

Window of Change

Extensions of Change-Point Analysis

Even Tronstad

Data Science
60 ECTS study points

Department of Mathematics
Faculty of Mathematics and Natural Sciences

Even Tronstad

Window of Change

Extensions of Change-Point Analysis

Supervisors:
Nils Lid Hjort
Céline Cunen
Ingrid Kristine Glad

Abstract

We consider a sequence of observations drawn from a parametric distribution that is stationary both before and after a period of transition. Within this transitional window, the distribution changes gradually from the initial to the final state. To model this behaviour, we introduce a novel statistical framework, the *Window of Change* model, and contrast it with traditional change-point models that assume abrupt shifts. We show that estimating the parameters governing the transition can be unexpectedly challenging and, in some cases, subject to bias. We develop estimation algorithms for the Window of Change model applicable to a broad class of parametric distributions in both the frequentist and Bayesian paradigms. The proposed methodology is evaluated through application to several real-world data sets, illustrating both its flexibility and practical relevance.

Contents

Preface	xiii
1 Introduction	1
1.1 Motivation	1
1.2 Related work	1
1.3 Overview	2
1.4 Background and notation	4
1.5 Frequentist and Bayesian frameworks	7
1.6 Change-Point analysis	8
1.7 An inviting example	9
1.8 The new twist	12
2 Preliminaries	19
2.1 Markov chain Monte Carlo	19
2.2 Mixture distributions and expectation maximization	21
3 The Window of Change	23
3.1 The weight function	23
3.2 Parameter transition	24
3.3 Mixture transition	24
3.4 Comparison	25
3.5 Extensions	28
3.6 Estimation and bias	29
3.7 Confidence curves	31
3.8 The problem of priors	32
3.9 Prior matching	34
3.10 Random walk representation	36
4 The simplest interesting example	39
4.1 Estimating the window length when the beginning is known	39
4.2 Maximum likelihood estimation	40
4.3 Confidence curves	42
4.4 Bayesian analysis	44
4.5 The unknown window	45
4.6 A summary	48
5 Normal simulation	53
5.1 Normal data with parameter transition	53
5.2 Normal data with mixture transition	58
5.3 Discussion	59
6 Computation	63
6.1 Gradient descent - parameter transition	63
6.2 Expectation maximization in mixture WoC models	66
6.3 MCMC for WoC models	68

Contents

7	Hypothesis testing	75
7.1	Defining a test statistic.	75
7.2	The null distribution	76
7.3	Power	78
7.4	The model selection perspective	80
8	Application: Bjørnholt	83
8.1	The WoC model	83
8.2	Results	85
9	Application: Tirant lo Blanch	89
9.1	A tale of two authors	89
9.2	Two WoC models.	89
9.3	The Bayesian analysis.	91
9.4	Discussion	95
10	Concluding remarks	97
10.1	Summary.	97
10.2	Future work.	97
	References	99
A	Sample code	101
A.1	Gibbs sampler for mixture of normal observations	101
B	Data sets	105
B.1	British mining disaster	105
B.2	Skiing days at Bjørnholt	107
B.3	Tirant lo Blanch	110

List of Figures

- 1.1 Yearly accidents counts in British mines (dots) and the profiled likelihood function for the change point using a Poisson model (bar plot). The likelihood function is scaled such that its maximum is 6 making it comparable to the data. 10
- 1.2 Inverted confidence curve for the change point τ . Notice how it clearly shows the maximum likelihood estimate with its single peak. It also slightly follow the shape of the likelihood function, with what appears to be three modes. The fact that it has three modes shows how a confidence curve does not necessarily imply a confidence distribution, i.e. cumulative confidence, and is why we choose to work with the cc. 11
- 1.3 The posterior distribution for the change point. This too closely mimics the profile likelihood function, suggesting that there is little posterior uncertainty in θ_L and θ_R . In fact, $\mathbb{E}[\theta_L | y] = 3.13$, $\text{Var}[\theta_L | y] = 0.086$ and $\mathbb{E}[\theta_R | y] = 0.915$, $\text{Var}[\theta_R | y] = 0.013$ 12
- 1.4 The profiled Likelihood with the maximum likelihood estimate marked in red. It is scaled such that white is zero and black is the maximum likelihood. The maximum likelihood estimate is $\hat{a} = 32$ and $\hat{b} = 47$, which is a sizeable window. The diagonal border, vaguely visible to the lower left, is the line representing $b = a + 1$, which is the edge of the permissible parameters. 14
- 1.5 The two-dimensional confidence curve for the joint confidence for a and b calculated with 100 samples. The maximum likelihood estimate is marked in red. Note the elongated shape of the region, which shows how we have less confidence in the duration of the change compared to the location. 15
- 1.6 The same confidence curve from 1.5, but parametrized with $\tau = (a + b)/2$ and $d = b - a$. The white squares represent impossible parameter combination (like for example $\tau = 1900$ and $d = 1$ which would require $a = 1899.5$ and $b = 1900.5$). The figure clearly shows that there is more uncertainty when estimating d than τ as the region is narrow and tall. 16
- 1.7 100 posterior model samples (red) alongside the MLE (blue). We see that they all concur on τ , and vary more in the d dimension. The posteriors for θ_L and θ_R are approximately normal, and the variance of θ_L is larger due to there being fewer observations before the window. 17

1.8	Top row: the confidence curve (left) and the corresponding following 95% confidence region (right) calculated using 100 samples for each a, b . Here, as in all other figures of this kind, is the y -axis a and the x -axis b . The confidence curve is somewhat noisy since the number of samples is small, but the computational burden was too large to justify more samples. The search space was limited to an ellipse, which explains the hard boundary around them. Bottom row: the Bayesian posterior credibility distribution (left) and the corresponding 95% credibility region (right)	18
3.1	Sample data sets for a parameter and a mixture transition for normal distributions with constant variance equal to 1. The true parameters where $\mu_L = 0$ and $\mu_R = 8$, and the window begins at $a = 30$ and ends at $b = 70$	26
3.2	Quantiles at the 0.1 and 0.9 (dashed); 0.2 and 0.8 (dashdotted); and 0.3 and 0.7 (dotted) levels for parameter (blue) mixture (red) transition. The same parameters from figure 3.1 where used, i.e. $\mu_L = 0$, $\mu_R = 8$, $\sigma = 1$, $a = 30$ and $b = 70$	27
3.3	Example of overdispersion with normal data. The x axis is the true standard deviation, while the y axis is the expected value of $\hat{d}$ when σ is assumed to be 1. The true parameters are $d_0 = 10$, $\theta_L = (0, \sigma^2)$, and $\theta_R = (2, \sigma^2)$. Clearly, both estimators are strongly negatively biased, but this is due to another phenomenon which we will discuss later.	28
3.4	The inherent bias in Bayesian analysis for WoC models (solid lines) compared to CP models (dashed lines). Left: $d = 20$ is held fixed. Right: $\tau = 50$ is held fixed (the dashed line is flat because d is held constant). For both, $n = 100$	36
4.1	Bias for the four estimators. Notice the bias for $\hat{d}$, and that $\hat{d}'$ quite fast returns to zero. $\hat{d}''$ is $\hat{d}$ shifted by one. The posterior means, $\mathbb{E}[\pi(d y)]$ and $\mathbb{E}[\pi'(d y)]$ are both positively biased. The latter, which has a decreasing prior for d , is less skewed.	41
4.2	Likelihood (left), inverted confidence curve (middle) and inverted confidence curve with half correction (right) for example data (4.3). Note that both are zero for $d < 5$ because $A = 4$. The half corrected cc does not have the prominent peak at the MLE, but is somewhat tighter in the sense that $d = 11$ is excluded from its 95% confidence set.	42
4.3	Cumulative sample distribution for $cc(d_0, Y)$ (left) and the corrected version (right) for $d_0 = 20$ and $n = 30$. The diagonal is dotted. 10000 samples where used.	43
4.4	Root Mean Square Error (RMSE) for the estimators. It shows that all corrections are comparatively effective at reducing the error.	44
4.5	Bias for when estimating $d = b - a$ with $a = 10$ and $n = 50$ as a function of d_0 . We see that the MLE is strongly negatively biased, while the posterior mean is slightly positively biased, but recedes towards zero. The first correction, $\hat{d}'$ is an improvement over the MLE, but not as strongly as when a was known. The final correction, $\hat{d}''$, which finds the set $\hat{a}'', \hat{b}''$ such that the expected bias matches the MLE, seems to work very well. Calculated using $n_{sim} = 10000$ samples	47
4.6	RMSE for the four estimators. Contrary to only d unknown does the corrections not universally decrease the error. When $12 < d$, they are all an improvement, but the posterior mean is clearly the estimator with the smallest error.	47

4.7	Bias for b as a function of d_0 . It is similar to figure 4.1, but the bias is stronger, and	49
4.8	RMSE as a function of d . The chaotic nature is not due to using only a few samples, but truly reflect the RMSE.	49
4.9	Upper left: Confidence curve with $a_0 = 20$, $b_0 = 40$, and $n = 100$. Upper right: 95% confidence region Lower left: Posterior $\pi(a, b y)$ using flat prior on a, b Lower right: 95% highest posterior credibility interval.	50
4.10	Confidence for τ (blue solid), confidence for d (blue dashed), credibility for τ (red solid), and credibility for d (red dashed) for the same data as in figure 4.9. The true values are $d_0 = 20$ and $\tau_0 = 30$	50
5.1	Left: Expected value for the ML estimator $\hat{d} = \hat{b} - \hat{a}$ as a function of the true d_0 for a selection of $\Delta\theta \in \left(\frac{1}{2}, \frac{2}{2}, \frac{3}{2}, \frac{4}{2}, \frac{5}{2}, \frac{6}{2}, \frac{7}{2}, \frac{8}{2}\right)$. The deepest purple is $\Delta\theta = 1/2$, and the brighter colour are higher values with yellow being $\Delta\theta = 4$. The black line is the diagonal $\mathbb{E}[\hat{d}] = d$, which represents an unbiased estimator. Initially, the MLE overestimates d , but eventually all underestimates it. Curiously, for the small $\Delta\theta = 0.5$, seems $\hat{d}$ invariant under d_0 . Right: Expected Root Mean Square Error for the same $\Delta\theta$. As we might expect, the greater the level difference, i.e. $\Delta\theta$, the easier is it to identity it, and thus the smaller the error.	54
5.2	Bias for d as a function of d_0 . Here, we see that the corrections are weaker. The posterior mean is initially very strongly positively biased, and decreases almost linearly until it again is negatively biased. Both $\hat{d}'$ and $\hat{d}''$ correct some bias, but not entirely.	56
5.3	RMSE as a function of d . We see that the posterior does indeed give the smallest error for $5 < d_0$, and curiously, $\hat{d}'$ and $\hat{d}''$ gives a larger error than simply the MLE.	56
5.4	Upper left: Confidence curve with $a_0 = 20$, $b_0 = 40$, and $n = 100$. Upper right: 95% confidence region Lower left: Posterior $\pi(a, b y)$ using flat prior on a, b Lower right: 95% highest posterior credibility interval. The black line is the diagonal $a = b$	57
5.5	Confidence for τ (blue solid), confidence for d (blue dashed), credibility for τ (red solid), and credibility for d (red dashed) for the same data as in figure 5.4. True values are marked in vertical lines.	57
5.6	Left: Expected value for the ML estimator $\hat{d} = \hat{b} - \hat{a}$ as a function of the true d_0 for a selection of $\Delta\theta$. The black line is the diagonal $\mathbb{E}[\hat{d}] = d$, which represents an unbiased estimator. Right: Expected Root Mean Square Error for the same $\Delta\theta$	59
5.7	Bias for b as a function of d_0 . The de-biasing is slightly more effective, with $\hat{d}''$ getting quite close. The posterior mean is again first quite positively biased, then negatively biased.	60
5.8	RMSE as a function of d . Again, the posterior mean has the lowest RMSE, while the two corrections have larger RMSE than the MLE, but the difference seems to diminish as d_0 increases.	60
5.9	Upper left: Confidence curve with $a_0 = 20$, $b_0 = 40$, and $n = 100$. Upper right: The 95% confidence region. Lower left: Posterior $\pi(a, b y)$ using flat prior on a, b . Lower right: The 95% highest posterior credibility region.	61

5.10	Confidence for τ (blue solid), confidence for d (blue dashed), credibility for τ (red solid), and credibility for d (red dashed) for the same data as in figure 5.9. True values are marked in vertical lines.	61
6.1	Traces of a (red) and b (blue) from 5 chains constructed as in example 6.3.2. The horizontal black lines are the true parameter values. The traces show good mixing, with a total number of samples (across all five runs), of 316 and 189 respectively, however, considering the total number of samples where $5 \cdot 1000$, this mean that we only retrieve 6.3% and 3.8% of the samples. For θ_L, θ_R where the effectiveness 30% and 24% respectively, again illustrating the relative difficulty of estimating a, b . The Gelman-Rubin statistics are almost exactly one, showing that the stationary distribution has been reached. The sampling took 12 seconds on an M1 MacBook Air.	71
6.2	Traces for the Metropolis-Hastings algorithm using the same data parameters as above, but 10 times the number of samples. The window parameters a, b where proposed using a symmetric shift, and the distribution parameter θ_L and θ_R where proposed with a symmetric normal sift with standard deviation of 1. The acceptance rate was about 0.7. The traces show stronger autocorrelation and thus worse mixing. Still, the Gelman-Rubin statistic suggest that the stationary distribution has been reached. The sampling took 1.1 seconds on an M1 MacBook Air.	71
7.1	CDF of null distribution for T_{LR} for a normal parameter transition WoC model given different level differences. The curve was constructed from 10000 samples, and the 0.95 level is marked by the black line.	77
7.2	CDF for the MLE test statistic for a normal parameter transition WoC model given different level differences. The curve was constructed from 10000 samples, and the 0.95 level is marked by the black line.	77
7.3	Empirical CDF for the Likelihood ratio test statistic. The significance level, alongside its critical value, is marked by the dashed line. The observed value is the full line.	78
7.4	The power of the statistics in figure 7.1	79
7.5	The power of the statistics in figure 7.2	79
7.6	CDF for $\pi(d = 1 Y)$. If $\pi(d = 1 Y)$ were to have correct coverage for any levels, the CFDs follow the diagonal, which they seem to approach as $\Delta\theta$ increases.	82
7.7	Power for the statistics. Note that none of them start at 0.05, which means that they do not have correct coverage, and we can therefore hardly compare them to the other test statistics. When $\Delta\theta = 4$, $\Pr(\text{Reject } H_0 H_0) = 0.084$ thus still not correct.	82
8.1	Skiing days at Bjørnholt (circles) with best fit CP model (black line) and best fit WoC model (red line). The two regression lines almost overlap.	86
8.2	Null distribution for $H_0 : d = 1$. H_0 is rejected with a p -value of 0.03 if f_L and f_R has unique variance, and with a p -value of 0.045 with constant variance. Computed using 1000 samples.	87
9.1	Profiled log-likelihood for the multivariate normal model. The MLE is marked in red. The edge in the upper left corner is the edge of the search space, and should not be interpreted that the likelihood is zero outside it.	92

9.2	Posterior distributions for d and τ calculated by normalizing the profiled log likelihood then marginalizing.	92
9.3	Traces from the five MCMC runs for the multivariate normal model. The red traces are a , and the blue traces are b . Starts where randomized but close to their mean value. Starting further away, i.e. letting say $a < 200$ only resulted in slower convergence, while having the same stationary distribution. Using $n_{sim} = 1000$ samples.	94
9.4	Posterior probability that chapter i was written by Joan de Galba. It shows clearly that Martorell was solely writing the book until about $i = 320$, then more and more chapters were written by Joan de Galba, but Martorell was influential until late into the book. At chapter 378 was he still overwhelmingly the most likely author (red dot). The first chapter almost certainly written by Joan de Galba was chapter 322 (blue dot)	94

List of Figures

List of Tables

1.1	Posterior summary for British Mining Disasters	16
6.1	Chain diagnostics for the Gibbs sampler using five chains and 1100 samples where the first 100 samples where discarded. Except the Gelman-Rubin statistic where all values calculated using the combined chains sans burn in.	70
6.2	Chain diagnostics for the Metropolis-Hastings sampler using five chains and 10100 samples where the first 100 samples where discarded. That is, 10 times the number of samples compared to the Gibbs sampler.	70
7.1	Critical values for $\alpha = 0.05$	80
8.1	Table of summary statistic of Bjørnholt data set for different models. All assume normal deviates. The Stationary assumes no change, the linear assumes a single linear trend, CP_{CV} assumes two linear trends which intersect at the change point with constant variance, CP assumes the same, but allow the variance to change as well. WoC_{CV} assumes the two linear trends with a WoC and constant variance, while WoC allows the variance to change as well.	85
9.1	Chain diagnostics for the MCMC.	93

List of Tables

Preface

Once, I was presented with an idea, a model, and a problem from my supervisors, which eventually became this thesis. The model in question became known as a ‘Window of Change’ model. Now, I am left with even more ideas, a family of models, and a staggering number of new questions.

