense, one could argue that epistemic statements are subjective and that one should therefore be allowed to tie them to reasonable intervals both of frequentist and Bayesian origin if one wishes to. Nonetheless, parts of the literature explicitly link post-data interpretation to acceptable conditional behaviour (Cox, 1958; Fisher, 1956; Reid, 1995; Efron, 1998; Cox, 1998). These authors typically define acceptable conditional behaviour in a less formal way than the relevant subset framework described in Section 5.1.2. Still, they share the same general insight that a procedure for post-data inference should not experience widely different coverage probabilities in important subsets of the sample space.

The link between post-data interpretation and acceptable conditional behaviour is clearly expressed in Casella (1992),

> Inference made conditional on the data must, necessarily, connect statement about the unknown parameters to the data actually observed. This fact separates conditional confidence inference from unconditional, or pre-data, confidence inference.

The general intuition is illustrated in the example in Section 5.1.3, only the blue curve can be considered to provide post-data inference, since it is the only one that takes into account the observed distance between the observations. Casella (1992) considers relevant subsets as defined in (5.1) to be a framework for the assessment of post-data validity.

Different authors differ in whether they consider good conditional properties as more or less important than global coverage properties. On a related note, Liu & Meng (2016) argue that the whole Bayesian versus frequentist discussion is best understood as differences in the emphasis placed on relevance versus robustness. Subjective Bayesian inference is considered maximally relevant to the observed data since it conditions on the full dataset, but has little robustness against misspecification of its prior. Most frequentist method on the other hand, are considered robust, but may have little relevance to the particular data, as exemplified by the red confidence curve in Figure 5.1. Lack of relevance in the sense of Liu & Meng (2016) is closely related to poor conditional coverage properties. The authors define an axis between relevance and robustness and place various statistical frameworks along this axis. Around the middle of the axis they place conditional frequentists, which explicitly condition on some aspects of the data, usually through ancillary statistics, or asymptotic ancillaries, see for instance Reid (1995) and Fraser (2004). Also near the middle, we find fiducial methods, and confidence distributions.

To what extent can we claim that CDs constitute a compromise between relevance and robustness as defined by Liu & Meng (2016)? CDs have historical ties to fiducial methods, and are sometimes understood to represent epistemic probabilities (Schweder & Hjort, 2016; Schweder, 2018), but CDs are defined solely through their global coverage properties, and there is therefore nothing in the CD definition which guarantees relevance to the particular data, in the sense of Liu & Meng (2016). However, some of the methods commonly used in order to construct

CDs are conditional frequentist methods. Particularly, the optimal CD for exponential families described in (2.8), is derived through conditioning arguments. In Paper IV that method is used to construct CDs for the log-odds ratios in 2×2 tables. There, the CD is constructed based on the conditional distribution of the number of successes in the treatment group given the total number of successes. Thus, these types of CDs will have good conditional coverage properties with respect to subsets of the sample space defined by the total number of successes. Our default method for constructing confidence curves described in (2.6) has no general conditional coverage guarantees, but there is an extensive literature exploring connections between second-order corrections of likelihood-based methods and conditional guarantees with respect to ancillary and asymptotically ancillary statistics, see for instance Severini (1990) and DiCiccio et al. (2015). Future development of the CD methodology could benefit from exploring conditional properties, particularly in connection with higher-order asymptotics of likelihood methods (Reid, 1995; Efron, 1998).

Finally, while the lack of conditional coverage guarantees might seem disheartening for frequentist methods in general, and CDs in particular, it is important to note that there is no agreement in the literature concerning the appropriate degree of conditioning, nor on what statistics one should condition on if one aims at some partial conditioning (rather than the full conditioning of Bayesian methods). Some authors even argue against conditioning, in general (Welch, 1939) or in particular cases (Robins & Wasserman, 2000). At any rate, as mentioned in Section 5.1.1, Bayesian and frequentist method are often in reasonable agreement.

5.2 Model selection with FIC

The motivation behind FIC, to select the model best suited to a certain task, is admirable and natural. In the application in Paper III the particular task was very clear: to estimate the yearly change in fat weight as precisely as possible. In this section I will explore whether that task was fulfilled. Before investigating this question I will present some general issues that can be raised about FIC methods. Some of these issues are specific to the FIC type III that we have used, but most apply to the FIC idea more generally. I begin with some comments concerning the interpretation of the model selected by FIC. Then, I discuss challenges concerning inference after a model has been selected by FIC. Some of these challenges are illustrated in the context of the application from Paper III.

5.2.1 Interpreting FIC

Model selection constitutes a large and diverse sub-field within Statistics, with an abundance of different methods which aim at solving a number of different problems. Model selection methods are sometimes placed along an axis based on the overall goal of the data-analysis: "causal" explanations of a phenomenon or empirical prediction (Shmueli, 2010). In general, model selection methods suited to the latter goal are aiming at maximising predictive performance on future data. Methods suited to the former goal, however, aim at finding the true data-generating mechanism (BIC), or at least choosing a statistical model which is as close as possible to this truth (AIC).

It is not apparent where to place FIC along this explain-predict axis. Naturally, that would depend on the chosen focus parameter. In the applications I have been involved with, for example in Paper III, the primary goal has not been prediction, but to provide understanding and knowledge about the system under study. In the rest of this discussion I will therefore discuss FIC in this type of "explain" setting, but bear in mind that FIC may be used for prediction tasks as well.

Unlike other model selection methods used when explanations are the goal, FIC does not aim at coming close to the true data-generating mechanism, but only at finding a model which successfully replicates certain aspects of the truth; the ones pertaining to the focus parameter. Despite being clearly stated, this feature of FIC is sometimes a source of confusion. For example in the discussions in the IWC concerning the FIC methodology in Paper III, there were comments along the line "the selected model is biologically implausible". When FIC selects a model with fewer predictor variables than the wide model, this does not mean that the discarded predictors should be understood as unimportant for the phenomenon we are studying; it only means that these variables are not deemed necessary when it comes to estimation of the focus parameter.

For FIC type III, the candidate models do not necessarily need to be considered plausible, even if we are in a situation where we are interested in explaining. While the wide model should be given a sound motivation (since it is assumed to be true), the candidate models can be understood as "working models" or estimator-producing machines, which are evaluated with respect to their ability to mimic relevant aspects of the wide model. In fact, all three FIC frameworks may primarily be considered as methods for *estimator selection*, rather than model selection.

In this way, FIC could be considered as having a less ambitious goal than model selection methods aiming at finding the "true model". This might sound like a drawback, but in some situations it could be deemed advantageous, realistic and pragmatic (Wit, Heuvel & Romeijn, 2012). Some authors advice against the use of model selection aiming at the "true model", especially in situations where there is substantive domain knowledge, see Gelman & Rubin (1995). The authors encourage spending efforts to build realistic models and criticise model selection, using BIC specifically, for "promising the impossible". FIC, of type III particularly, requires the users to spend energy on building and validating the wide model. If the user has a well-motivated, possibly causal, model in mind, FIC type III can help the user find an estimator which will give more precise estimates of the quantity of interest.

5.2.2 Inference after FIC

When applying model selection methods in settings where explanations or causal relationships are of interest, the user often wants more than just a model and a point-estimate. After having selected a suitable model, the user often wants to perform hypothesis tests or for instance compute a full confidence curve for the focus parameter. In this subsection and the next, I discuss two issues concerning inference after model selection. The first one is particular to model selection with FIC: under which model should the inference after model selection be conducted?

Assume we have a focus parameter γ, and that we used FIC, which selected the candidate model

M^*. Further, assume that we are in a setting where the estimator $\widehat{\gamma}_{M^*}$ is approximately normal. Then, the construction of a confidence curve for γ using the selected model M^*, requires an estimate of the variance of $\widehat{\gamma}_{M^*}$. After model selection with FIC, we have the choice between two different variances, $\mathrm{Var}_{M^*}(\widehat{\gamma}_{M^*})$ and $\mathrm{Var}_{\mathrm{true}}(\widehat{\gamma}_{M^*})$. The first one is the variance of $\widehat{\gamma}_{M^*}$ evaluated under the selected model M^*. Estimates of this quantify are immediately available after fitting model M^*. The alternative choice is the variance of $\widehat{\gamma}_{M^*}$ evaluated under the assumed true data-generating mechanism. When using FIC type III this is the wide model, and estimates of this quantity are available via the FIC formulas. In many situations, for example for the LME models in Paper III, these two alternatives are considerably different.

For other information criteria, like AIC or BIC, this issue does not arise, since no true model is assumed in these frameworks and $\mathrm{Var}_{M^*}(\widehat{\gamma}_{M^*})$ is therefore the natural choice. After model selection with FIC however, that choice seems a bit inconsistent with the general FIC narrative where the candidate models are primarily meant as estimator-producing machines and do not reflect our assumptions about reality. In all three FIC frameworks, it therefore seems to me that the natural choice is to use the estimated $\mathrm{Var}_{\mathrm{true}}(\widehat{\gamma}_{M^*})$ to compute for instance confidence curves after model selection. Thus, the confidence curves are computed under the assumed true model, in FIC type III that would be the wide model, in FIC type I that is the n dependent true model, in FIC type II it is the non-parametric model.

5.2.3 Post-selection issues

There are other more general issues concerning inference after model selection. The problem of post-selection inference is not specific to FIC methods, and applies to all model selection frameworks.