If I were to do this project again, I would do everything differently. This is because this thesis was actually written in reverse. When I first started this project, I was presented with an interesting model, an idea my dear supervisors had conceived, along with a suitable data set, and I was tasked with trying to understand the model. I began writing algorithms, estimating parameters and getting comfortable with the model, but there was something off. My estimates seemed wrong; wildly changing depending on parameters which initially seemed inconsequential. I started simulating data, so I could know the true parameters, and the estimates were far too low! ‘How strange’ I thought. ‘Where is the bias coming from?’ became the dominating research question throughout the thesis, and much effort was expended trying to isolate it. Eventually, in this pursuit, the model became so simple that no reasonable real data could possibly be distributed according to it, yet the bias remained.

At this point, we had worked backward, starting from what became this manuscript’s last chapter to what we finally baptized ‘The simplest interesting example.’ This example became the locus of the thesis, and we developed ideas and theory trying to understand it.

If I were to write this thesis again, I would begin here, with this seemingly simple but rich example, and I consider having found this example one of my chief contributions. It is a fitting locus since it manages to remove all aspects of the model in question that it shares with its close cousin, Change-Point models. Furthermore, it also unifies the two different Window of Change models we are concerned with, colloquially known as the ‘Nils model’ and the ‘Céline model’ (later professionalized to the ‘parameter transition’ and the ‘mixture transition’), making it almost a poetic start.

I would like to deeply thank Nils Lid Hjort and Céline Cunen for the great support and guidance they have offered me. In addition to inspiring and enlightening discussions, you have provided fun, thoughtful, and compassionate conversations.

Chapter 1

Introduction

1.1 Motivation

A common task for a statistician is to decide whether something has occurred, which amounts to a change of some kind. The related problem of identifying when a change has occurred is the locus of this thesis. Change-Point analysis has been developed to address this in the specific case where something is first constant and then abruptly changes to a new constant. In the canonical example, the change occurs at a point in time, and we ask: ‘When did the new normal come to be?’ With this question in mind, we can construct models, perform inference, and calculate the corresponding uncertainties.

In the case where one stationary distribution suddenly changes into a new stationary distribution, a Change-Point model is appropriate. Even when this assumption is not immediately justifiable, Change-Point models have the advantage of being simple both conceptually and operationally. In particular, the interpretation of the Change-Point itself is very clear and intuitive. This makes them favourable when inference of this character is demanded.

There is, however, a substantial problem with Change-Point analysis, namely that it postulates a discontinuous jump from old to new. In many cases, it is absurd to assume that such a transition is possible, which motivates extensions that solve the deficiency. The Window of Change model developed in this thesis is designed to include a gradual change, but still retain the stationary properties of Change-Point models before and after the change point itself. In other words, the distribution is initially stationary and also eventually stationary. Only in what we call ‘the window’ is the distribution allowed to change. Thus, it eliminates the discontinuous jump while retaining a stationary distribution before and after the window. Even though the model is constructed as an extension of Change-Point analysis, this alteration fundamentally changes the nature of the model, and we will show this in this thesis. The question which Window of Change models answer is thus a slight adjustment of the one for Change-Point analysis: ‘When did the old normal end, and when did the new normal begin?’

1.2 Related work

This thesis is situated in the literary context of Change-Point analysis. The reason is that it is from this context that the Window of Change model was conceived. In a fundamental way, Window of Change analysis is different from Change-Point analysis, and there are other avenues from which one could have developed a Window of Change model instead. There are still good reasons to present it this way: the nature of the

inference is similar, and accordingly, the datasets we will apply it to have been analysed with Change-Point models by others.

Cunen, Hermansen and Hjort (2018) constructed confidence curves for the change point. They primarily used two methods to do so, namely homogeneity tests and profiled log-likelihood functions. The profiled log-likelihood methods yielded good results and form the foundation upon which we shall build. Another contribution they offer is a comparison between Bayesian and frequentist analyses in Change-Point problems. The comparison, based on a posterior approximation of the left and right distribution, demonstrates a certain bias towards the extremities for the Bayesian when estimating the change point.

Carlin, Gelfand and Smith (1992) provides the Bayesian framework that we extend. The crucial insight in their paper is the construction of an efficient and general Gibbs sampler for Change-Point problems. It is so general that as long as one can sample from the left and right stationary distributions adequately well, the entire Change-Point model can be sampled with ease. The assumptions made in this thesis are that the stationary distributions themselves are easy to work with and that only the matter of estimating the change point is of any particular difficulty. The precise meaning of this will be clarified later. The reason for this assumption is that they were first and foremost interested in the specific kind of inference that Change-Point analysis enables. The change point itself is of primary interest, that is, when the data stops being generated by the first model, and starts being generated by the second model. There is also a technical reason for this assumption, namely that inference on the left and right distributions is ‘standard’. What this means will be clear in Chapter 3. Correspondingly, we are most concerned with inference on the window, and assume that estimating the left and right distribution is ‘standard’.

Finally, Hinkley (1970) formulates the Change-Point problem via random walks, and with this representation finds the asymptotic sampling distribution of the maximum likelihood estimator for the change point. In addition, he shows that the change point is a location parameter, symmetric under shifts, as remarked by Cobb (1978). Cobb constructed an expression of the sample distribution for the change point with an approximate ancillary statistic. The approximation can be made arbitrarily good, and this means that virtually all information regarding the sampling distribution for the change point is contained in the statistic. In other words, nuisance parameters, like the parameters of the left and right distributions, do not influence the distribution of the statistic, thereby avoiding the complexity of the test statistic depending on some estimated quantities. Furthermore, he shows that this statistic is numerically identical to the Bayesian posterior, and thus provides a different comparison between the frequentist and Bayesian.

1.3 Overview

This thesis concerns Window of Change models from multiple perspectives.

Chapter 1 contains background information, notation, and central definitions. It also contains an illustrative example analysis of a data set, which serves as a brief introduction to both Change-Point analysis, and Window of Change analysis. It shows how the methods in Change-Point models can be extended and what the implications are. Finally, this chapter discusses some philosophical and mathematical differences between frequentist and Bayesian analysis, which are of greater-than-normal importance in this thesis.

Chapter 2 covers some preliminary notation and concepts that are necessary for discussing two estimation algorithms, but are otherwise not central to the thesis.

Chapter 3 formally introduces the Window of Change model constructed via a linear weight function. This weight function allows for many ways of defining transition models, but we mainly look at two natural, yet quite different variants. These two transition models are defined, compared, and contrasted, and some elementary properties of them are derived. We then show how estimating the left and right distribution differs from estimating when the window begins and ends. Which serves as the justification of how the former is easy, and the latter is difficult. Estimation of the window length is then extensively discussed, and bias correction procedures are developed. We also discuss how to select priors for the location of the window, and how this might not be as trivial as it first appears. The Bayesian discussion continues with a posterior approximation, which serves as a good comparison base for comparing frequentist and Bayesian analyses. Additionally, we deduce a random walk representation for Window of Change models. Finally, we briefly show how it is possible to extend these transitions to a large family of transition models through a transformation of variables, which shows that the model is more general than simply linear change.

In Chapter 4, we develop a very simple instance of a Window of Change model, which still has some surprising and interesting properties. It is simple in the sense that all the randomness is happening in the window, and everything is deterministic outside it. One could then think that estimation would be simple, in the sense that estimation can be done with little error, but it becomes apparent that this is not the case. Furthermore, a surprising bias is shown to exist, and we develop methods of correcting this bias.

Since real-world data also exhibit stochasticity outside the window, we must consider the case where the left and right distributions are more complicated. Therefore, Chapter 5 extends the very simple example with data sampled from two known normal distributions. With this extension, the different properties of the two transitions also manifest, and we consider them separately. We see that the corrections can be extended, but they are less successful than in the very simple case.

Chapter 6 develops methods for estimating the left and right distribution when the parameters are unknown. This is, of course, a necessity for using the model on real data. It considers maximum likelihood estimation and Bayesian sampling algorithms, and it shows how to implement these methods effectively.

Chapter 7 concerns the problem of identifying when a Window of Change model is preferable to a Change-Point model. We test against Change-Point models as they are the natural contenders to Window of Change models. For the frequentist, this is done through hypothesis testing, while for the Bayesian, it is done by the posterior probability. We then compare their power and coverage, i.e. whether they reject the null hypothesis at the prescribed level. Finally, we compare hypothesis testing with model selection frameworks, which are a common substitutes for formal hypothesis testing. Normal model selection criteria, like the Akaike information criterion and the Bayesian information criterion, are, however, not immediately applicable when comparing Change-Point models and Window of Change models.

Window of Change models are applied to two real data sets: skiing days at Bjørnholt and word counts in *Tirant lo Blanch*. In Chapter 8, we analyse the number of skiing days at Bjørnholt using a regression model. The analysis demonstrates how Window of Change models can be used for regression models where the change occurs through a dependent variable, and not only the index.

In Chapter 9, we analyse the novel *Tirant lo Blanch*, where the observations are

10-vectors of word counts per chapter. This analysis, in particular, shows how the estimated window length can be highly dependent on the data model as well as the transition model. It also demonstrates how even high-dimensional estimation can be performed efficiently with the tools developed in this thesis.

Finally, we conclude in Chapter 10, summarizing the findings and commenting on future work.

1.4 Background and notation

This section serves three purposes: It introduces the notation we will use throughout the thesis; in doing so, it also presents some elementary definitions in statistics and probability; and finally, it presents some non-standard constructions that are central to this thesis.

Since proper statistical notation can be very detailed, it can be cumbersome to use, especially for simple problems. We will therefore also introduce some shorthand notation that will be used when there is no ambiguity.

We work with observation (or data) vectors $y = y_1, \dots, y_n$ of a given size n . Here y represents the observation vector, and n always denotes the length of such an observation vector. While the notation suggests that y_i is a scalar, this is not necessarily true, and we allow it to be a vector, which in turn means that y is a matrix. It is common to treat y_i notationally as a scalar and y notationally as a vector, and we will do the same. Sometimes, observations come with covariates, which we denote as $\mathbb{X}$. As before, this notation suggests that the covariates are an $n \times m$ matrix, which is commonly the case, but we also include $m = 1$, in which case there is only one covariate x_i for each observation y_i .

The Greek letter θ is reserved for representing parameters of a distribution, while Θ is the set of possible parameter values. It may, and often is, a vector. We use the subscripts θ_L and θ_R to indicate that the former is the parameter for the distribution left of the change point (or window) and right of the change point (or window). Furthermore, we call these the *left* and the *right* distribution, which we denote f_L and f_R . On the other hand, if there are no subscripts, θ denotes the complete set of parameters for a given distribution or model. For example, $\theta = (\tau, \theta_L, \theta_R)$ is the complete parameter vector for a Change-Point problem. Furthermore, if θ_L and θ_R are of dimension m , then this θ is of dimension $2m + 1$.

If a random X variable is distributed according to a function f_X , parametrized by θ , we write

$$X \sim f_X(x, \theta).$$

When it is clear which random variable the function concerns, we omit the subscript and let the function arguments (i.e. x in the equation above) indicate which random variable it belongs to. If X is discrete and x is a sample, then $f(x; \theta) = \Pr(X = x \mid \theta)$ and f are called a probability mass function (PMF), while if X is continuous, $f(x; \theta) = dF(x; \theta) / dx$ where $F(x) = \Pr(X \leq x \mid \theta)$, and f is called a probability density function (PDF).

We will use the clearer, but more cumbersome notation if the intricacies of the problem require careful treatment. When doing Bayesian analysis, the letter π is often used instead of f to denote probability functions, and the function arguments indicate which parameter the distribution corresponds to. For example, $\pi(x) = f_X(x)$ and $\pi(x \mid \theta) = f_X(x; \theta)$. The posterior distribution for a parameter θ , given some observations y is $\pi(\theta \mid y)$.

We continue by defining the likelihood function, which is central to all methods in this thesis. Introduced by Fisher (Fisher & Russell, 1997) it serves as the primary mathematical object through which inference is conducted. The advantage of this construction is that it fixes the random observations, Y , as ‘parameters’ of the joint probability distribution, and treats the old parameters, θ , as the functional argument. It inverts the relationship between observations and parameters.

Thus, we can quantify the probability of the specific observed outcome as a function of the possible parameters, but this is not a true probability; we never observe samples of the parameters. To differentiate true probability from this inversion, the term *likelihood* is used instead.

Definition 1.4.1 (Likelihood). Given a set of independent observations y with a probability function (e.g. PMF or PDF) f_Y and parameter θ , the likelihood of the observation is

$$L(\theta; y) = \prod_{i=1}^n f_{Y_i}(y_i; \theta).$$

If Y_i are identically distributed, we may write

$$L(\theta; y) = \prod_{i=1}^n f(y_i; \theta).$$

Often, working with the logarithm of the likelihood is more convenient.

$$\ell(\theta; y) = \log L(\theta; y).$$

We let $\log 0 = -\infty$ in this case since we allow our likelihood to be zero. The likelihood being zero for some given value of θ simply implies that the observations are impossible for this value.

When certain parameters are of greater interest than others, we can work with *profile log-likelihoods*, which reduces the domain of the likelihood function to just the set that concerns us.

Definition 1.4.2 (Profiled log-likelihood). Given a parameter vector $\theta = (\nu, \phi)$, where ν is a nuisance parameter and ϕ is the focus parameter, define the profile log-likelihood function as

$$\ell_{\text{prof}}(\phi; y) = \max_{\nu} \ell(\phi, \nu; y).$$

Both ν and ϕ can, and often are, vectors.

In Change-Point problems, as well as Window of Change problems, the change point itself is the focus parameter, while parameters of the distributions themselves can be treated as nuisance parameters. They are, of course, still of interest, but inference on them is easier and of a different nature than that of the change point.

The deviance function is used in many of our tools.

$$D(\phi, y) = 2(\ell_{\text{prof}}(\hat{\phi}; y) - \ell_{\text{prof}}(\phi; y)).$$

The first term, $\ell_{\text{prof}}(\hat{\phi}; y)$, is simply the maximum log-likelihood, and the second term $\ell_{\text{prof}}(\phi; y)$ is the profiled log-likelihood of the focus parameter ϕ .

In particular, the deviance function is used to construct confidence curves.

Definition 1.4.3 (Confidence curve (Cunen, Hermansen & Hjort, 2018)). A confidence curve for ϕ , $\text{cc}(\phi)$, is a function over a suitable parameter space Φ where the function value is at what significance level $\text{cc}(\phi) = 1 - \alpha$, we can reject the hypothesis test $H_0 : \phi = \hat{\phi}$.

We will only use the deviance to construct confidence curves, and may therefore let cc refer only to the following expression

$$\text{cc}(\phi) = \Pr_{\nu}(D(\phi, Y) \leq D(\phi, y))$$

Letting the parameter be ϕ rather than θ highlights that we are only interested in confidence for certain parameters, i.e. the focus parameters. We prefer to work with significance, and not coverage throughout this thesis. What this means is that $\alpha = 0.05$ controls the type 1 error, i.e. the probability of erroneously rejecting the null hypothesis. Then, $1 - \alpha$ is the coverage of e.g. confidence sets. The cc can be used to make confidence sets directly:

Definition 1.4.4 (Confidence set). The random set $C_{1-\alpha}$ is a $1 - \alpha$ -confidence set for θ if

$$\min_{\theta \in \Theta} \Pr(\theta \in C_{1-\alpha}) \geq 1 - \alpha$$

The minimum in the definition shrinks the set to the smallest possible set, which in turn makes the confidence set the strongest possible claim. (Clearly, if $\Pr(\theta \in C) \geq 1 - \alpha$, then $\Pr(\theta \in C \cup A) \geq 1 - \alpha$ for any A . Thus, we can always enlarge our set while fixing confidence.)