In the previous subsection I discussed how to compute confidence curves after model selection with FIC. Usually, the same data are used in both the model selection and the inference step, which entails that the resulting confidence curves will be invalid, i.e. they will not obtain correct coverage probabilities (Claeskens & Hjort, 2008, Chapter 7). Specifically, the confidence curves will be too narrow, giving a too optimistic outlook on the actual precision. When we compute a confidence curve after the model M has been selected, we implicitly condition on the event that M was selected. The confidence curve is therefore only valid conditional on this event, but in repeated applications there can be considerable variation concerning which model is selected, and the final confidence curve will therefore not be valid unconditionally. In other words, the uncertainty in the model selection step is not accounted for in the computation of the confidence curve.

The post-selection estimator takes into account the uncertainty in the model-selection. Sometimes it is denoted by $\widehat{\gamma}_{\widehat{M}}$ to highlight that the selected model $\widehat{M}$ is random and simply an estimate of the truly most suitable model (according to the criterion used). This post-selection estimator has a very complicated distribution, typically far from normal and with much larger variance than the estimator which conditions on a particular model being selected (Claeskens & Hjort, 2008, Chapter 7).

This phenomenon has been known for a long time, see for instance Hurvich & Tsai (1990);

Claeskens & Hjort (2003); Leeb & Pötscher (2005). The failure to take into account the uncertainty in the model selection is sometimes referred to as the "quiet scandal of statistics" (Breiman, 1992). Despite being well-known, the issue is often disregarded in practice; partly because it is not routinely taught in basic statistics courses, and also because it is difficult to construct corrected confidence intervals (Leeb & Pötscher, 2005). There are some efforts in that direction in for instance Claeskens & Hjort (2008, Chapter 7), see also Lee et al. (2016). The problem can be bypassed by splitting up the available data and use one part for model selection and the other part for inference (Hurvich & Tsai, 1990). This solution entails a considerable loss of power and the final confidence curves can be considered conservative.

Inference after model selection with FIC suffers from the same post-selection issue as with other criteria, see an illustration in the next section. In our IWC report on the Minke whale dataset (Cunen, Walløe & Hjort, 2017), we split the dataset in two in order to avoid the issue, but in that case we had a large dataset.

5.2.4 Revisiting the illustration in Paper III: Did we gain from FIC?

In light of the discussion above, I revisit the illustration in Paper III. By bootstrapping the full FIC selection procedure, one can investigate the uncertainty in the selected model, which was M_5 (see the FIC plot in Section 3.1), and assess the realised coverage of the final confidence curve.

Specifically, I generated 1000 new datasets from the fitted wide model and ran the full FIC selection procedure on these datasets. From Figure 5.2 it is clear that there is considerable variation in the FIC scores. There is nonetheless some stability in the ranking of the seven models which is not apparent from the figure. The winning model M_5 was selected in 36% of the simulations runs, the most of all the seven models. The second most selected model was M_1 in 22% of the rounds. The wide model was selected 8% of the time. The model with the lowest FIC score, the one without year effect M_6, was only selected in 0.6% of the rounds.

In order to assess the realised coverage of the final confidence curve, I computed a confidence curve for the focus parameter in the selected model *under* the wide model, in each of the simulation runs. Since the data were generated from a known truth, I could calculate the realised coverage probabilities for intervals of various levels. In this case, the realised coverage of the 95% interval was only 61%. This indicates that the final confidence curve, displayed in black in Figure 5.3, is seriously affected by post-selection issues and clearly far too narrow.

To what extent did we actually gain something by using FIC in this illustration? First of all, but rather trivially, we gained an estimate of the year effect which under some assumptions can be considered the most precise estimate of this effect among those considered. In this case the estimate from the winning model M_5 was -0.0071 tons while the original estimate was -0.0069 tons.

Secondly, we gained some knowledge about which covariates were necessary to include in a model when seeking to estimate the year effect. The winning model M_5 was considerably smaller than the wide model M_0.

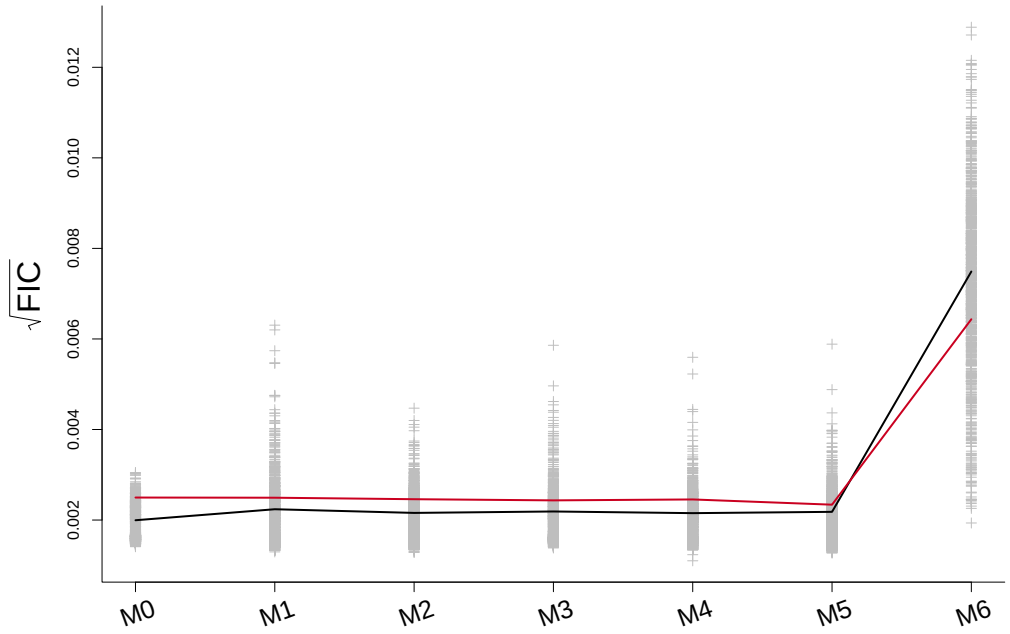

Figure 5.2: *The red lines are the root-FIC values in the original dataset, the grey crosses are the root-FIC scores on 1000 simulated datasets and the black lines are the average root-FIC scores in the simulated datasets.*

Furthermore, we wished to do inference after the model selection step. Specifically, we computed a confidence curve for the focus parameter. In Figure 3.1, the confidence curve for μ from M_5 *under* the wide model is displayed (black line), along with the confidence curve for μ from the wide model directly, without any model selection step (blue line). Apparently, for this example, the increase in precision using model M_5 is very slight. Also, due to the post-selection issues discussed above, we know the confidence curve after the model selection step actually should have been wider in order to accommodate for the uncertainty in the model selection step. Clearly, almost any widening of the curve will cancel out the increased precision due to the FIC procedure. We could of course have split the dataset in two to avoid the post-selection issue, but then the final confidence curve would in all likelihood be considerably wider due to the reduced sample size.

Therefore, when it comes to the computation of confidence curves, we cannot say that we have gained anything from the FIC procedure in this case. Of course the situation could be different for another dataset. It is conceivable that situations exist where the winning model is so dominant compared to the wide model that the increase in precision in the selected model outweighs the increased variance due to post-selection issues.

Finally, FIC also served a role as an implicit test of the year effect. There is an intimate connections between many model selection methods and testing, see Leeb & Pötscher (2005). In this way, model selection is a way to decide between rival theories. Here the two theories were that there is an effect of year (in models 0 to 5) and there is no effect of year (in model 6). Let us consider only model M_0 and M_6, the FIC procedure can be seen as a test of the null hypothesis that M_6 offers an adequate estimator for the focus parameter, i.e. that there is no effect of year.

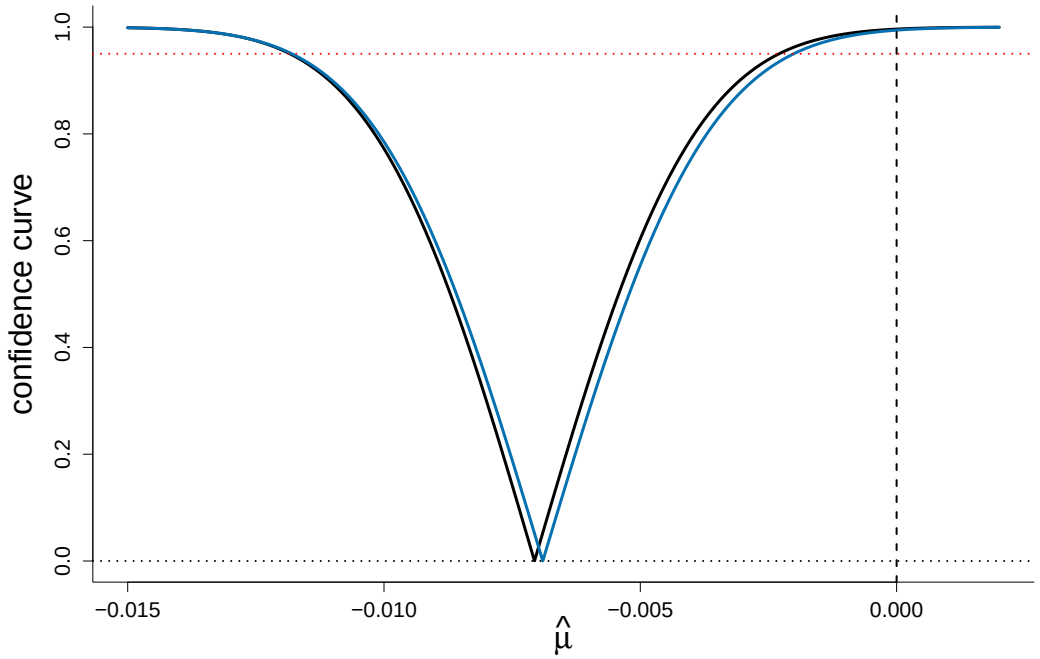

Figure 5.3: *Confidence curves for the year effect (see Paper III). In blue, the curve from the wide model M_0, in black the curve from the model selected by FIC, M_5 (computed under the wide model). There is a very slight increase in precision when using M_5 compared with M_0.*

The implied test level is then (using notation from Paper III),

$$\begin{aligned}\alpha_n &= P(\text{selecting } M_0 \text{ when } M_6 \text{ is correct abour the effect of year})\\ &= P(\text{fic}_0 < \text{fic}_6 \,|\, \mu = 0)\\ &= P(\widehat{\nu}_0/n < (0-\widehat{\mu}_0)^2 - \widehat{\nu}_0/n \,|\, \mu = 0)\\ &= P(|\widehat{\mu}_0/\sqrt{\widehat{\nu}_0/n}| > \sqrt{2} \,|\, \mu = 0).\end{aligned}$$

Unsurprisingly, in the case of LME models, the test actually corresponds to the ordinary Wald type test of the effect of year in the wide model. When n is large this probability approaches the probability of a absolute value of a standard normal exceeding $\sqrt{2}$ which is 0.157. This number is then a fixed test level given by the FIC theory, see Jullum & Hjort (2017).

5.3 Statistics and scientific discovery

Statistics is a field in its own right, with its own joys and challenges. Nevertheless, it is important to remember that the sciences, and applications in general, are the *raisons d'être* of statistics. In the two first sections of this discussion, I treated some issues concerning inference with confidence curves and model selection with FIC and these discussions were kept within the abstract assumptions in which statisticians usually operate. This is the "theoretical world" (in the words of Kass (2011)), the world of statistical models, random variables and confidence intervals. Ultimately, it is not sufficient to examine the methods and tools we develop *within* this theoretical world, we should also strive to examine the extent to which we have answered a scientific, or real world, question. Do our statistical statements translate into scientific dis-

coveries? In this final section, I will first provide a brief overview of some statistical literature on this topic, before I examine the two main applications which are included in this thesis: the wars and the whales.

First, I need to address the term "scientific discovery". Here, I will simply use it to denote any true, non-trivial scientific finding. This is a significantly less ambitious definition than some authors, which reserve "scientific discovery" for genuine breakthroughs; paradigm-shifts or "revolutionary science" in the words of Kuhn (as referred to by Lehmann (1990)). The question on whether statistics has anything to offer "revolutionary science" is addressed in Nelder (1986) and Lehmann (1990), and will not be treated here. I will content myself with discussing the role of statistics in the more everyday scientific activities of incremental advances within the current paradigm of a given field.

Cox (1958) briefly discusses the relationship between statistical inference and scientific conclusions. He acknowledges that statistics play an important role in the process from observations to knowledge, but highlights that the final step from a statistical statement to a scientific conclusion requires additional input, typically in the form of domain knowledge. Importantly, he states

> the statistical uncertainty is only a part, sometimes small, of the uncertainty of the final inference.

To most statisticians these statements might seem self-evident, and completely consistent with the practice of our field. There are many reasons to believe that these notions have not been successfully communicated to the scientific community, however. I will elaborate on this point in the next paragraph, but first I will make a short comment on the data-generating mechanism in Cox (1958). The line of arguments in that paper concerns the traditional story of statistical inference where a random sample has been drawn from an actual population. Most classical (or otherwise) methods were developed within this framework, but these days statistics is involved with many applications where the data are not obtained in this fashion. For example in Paper II, we model the number of battle deaths in wars using probability distributions, but we certainly do not believe that our 95 wars are a random sample of wars from a larger population. In cases like these, the statistical model is not meant to be a description of the real data-generating mechanism, but simply a tool for uncertainty quantification of our statements (Liu & Meng, 2016). But what exactly do these uncertainty measurement mean in the absence of actual randomness? According to Kass (2011), we are in that case required to make a "leap of faith", which might be feel a bit disappointing, but is very useful for making statistical reasoning possible in many applications.

In the quote above, Cox (1958) warns against equating statistical and scientific uncertainty, but this is a widespread practice in many fields. Often, scientists only express the statistical uncertainty, i.e. the uncertainties obtained within their statistical models (Greenland, 2017). Naturally, the other uncertainties, primarily relating to establishing a link between the real world question and the statistical framework, are much harder to express, at least in a formal way (although model validation and critique are examples of such efforts). A typical illustration of the apparent muddling of statistical and scientific uncertainty is the widespread practice of stating a real-world conclusion once a p-value smaller than 0.05 has been obtained, or when

the 95% confidence interval does not contain some value of interest. This is also common among statisticians, and there are examples in this thesis too. Of course statisticians know that these statements are conditional on all the modelling assumptions, but they will often be taken as truths outside statistics, or at least communicated as such. First rather carefully in the context of a scientific publication, then much more forcefully, and often distorted, by media and communications offices at one's university.

In this light, statistical methods are used in the proliferation of false findings, and thus play a role in the much discussed replication crisis in science (Fraser & Reid, 2016). On the other hand, there are also instances where the statistical uncertainty is unnecessarily enlarged in order to obtain some non-significant results, see Greenland (2017) and Schweder (2001). In both cases, statistics is used, either unintentionally or intentionally, to obtain the answer the analyst wants. These and other problematic aspects have led some statisticians to the conclusion that the field of statistics, in its current form, might be harmful to good scientific practice (Tukey, 1962; Box, 1990; Nelder, 1999; Kass, 2011; Greenland, 2017; Gelman, 2018). These authors offer partly overlapping explanations. Some of the authors state that the field focuses too much on the mathematical aspects of statistics and too little on applications, for instance in Box (1990) and Nelder (1999). Tukey (1962) expresses a similar sentiment when he states that contributions in statistics should either be justified through their utility for data-analysis, or as pieces of pure mathematics. It is implied that most contributions fall somewhere in between.

Other explanations concern deficiencies in communication and teaching. In the opinion of Kass (2011), statistical courses do not place enough emphasis on the role of theoretical assumptions and understanding their match or mis-match with the real world. Greenland (2017) and Nelder (1999) criticise the typical courses in applied statistics for scientists, which focus on significance testing, p-values and other more-or-less automatic procedures which hinder critical thinking. Finally, statistical methods seem sophisticated and difficult and therefore transmit a false sense of confidence to their users and to the readers of scientific papers (Greenland, 2017; Donoho, 2017).

None of the authors propose that statistical methods should be abandoned, but rather call for changes in statistical practice, communication and teaching. Primarily they warn both statisticians and scientist to be cautious and to communicate all assumptions clearly. Kass (2011) advice statisticians to take a cautionary attitude to science and advocates that all inferential statements should be made subjunctive, i.e. of the type "If data were generated by ..., then there would be a 95% probability of ..." . Also, scientists and statisticians are encouraged to be more critical of their contributions, both methodological and applied (Gelman, 2018). Take this quote from Greenland (2017);

> A cautious scientist will thus reserve judgement and treat no methodology as correct or absolute, but will instead examine data from multiple perspectives, taking statistical methods for what they are: semi-automated algorithms which are highly error-prone without extremely sophisticated human input from both methodologists and content-experts.

Inspired by these admonitions, I have returned to the applications I have been involved with, and tried to evaluate them critically. In particular I have tried to clarify what questions we were

trying to answer, and examined whether the models used provide a suitable and realistic enough description of the data. Further, I have considered to what extent the statistical statements can be followed by scientific conclusions. Naturally, what constitutes a suitable model will depend on the purpose of the analysis. Lehmann (1990) distinguishes between empirical, descriptive and explanatory models, see also Breiman (2001). Explanatory models aim at providing the mechanism underlying the system under study, i.e. offer understanding. Empirical models on the other hand, do not seek understanding of the underlying system, but are meant to be used as tools for certain aims, typically prediction. Descriptive models might be consider to lie between the other two classes. They provide a lower-dimensional representation of the data, usually with the aim of uncertainty quantification, but without claiming any mechanistic understanding of the system, nor with the aim of providing good predictions.

5.3.1 Wars

Contrary to much of the previous discussion, which investigated the relationship between statistical statements and real-world answers, Paper II actually examines a purely statistical question and primarily provides a purely statistical answer. The question treated in Clauset (2017, 2018) and which we also pursued, starts with a known empirical pattern, the long peace, and asks whether this pattern can be considered "statistically significant" given the inherent variability in the data. Unlike the situation in the next section, there is no doubt that the long peace exists, at least in the limited sense that there *were* fewer large interstate wars in the period after world war II. The remaining question is thus whether this change was considerable enough and has lasted long enough to be considered a genuine change in some sort of statistical sense. Mechanisms or explanations are not actively modelled in Paper II nor in the papers by Clauset, but they are discussed. The underlying motivation behind the papers seems to be two-fold

1. the observed pattern should be deemed "statistically significant" in order to be worthy of further investigations; and
2. whether there has been a "significant" change or not has consequences for what type of wars we should expect in the future.