We could, of course, construct the cc from any other statistic, but choosing the deviance has the nice property of being the uniformly most powerful test in many cases. This means that the deviance test correctly rejects the null hypothesis most often. This is proven in the famous Neyman-Pearson lemma (Neyman, Pearson & Pearson, 1997). However, we are not in one of these cases, which we will examine later.

We conclude this section by introducing some Bayesian constructions that serve as a natural counterpart to confidence curves. To do this, we need the term *credibility*, which is the Bayesian analogue to *confidence*. For example, a 95% credibility interval for, say, θ is a fixed interval $C_{B,0.95}(\theta | y)$ for which the probability that θ is contained within it is 0.95:

$$\Pr(\theta \in C_{B,0.95}(\theta | y)) = 0.95.$$

In contrast, a 95% confidence interval for θ , $C_{0.95}(\theta)$, is a random interval satisfying the same identity

$$\Pr(\theta \in C_{0.95}(\theta)) = 0.95.$$

We wish to construct such sets and suggest using the *credibility distribution*, which is the cumulative posterior. Later, we will also consider using the highest posterior density region, which is the smallest region containing at least $1 - \alpha$ of the posterior density/mass. For a one-dimensional parameter θ , the cumulative posterior is

$$CD_B(\theta, y) = \int_{-\infty}^{\theta} \pi(\theta' | y) d\theta'$$

where $\pi(\theta' | y)$ is the posterior for θ given the data, but rewritten with θ' to differentiate the function argument from the integration limit. This distribution is, in fact, equivalent to the frequentist confidence distribution (in the sense that confidence and credibility

are equivalent); hence, we call it CD_B . With this, we can construct the fully Bayesian credibility curve:

$$cc_B(\phi) = |1 - 2CD_B(\phi, y)|.$$

Again, letting the parameter in question be called ϕ highlights that we are only concerned with the focus parameters. The credibility curve indicates at which level of credibility a value $cc_B(\theta)$, θ is excluded from a central credibility interval. The intervals are centered around the median, which is readily apparent as $cc_B(\theta_{median}) = 0$, i.e. never rejected.

1.5 Frequentist and Bayesian frameworks

This section briefly summarizes Bayesian and frequentist statistics.

Frequentist and Bayesian analyses are frequently contrasted in their interpretation and methods. The early history of statistics can be seen through the development of this debate. While Reverend Bayes pioneered the field of mathematical probability calculations, many were unsatisfied with certain assumptions of prior distributions. These assumptions are necessary in the mathematical formalism, but their choice is of consequence in the final inference, and even worse, no choice seems to be more correct, the least informative, or somehow more privileged than many others.

Example 1.5.1 (A coin toss and prior belief). Consider the case of a coin toss with probability p of getting heads. That is, $\Pr(X = 1) = p$ if we let X denote the coin toss and 1 be the outcome ‘heads’. Then, the Bayesian analyst requires a prior distribution on p , $\pi(p)$. Any probability distribution is of course permissible, but if we wish to say as little as possible about the coin before observing it, we may employ some principles of ignorance. A very common choice is built upon the notion that if we do not know anything at all about p beforehand, then any value for p should be equally likely. Thus, we conclude that the least assumptious prior for p should be flat, that is, $\pi(p) = 1$. This was Thomas Bayes’ original argument, and the result is that the best estimate for p after having observed say $x = 1$, is $\hat{p}_B = 0.75$. We see that even though only heads have been observed, $p = 1$ is not the resulting ‘belief’, but we still favour a large p . This resonates with the fact that it is common to assume that a coin can land on both sides ($p = 1$ would forbid this, which is unreasonable), and the fact that having observed more heads than tails, we are inclined to believe that $p > 0.5$.

But there is a fatal flaw in this argument. Namely, what is considered constant *depends on the parametrization of p* . A ‘flat’ belief about the *probability* of heads results, for instance, in a ‘curved’ belief about the *odds*, of heads. Given the above case, a flat prior on the odds would not yield any inference at all, as it results in an ill-defined posterior distribution.

When we construct, formalize and work with mathematical formulae, we frequently choose parameters which are convenient to work with to do a specific task. For example, we often choose to work with *mean* and *variance* for computations in the normal distribution, while perhaps standard deviation is more commonly chosen when presenting the results. This is unproblematic because one can show, usually with some elementary algebra, that the formulations are equivalent. The problem occurs when the choice of otherwise equivalent mathematical expressions somehow *informs our belief* about which it models. Should our belief about p be chosen based on the fact that probability is more convenient to compute with? The reasonable answer is ‘no’, mathematical convenience is not a justification for one’s belief. The faithful Bayesian could still defend her prior belief by affirming that $\pi(p)$ in fact is her prior belief. Furthermore, she could interject

that since choosing a prior also means choosing a parametrization, is she justified in making *a* choice.

The problem of priors is not that the parametrization affects the posterior, but rather to choose a prior without regard to this very fact. ▲

The consequence of this necessary seemingly arbitrary choice of priors led to the development of *frequentist statistics*. This development interpreted probability not as a subjective belief, but rather as the long-term frequency of outcomes, hence the name. The shift in interpretation made the problematic ‘prior’ beliefs obsolete, but it inadvertently also reversed the roles of data and parameters. In the frequentist paradigm, parameters are *fixed*, while data is *random*. All data is assumed to come from some typically abstract super-population, and the repeated frequencies give rise to notions of probability.

1.6 Change-Point analysis

Change-Point analysis has been extensively studied, and one way of defining the model is the following:

Definition 1.6.1 (Change-Point Model). Let f_L and f_R be two probability distributions, possibly parametrized by θ_L and θ_R . If there exists a τ such that

$$\begin{aligned} Y_i &\stackrel{i.i.d.}{\sim} f_L, & i \leq \tau, \\ Y_i &\stackrel{i.i.d.}{\sim} f_R, & \tau < i, \end{aligned}$$

(where ‘i.i.d.’ stands for ‘independent identically distributed’) then Y_i follows a Change-Point model with τ being the change point.

We use the term ‘Change-Point’ or the abbreviation CP when referring to the model, and the term ‘change point’ when referring to the point itself.

CP models have certain interesting properties. Firstly, the observations are dependent, which means that their order matters and that the likelihood cannot be factored as one is used to. The observations are, however, conditionally independent, which means that if you know τ , the likelihood can be factored. Dependent data is common in time series analysis; one might call it a defining property, but the particular conditional independence is not. In practice, τ divides the data sets into two separate data sets, which are suddenly independent and can be treated as i.i.d. samples from separate probability distributions.

Another very interesting and relevant property, which showcases the non-standard nature of CP problems, is that any estimator for $\hat{\tau}$ cannot be consistent. Consistent in this context means that as the number of observations arbitrarily increases (such that $\tau \rightarrow \infty$ and $n - \tau \rightarrow \infty$), $\Pr(\hat{\tau} = \tau_0) < 1$. This implies that regardless of how much information we have, there will always be a positive probability that the maximum likelihood estimate is not the true value. The reason for this is that observations far from the change point contribute increasingly little information about its location. The story is different for the parameters θ_L and θ_R , which can be consistently estimated under the conditions above.

Furthermore, in standard settings, we are used to the fact that the asymptotic distribution of the maximum likelihood estimator (MLE) is distributed according to a normal distribution centered around the true parameter and with a particular covariance matrix. In Bayesian analysis, we are conversely used to the Bernstein-Von Mises theorem,

which shows that under standard conditions the posterior $\pi(\theta \mid y)$ also converges to a certain normal distribution, and furthermore, that this posterior converges (in distribution) to the frequentist MLE distribution. In particular, for standard settings, we are used to the fact that frequentist confidence intervals and Bayesian credibility intervals concur if n is large enough. If they did not agree even in the limit, being a ‘frequentist’ or a ‘Bayesian’ statistician would be a much more meaningful decision. None of these results hold for estimating the change point, and consequently, choosing the frequentist or the Bayesian paradigm has lasting consequences.

Since it is an uncommon feature for the parameter of interest to be an index, we might wonder what the natural mathematical representation is. Hinkley (1970) showed that it can be represented as random walks.

Lemma 1.6.2 (Hinkley (1970)). *The log-likelihood of a CP model can be expressed as*

$$\ell(\tau; y) = \sum_{i=1}^{\tau} (\log f_L(y_i) - \log f_R(y_i)) + \text{const}$$

where $\log f_L(y_i)$ and $\log f_R(y_i)$ are entirely independent of τ .

Proof. Clearly, the distributions f_L and f_R are independent of τ . We continue by rearranging the log-likelihood function:

$$\ell(\tau; y) = \sum_{i=1}^{\tau} \log f_L(y_i) + \sum_{i=\tau+1}^n \log f_R(y_i)$$

And manipulate accordingly

$$\begin{aligned} &= \sum_{i=1}^{\tau} \log f_L(y_i) - \sum_{i=1}^{\tau} \log f_R(y_i) + \sum_{i=1}^{\tau} \log f_R(y_i) + \sum_{i=\tau+1}^n \log f_R(y_i) \\ &= \sum_{i=1}^{\tau} (\log f_L(y_i) - \log f_R(y_i)) + \sum_{i=1}^n \log f_R(y_i) \\ &= \sum_{i=1}^{\tau} (\log f_L(y_i) - \log f_R(y_i)) + \text{const}. \end{aligned}$$

Which is the desired result. $\square$

Clearly, this formulation also applies to profiled log-likelihoods. It is important to note that τ only appears in the limits of the sum and not in the summands. This means that the log-likelihood is equivalent to a random walk, where the increments are distributed according to the ratio of log-likelihoods for any given observation of the left and right distributions.

1.7 An inviting example

We now conclude the introduction with a Window of Change (WoC) analysis of an appropriate data set, highlighting the differences between CP and WoC analysis. We begin by introducing the data, reiterating the CP analysis, and in the next section, we present the WoC analysis.

An interesting data set, sometimes used in the CP literature, is the British Mining Disasters data set (Jarrett, 1979). It contains the intervals between coal-mining disasters

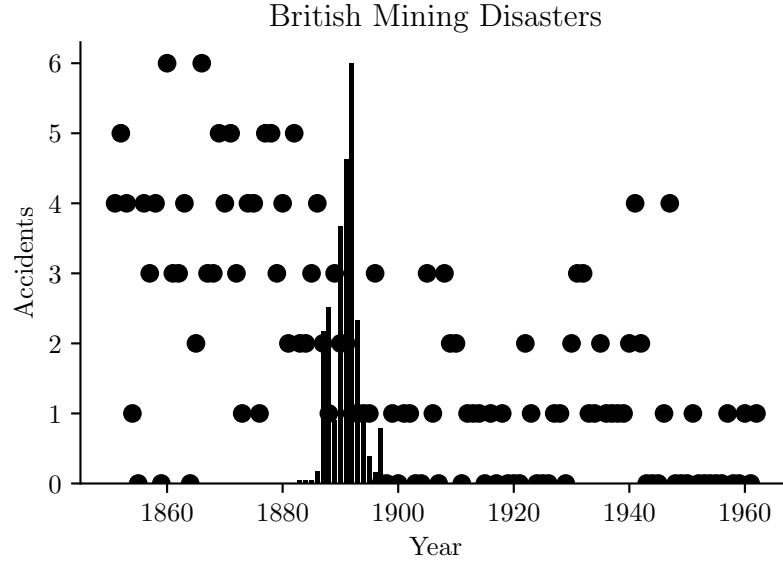


Figure 1.1: Yearly accidents counts in British mines (dots) and the profiled likelihood function for the change point using a Poisson model (bar plot). The likelihood function is scaled such that its maximum is 6 making it comparable to the data.

from 15th March 1851 to 22nd March 1962. It is, however, common to consider the count data for each year instead, as we shall do now.

The typical modelling is something like this (see Carlin, Gelfand and Smith (1992), Cunen, Hermansen and Hjort (2018)): Let $f_L = \text{Poisson}(\theta_L)$, $f_R = \text{Poisson}(\theta_R)$, and $\tau \in \{1, \dots, 112\}$ be the possible change points counting from the starting year 1851. The choice of a Poisson model is natural for positive count data. It implies that there is a constant probability of an accident at any given time. The Poisson PMF is as follows:

$$f(y; \theta) = \frac{\theta^y \exp\{-\theta\}}{y!}$$

and it has expectation θ . Alternatively, the CP model can be expressed as $Y_i \sim \text{Poisson}(\theta_i)$ where

$$\theta_i = \begin{cases} \theta_L, & i \leq \tau \\ \theta_R, & \tau < i. \end{cases} \quad (1.1)$$

The likelihood function becomes:

$$\begin{aligned} L(\tau, \theta_L, \theta_R; y) &= \prod_{i=1}^{\tau} f(y_i; \theta_L) \prod_{i=\tau+1}^n f(y_i; \theta_R) \\ &= \prod_{i=1}^{\tau} \frac{\theta_L^{y_i} \exp\{-\theta_L\}}{y_i!} \prod_{i=\tau+1}^n \frac{\theta_R^{y_i} \exp\{-\theta_R\}}{y_i!} \\ &= \theta_L^{\sum_{i=1}^{\tau} y_i} \theta_R^{\sum_{i=\tau+1}^n y_i} \exp\{-\tau\theta_L - (n - \tau)\theta_R\} \prod_{i=1}^n \frac{1}{y_i!} \end{aligned}$$

Then, the maximum likelihood estimate for the change point is $\hat{\tau} = 41$, which equates to the year 1891, and for the nuisance parameters $\hat{\theta}_L = 3.1$, and $\hat{\theta}_R = 0.9$.

We construct a confidence curve for τ using Cunen, Hermansen and Hjort (2018) recipe. The resulting confidence set for $\alpha = 0.05$ is $C_{0.95}(\tau) = \{1887, 1888, 1889, 1890, 1891, 1892, 1893, 1894, 1897\}$, meaning we cannot reject the hypothesis that τ_0 is any of the listed values. Here, we parametrize the confidence set in years for clarity, rather than using the index, which is strictly correct given our notation.

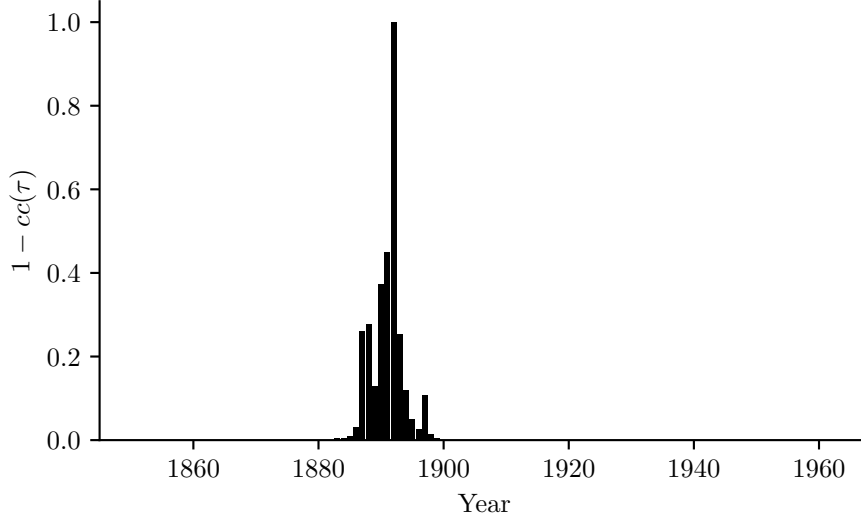


Figure 1.2: Inverted confidence curve for the change point τ . Notice how it clearly shows the maximum likelihood estimate with its single peak. It also slightly follow the shape of the likelihood function, with what appears to be three modes. The fact that it has three modes shows how a confidence curve does not necessarily imply a confidence distribution, i.e. cumulative confidence, and is why we choose to work with the cc.

A Bayesian analysis is in order. Under the same data-generating model, i.e. Poisson, the model uses the same likelihood. While we are free to choose any data model in the Bayesian analysis, using the Poisson ensures that the likelihood is identical to the frequentist analysis, and furthermore that the parameters have the same interpretation (although the interpretation of what *a parameter is* still differs). In addition to selecting a data model, we must elicit a prior for all parameters, which in this case is τ, θ_L, θ_R . We assume, as is common in CP analysis, that it can be factored accordingly $\pi(\tau, \theta_L, \theta_R) = \pi(\tau)\pi(\theta_L)\pi(\theta_R)$. Constructing a prior is, of course, a good opportunity to include prior information, but we are currently concerned with little- or non-informative priors. The reason is that we currently wish to let the data ‘speak for itself’, which makes the comparison with the frequentist analysis more appropriate.