Both of these motivations seem convincing, but regarding the second one, the potential limitations in the modelling framework should be considered. Both Clauset and Paper II use power-law models for the distribution of battle deaths. Richardson (1948, 1960) demonstrated that deaths in wars are distributed as independent draws from a power-law distribution. This finding has since been studied extensively and been shown to hold for various war datasets (Pinker, 2011; Cirillo & Taleb, 2016). The power-law distribution of war deaths must be considered primarily a descriptive model, which has proven useful to model historical data, but which makes no claim at being the true data-generating mechanism (and therefore certainly not explanatory). These models can be used to produce predictions, for example the probability of a "large" war in the next 100 years, but they were not constructed for this purpose. Introducing relevant covariates may allow these models to produce more precise predictions.

Motivations aside, the statistical question of whether the long peace is "statistically significant" can be formalised in various ways. In Paper II we chose to model the question as a change-

point problem. This immediately establishes a constraint on the type of change we believe in: a single abrupt change in the parameters governing the system. This framework does not open for gradual changes, cycles or multiple change-points. However, a single change-point model could be considered a starting point, as it is quite simple. Also, a one-change-point model could be a reasonable approximations to various other patterns.

Once we have decided to search for a change-point, we have to make a choice between different methods. We used the so-called method B from Paper I and as we discuss in Section 4.1, this methods explicitly assumes that there really is a change-point in the sequence of observations. In Section 4.1, I made some recommendations in order to avoid the unfortunate impression that the results from using method B are simply due to the initial assumption. For example, one should always construct confidence curves for the degree of change as well. If these curves indicate a large and significant change, the evidence in favour of the existence of the change-point is strengthened. Some of the confidence curves for the degree of change computed in Paper II, indicated such large and significant changes, notably the ratio between the 75% quantiles of battle deaths before and after 1950. In addition, we should perhaps have performed some general tests of homogeneity for the entire sequence, before embarking on the change-point search. There are a large number of such tests to choose between.

Clauset (2018) mentions change-points briefly, but his main results are based on a homogeneity test. He investigates the extent that the observed sequence of wars (both their size and timing) is unusual under the null hypothesis of no change. This is achieved by generating multiple sequences of wars from such null models, some non-parametric and other semi-parametric. Before our resubmission of Paper II, I plan to investigate Clauset's homogeneity tests, and others, using our favoured parametric models. This would indicate whether the differences in conclusions between Paper II and Clauset (2018) are due to different models or different methods (Clauset's homogeneity tests versus method B).

5.3.2 Whales

In Paper III the question on body conditions of Minke Whales was used as an illustration. The same question was examined in more details in a report by Cunen, Walløe & Hjort (2017). Here the subject matter question was clear: was the body condition of Antarctic Minke Whales reduced in the period from 1988 to 2005? This is a question concerning the entire populations of Minke whales around the Antarctic, or at least the populations in the areas sampled (which correspond to the oceans on one side of the Antarctic continent).

An even more interesting question would perhaps be: what caused the decline in body condition? But although explanations are briefly discussed in the literature, the Minke whale dataset used in Paper III contained no covariates related to these explanations. Thus, the analysis did not aim at providing explanations of the potential decline, but an appropriate model would still need to be explanatory in the sense of Lehmann (1990). The model should account for the various mechanisms explaining the body condition of a particular whale, sampled in a particular year, for instance the effect of sex and of date within each season. As we have seen in the previous discussion, the realism or plausibility of the statistical model is of prime importance when trying to establish the link between the statistical statement and the subject matter question.

As mentioned in Section 4.3, the Antarctic Minke whale question has been discussed extensively in the IWC. In my opinion a lot of the discussion concerned points that were of second order importance to the conclusion, for example extensive discussions on the performance of various information criteria. In my view, the first task should have been to discuss the sampling mechanisms and biology with the aim at agreeing on which covariates influence the observed body condition and in what manner. In Cunen, Walløe & Hjort (2017) and also in various reports by Australian statisticians in the IWC, biological mechanisms were invoked when building statistical models. This should have been done in a more systematic matter, and points of agreement and disagreement concerning these mechanisms should have been sorted out prior to looking at the statistical conclusions. Once a set of relevant covariates, including interactions, had been agreed on, model selection techniques could play a role in selecting the appropriate shape of the effects (linear, or more complex?) and perhaps the distribution of the error term. Goodness-of-fit methods and other model validation techniques could then have been used in order to check that the final model had a reasonable agreement with the observed data. FIC methods could play a role in order to simplify the wide model, as discussed in Section 5.2

It is easy to have lofty ambitions and ideals in hindsight. Still, good statistical practice requires the analyst to think deeply on the appropriate statistical model when real world answers are sought (ideally these considerations should already be present in the design of the data collection). In a forum where there is considerable amount of prestige attached to the conclusions, like in the IWC, it would be particularly fruitful to discuss aspects of the modelling before looking at the statistical conclusions. This is not meant to imply any dishonesty or intentional malpractice on the parts of any of the scientists, but simply an acknowledgement of various psychological and social factors present in all of us: we all have certain biases, most people find it difficult to admit mistakes, and all scientists are under various pressures to produce interesting and publishable results.

References

BAYARRI, M. J. & BERGER, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science* **19**, 58–80.

BERGER, J. O. (1985). The frequentist viewpoint and conditioning. In *Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, L. LeCam & R. Olshen, eds., vol. 1. Monterey, CA.

BERGER, J. O. (2006). The case for objective Bayesian analysis. *Bayesian analysis* **1**, 385–402.

BERGER, J. O. & WOLPERT, R. L. (1988). *The Likelihood Principle*. Hayward, CA: Institute of Mathematical Statistics.

BOX, G. (1990). Comment to "The unity and diversity of probability" by G. Shafer. *Statistical Science* **5**, 448–449.

BREIMAN, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association* **87**, 738–754.

BREIMAN, L. (2001). Statistical modeling: the two cultures [with discussion and a rejoinder]. *Statistical science* **16**, 199–231.

BUEHLER, R. J. (1959). Some validity criteria for statistical inferences. *The Annals of Mathematical Statistics* **30**, 845–863.

CASELLA, G. (1992). Conditional inference from confidence sets. *Lecture Notes-Monograph Series* **17**, 1–12.

CIRILLO, P. & TALEB, N. N. (2016). On the statistical properties and tail risk of violent conflicts. *Physica A: Statistical Mechanics and its Applications* **452**, 29–45.

CLAESKENS, G. & HJORT, N. L. (2003). The focused information criterion [with discussion and a rejoinder]. *Journal of the American Statistical Association* **98**, 900–916.

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

CLAUSET, A. (2017). The enduring threat of a large interstate war. Tech. rep., One Earth Foundation.

CLAUSET, A. (2018). Trends and fluctuations in the severity of interstate wars. *Science Advances* **4**, 1–9.

COX, D. R. (1958). Problems connected with statistical inference. *The Annals of Mathematical Statistics* **29**, 357–372.

COX, D. R. (1998). Comment to "R.A. Fisher in the 21st century" by B. Efron. *Statistical Science* **13**, 114–115.

CUNEN, C., WALLØE, L. & HJORT, N. L. (2017). Decline in energy storage in Antarctic minke whales during the JARPA period: Assessment via the focused information criterion (FIC). *IWC/SC/67A/EM04* , 1–55.

DICICCIO, T. J., KUFFNER, T. A., YOUNG, G. A. & ZARETZKI, R. (2015). Stability and uniqueness of p-values for likelihood-based inference. *Statistica Sinica* **25**, 1355–1376.

DONOHO, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics* **26**, 745–766.

EFRON, B. (1998). R.A. Fisher in the 21st century. *Statistical Science* **13**, 95–114.

EFRON, B. (2010). The future of indirect evidence. *Statistical Science* **25**, 145–157.

EFRON, B. & HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65**, 457–483.

FISHER, R. A. (1930). Inverse probability. In *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 26. Cambridge University Press.

FISHER, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics* **6**, 391–398.

FISHER, R. A. (1956). *Statistical Methods and Scientific Inference.* Oxford, England: Hafner Publishing Co.

FRASER, D. & REID, N. (2016). Crisis in science? or Crisis in statistics! Mixed messages in statistics with impact on science. *Journal of Statistical Research* **48**, 50–64.

FRASER, D. A. (2004). Ancillaries and conditional inference. *Statistical Science* **19**, 333–369.

FRASER, D. A. (2011). Is Bayes posterior just quick and dirty confidence? *Statistical Science* **26**, 299–316.

GELMAN, A. (2018). Ethics in statistical practice and communication: Eight recommendations. *Manuscript* .

GELMAN, A. & HENNIG, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **180**, 967–1033.

GELMAN, A. & RUBIN, D. B. (1995). Avoiding model selection in Bayesian social research. *Sociological methodology* **25**, 165–173.

GOUTIS, C. & CASELLA, G. (1995). Frequentist post-data inference. *International Statistical Review* **63**, 325–344.

GREENLAND, S. (2017). For and against methodologies: some perspectives on recent causal and statistical inference debates. *European Journal of Epidemiology* **32**, 3–20.

GUSTAFSON, P. & GREENLAND, S. (2009). Interval estimation for messy observational data. *Statistical Science* **24**, 328–342.

HANNIG, J. (2009). On generalized fiducial inference. *Statistica Sinica* **19**, 491–544.

HANNIG, J., IYER, H., LAI, R. C. & LEE, T. C. (2016). Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association* **111**, 1346–1361.

HERMANSEN, G., HJORT, N. L. & JULLUM, M. (2015). Parametric or nonparametric: The FIC approach for time series. Tech. rep., University of Oslo.

HJORT, N. L., HOLMES, C., MÜLLER, P. & WALKER, S. G. (2010). *Bayesian Nonparametrics*. Cambridge University Press.

HJORT, N. L. & SCHWEDER, T. (2018). Confidence distributions and related themes. *Journal of Statistical Planning and Inference* **195**, 1–13.

HURVICH, C. M. & TSAI, C. (1990). The impact of model selection on inference in linear regression. *The American Statistician* **44**, 214–217.

JULLUM, M. (2015). *New focused approaches to topics within model selection and approximate Bayesian inversion*. Ph.D. thesis, University of Oslo.

JULLUM, M. & HJORT, N. L. (2017). Parametric or nonparametric: The FIC approach. *Statistica Sinica* **27**, 951–981.

JULLUM, M. & HJORT, N. L. (2018). What price semiparametric Cox regression? *Lifetime Data Analysis* , 1–33.

KASS, R. E. (2011). Statistical inference: The big picture. *Statistical Science* **26**, 1–9.

KASS, R. E. & WASSERMAN, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association* **91**, 1343–1370.

LEE, J. D., SUN, D. L., SUN, Y. & TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics* **44**, 907–927.

LEEB, H. & PÖTSCHER, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21**, 21–59.

LEHMANN, E. L. (1990). Model specification: the views of Fisher and Neyman, and later developments. *Statistical Science* **5**, 160–168.

LEHMANN, E. L. & ROMANO, J. P. (2006). *Testing Statistical Hypotheses*. Springer Science & Business Media.

LIU, K. & MENG, X.-L. (2016). There is individualized treatment. Why not individualized inference? *Annual Review of Statistics and Its Application* **3**, 79–111.

MARTIN, R. & LIU, C. (2013). Inferential models: A framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association* **108**, 301–313.

MARTIN, R. & LIU, C. (2015). *Inferential Models: Reasoning with Uncertainty*. Chapman and Hall/CRC.

MOREY, R. D., HOEKSTRA, R., ROUDER, J. N., LEE, M. D. & WAGENMAKERS, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic bulletin & review* **23**, 103–123.

NELDER, J. A. (1986). Statistics, science and technology. *Journal of the Royal Statistical Society. Series A* **48**, 109–121.

NELDER, J. A. (1999). From statistics to statistical science. *Journal of the Royal Statistical Society: Series D (The Statistician)* **48**, 257–269.

NEYMAN, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* **97**, 558–625.

PINKER, S. (2011). *The Better Angels of Our Nature: Why Violence Has Declined*. Toronto: Viking Books.

REID, N. (1995). The roles of conditioning in inference. *Statistical Science* **10**, 138–157.

REID, N. (2015). Approximate likelihoods. In *Proceedings of the ICIAM, Beijing, China*. Cambridge University Press.

RICHARDSON, L. F. (1948). Variation of the frequency of fatal quarrels with magnitude. *Journal of the American Statistical Association* **43**, 523–546.

RICHARDSON, L. F. (1960). *Statistics of Deadly Quarrels*. Los Angeles: Boxwood Press.

ROBERT, C. P. (2011). Discussion of "Bayes posterior just quick and dirty confidence?" by DAS Fraser. *Statistical science* **26**, 1–3.

ROBINS, J. & WASSERMAN, L. (2000). Conditioning, likelihood, and coherence: A review of some foundational concepts. *Journal of the American Statistical Association* **95**, 1340–1346.

ROBINSON, G. (1979). Conditional properties of statistical procedures. *The Annals of Statistics* **7**, 742–755.

SCHWEDER, T. (2001). Protecting whales by distorting uncertainty: non-precautionary mismanagement? *Fisheries Research* **52**, 217–225.

SCHWEDER, T. (2018). Confidence is epistemic probability for empirical science. *Journal of Statistical Planning and Inference* **195**, 116–125.

SCHWEDER, T. & HJORT, N. L. (2002). Confidence and likelihood. *Scandinavian Journal of Statistics* **29**, 309–332.

SCHWEDER, T. & HJORT, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge: Cambridge University Press.

SEVERINI, T. A. (1990). Conditional properties of likelihood-based significance tests. *Biometrika* **77**, 343–352.

SHMUELI, G. (2010). To explain or to predict? *Statistical Science* **25**, 289–310.

TUKEY, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics* **33**, 1–67.

WASSERMAN, L. (2011). Frasian inference. *Statistical Science* **26**, 322–325.

WELCH, B. (1939). On confidence limits and sufficiency, with particular reference to parameters of location. *The Annals of Mathematical Statistics* **10**, 58–69.

WIT, E., HEUVEL, E. V. D. & ROMEIJN, J.-W. (2012). 'All models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica* **66**, 217–236.

XIE, M.-G. & SINGH, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review [with discussion and a rejoinder]. *International Statistical Review* **81**, 3–39.

I

Journal of Statistical Planning and Inference 195 (2018) 14–34

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Confidence distributions for change-points and regime shifts

Céline Cunen, Gudmund Hermansen, Nils Lid Hjort *

Department of Mathematics, University of Oslo, Norway

ARTICLE INFO

Article history:
Available online 2 October 2017

Keywords:
Change-points
Confidence distributions
Homogeneity testing
Log-likelihood profiling
Monitoring bridges
Regime shifts
Tirant lo Blanch

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.

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 in several areas of application, from engineering (see e.g. Frick et al., 2014), economics and finance, to biology (e.g. Gould and 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 not only 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,

* Corresponding author.
E-mail addresses: cmlcunen@math.uio.no (C. Cunen), nils@math.uio.no (N.L. Hjort).

https://doi.org/10.1016/j.jspi.2017.09.009
0378-3758/

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 $R(\alpha) = \{\tau : \mathrm{cc}(\tau, Y) \le \alpha\}$ ought to have probability approximately equal to α, with Y denoting a random dataset drawn from the model; see Schweder and 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 and 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 lengths, 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 skiing 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 and 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) and Fearnhead (2006), and also Section 5.3.

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 $H_{1,n} : \theta_1 = \cdots = \theta_n$, 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

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

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. Fig. 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

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

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, ..., 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,

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

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 (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 and Zhang (2015), which may be put to work as long as a similarity measure on the sample space can be given. In fact Chen and Zhang (2015) utilise an idea similar to our (2.1), 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 and Spjøtvoll (1982) and 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 $\frac{1}{2} + \frac{1}{2}\alpha$ replacing $\sqrt{\alpha}$ in (2.1), yielding an alternative set $R_b(\alpha)$; this secures a conservative $P_\tau\{\tau \in R_b(\alpha)\} \ge \alpha$. The difference is actually slight for confidence levels $\alpha > \frac{1}{2}$ 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

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

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

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

involving the maximisers of $\ell(\tau, \theta_L, \theta_R)$ over θ_L and θ_R for given τ. The maximiser of ℓ_{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

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

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

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

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

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

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

$$\ell_{\mathrm{prof}}(\tau) = -\tau \log \widehat{\sigma}_L(\tau) - (n - \tau) \log \widehat{\sigma}_R(\tau),$$

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

$$-n \log \sigma - \frac{1}{2}(1/\sigma^2)[Q_L(\tau) + \tau\{\bar{y}_L(\tau) - \xi_L\}^2 + Q_R(\tau) + (n - \tau)\{\bar{y}_R(\tau) - \xi_R\}^2],$$

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 $\ell_{\mathrm{prof}}(\tau) = -n \log \widehat{\sigma}(\tau)$. 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

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

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

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

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

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

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

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

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 and 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 and 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 $x : [0, 1] \to \mathbb{R}$, right continuous with limits from the left, equipped with the Skorokhod 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 $\ell_j(\theta_0) = \ell_j(\widehat{\theta}_j) + \frac{1}{2}(\theta_0 - \widehat{\theta}_j)^{\mathrm{t}}\ell_j''(\widehat{\theta}_j)(\theta_0 - \widehat{\theta}_j) + o_{\mathrm{pr}}(1)$, which leads to

$$\widehat{\ell}_j = \ell_j(\theta_0) + \frac{1}{2}W_j + o_{\mathrm{pr}}(1), \tag{4.5}$$

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

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

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 $\{\tau : \mathrm{cc}(\tau, y) \le \alpha\}$, 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 and 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 and 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

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.

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

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 $\widehat{b} = \sum_{i=1}^{\tau}(i - \bar{i})y_i/K(\tau)$, where we know that $K(\tau)\widehat{b}^2 \sim \chi^2_1$, under the homogeneity hypothesis; here $K(\tau) = \sum_{i\le\tau}(i - \bar{i})^2$ 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 and 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 and Koning (2002) implying that these are asymptotically distributed as maxima of Brownian bridges; cf. (4.2).