A common choice of prior for Poisson data is a gamma prior since it is conjugate. The PDF is

$$f(\theta; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}$$

for positive parameters α and β . Conjugacy means that the prior and posterior are of the same probability distribution, but with different parameters. In a Gamma-Poisson model, if $Y \sim \text{Poisson}(\theta)$ and $\theta \sim \text{Gamma}(a, b)$, then $\theta \mid Y \sim \text{Gamma}(a+n, b+n\bar{Y})$. The conjugacy is not immediately advantageous since the full CP model is not conjugate, but we shall see that it does indeed simplify things later. The prior on τ is of more interest

because it will not be 'washed away' by the likelihood. This is due to the limited information the observations contain regarding τ . We still suggest that a flat prior is adequate and let $\pi(\tau) = \text{const.}$ Using $\mathbb{E}[\theta_L] = \mathbb{E}[\theta_R] = 2$ and $\text{Var}[\theta_L] = \text{Var}[\theta_R] = 5$, we find that $\alpha = 5$ and $\beta = 5/2$, and we have fully specified the model. We can now continue with the results.

The Bayesian analysis strongly concurs with the frequentist analysis. The posterior mode is identical to the ML estimate at $\hat{\tau} = 41$, while the posterior mean is slightly lower. The posterior median is $\hat{\tau} = 40$, due to the slight skew in the posterior distribution. Finally, the centred posterior credibility set for the significance level $\alpha = 0.05$, $C_{B,0.95}(\tau) = \{1887, 1888, 1889, 1890, 1891, 1892, 1893, 1894, 1895, 1896\}$, is very similar to the confidence set, but it is contiguous. This property is a result of constructing it from a credibility distribution centred around the median. The resulting inverted credibility curve is therefore guaranteed to be unimodal, and the confidence set is consequently guaranteed to be contiguous.

In other words, the conclusion of this simple but representative analysis suggests the following: Before 1891, the expected number of accidents was just over 3, and abruptly after, the expected number of accidents is reduced to just below 1. Thus, if we are to look for an explanation for this change, we should investigate causes around the year 1891, but possibly as early as 1887 and as late as 1897.

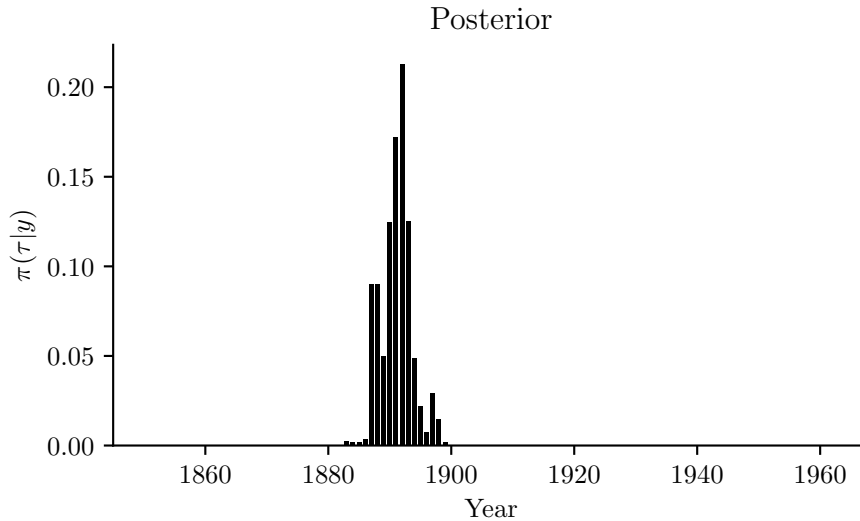


Figure 1.3: The posterior distribution for the change point. This too closely mimics the profile likelihood function, suggesting that there is little posterior uncertainty in θ_L and θ_R . In fact, $\mathbb{E}[\theta_L | y] = 3.13$, $\text{Var}[\theta_L | y] = 0.086$ and $\mathbb{E}[\theta_R | y] = 0.915$, $\text{Var}[\theta_R | y] = 0.013$.

1.8 The new twist

Although the accident rate clearly sharply declined (see figure 1.1), one might wonder if it is reasonable to believe, and consequently model, that all the change occurred from one year to the other. Some causal explanation for this decline is a decrease in labour productivity, an increase in unionizing and a reduction of overtime work (Raftery and Akman (1986), c.f. Taylor (1968)). Even with rapid sociological and political development, a change this abrupt seems unrealistic. Without further comment on the

historical material conditions of coal miners in Britain, we will note that similar kinds of explanations might be common when doing CP analysis. Ultimately, if the believed causes are incompatible with instantaneous change, the model does not reflect our belief, and is inappropriate.

This reflection justifies an amended model where the change is allowed to be gradual; however, it is still constant far from the WoC. In particular, for $i \leq a$, let $y_i \sim f_L$ and for $b < i$, let $y_i \sim f_R$, with the condition that $a < b$. Thus, the model still assumes independent identically distributed (i.i.d.) observations from a Poisson distribution for the early observations, and for the late observations (although with a different parameter). The indexes a and b then demarcate a window in which the parameters are allowed to change however we like.

A natural, as well as simple, manner of changing is letting θ_L change linearly to θ_R . Let therefore

$$\theta_i = \begin{cases} \theta_L, & i \leq a \\ \theta_L \left(\frac{b-i}{b-a} \right) + \theta_R \left(1 - \frac{b-i}{b-a} \right), & a < i \leq b \\ \theta_R, & b < i \end{cases} \quad (1.2)$$

and

$$Y_i \sim \text{Poisson}(\theta_i).$$

Later, we will discuss other natural alternatives to this transition model. This *weight function* is useful:

$$w_i = w(i; a, b) = \begin{cases} 1, & i \leq a \\ \frac{b-i}{b-a}, & a < i \leq b \\ 0, & b < i \end{cases}.$$

Equation (1.2) then becomes $\theta_i = w_i \theta_L + (1 - w_i) \theta_R$ for fixed a, b .

Before we continue the analysis, some comments on θ_i are in order. First, $\theta_a = \theta_L$ as $w_a = 1$. Similarly, $\theta_b = \theta_R$ since $w_b = (b-b)/(b-a) = 0$. Finally, the special case where $b-a=1$, equation (1.2) reduces to (1.1), so we see that a CP model is indeed a WoC model. This can be seen by noticing that if $b = a+1$, then $w_i = 1$ for $i \leq a$, and $w_i = 0$ for $a < i$. Thus, we see that we have a CP model where a serves as the change point. (In fact, $a+0.5$ would be the ‘change point’ in the WoC analysis parametrization, but the conclusion is the same.)

We may now construct the likelihood function:

$$\begin{aligned} L(a, b, \theta_L, \theta_R; y) &= \prod_{i=1}^n f(y_i; \theta_i) \\ &= \prod_{i=1}^a f(y_i; \theta_L) \prod_{i=a+1}^b f(y_i; w_i \theta_L + (1 - w_i) \theta_R) \prod_{i=b+1}^n f(y_i; \theta_R) \\ &= \theta_L^{\sum_{i=1}^a y_i} \prod_{i=a+1}^b (\theta_L + (1 - w_i) \theta_R)^{y_i} \theta_R^{\sum_{i=b+1}^n y_i} \\ &\quad \cdot \exp \left\{ -a\theta_L - \sum_{i=a+1}^b (w_i \theta_L + (1 - w_i) \theta_R) - (n - b) \theta_R \right\} \prod_{i=1}^n \frac{1}{y_i!} \end{aligned}$$

Which somewhat resembles the one for CP models. The chief difference is that instead of the change point parameter τ only appearing in the limits of the products, the window parameters a, b also influence the factors themselves. This change non-trivially

complicates the problem. For now, however, it suffices to say that the function is well-defined, and the maximizer is unique.

With this model, we obtain the maximum likelihood estimates $\hat{a} = 32$ and $\hat{b} = 47$, corresponding to year 1883 and 1898, which suggests that the change was far from immediate. The maximum likelihood was $\ell_{\text{prof}}(\hat{a}, \hat{b}) = -167.8$, which is about one more than for the CP model $\ell_{\text{prof}}(\hat{\tau}, \hat{\tau} + 1) = -168.86$. Solving for the profiled log-likelihood is,

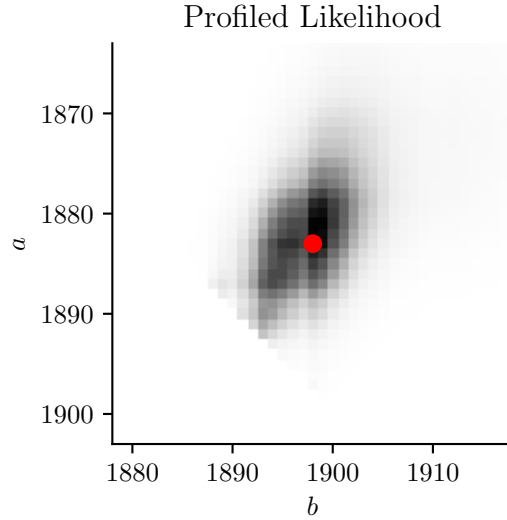


Figure 1.4: The profiled Likelihood with the maximum likelihood estimate marked in red. It is scaled such that white is zero and black is the maximum likelihood. The maximum likelihood estimate is $\hat{a} = 32$ and $\hat{b} = 47$, which is a sizeable window. The diagonal border, vaguely visible to the lower left, is the line representing $b = a + 1$, which is the edge of the permissible parameters.

however, not simple any more, as inference on θ_L is dependent on inference on θ_R , even when conditioned on a, b . We will discuss methods of achieving this later, but standard optimizer algorithms do work, though at the cost of speed. It can be seen in figure 1.4. Note that when we visualize functions of a, b , such as the profiled log-likelihood or the posterior for a, b , the x -axis represents b and the y -axis represents a . The figures can then be read ‘left to right’ for any combination of a, b . Furthermore, since $a < b$, the function is concentrated on the upper triangle and therefore sometimes exhibits a sharp border along the line $a = b$.

We now wish to extend our inferential tools from a change point τ to the window a, b . First, we wish to extend confidence curves to two dimensions. They can be defined as follows:

$$cc(a, b) = \Pr_{\theta_L, \theta_R}(D(a, b, Y) \leq D(a, b, y))$$

where $D(a, b, Y)$ is the random deviance of a, b , and $D(a, b, y)$ is the observed deviance. This is conceptually not very dissimilar to the one-dimensional case, but since we are now working on joint confidence, we must speak of confidence regions rather than intervals. These intervals are irregular and cumbersome to write out in text, and we therefore prefer to present them as images. The $cc(a, b)$ for British mining disasters is shown in figure 1.5. While the regions are irregular, we can still make some general observations. The width of the elongated region is about the same in the a and b dimensions, but along the re-parametrized axes $\tau = (a + b)/2$ and $d = b - a$, we see a dramatic difference. It is

much slimmer in the τ dimension compared to the d dimension. Thus, the uncertainty regarding τ is much smaller than the uncertainty of d .

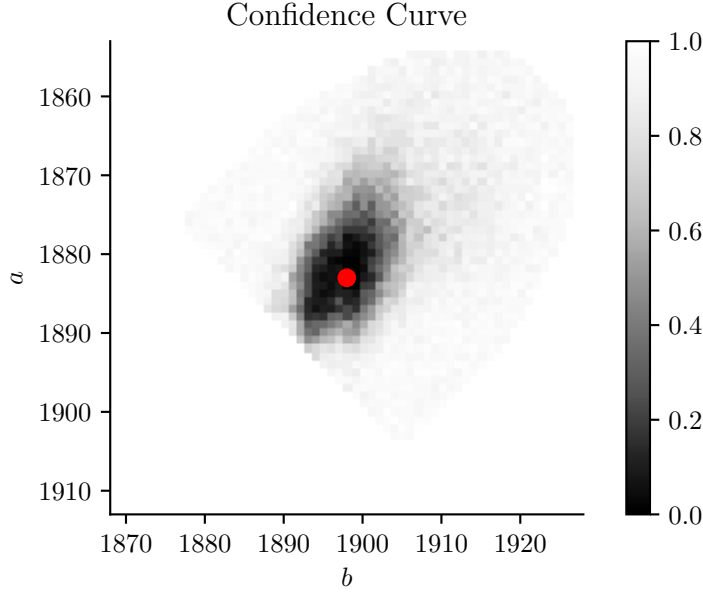


Figure 1.5: The two-dimensional confidence curve for the joint confidence for a and b calculated with 100 samples. The maximum likelihood estimate is marked in red. Note the elongated shape of the region, which shows how we have less confidence in the duration of the change compared to the location.

Motivated by this observation, we present the cc in this parametrization in figure 1.6. In general, a confidence curve cannot be transformed while preserving its statistical properties. This is because the confidence curve is not a probability distribution, and thus the normal transformation of variable methods need not apply. Finally, we might wish to test whether a WoC model is more suitable than a CP model for this data. We can do this by constructing the deviance test statistic

$$T = 2(\max_{a,b} \ell_{\text{prof}}(a, b; y) - \max_a \ell_{\text{prof}}(a, a + 1; y))$$

and finding its distribution. We will return to how we find this in Chapter 7, but we will already here note the similarity between this test and the confidence curves. In fact, the p -value for this statistic is asymptotically equal to $\text{cc}(a^*, b^*)$ where

$$a^* = \arg \max_a \ell_{\text{prof}}(a, a + 1; y)$$

and $b^* = a^* + 1$. Since, however, this particular value of the cc is of special interest, we wish to calculate it with more accuracy. Using 1000 samples, the 95% quantile for T is 1.47, while the observed statistic is 1.61; thus, we may reject the CP model in favour of the more complex WoC model.

The Bayesian analysis, which we will now do, is easier to extend, as there is no real conceptual elevation moving from a single change point to a window of change. The only minor challenges in extending the algorithm are the fact that a, b is two-dimensional, so we must normalize a much larger parameter space, and that inference on θ_L and θ_R is coupled. This means that we cannot find nice and simple conjugate priors for them and

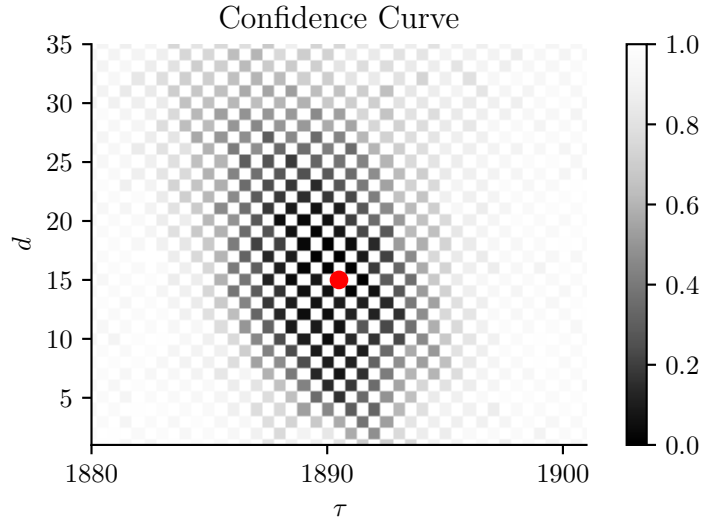


Figure 1.6: The same confidence curve from 1.5, but parametrized with $\tau = (a + b)/2$ and $d = b - a$. The white squares represent impossible parameter combination (like for example $\tau = 1900$ and $d = 1$ which would require $a = 1899.5$ and $b = 1900.5$). The figure clearly shows that there is more uncertainty when estimating d than τ as the region is narrow and tall.

	mean	std
a	1879.08	7.54
b	1899.69	6.55
d	20.61	12.24
θ_L	3.36	0.39
θ_R	0.90	0.12

Table 1.1: Posterior summary for British Mining Disasters

instead must use a Metropolis-Hastings proposal step. Since a, b is two-dimensional, we must also consider a prior for these parameters more carefully, which we will address in Chapter 3. For now, let $\pi(a, b) = \text{const}$. Again, choosing a prior is a good opportunity to include prior information, but we will let it remain constant for the time being. We will return to the discussion of priors later. We continue our analysis with the flat prior and the same prior for θ_L and θ_R from the CP analysis.

The results are summarized in table 1.1 and displayed in figure 1.7. We observe that the Bayesian analysis favours a slightly wider window, and furthermore, that the uncertainty in the window length estimate is about twice as large compared to a, b . Regarding θ_L and θ_R , we see that the estimates deviate slightly from the maximum likelihood estimate. Since the window is larger, we expect them to be further from their mean $(\theta_L + \theta_R)/2$. Finally, their uncertainties are quite small compared to the uncertainty of a, b .