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 tests 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 tend 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 $\widehat{b} = \sum(i - \bar{i})y_i/K$ with $K = \sum_{i=1}^n (i - \bar{i})^2$, and employ $t = K^{1/2}\widehat{b}/\widehat{\sigma}$, 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 and 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

$$V_L = \max\{\max_{t\le1} |M_{L,1}(t)|, \max_{t\le1} |M_{L,2}(t)|\},$$

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.

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

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 $\sigma_R = 2$.

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. 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

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

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 $\pi(\tau \mid \text{data})$ 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:

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

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 and Hjort, 2008, Ch. 4).

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

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

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 $\log\pi(\tau \mid \text{data})$ 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 not only 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

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

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 $U_\tau = \sum_{\tau_B<\tau}\pi(\tau_B \mid Y) + \frac{1}{2}\pi(\tau \mid Y)$, 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

$$\pi(\tau \mid y) \propto \exp\{\ell_{\mathrm{prof}}(\tau)\}$$

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 $i \le \tau$ and θ_R for $i \ge \tau + 1$. 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

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

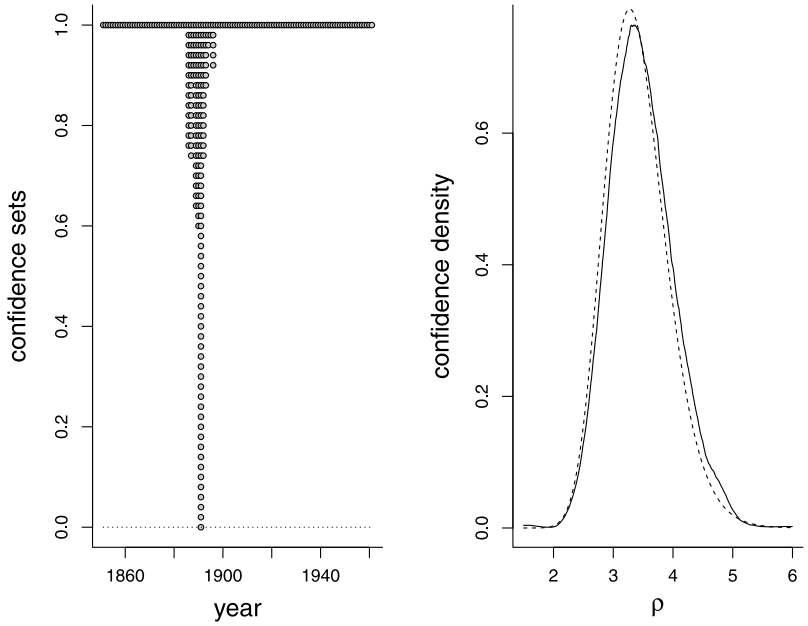

Fig. 6.1. Left panel: Confidence curve for the change-point τ, using the deviance based method B. Right panel: Confidence density for the degree of change $\rho = \theta_L/\theta_R$, via method B (full line), and the Bayesian method (dashed line).

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 Fig. 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.

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 $\text{cc}(\rho, y)$ can be converted to a cumulative confidence distribution $C(\rho, y)$, via $\text{cc}(\rho, y) = |1 - 2\,C(\rho, y)|$, which then via numerical derivation yields a confidence density, say $c(\rho, y)$, displayed in the right panel of Fig. 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 Fig. 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 $\text{cc}(\rho, y_{\text{obs}}) = \Gamma_1(D(\rho, y_{\text{obs}}))$, 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 Fig. 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

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.

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

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 and Ginebra (2005), Koziol (2014) and Chen and 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 chapters 350 and 400 (Chen and 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 and Zhang (2015), model the 425 word count vectors as multinomially distributed. Chen and 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 and 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, allowing 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 and Hjort (2008, Ch. 3), defined as AIC $= 2\,\ell_{\max} - 2\,\dim$, 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.

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

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

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 and Ginebra (2005), and also

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

	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

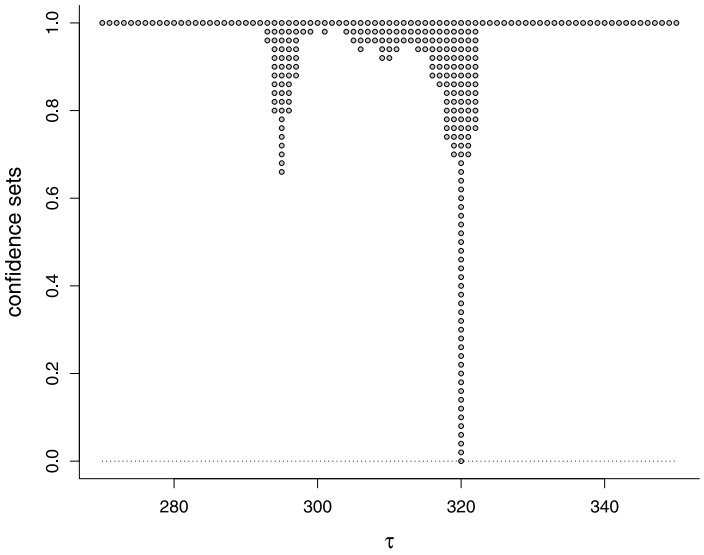

Fig. 7.1. Confidence curve for the change-point τ, using method B based on the multinormal model.

the mode of the change-point posterior distribution in Girón et al. (2005). The multinormal model led to ML estimates given in Table 7.2 for the mean of the word length proportions (the covariances matrices are not given).

By simulating the distribution of $D(\tau, Y)$ we obtain the confidence curve for τ, shown in Fig. 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.

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 and 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

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

and confidence curves can easily be constructed by method B; see Fig. 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 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 and Ginebra, 2005; Girón et al., 2005; Koziol, 2014; Chen and 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: skiing 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 Fig. 8.1 the number of such days at the particular location of Bjørnholt in Oslo's skiing and recreation area Nordmarka is plotted for the

[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.

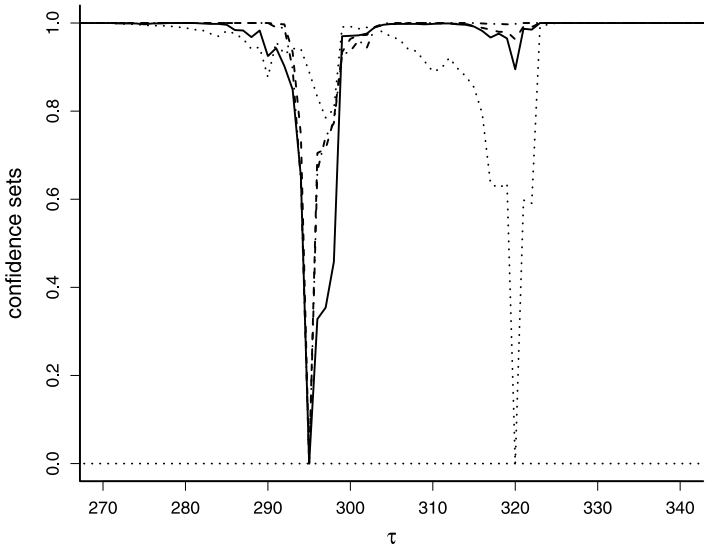

Fig. 7.2. Confidence curves for the change-point τ, using method B: Full line, based on the average word length per chapter; dashed line, based on the standard deviation in word lengths; dotted line, based on the proportions of words of length 3 letters or less; and dot-dashed line, based on the proportions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

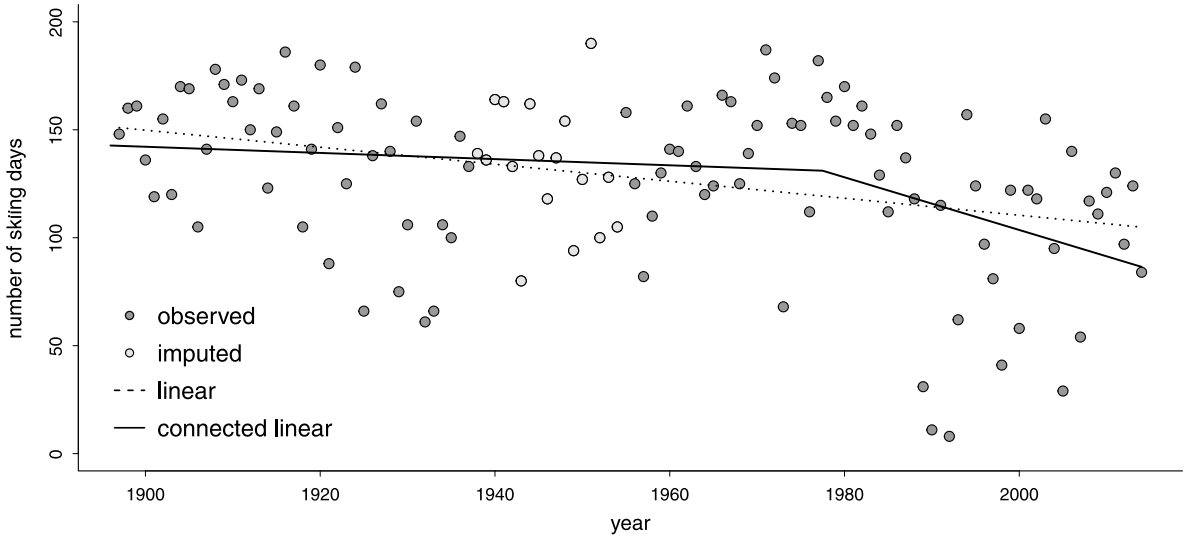

Fig. 8.1. The number of skiing days for the winter seasons 1896–97 to 2013–14 at Bjørnholt. The imputed data points are meteorologists' reconstructions making use of nearby locations. The global linear trend (dashed line) decreases with an estimated slope of about −0.40. Some time after 1960 there appears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

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 forms yet another potential source for studying climate change.

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

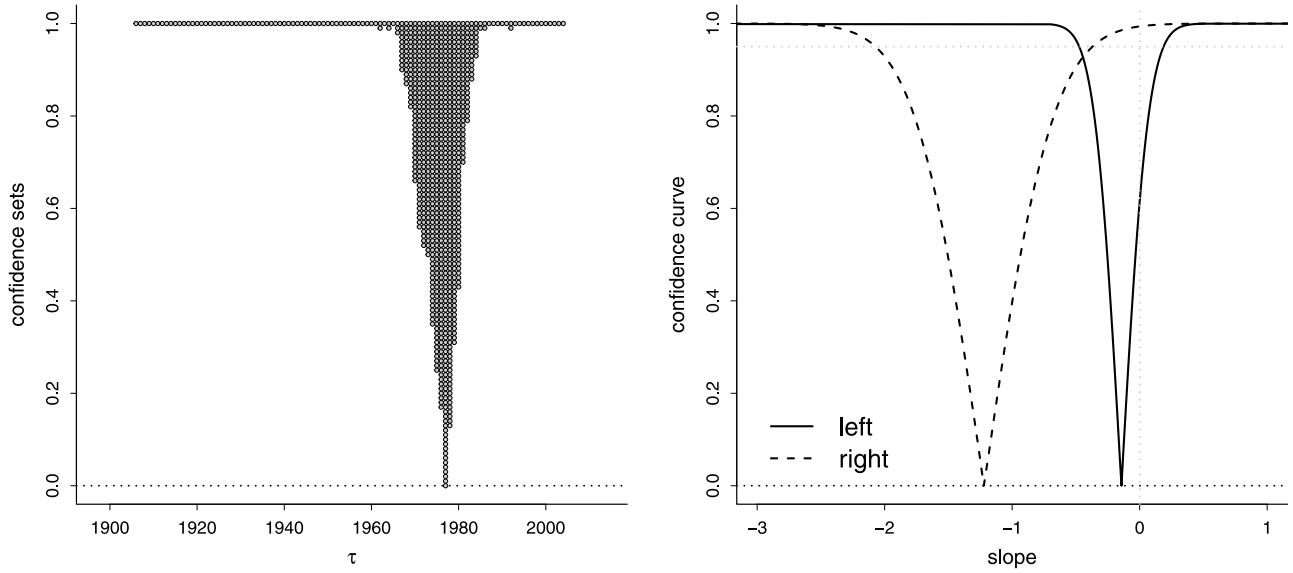

Fig. 8.2. The confidence sets for the change-point τ (left) and the confidence curves for the two slopes (right) in the connected linear model. The estimated change-point for the connected linear model (with its six parameters) is 1977, where the mean slope changes from -0.14 to -1.22.

$m(\beta_L, t) = \beta_L$ 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.

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 Fig. 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 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 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). Fig. 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 Fig. 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

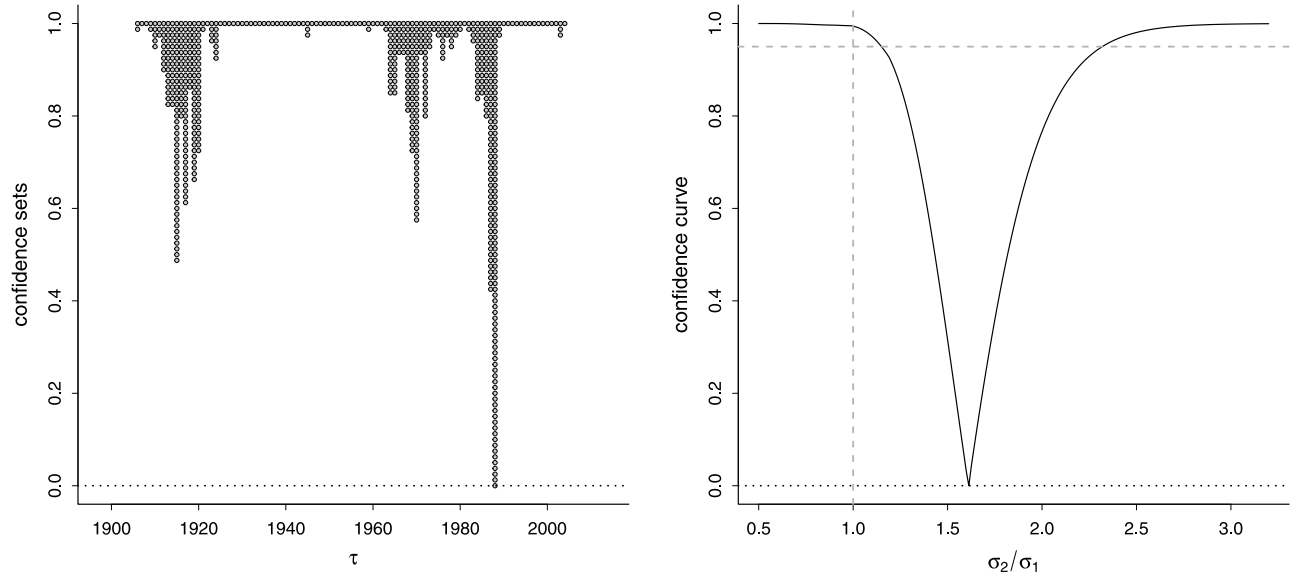

Fig. 8.3. The confidence curve (left) suggests three possible locations for a change in σ, viz. 1915, 1970 and 1988, with the latter given most confidence. The confidence curve for σ_R/σ_L (right) indicates that the σ of the AR(1) part of (8.1) has increased, around 1988, with a factor of about 1.61.

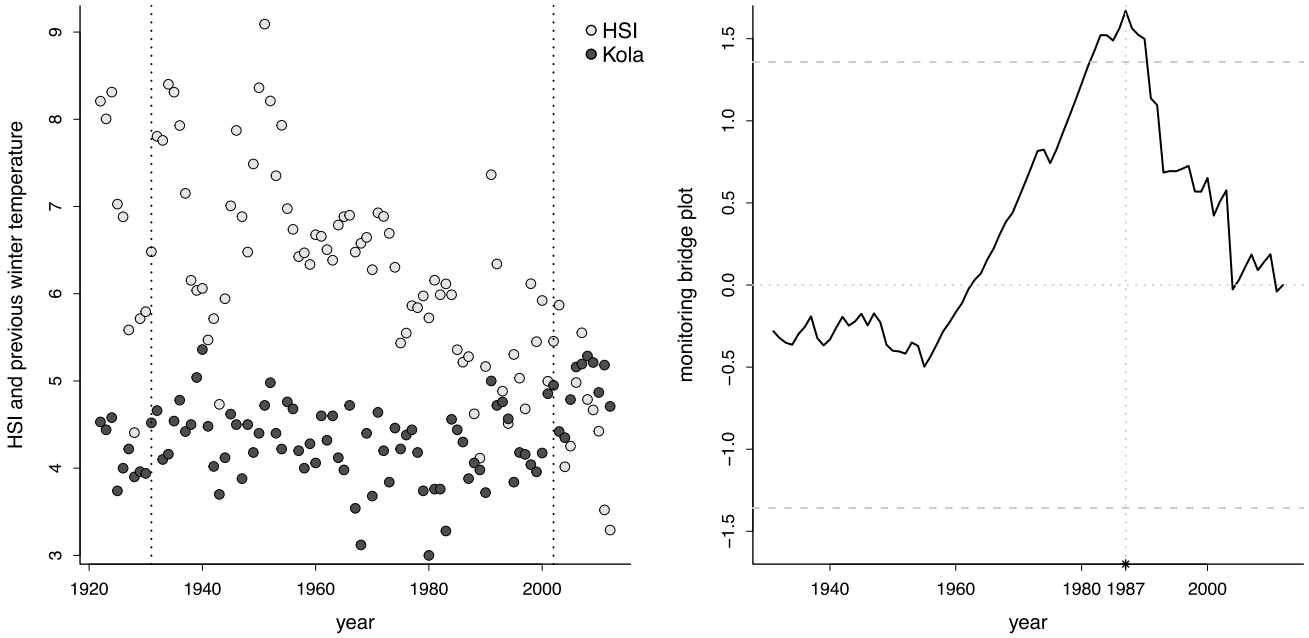

Fig. 9.1. Left panel: The Hjort liver quality index (HSI) series for 1921–2012 (grey), along with average Kola winter temperature (black, in degrees Celsius). The vertical lines indicate the range (1927–2006) where we are searching for a potential change; the boundary points are excluded due to instability 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.

typically studied as a time series; see Fig. 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) represents 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).

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.

Model	$\widehat{\beta}_0$	$\widehat{\beta}_{\text{kola}}$	$\widehat{\rho}$	$\widehat{\sigma}$
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)

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 and Marshall (2015) and 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 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 Fig. 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, . . . , 2002. The resulting confidence curves are presented in Fig. 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) and Vasilakopoulos and 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 spelt out at the end of Section 2, yields a fairly similar confidence curve, for all confidence levels above 0.60 (see Table 9.1).