Since the posterior distribution is, in fact, a probability distribution, we obtain the posterior probability that $d = 1$ from the samples as well. In this case, this is $\pi(d = 1 | y) = 0.0081$, clearly smaller than 0.05, and we can thus conclude that there is a window of change.

If we compare the Bayesian results to the frequentist results, we see that the

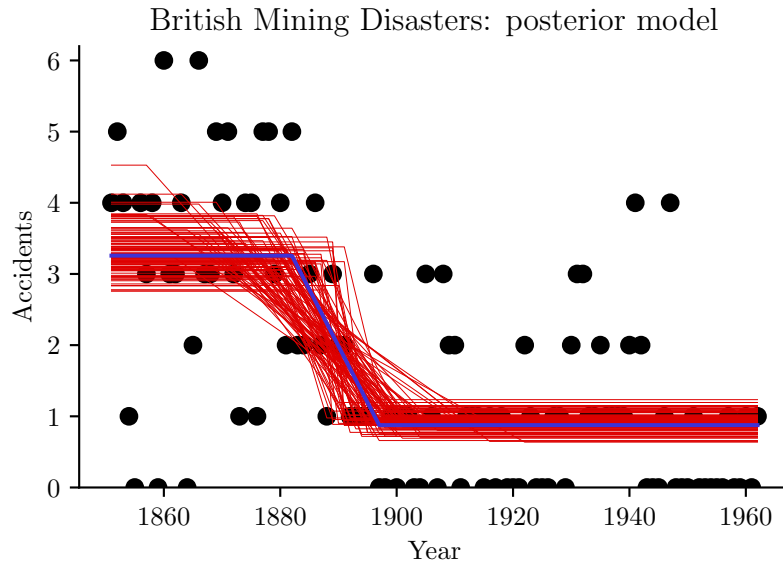


Figure 1.7: 100 posterior model samples (red) alongside the MLE (blue). We see that they all concur on τ , and vary more in the d dimension. The posteriors for θ_L and θ_R are approximately normal, and the variance of θ_L is larger due to there being fewer observations before the window.

confidence curve and the credibility curve largely concur (see figure 1.8). The posterior means of a and b are, however, slightly larger than the maximum likelihood estimates. This is a phenomenon that seems quite general, and we shall see many manifestations of it in the later chapters.

In conclusion, we have seen how the CP model can naturally be extended to include a window of change, and we have demonstrated that, in the case of British mining disasters, it is indeed the preferred model. What are the consequences of this result? Causal explanations for the decline in mining accidents should look at events happening around the year 1883, not 1891. Furthermore, the causes should have concluded their effects by 1898.

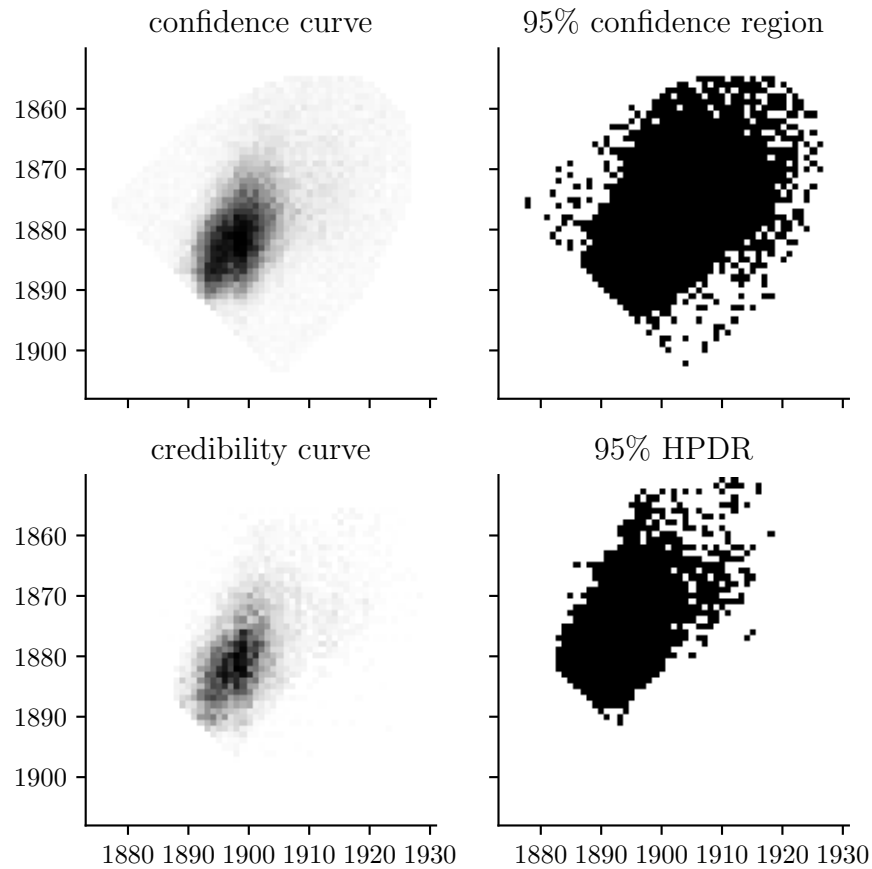


Figure 1.8: Top row: the confidence curve (left) and the corresponding following 95% confidence region (right) calculated using 100 samples for each a, b . Here, as in all other figures of this kind, is the y -axis a and the x -axis b . The confidence curve is somewhat noisy since the number of samples is small, but the computational burden was too large to justify more samples. The search space was limited to an ellipse, which explains the hard boundary around them. Bottom row: the Bayesian posterior credibility distribution (left) and the corresponding 95% credibility region (right)

Chapter 2

Preliminaries

In this chapter, we will cover some preliminaries for Markov chain Monte Carlo and the expectation maximization algorithm in mixture distributions. The reference book for this material is (Givens & Hoeting, 2013).

2.1 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) is perhaps the most powerful tool in Bayesian inference. This section will give a short introduction to the central concepts of MCMC theory, which will be necessary when we discuss how to compare MCMC algorithms.

Bayesian inference is *solved* in the sense that a valid model, which consists of a likelihood and a prior distribution on all the unknown parameters, completely dictates the inference (ignoring possible pathological cases). This is because it completely and uniquely defines the posterior distribution:

$$\pi(\theta | y) = \frac{\pi(y | \theta)\pi(\theta)}{\pi(y)}.$$

This does not mean that inference is always easy, as the challenge becomes apparent when we realize that simply identifying the posterior distribution does not make it mathematically tractable. In particular, the marginal likelihood $\int_{\Theta} \pi(y | \theta)\pi(\theta) d\theta = \pi(y)$, which serves to normalize $\pi(y | \theta)\pi(\theta)$ such that it becomes a proper probability distribution, is often not solvable analytically. While any numerical integration technique, in principle, will be able to compute the marginal likelihood, Monte Carlo simulations are effective because the precision of the approximation is independent of the dimensionality of the integral. The Monte Carlo integration integrates by drawing random samples and is therefore a sampling-based method. *Markov chain* Monte Carlo, in turn, is a method of drawing these samples. MCMC is both a developed and active field of research, and the literature is vast. We will only be touching upon it briefly.

MCMC gets its name because it has the Markovian property

$$\Pr(\theta^{(t+1)} | \theta^{(t)}, \theta^{(t-1)}, \dots, \theta^{(1)}) = \Pr(\theta^{(t+1)} | \theta^{(t)}),$$

where $\theta^{(t)}$ is a sample of θ at iteration t . The probability of sampling $\theta^{(t+1)}$ only depends on the current sample $\theta^{(t)}$. There is, however, a dependency of $\theta^{(t)}$ on all former samples. A correctly constructed Markov chain is a chain that has the correct stationary distribution, which for our purposes means that after a certain point, the chain ‘forgets’ its history. This means that the marginal distribution for $\theta^{(t)}$ and $\theta^{(t+1)}$

are identical (Givens & Hoeting, 2013, p. 16). If we have a chain at its stationary distribution, and this distribution is the posterior, then we see that we can simply iterate the chain any number of times and produce our samples. The samples, however, are not *independent*, but rather exchangeable. An exchangeable sequence is one for which the marginal likelihood is constant under permutations. More precisely, independence of a sequence $\{X\}=X_1, X_2, \dots, X_n$ implies that $\Pr(X_i = x) = \Pr(X_i = x \mid X_{i-1})$, while exchangeability only implies that $\Pr(X_1, \dots, X_n) = \Pr(X_{\sigma(1)}, \dots, X_{\sigma(n)})$ where $\sigma(\cdot)$ is any permutation.

While exchangeability is a weaker property, it is still convenient, as we can sometimes mathematically treat the samples as independent. For example, estimation with exchangeable samples is consistent in the sense that they are asymptotically identical to estimates of independent data. The sample mean of an exchangeable sequence is thus a consistent estimator for the mean, as it is in the independent case (assuming there exists a mean). Even more conveniently, any transformation of variables can be done. In other words, let $Y = g(X)$ and $\{X\}$ be an exchangeable sequence, $\{Y\}$ is an exchangeable sequence of Y , and $\mathbb{E}[Y] \approx \bar{Y}$.

There are, however, two major caveats when performing inference on exchangeable sequences. The first one is that each observation X_i contains less information than if it were independent. This means that if we have m exchangeable samples $\{X\}$, the uncertainty regarding $\mathbb{E}[X]$ would be higher than the same estimate calculated using an independent sequence of X . It would be similar to k i.i.d. samples for some $k < m$. It is therefore desirable to have a measure of the amount of information contained within an exchangeable sample, and a common choice is the *effective* sample size (Givens & Hoeting, 2013, p. 224).

$$M_{eff} = \frac{M}{1 + 2 \sum_{k=1}^{\infty} \rho(k)}$$

$\rho(k)$ is the autocorrelation between samples k iterations apart:

$$\rho(k) = \frac{\mathbb{E}[(X_{i+k} - \mu)(X_k - \mu)]}{\sigma^2}$$

where $\mathbb{E}[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2$. We see that if $2 \sum_{k=1}^{\infty} \rho(k) = 0$, meaning there is no autocorrelation and the sequence is, in fact, i.i.d. $M_{eff} = M$, i.e. the effective number of samples is equal to the actual number of samples. On the other hand, when the autocorrelation increases, the effective number of samples is reduced. A good Markov Chain is thus one with small autocorrelation, but if it is large, we can compensate by drawing more samples (clearly, doubling M would double M_{eff}). Of course, we do not know the autocorrelation of our chains and must estimate it from the data. As suggested by Givens and Hoeting (2013, p. 224), we use the empirical autocorrelation and truncate the sum if $\hat{\rho} < 0.1$. It must be noted that the effective sample size is only a measure of information and not exact.

Having considered the quality of the chain, we return to our second challenge, namely that the chain is only guaranteed to reach the stationary distribution *in the limit*. We can, however, only draw a finite number of samples and, in general, have no guarantee that our chain has reached its stationary distribution by the time we stop. Particularly if the starting values X_0 are very far from the mass of the distribution, the algorithm can spend a long time converging towards the stationary distribution. It is therefore common to discard the first iterations of the chain, which is called *burn-in*, and common choices are anything between 10% and 50% or a fixed number. Furthermore, multiple different starting values are recommended because if all of them converge to something similar,

we have more confidence that we have reached the stationary state. The Gelman-Rubin statistic (Givens & Hoeting, 2013, p. 221) can be used to assess if the burn-in period is sufficient. It compares the *within* chain variance with the *between* chain variance, and if it is larger than a certain value, the chains are unsatisfactory, necessitating either a longer burn-in period or more samples. There are, of course, more challenges and nuances in MCMC theory, and this is not a comprehensive overview, but this suffices to describe and compare the qualities of different chains against each other.

Finally, while the effective number of samples, burn-in, and multiple starts are all important techniques and diagnostics, it is advisable to visually inspect the chains as well. This is typically done through a *trace plot*, which simply plots the samples as a function of their index. A good chain plot will be noisy, with few large trends.

2.2 Mixture distributions and expectation maximization

Conceptually, a mixture model consists of a collection of probability distributions, and a sample is drawn from one of them at random. For example, if we were to sample a random person's height, we can imagine that we either sample a male, which has a certain normal distribution associated with him, or a female, which has a different normal distribution. While the distribution is simply a normal distribution when knowing the sex, in the case where this information is hidden, we have to treat the height distribution as a mixture of the two normal distributions, with weights according to our belief about the proportion of the sexes.

A mixture distribution is a distribution whose likelihood is a mixture of other likelihoods, known as components. Given two distributions $f(y; \theta_L)$ and $f(y; \theta_R)$, the mixture of the two has the following likelihood:

$$L(\theta_L, \theta_R; y) = \prod_{i=1}^n [p_i f(y; \theta_L) + (1 - p_i) f(y; \theta_R)],$$

for some *class probability* $p_i \in [0, 1]$. The likelihood is difficult to work with since there is a sum inside the product, and MLEs for θ_L and θ_R rarely have analytic solutions. The expectation-maximization (EM) algorithm can be used to optimize the log-likelihood, but before it can be applied, it is necessary to introduce some notation.

We introduce a latent, or hidden, variable Z_i for each observation

$$Z_i = \begin{cases} 0, & \text{if } Y_i \sim f_L \\ 1, & \text{if } Y_i \sim f_R \end{cases}$$

which indicates from which distribution y_i is sampled. We can now write what is called the full likelihood

$$L(\theta_L, \theta_R; y, z) = \prod_{i=1}^n [(p_i f(y; \theta_L))^{1-z_i} ((1 - p_i) f(y; \theta_R))^{z_i}],$$

which is not plagued by the product of sums. Unfortunately, we do not know z , but we can use the EM algorithm to estimate them, jointly with θ_L and θ_R . It does so by maximizing a *minorizing* function Q , which has the property that when Q is maximized, then so is $L(\theta_L, \theta_R; y)$. Let $\theta^{(t)}$ be the current estimate for θ at iteration t , then the minimizer function Q is

$$Q(\theta \mid \theta^{(t)}) = \mathbb{E} [\ell(\theta; Y, Z) \mid y, \theta^{(t)}].$$

The function is the expected value of the observed log-likelihood given the current estimate for the parameters $\theta^{(t)}$. Note that conditioned on $\theta^{(t)}$, the only randomness left is the class variables z . The EM algorithm then alternates between calculating this Q , then optimizing it for $\theta^{(t+1)}$.

The introduction of z decouples θ_L and θ_R in the log-likelihood, and they can now be estimated separately. Often, the separation reduces the full estimation to well-known problems, and is the main advantage of this approach. With a good start, the algorithm is guaranteed to converge as it can be shown that every iteration increases the observed log-likelihood (Givens & Hoeting, 2013, p. 103). The convergence speed is at best linear, and depends on the amount of missing information (Givens & Hoeting, 2013, p. 103). The amount of missing information is, in general, low in our case, since there is no missing information outside the window.

Chapter 3

The Window of Change

In this chapter, we define and develop two transition models. By transition models, we mean the exact way in which f_L transforms to f_R within the window.

We begin by defining a weight function, which is a constant function outside the Window of Change, but a linear decrease from zero to one within the window. With this function, we are then able to define the two different transition models. The two transition models have different natural applications. We then compare them in order to highlight some key differences, which indicate when each model is appropriate. This might serve as a test when contemplation yields no obvious answer to which transition model is ‘correct’. Finally, we show how this simple linear transition function can be transformed to include non-linear transitions.

3.1 The weight function

There are many possible and imaginable ways a thing can change, but we cannot consider them all at once. We are also mostly concerned with how introducing a gradual change affects the analysis, without analysing any particular way of changing. Therefore, we must limit ourselves in our pursuit of knowledge to a start, and a natural start is *linear change*.

Definition 3.1.1 (Weight function). Let $a \in 1, 2, \dots$, $b \in a + 1, a + 2, \dots, N - 1$, we define the (linear) weight function as a function of i

$$w(i; a, b) = \begin{cases} 1, & i \leq a \\ \frac{b-i}{b-a}, & a < i \leq b \\ 0, & b < i \end{cases}$$

Often, when it is clear from context, we simply write $w_i = w(i; a, b)$.

The weight function therefore constant outside the window and linearly decreases within it. It is worth pointing out that if $d = b - a = 1$, then there are zero observations from ‘within’ the window, and the weight function is either constantly one before $i = a$, and constantly zero after $i = 1$. There are therefore $d - 1$ observations which are within the window. Furthermore, if $d = 1$ then any of the WoC models which we will present collapses to a CP model.

While we have now decided to model the change as linear, we still must determine the particular manner in which the change is linear. There are two natural contenders: the parameters might change linearly, or the distributions might mix linearly. We will, in turn, define these two ways of transitioning.

3.2 Parameter transition

The parameter transition is characterized by transforming the parameters themselves. We may write it as

$$Y_i \sim f(y_i, \theta_i)$$

where $\theta_i = w_i\theta_L + (1 - w_i)\theta_R$.

This would be a natural model choice if we believe that some property gradually changes, such as the expected value of a phenomenon or perhaps the variance. Consider, for example, the case of climate change and global warming, where we observe the yearly average temperature. Assuming that the temperature had been stable and will be stable in the future, if this temperature is to increase, then we expect it to do so gradually, perhaps over a few years. A parameter transition is then the natural WoC model choice. The likelihood for a parameter transition model is the following:

$$L(\theta_L, \theta_R, a, b; y) = \prod_{i=1}^n f(y_i; w_i\theta_L + (1 - w_i)\theta_R)$$

Note that if $\mathbb{E}[Y] = \theta$, then $\mathbb{E}[Y_i] = \theta_i = w_i\theta_L + (1 - w_i)\theta_R$. In this regard, a parameter transition, conditioned on a, b , can sometimes be reduced to a regression model:

$$\mathbb{E}[Y_i] = f(i),$$

subject to $\mathbb{E}[Y_i] - \mathbb{E}[Y_{i-1}] = 0$ for $i \notin a + 1, a + 2, \dots, b$. In other words, the expected value of Y_i is constant outside the window and changes linearly within it. If we do not condition on a, b , it cannot be interpreted as a regression model. Still, this suggests that a parameter transition WoC model can be used as an alternative to a regression model if we desire that θ_L and θ_R do not change outside some confined window. Since WoC models are broader, they allow for more complicated changes that are typical for regression models. For instance, the changing parameter could affect the variance or skew, resulting in heteroscedasticity.

As we saw in Chapter 1, the British mining disasters data is an example of where this model is applicable, which is why a parameter transition was the chosen model. Since $\mathbb{E}[Y_i] = \theta_i$ for Poisson data, this can be considered a special case of Poisson regression when a, b is fixed with the extra conditions that WoC models enforce: namely, the observations are i.i.d. to the left and right of the window, and change linearly in the middle. Since we do not know a, b , estimating them is similar to a model selection problem, where each combination of a, b has an associated regression model. The similarity could still be useful for interpreting the parameters and efficiently estimating θ_L and θ_R .

3.3 Mixture transition

The mixture transition is named so because, in the window of change, it behaves as a mixture model with w_i being the mixture parameter.

$$f(y_i) = w_i f_L(y_i) + (1 - w_i) f_R(y_i).$$

The likelihood under this model is the following:

$$L(\theta_L, \theta_R; y) = \prod_{i=1}^n [w_i f(y_i; \theta_L) + (1 - w_i) f(y_i; \theta_R)]$$

Conceptually, we can imagine that we flip a coin with the probability of heads equal to w_i , and if it lands on heads, we sample y_i from f_L , and if it lands on tails, we sample y_i from f_R .

This model is suitable if one believes that the observations are a sample from a bi-stable WoC model. As an example, consider the problem of tracking virus mutation. We imagine that a new random patient has her temperature measured every day and tracks the changes over time. As is often the case, new mutations are more infectious while giving milder symptoms. Imagine then that patients with the original virus have a certain amount of fever, say $T \sim N(\theta_L, \sigma^2)$, while the patients with the mutated virus have slightly milder fever, $T \sim N(\theta_R, \sigma^2)$, $\theta_L > \theta_R$. Since the temperature is otherwise random, detecting when the mutation appears, and finally dominates the original virus is just a case of a WoC model.

The construction of the example reveals a characteristic of mixture transitions, namely that it is often the appropriate model when y_i is a single observation, and not somehow an aggregate. Compare British Mining Disasters to the example above. It is unnatural to imagine that the yearly disaster counts are bi-stable, such that it either is a ‘deadly’ year or a ‘safe’ year. One of the reasons is that the accident counts are the sum of accidents across all coal mines in Britain. If the changes we observe in the accident counts are a result of e.g. new regulations, we expect that individual mines would transition from ‘deadly’ to ‘safe’, and transition at individual times. There is therefore expected to be some years that are neither ‘deadly’ nor ‘safe’.

3.4 Comparison

The parameter transition and the mixture transition both transition linearly from f_L to f_R , and are, in this regard, similar. Yet, they have some important distinctions which we will now examine. We have already seen some examples of when a parameter model is the natural choice and vice versa, but if we are in a situation where there is no intuitive choice between the two, there is a need to distinguish them. In the case of British mining disasters, we argued that a steady decline in accident counts was a reasonable model. The justification relied on certain ideas of how such a change would occur, like that the accident counts were aggregate quantities, and thus we expect them to steadily decline. However, we do not always have solid notions of how a change could or should occur, and we therefore need some mathematical justification. For this reason will we now do a simple analysis of how the probability function (PDF or PMF) behaves for both the parameter transition and the mixture transition. This could serve as a prior justification for choosing one transition over the other, in the event that we have no strong ideas otherwise.

A crucial difference between parameter and mixture transitions is that a mixture transition induces an increased variance within the window, while a parameter transition does not induce this. By ‘induce’ we mean that the extra variance is a consequence of the WoC construction, and not that e.g. f_L and f_R having different variance. This is easily shown through the law of total variance, but to do so we need to introduce some notation. Let $Z_i = 0$ denote that $Y_i \sim f_L$ and correspondingly $Z_i = 1$ denote that $Y_i \sim f_R$. Furthermore, let $\mathbb{E}[Y_i | Z_i = 0] = \mu_L$ and $\mathbb{E}[Y_i | Z_i = 1] = \mu_R$. We can then first see that the *mean* transitions linearly from μ_L to μ_R :

$$\mu_i = \mathbb{E}[Y_i] = \mathbb{E}[\mathbb{E}[Y_i | Z_i]] = w_i \mu_L + (1 - w_i) \mu_R,$$

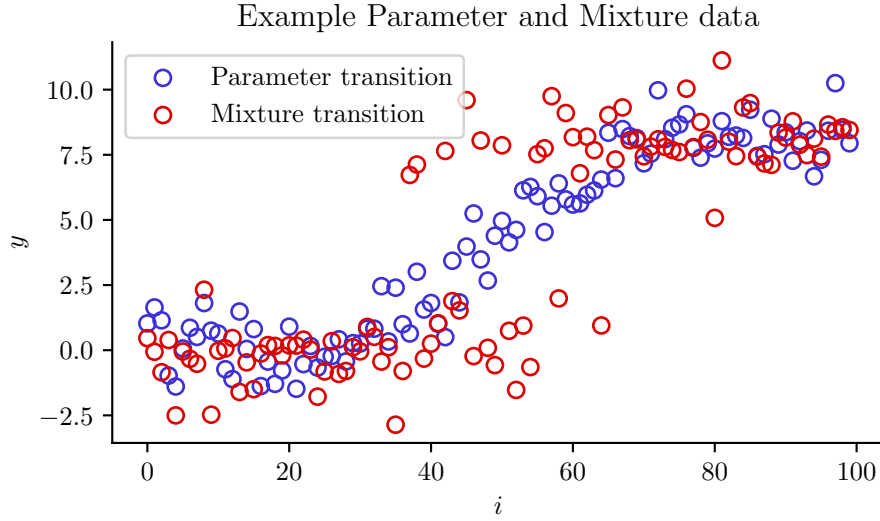


Figure 3.1: Sample data sets for a parameter and a mixture transition for normal distributions with constant variance equal to 1. The true parameters where $\mu_L = 0$ and $\mu_R = 8$, and the window begins at $a = 30$ and ends at $b = 70$.

but the variance increases regardless

$$\text{Var}[Y_i] = \mathbb{E}[\text{Var}[Y_i | Z_i]] + \text{Var}[\mathbb{E}[Y_i | Z_i]].$$

The first term is a linear mixture of the left and right variance, while the second term is the extra variance induced by the mixture. It can be further decomposed

$$\begin{aligned} \text{Var}[\mathbb{E}[Y_i | Z_i]] &= \mathbb{E}[\mathbb{E}[Y_i | Z_i]^2] - \mathbb{E}[\mathbb{E}[Y_i | Z_i]]^2 \\ &= w_i \mu_L^2 + (1 - w_i) \mu_R^2 - (w_i \mu_L + (1 - w_i) \mu_R)^2. \end{aligned}$$

Thus, the extra variance follows a parabola. If $\text{Var}[Y_i | Z_i = 0] = \text{Var}[Y_i | Z_i = 1] = \sigma^2$, then the maximum variance is $\sigma^2 + 1/4(\mu_L - \mu_R)^2$. This means that as the difference in means increases, then the variance within the window increases. A simple plot, figure 3.1, shows this very clearly, and highlights the nature of the two models.

This fact could serve as a test for identifying which transition should be used, but more importantly, it shows how the modelling of f_L and f_R is of greater importance compared to Change-Point analysis. The problem is overdispersion (more-than-expected variance) could be ‘explained’ by a wider window. Misspecifying f_L and f_R could thus have a very large impact on the estimates for a and b . The reverse phenomenon is also a problem; an underdispersed model would favour very small windows (see figure 3.3). This is a problem because in this thesis we assume that the distributions f_L and f_R can be reasonably estimated such that we can turn our attention to the window itself. What this example shows is that the modelling of f_L and f_R has a strong, unintended impact on the estimation of our focus parameters. The direct dependence of the window length and variance then requires us to very carefully model them. We will not delve into testing for overdispersion.

Example 3.4.1 (Overdispersion in normal model). Let $f_L = N(0, \sigma^2)$ and $f_R = N(2, \sigma^2)$ for some σ^2 , but we (erroneously) think that $\sigma^2 = 1$. How does this affect the MLE for

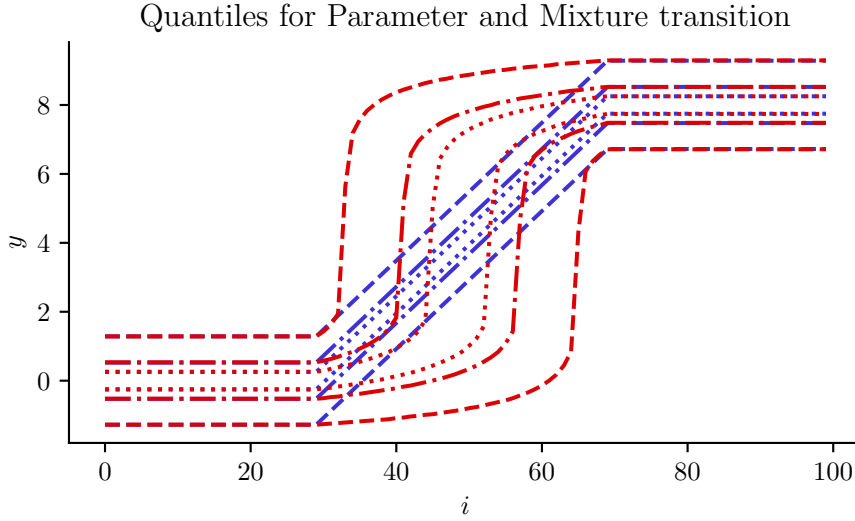


Figure 3.2: Quantiles at the 0.1 and 0.9 (dashed); 0.2 and 0.8 (dashdotted); and 0.3 and 0.7 (dotted) levels for parameter (blue) mixture (red) transition. The same parameters from figure 3.1 where used, i.e. $\mu_L = 0$, $\mu_R = 8$, $\sigma = 1$, $a = 30$ and $b = 70$.

d ? In the parameter transition, the log-likelihood for a fixed a, b is

$$l(\theta; y, a, b) = -\frac{1}{2} \sum_{i=1}^n [y - (w_i \mu_L + (1 - w_i) \mu_R)]^2 / \sigma^2 - \frac{1}{2} \log \sigma^2. \quad (3.1)$$

The maximizer for μ_L, μ_R is therefore independent of σ^2 , and consequently the misspecification has no impact on the profiled log-likelihood.

In the mixture model, estimation of μ_L and μ_R is not decoupled from σ^2 , and as discussed, overdispersion can be compensated for with larger windows. Figure 3.3 illustrates this for $\sigma^2 \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1., 1.1, 1.2, 1.3, 1.4, 1.5\}$.

$\hat{\tau}$ was unaffected by the misspecification, and remained unbiased (not shown).

▲

The parameter transition also has some challenges. First, the parametrization has an impact on the interpretation of linear change. This is obvious when considering the case where we wish to model an increase in variance for a normal-normal model. Should the variance or the standard deviation increase linearly? What is linear in one case is then quadratic or square root in the other. In many cases, we are used to working with certain parametrizations for their convenience or interpretations, but in our case, we must also consider the impact it has on the model. For example, we might want to work with standard deviation in a certain model, as the standard deviation is in the same units as the mean for a normal-normal model. If we model money, for instance, measuring variance in money squared is curious and unintuitive for many. On the other hand, one might want to model with this particular parametrization for interpretability purposes.

The second challenge with parameter transitions is that f_L and f_R must be of the same parametric family, i.e. $f_L = f(\theta_L)$ and $f_R = f(\theta_R)$. This also implies that any given parameter cannot change interpretation either. This is not the case for mixture transitions and is furthermore not a limitation of normal change point analysis. When

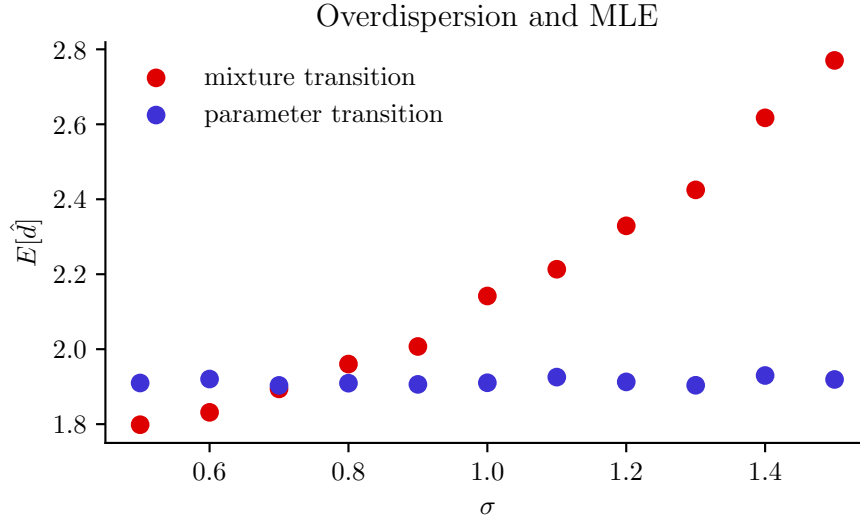


Figure 3.3: Example of overdispersion with normal data. The x axis is the true standard deviation, while the y axis is the expected value of $\hat{d}$ when σ is assumed to be 1. The true parameters are $d_0 = 10$, $\theta_L = (0, \sigma^2)$, and $\theta_R = (2, \sigma^2)$. Clearly, both estimators are strongly negatively biased, but this is due to another phenomenon which we will discuss later.

conducting Change-Point analysis, it is uncommon to change the parametric family at the change point. The reasons for this are often justified by convenience and clarity; papers showcasing methods and properties of change point models favour simple models. A statistics practitioner, however, might not have this luxury, and this restriction enforced by the parameter transition might be a real limitation.

Using the analogy of a regression model, we see that a parameter transition model is incapable of online inference. In online inference, we receive a stream of data and continuously fit our model to it. The problem is that the regression model is unidentifiable until $i = b$. In particular, for any given estimate for θ_R and d , doubling the estimates would be mathematically equivalent. Furthermore, CP models are frequently used to detect anomalies in sensor readings, and we wish to intervene as soon as something seems broken. If this is the goal and we use a WoC model, then observing a , i.e. the beginning of the change, is sufficient to warrant intervention. There is no need to wait for the end of the window if the goal is to detect whether a change has occurred at all. The next transition that we consider is not indeterminate until $i = b$ and could, in principle, be used online. However, the latter argument still holds: namely, if we detect a , then there is no reason to wait for b before we intervene. For this reason, WoC analysis is firmly placed in the world of offline inference.