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 in beginning of the 1980s. Also, Vasilakopoulos and 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.

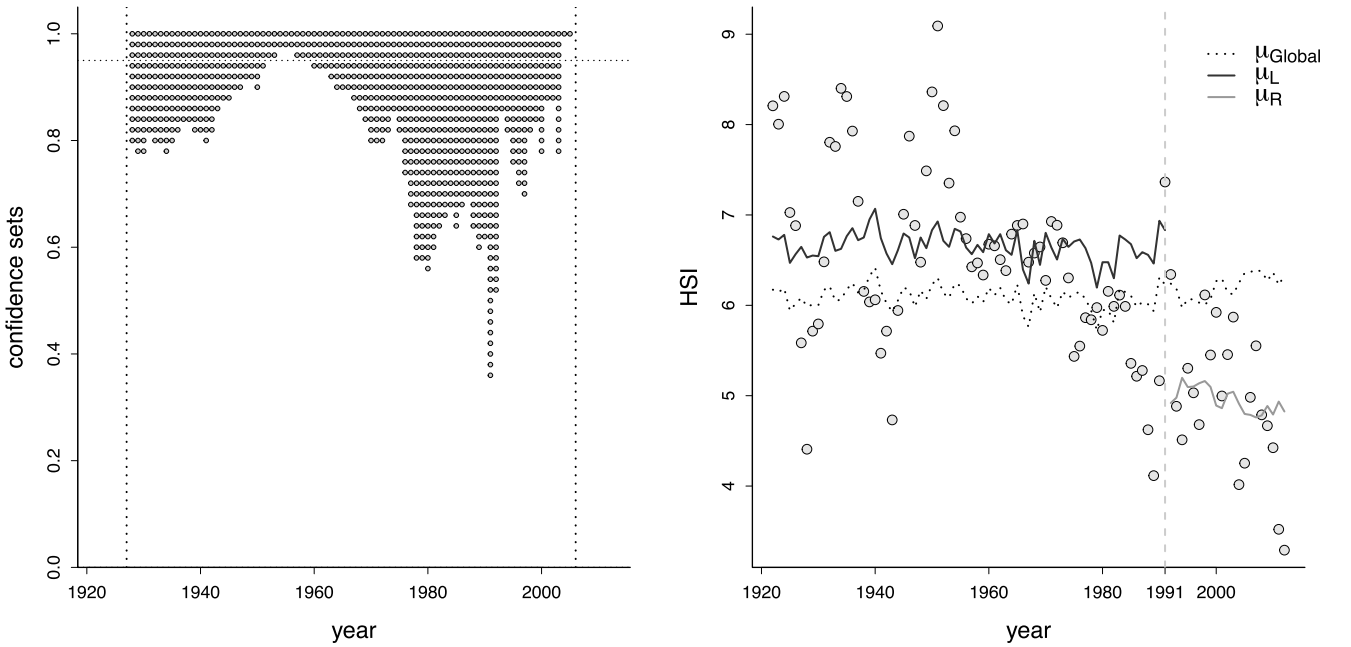

Fig. 9.2. Left panel: The confidence curve for a regime shift τ obtained via method A, with absolute maxima of monitoring log-likelihood bridges, as in Section 4.2. The curve indicates two plausible regions for τ; one right before 1980 (which may be related to a decrease in variance, as suggested in Fig. 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.

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-squaredness 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) and Hinkley and Hinkley (1970) and later on by Cobb (1978) and 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 $n-\tau$ is 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.

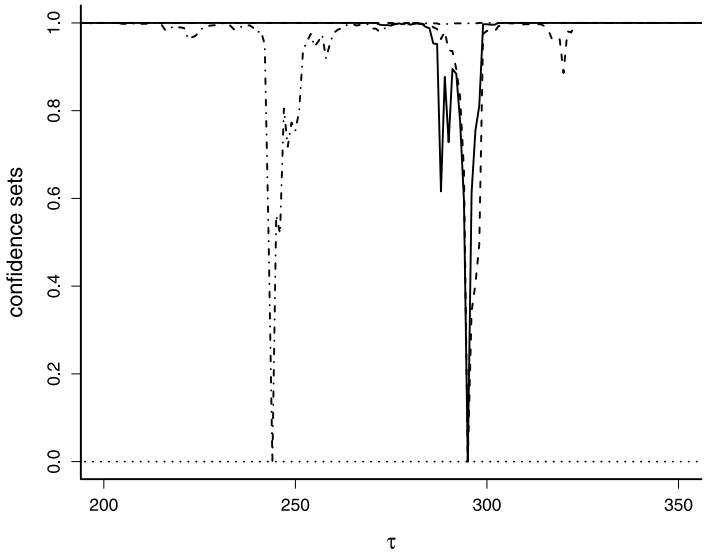

Fig. 10.1. Confidence curves for the change-of-author-point τ, the dot-dashed line is the curve based on the mean sentence length in each chapter, the dashed line is the curve based on the mean word length in each chapter, and the full line is the combined confidence curve.

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 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 and Hjort (2015) and Cunen and Hjort (2016). The parallel for the present case is to stay with the log-likelihood profiles, naturally extending method B. Let $\ell_{\text{prof},1}(\tau)$ and $\ell_{\text{prof},2}$ be the profiled log-likelihoods function from information sources 1 and 2. These can be summed to $\ell_{\text{prof,comb}}(\tau) = \ell_{\text{prof},1}(\tau) + \ell_{\text{prof},2}(\tau)$, from which we can find the combined maximum likelihood estimator $\widehat{\tau}$ and construct the combined deviance function, say $D_{\text{comb}}(\tau, Y) = 2\{\ell_{\text{prof,comb}}(\widehat{\tau}) - \ell_{\text{prof,comb}}(\tau)\}$. 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 Fig. 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 confidence regions, at level α, in the following way (see corresponding formula (2.1))

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

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 not only 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 and Au (1989), Bai and Perron (1998) and Braun et al. (2000). However, these articles do not treat change-points very generally. Yao and Au (1989) and Braun et al. (2000) consider segmentation problems, while Schweder (1976) and Bai and 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. Wiley, New York.

Boitsov, V.D., Karsakov, A.L., Trofimov, A.G., 2012. Atlantic water temperature and climate in the Barents Sea, 2000–2009. ICES J. Mar. Sci. 69, 833–840.

Brander, K.M., 2010. Impacts of climate change on fisheries. J. Mar. Syst. 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. Appl. Stat. 41, 389–406.

Carlstein, E., Müller, H., Siegmund, D., 1994. Change-Point Problems. Institute of Mathematical Statistics, New York.

Chen, H., Zhang, N., 2015. Graph-based change-point detection. Ann. Statist. 43, 139–176.

Claeskens, G., Hjort, N.L., 2008. Model Selection and Model Averaging. Cambridge University Press, Cambridge.

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. Scand. J. Stat. 9, 147–152.

Csörgő, S., Faraway, J., 1996. The exact and asymptotic distributions of Cramér–von Mises statistics. J. R. Stat. Soc. Ser. B Stat. Methodol. 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: Samaniego, F. (Ed.), Proceedings Ofthe 60th World Statistics Congress. International Statistical Institute, Rio de Janeiro, pp. 3581–3586.

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

Efron, B., 2015. Frequentist accuracy of Bayes estimators. J. R. Stat. Soc. Ser. B Stat. Methodol. 77, 617–646.

Fearnhead, P., 2006. Exact and efficient Bayesian inference for multiple changepoint problems. Stat. Comput. 16, 203–213.

Fraser, D.A.S., 2011. Is Bayes posterior just quick and dirty confidence? [with discussion and a rejoinder]. Statist. Sci. 26, 249–316.

Frick, S., Munk, A., Sieling, H., 2014. Multiscale change-point inference [with discussion contributions]. J. R. Stat. Soc. Ser. B Stat. Methodol. 76, 495–580.

Frigessi, A., Hjort, N.L., 2002. Statistical models and methods for discontinuous phenomena. J. Nonparametr. Stat. 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. Amer. Statist. 59, 19–30.

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

Gould, S.J., Eldredge, N., 1977. Punctuated equilibria: the tempo and mode of evolution reconsidered. Paleobiology 3 (02), 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. Can. J. Fish. Aquat. Sci. 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. J. Nonparametr. Stat. 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 J. Mar. Sci. 71, 2053–2063.

Koziol, J.A., 2014. A note on change-point estimation in a multinomial sequence. Enliven: Biostat. 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. J. Amer. Statist. Assoc. 109, 1450–1465.
Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press, New York.
Riba, A., Ginebra, J., 2005. Change-point estimation in a multinomial sequence and homogeneity of literary style. J. Appl. Stat. 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. J. Amer. Statist. Assoc. 71, 491–501.
Schweder, T., Hjort, N.L., 2016. Confidence, Likelihood, Probability. Cambridge University Press, Cambridge.
Siegmund, D., 1988. Confidence sets in change-point problems. Int. Statist. Rev./Rev. Int. Statist. 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 Biol. 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. Internat. Statist. Rev. 81 (1), 3–39.
Yao, Y.-C., Au, S., 1989. Least-squares estimation of a step function. Sankhyā Ser. A 370–381.

II

III

IV