3.5 Extensions

Any monotonic function g , $w_i^* = g(w_i)$ satisfying $g(0) = 0$ and $g(1) = 1$ yields a new weight function, which in many cases can readily be substituted for w_i . A linear change is, therefore, not as restrictive as it seems at first glance.

Non-linear change can also be introduced in the parameter through transformation of variables. As already noted, the parametrization affects in what regard the change is linear, and this property can be used to introduce a suitable transition function. A pertinent question is the following: say we want to model a linear transition in γ_L , but

we prefer to mathematically work with θ_L . Is it then possible to transform the weights such that we can scale θ_L with w_i^* ? The answer is yes, but it introduces yet another set of non-linear equations.

Lemma 3.5.1. *For any strictly increasing function φ , let $\gamma = \varphi(\theta)$. A linear change through a parametrization γ , that is $\gamma_i = w_i\gamma_L + (1 - w_i)\gamma_R$, can be achieved through a transformation of weights in θ space. Furthermore, this rescaling is*

$$w_i^* = \frac{\varphi^{-1}(\gamma_i) - \theta_R}{\theta_L - \theta_R}. \quad (3.2)$$

Proof. The proof is simple algebraic manipulation. Note first that $\varphi^{-1}(\cdot)$ exists by assumption. The condition of linear change in γ space is obtained when

$$\varphi(w_i^*\theta_L + (1 - w_i^*)\theta_R) = w_i\varphi(\theta_L) + (1 - w_i)\varphi(\theta_R).$$

Letting $\gamma_i = w_i\varphi(\theta_L) + (1 - w_i)\varphi(\theta_R)$, we apply the inverse transformation on both sides and simplify:

$$w_i^*(\theta_L - \theta_R) + \theta_R = \varphi^{-1}(\gamma_i),$$

Thus giving the desired formula after some manipulation:

$$w_i^* = \frac{\varphi^{-1}(\gamma_i) - \theta_R}{\theta_L - \theta_R}.$$

□

We see that it is possible to use transformation of variables to introduce non-linear change, but that the transformation depends on the true parameter. In this regard, this result is not very satisfactory, as solving the equations for $\hat{\theta}_L$, $\hat{\theta}_R$ jointly with w_i^* would yield a far more complex set of equations. The problem, however, is not as grave as it sounds, because only observations within the window contribute meaningfully to the equations. For instance, if we disregard this window and consider only the case of ML estimation for γ_L using $y_1, \dots, y_a$, then the parametrization does not matter, and estimation is simple. This means that the estimate for, say, $\hat{\gamma}_L$ is approximated by $\varphi(\hat{\theta}_L)$, and as $a \rightarrow \infty$ and $n - b \rightarrow \infty$, it becomes exact.

3.6 Estimation and bias

A surprising property of the WoC model is that certain maximum likelihood estimates are biased:

$$\text{Bias}(\hat{\theta}) = \mathbb{E}_{\theta} [\hat{\theta}] - \theta.$$

We are particularly interested in the parameter $d = b - a$, but we will also discuss the estimation of $\tau = (a + b)/2$ and the nuisance parameters θ_L and θ_R . However, we begin by discussing the properties of the simpler estimates, namely the nuisance parameters.

Throughout this thesis, we gloss over the properties of $\hat{\theta}_L$ and θ_R because their characteristics are standard. By standard, we mean that estimating θ_L and θ_R is comparable to the i.i.d. case, which we assume is well understood. In this sense, it is ‘easy’ to estimate θ_L and θ_R . For example, in most cases we have that $\hat{\theta}_L$ is asymptotically normal

$$\sqrt{a}(\hat{\theta}_L - \theta_L) \rightarrow^d N(0, I(\theta_L)^{-1})$$

where $I(\theta_L)^{-1}$ is the inverse Fisher information matrix. The reason for this is quite simple, namely that as $n \rightarrow \infty$, we may simply ignore the observations within the window, as they contribute vanishingly little to the average Fisher information. This assumes that d is bounded, which we always assume in this thesis, whether implicit or not. The converse argument shows that θ_R also converges, but only when $n - b \rightarrow \infty$. Therefore, in our 'easy' settings, we assume that a and b are sufficiently far from the borders.

The parameter $\tau = (a + b)/2$ is interesting in comparison to the change point in CP models, as it has a very similar interpretation. For symmetric distributions f_L and f_R that only differ in mean, we can show that $\hat{\tau}$ is unbiased, just as its CP counterpart is. In short, we can 'reverse' the order of y and show that this too is a WoC model, whose estimator $\hat{\tau}$ must have the same bias, and the only possible bias is then zero.

Lemma 3.6.1 (Bias for $\hat{\tau}$). *Let a and n be so large such that we can disregard edge conditions. For a symmetric distribution $f(y; \theta)$ where θ is a location parameter, for the WoC model $f_L = f(y; \theta_L)$ and $f_R = f(y; \theta_R)$, $\text{Bias}(\hat{\tau}) = 0$*

Proof. Fix $a, b < n$ and let Y be a random observation from the WoC model. Since f_L and f_R are symmetric and only differ by their means, there exists a constant c such that $c - Y_i \sim f_R$. Let $X_i = c - Y_{n-i}$, then X follows a WoC model with the new window $a_x = n - b$ and $b_x = n - a$. Observe that $\tau_x = (a_x + b_x)/2 = n - \tau$, and finally note that $\text{Bias}(\hat{\tau}_x) = -\text{Bias}(\hat{\tau})$ because we disregard edge conditions. Thus, $\text{Bias}(\hat{\tau}) = 0$. $\square$

While the conditions in the proof seem restrictive, they are, in fact, still general enough to highlight interesting properties about our final parameter d . Even when the conditions do not hold, we will see that inference on τ is still a simpler problem than inference on d . For example, in the analysis of British Mining Disasters in the introduction, we clearly see that the uncertainty regarding τ is smaller.

Consider the following correction function

$$c(d) = \frac{\mathbb{E}_d[\hat{d}]}{d},$$

which is the expected value of $\hat{d}/d$ as a function of the true d . Importantly, this function is not random, and although we cannot find a closed form expression for the bias, it is possible to approximate the curve with simulation. The correction is achieved by dividing the estimate $\hat{d}$ by the functional value of $c(\hat{d})$:

$$\hat{d}' = \frac{\hat{d}}{c(\hat{d})}. \quad (3.3)$$

This correction has the unfortunate property of producing non-integer estimates for d . In certain cases, this might be admissible, but clearly not in all. Therefore, we construct a second correction such that it only produces integer values.

$$\begin{aligned} \hat{d}'' &= \arg \min_d \left[|\hat{d} - \mathbb{E}_d[\hat{d}]| \right] \\ &= \hat{d} - \text{round}(\text{Bias}(\hat{d})), \end{aligned}$$

where $\text{round}(x)$ is the nearest integer to x . The rationale behind this correction is that having observed the MLE $\hat{d}$, we find the d_0 which results in an estimator that has the

expected value closest to the observed estimate. It can be considered a moment estimator for $\hat{d}$.

The two corrections just presented are free from nuisance parameters, which we do not always have the luxury of ignoring. The appropriate correction curve is

$$c(d; \theta_L, \theta_R) = \frac{\mathbb{E}_{d, \theta_L, \theta_R} [\hat{d}]}{d}. \quad (3.4)$$

When we analyse a real data set, we almost always must estimate θ_L and θ_R , and our corrections will depend upon those estimates. This, of course, introduces both uncertainty and possible biases in computing $\text{Bias}(\hat{d})$. We propose estimating $\hat{\theta}_L$ and $\hat{\theta}_R$ first from the data, then simulating with these parameters, i.e. substituting $\hat{\theta}_L$ and $\hat{\theta}_R$ in equation (3.4). The justification for this is, as we already mentioned, that we can estimate the nuisance parameters with arbitrary precision as $n \rightarrow \infty$ and $n - b \rightarrow \infty$. The extra uncertainty can be partially accounted for by a ‘double’ simulation: First, estimate θ_L and θ_R , then in the outer simulation loop, sample y^* with $\hat{\theta}_L$ and $\hat{\theta}_R$, and re-estimate $\hat{\theta}_L^*$ and $\hat{\theta}_R^*$ and use these values in equation (3.4). The procedure is very slow, and some light simulation verified that the ‘single’ simulation loop yielded similar results.

3.7 Confidence curves

Extending confidence curves from CP analysis to WoC analysis is non-trivial. One alternative is to consider the joint confidence for a, b , and since the definition ?? already accommodates vector parameters, may we use it directly.

$$cc(a, b) = \Pr(D(a, b; Y) \leq D(a, b; y)), \quad (3.5)$$

which yields the two-dimensional confidence surface. This construction can be used to find confidence sets for any a, b , but say we are more interested in d or τ . Then the joint construction is inadequate, and furthermore, it cannot simply be transformed to yield the singular confidence (in contrast to Bayesian posteriors, which can be marginalized).

Instead, if we want confidence for d alone, we must use the construction

$$cc(d) = \Pr(D(d; Y) \leq D(d; y)),$$

where

$$\begin{aligned} D(d; y) &= 2(\max_{a, b} \ell_{\text{prof}}(a, b; y) - \max_{\tau} \ell_{\text{prof}}(\tau - d/2, \tau + d/2; y)) \\ &= 2(\max_{a, b} \ell_{\text{prof}}(a, b; y) - \ell_{\text{prof}}(d; y)) \end{aligned} \quad (3.6)$$

subject to $\tau + d/2 \in \mathbb{N}$. The expression $\ell_{\text{prof}}(d; y)$ is the log-likelihood function profiled over τ , treating τ as a nuisance parameter. Similarly, if we want confidence for τ alone, we must use the following construction

$$cc(\tau) = \Pr(D(\tau; Y) \leq D(\tau; y)) \quad (3.7)$$

and similarly defining

$$D(\tau; y) = 2(\max_{a, b} \ell_{\text{prof}}(a, b; y) - \max_d \ell_{\text{prof}}(\tau - d/2, \tau + d/2; y)) \quad (3.8)$$

$$= 2(\max_{a, b} \ell_{\text{prof}}(a, b; y) - \ell_{\text{prof}}(\tau; y)) \quad (3.9)$$

again subject to $\tau + d/2 \in \mathbb{N}$, $\ell_{\text{prof}}(\tau; y)$ is the log-likelihood profiled over d , and d in turn is considered a nuisance parameter.

We do not know the nuisance parameters, but we need them to sample data in the simulation. We will treat them as Cunen, Hermansen and Hjort (2018) recommend, that is, we simulate data under their maximum likelihood estimates, but treat them as unknown in the deviance calculation. Furthermore, the implementation of the sampling algorithm is based on Hagen (2020, p. 56)’s master thesis. This approach certainly introduces some bias, but as $a \rightarrow \infty$ and $n - b \rightarrow \infty$, θ_L , and θ_R can be well estimated. The associated bias is therefore expected to diminish with more data. This is not the case for τ and d when we treat either one of them as nuisance parameters. The reason is that they cannot be well estimated in the sense that with more data, the uncertainty reduces arbitrarily. Particularly, estimating d is difficult, and d is the parameter with which we are often the most concerned in this thesis. However, when constructing confidence curves for d , we need the estimate for τ , which is comparatively easier to work with.

There is a further problem with our proposed approach, namely that there are impossible combinations of d and τ . For example, say that we want to make a *cc* for τ , and simulate data under the nuisance parameter $\hat{d}$, and assume that d is odd, τ cannot be an integer value. This is problematic because while $\hat{\tau}$ lies halfway between two integers, it is still very possible that τ_0 actually is an integer. In these cases, we simulate under $\hat{d}$ for the values of τ that yield integer values for a, b , and simulate under $\hat{d} + 1$ for the others. The choice to add one instead of subtracting one is because $\hat{d}$ often is negatively biased. One could argue that we should randomly simulate under $\hat{d} - 1$ and $\hat{d} + 1$, but we will not consider this case. When making a *cc* for d , we add or subtract $1/2$ to $\hat{\tau}$ such that it is closer to the centre. Again, one could argue that we should randomly increase or decrease the value instead.

Finally, while constructing a *cc* for, say, τ only requires testing for at most n different values of τ , testing all combinations of a, b amounts to testing n^2/w different combinations, many of which are clearly very unlikely, i.e. most combinations are $cc(a, b) \approx 1$. To lessen the computational burden, we therefore only check values around $\hat{a}, \hat{b}$. Finding the ‘radius’ of this search space requires some testing, but inspecting the profiled log-likelihood gives a good indication of the main mass of the *cc*. Preliminary testing has shown quite generally that the uncertainty is larger in the d dimension compared to the τ dimension, and we therefore often prefer testing in an ellipse around $\hat{a}, \hat{b}$.

3.8 The problem of priors

We now discuss how we are to construct priors for the window parameters a, b . We focus on non-informative priors for two reasons. The first reason is that it makes comparisons between frequentist and Bayesian analyses the simplest. In other words, we want to let the data ‘speak for itself’. It would be difficult to compare the Bayesian and frequentist analyses if we used a prior with undue influence on the posterior, as the difference in results could not be attributed to either the nature of the Bayesian analysis or the choice of prior. Since there is a limited amount of information regarding the location of a, b in WoC analysis, this reason is extra important. The reason is that the prior would never ‘wash away’, and always have some influence on the posterior. The second reason is that if prior information is to be included, we must still investigate how even this informed choice influences the posterior. Extra information can always be included in a prior, and the information we do not intend to include should ideally be minimized.

The first candidate for a non-informative prior is

$$pi(a, b) = \text{const},$$

i.e. the flat prior. This is non-informative in the sense that any combination of a, b is equally likely. A prior on a, b induces a prior on the alternative parametrization d and τ , which we now deduce.

Given the flat prior on a, b , the prior on d turns approximately left triangular:

$$\pi(d) \propto \text{const} - d \propto 1 - d/n.$$

The simplest way to see this is that for any given n , there are $n - 1$ unique a, b such that $b - a = 1$, $n - 2$ unique combinations such that $b - a = 2$, etc. The prior on τ becomes triangular with a peak at $n/2$.

$$\pi(\tau) \propto |n/2 - \lceil \tau \rceil|$$

A similar diagonalization argument works for τ : there is only one possible combination of a, b such that $\tau = 1.5$, namely $a = 1$ and $b = 2$, one possibility for $\tau = 2$, $a = 1$ and $b = 3$, two combinations for $\tau = 2.5$, $a = 1$ and $b = 4$, and $a = 2$ and $b = 3$, etc.

Thus, a flat prior on a, b implies a preference for small values of d , and middle values for τ . In principle, this is no problem, and, as we have seen, a consequence of believing prior to having observed any data that all a, b are equally probable. However, if we are particularly interested in d and τ , we might want to put a flat prior on them instead. If we put a flat prior on d , then we believe that any window length is equally probable. Since there is one combination of a, b which has a length of $n - 1$, namely $a = 1$ and $b = n$. For $d = n - 1$, there are two combinations, $a = 1$ and $b = n - 1$ and $a = 2$ and $b = n$, etc. Thus, we see that small a and large b are very strongly favoured. If we are looking for a non-informative prior, it seems a bit unintuitive to place so much prior weight on $a = 1$ and $b = n$.

With the analysis of what priors on d, τ a flat prior on a, b induces, we continue by discussing what kind of priors on d, τ might be desirable, and why. The flat priors push a, b far towards the extremities. Furthermore, a flat prior on d does not seem like a universally desirable property. For instance, this prior depends strongly on the length of the observation vector y . Is it reasonable to change your prior opinion about an event, far in the past (i.e. $b - n$ large), by observing y_{i+1} ? It might seem more reasonable to confine $d < d_{max}$. Similarly, believing that τ is very close to the edge might also seem unreasonable, and if not impractical, as it would make estimation very difficult.

Since a, b are discrete, common approaches to construct non-informative priors, like Jeffreys' priors, do not work. We could instead embed a, b in a continuous parameter space. I.e. let $u = a/n$ and $v = b/n$. Then the random variable U, V would be approximately continuous as n increases, and we could try to find e.g. Jeffreys' prior as Berger, Bernardo and Sun (2012) recommends. However, since this would force $d/n \rightarrow 0$, the resulting construction could behave poorly. Another common choice for priors for positive parameters is its reciprocal. This prior has the advantage of being improper, which means that it places a sizeable prior mass on all possible values of d . It is also less dependent on n and changes comparatively little if we observe y_{n+1} . However, it places very much prior mass on small d , which seems neither non-informative, nor 'in the spirit' of using a WoC model.

Concluding, we can say that $\pi(d) \propto n - d$ in fact seems reasonable, and perhaps even more so if we restrict the prior by d_{max} . $\pi(d) \propto 1/d$ prefers small d too strongly,

and induces a bias, which we will see later, but it has the convenient property of not being very dependent on n as n grows, which is suitable in the case where we cannot assume that there is some d_{max} such that $d < d_{max}$. The extreme preference for small d , however, can be undesirable, and perhaps letting $\pi(d)$ initially be flat, then decrease, could be a good compromise.

3.9 Prior matching

As we have seen, procuring a prior for a, b requires some care. Another principle of constructing priors is to consider what priors need to be used to ‘match’ the frequentist analysis. Of course, if one insists on using a prior on a, b that forces the posterior inference to match the frequentist analysis, one could simply forgo the Bayesian analysis altogether and use the frequentist analysis from the beginning. However, finding matching priors sheds light on the different properties of the analyses.

Cunen, Hermansen and Hjort (2018) showed that the Bayesian analysis of CP problems has a certain bias when estimating τ on the scale of

$$\frac{1}{(\tau(n - \tau))^{p/2}}, \quad (3.10)$$

where p is the dimensionality of θ_L and θ_R . Thus, letting $\pi(\tau) \propto (\tau(n - \tau))^{p/2}$ results in the posterior being approximately proportional to the profiled log-likelihood

$$\pi(\tau | y) \propto \pi(y | \tau)(\tau(n - \tau))^{p/2} \approx \ell_{\text{prof}}(\tau; y) + \text{const}$$

This prior prefers middle values for τ . They use the posterior normal approximation, also known as the Laplace approximation, to factor the posterior distribution into the profiled likelihood and the Bayesian ‘bias’, which is the factor in (3.10).

We need the Laplace approximation, which is a second-degree Taylor expansion of the posterior mean around the maximum a posteriori (MAP) estimate

$$\hat{\theta}_{MAP} = \arg \max_{\theta} \pi(\theta | y) \approx \arg \max_{\theta} \pi(y; \theta).$$

The Taylor expansion is then

$$\log \pi(\theta | y) \approx \log \pi(\hat{\theta}_{MAP} | y) + \frac{1}{2}(\theta - \hat{\theta}_{MAP})^T H(\theta - \hat{\theta}_{MAP})$$

where $H = -\left[\frac{d^2}{d\theta_i d\theta_j} \log \pi(\theta | y)\right]_{\theta=\hat{\theta}_{MAP}}$. This implies that the posterior is approximately normal, centred around the MAP and with a particular covariance matrix. We are justified in approximating the posterior, as the number of observations increases (subject to a and $n - b$ increasing) the typical limiting theorem holds, as we discussed previously. Applying this approximation only requires that we find the MAP and the covariance matrix.

With a WoC model, we want to find the matching prior for a, b using this approximation, which means finding this quantity:

$$\begin{aligned} \pi(a, b | y) &= \int \int \pi(a, b, \theta_L, \theta_R | y) d\theta_L d\theta_R \\ &= \frac{\pi(a, b)}{\pi(y)} \int \int \pi(y | a, b, \theta_L, \theta_R) \pi(\theta_L) \pi(\theta_R) d\theta_L d\theta_R \end{aligned}$$

In the CP setting, the double integral could be split into two, one for each θ_L and θ_R , but we must treat them jointly. Recognize that the integrand is proportional to the posterior distribution of θ_L, θ_R with a, b and y known. Thus, we can approximate $\pi(a, b \mid y, a, b)$ with the multivariate normal distribution.

We will now find the approximation in the univariate normal case using a parameter transition. In this particular case, this approximation is, in fact, exact. Let Y_i be distributed according to the parameter transition

$$Y_i \sim N(w_i \theta_L + (1 - w_i) \theta_R, 1). \quad (3.11)$$

And consider the case where a, b are known. We have shown how $\hat{\theta}_L, \hat{\theta}_R$ can be found analytically, but we derive it again using a different approach. Notice that (3.11) is equivalent to

$$Y_i = \beta_0 + X_i \beta_1 + \epsilon,$$

letting $\beta_0 = \theta_L$, $\beta_1 = \theta_R - \theta_L$, $X_i = (1 - w_i)$, and $\epsilon \sim N(0, 1)$. Then, we know that the MLE is

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

for a given observation vector y . Thus, $\theta_{\text{MAP}}^T = (\hat{\beta}_0, \hat{\beta}_1 - \hat{\beta}_0)$. We turn to the second step, finding H . Let $\ell = \pi(y \mid a, b, \theta_L, \theta_R)$, and we have that

$$\begin{aligned} \frac{d^2 \ell}{d\theta_L^2} &= - \sum_{i=1}^n w_i^2 \\ \frac{d^2 \ell}{d\theta_R^2} &= - \sum_{i=1}^n (1 - w_i)^2 \\ \frac{d^2 \ell}{d\theta_L d\theta_R} &= - \sum_{i=1}^n w_i (1 - w_i) \\ \frac{d^2 \ell}{d\theta_R d\theta_L} &= - \sum_{i=1}^n w_i (1 - w_i). \end{aligned}$$

The Hessian is independent of θ_L and θ_R because we chose a normal model. In general, this is not true. We can find analytic solutions to the sums,

$$\sum_{i=1}^n w_i = a - 1 + \frac{1}{6} d(d+1)(2d+1) = \Sigma w w \quad (3.12)$$

Returning to the approximation (and abbreviating the sums in the obvious manner), we get

$$\pi(y \mid a, b, \theta_L, \theta_R) \approx N \left(\begin{pmatrix} \hat{\theta}_{R, \text{MAP}} \\ \hat{\theta}_{L, \text{MAP}} \end{pmatrix}, \begin{pmatrix} \Sigma w w & \Sigma w (1 - w) \\ \Sigma w (1 - w) & \Sigma (1 - w)(1 - w) \end{pmatrix}^{-1} \right),$$

The resulting discrepancy from a purely likelihood-based inference is then on the scale of

$$|H|^{-1} = \frac{1}{(\Sigma w w \cdot \Sigma (1 - w)(1 - w) - [\Sigma w (1 - w)]^2)^{1/2}} \quad (3.13)$$

We obtain the bias form CP models by letting $d = 1$, and $\Sigma w w = a$, $\Sigma (1 - w)(1 - w) = n - a$, and $\Sigma w (1 - w) = 0$.

The factor in (3.13) pushes $\tau = (a+b)/2$ towards the extremities, as in the CP model, but more strongly (see figure 3.4). Holding τ constant, d is pushed towards larger values

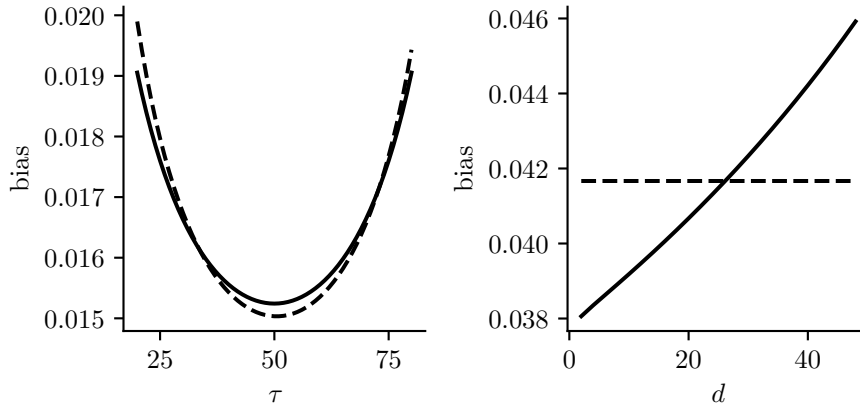


Figure 3.4: The inherent bias in Bayesian analysis for WoC models (solid lines) compared to CP models (dashed lines). Left: $d = 20$ is held fixed. Right: $\tau = 50$ is held fixed (the dashed line is flat because d is held constant). For both, $n = 100$.

approximately linearly (of course, there is never a bias for $\hat{d}$ in a CP model as it is fixed at 1). Therefore, if we think that our Bayesian posterior for a, b should be proportional to $\ell_{\text{prof}}(a, b)$, the prior must have the inverse relationship, i.e. favour middle value τ and small d .

In light of the discussion of priors in the previous section, we see that $\pi(a, b) = \text{const}$ is a reasonable prior for τ since it favours middle values for τ . Therefore, the inherent bias in the Bayesian procedure somewhat rectifies this by pushing τ towards the edges. It does not, however, compensate completely, and the resulting posterior distribution is dominated by the implicit prior in $\pi(a, b) = \text{const}$. The same applies to d , where $\pi(a, b) = \text{const}$ implies that $\pi(d) \propto n - d$ overcompensates for the bias in the profiled log-likelihood. In reality, the posterior mean for d is often larger than the MLE, which is due to the fact that $\pi(d | y)$ is often skewed, and favouring small values for d actually makes the frequentist and Bayesian conclusions more similar.

While ‘debiasing’ the Bayesian procedure, now understood as choosing a prior such that the posterior distribution is proportional to the likelihood alone, is interesting for comparing the frequentist and Bayesian analyses, we must not forget that the frequentist and Bayesian use the likelihood differently. While the frequentist is usually mostly concerned with the maximum of the likelihood, the Bayesian is most often concerned with the mean. With symmetric distributions, this is the same, and the prior matching has convincing merit. WoC models, however, exhibits skewed likelihoods. It is therefore not universally bad that the posterior distribution favours small d , if the mean of the profiled log-likelihood is rightly skewed.

3.10 Random walk representation

For CP problems, the likelihood function is naturally represented as a random walk. Something similar is possible for WoC models, but it becomes more complicated. For the current purpose, it is advantageous to reparametrize the problem to d and τ again.

This parametrization, however, is cumbersome. Since $a \in \{1, \dots, n - 1\}$ and $b \in \{2, \dots, n\}$, $\tau \in \{3/2, 4/2, \dots, (2n - 1)/2\}$, while $d \in \{1, \dots, n - 2\}$. This means that there are combinations of τ and d which do not result in an integer value for a and b (let, for example, $\tau = 3/2$ and $d = 2$, then $a = 1/2$ and $b = 5/2$). For simplicity, we

restrict the spaces such that τ is an integer and d is an even integer. The effect of this is that d can only be even. This restriction allows for clearer notation without loss of meaningful generality.

Continuing with the random walk representation, let the log-likelihood function be expressed in these new parameters:

$$\ell(\tau, d; y) = \sum_{i=1}^{\tau-d/2} \log f_L(y_i) + \sum_{i=\tau-d/2+1}^{\tau+d/2} \log f_T(y_i; d) + \sum_{i=\tau+d/2+1}^n \log f_R(y_i)$$

By adding and subtracting terms, we get:

$$\begin{aligned} \ell(\tau, d; y) &= \sum_{i=1}^{\tau-d/2} \log f_L(y_i) + \sum_{i=\tau-d/2+1}^{\tau} \log f_L(y_i) - \sum_{i=\tau-d/2+1}^{\tau} \log f_L(y_i) \\ &\quad + \sum_{i=\tau-d/2+1}^{\tau+d/2} \log f_T(y_i; d) \\ &\quad + \sum_{i=\tau+d/2+1}^n \log f_R(y_i) + \sum_{i=\tau+1}^{\tau+d/2} \log f_R(y_i) - \sum_{i=\tau+1}^{\tau+d/2} \log f_R(y_i) \end{aligned}$$

Finally, rearranging:

$$\begin{aligned} \ell(\tau, d; y) &= \sum_{i=1}^{\tau} \log f_L(y_i) + \sum_{i=\tau+1}^n \log f_R(y_i) \\ &\quad + \sum_{i=\tau-d/2+1}^{\tau} (\log f_T(y_i; d) - \log f_L(y_i)) + \sum_{i=\tau+1}^{\tau+d/2} (\log f_T(y_i; d) - \log f_R(y_i)) \\ &= \Sigma_{CP}(\tau) + \Sigma_{WoC}(\tau, d) + \text{const} \end{aligned}$$

The first two sums constitute the normal change point likelihood, while the second two sums represent the additional WoC contributions. The first term being the CP likelihood means that we can express the first terms

$$\Sigma_{CP} = \sum_{i=1}^{\tau} \log f_L(y_i) + \sum_{i=\tau+1}^n \log f_R(y_i)$$

as

$$= \sum_{i=1}^{\tau} (\log f_L(y_i) - \log f_R(y_i)) + \text{const.}$$

Having retrieved this familiar expression, we continue to study the extra terms that the window model induces. First, when $\tau \rightarrow \infty$ and $n - \tau \rightarrow \infty$, then τ is a location parameter, as shown by Hinkley (1970). We can therefore centre the τ at zero,

$$\Sigma_{WoC}(d) = \sum_{i=-d/2+1}^0 (\log f_T(y_i; d) - \log f_L(y_i)) + \sum_{i=1}^{d/2} (\log f_T(y_i; d) - \log f_R(y_i)) \quad (3.14)$$

$$= \sum_{i=1}^{d/2} (\log f_T(y_{-i+1}; d) - \log f_L(y_{-i+1})) + \sum_{i=1}^{d/2} (\log f_T(y_i; d) - \log f_R(y_i)) \quad (3.15)$$

While this representation might be useful for further analysis, and perhaps an avenue to simulation, the prospect is unfavourable. Even for the CP random walk, $\Sigma_{CP}(\tau)$ there are no closed-form solutions for the MLE, as Hinkley (1970) showed. However, what this representation shows is that $\tau = (a + b)/2$ is a location parameter, given a and $n - b$ are sufficiently large. This means that $\Sigma_{CP}(\tau)$ can be recentred around τ_0 , preserving the properties of $\hat{\theta}$ (although shifted). This is not the case for Σ_{WoC} , showing the extra difficulty the WoC model induces. Another observation of note is that while Σ_{WoC} can be considered a random walk, the increments are dependent on d . One could be inclined to believe that a WoC model is equivalent to a CP model with two change points, but this observation shows that a WoC model is, in fact, generally different.

Interestingly, Σ_{WoC} , as a function of d , is the log-likelihood ratio for a CP model and a WoC model with window length d under a_0, b_0 .

$$\ell(\tau_0 - d_0/2, \tau_0 + d_0/2) - \ell(\tau_0, \tau_0 + 1) = \Sigma_{WoC}(\tau_0, d_0). \quad (3.16)$$

Thus, if Wilks' theorem were to hold, twice this ratio would be approximately χ^2 . Simulations, which we will do later, show that this is not the case.

Chapter 4

The simplest interesting example

In this chapter, we will investigate a very simple Window of Change model that exhibits a bias in estimating the window length $d = b - a$. It is simple in the sense that outside the window, every event is deterministic. The model is

$$\Pr_{a,b}(Y_i = 1) = 1 - w_i \quad (4.1)$$

where $1 < i < n$ and w_i is the familiar weight function. We let n be sufficiently large so that we can safely disregard edge conditions. This model enables us to analyse the effect of randomness that the window contributes.

First, we are interested in estimating b when a is known. We show that the maximum likelihood estimator $\hat{d}$ is biased, and we apply some corrections. The corrections are compared based on their success in de-biasing, their mean square error, and finally, whether they produce integer estimates or not. This is perhaps the most tractable case of estimating a random window length, yet it is still neither simple to show nor intuitive that the estimates are biased.

Then, we turn to the full model, where both a and b are unknown, and we compare the biases and estimates to those of the one-dimensional case.

4.1 Estimating the window length when the beginning is known

In the case where the a is known, we may translate the problem so that $a = 1$, and the data model is then

$$\Pr_{a,b}(Y_i = 1) = \min\left(1, \frac{i}{d}\right) \quad (4.2)$$

This represents a sequence of Bernoulli trials with the probability of success initially being low and increasing to one linearly. The likelihood function is the following

$$L(d; y) = \prod_{i=1}^n \min\left(1, \frac{i}{d}\right)^{y_i} \min\left(1, \frac{d-i}{d}\right)^{1-y_i},$$

where we must interpret the indeterminate form 0^0 as one, which is intuitive since not observing an impossible outcome should not influence our belief about anything. Furthermore, the possibility of a zero likelihood also implies a possibility that the log-likelihood is negative infinity. This fact is unproblematic as the likelihood is the appropriate function for all computations, while the log-likelihood is simply a formulation that is easier to work with.

The goal is estimating d from the observed values of y . Define A as the index of the last observed 0:

$$A = \max \{k : y_k = 0\}$$

and note that the probability that $d < A$ is zero, which in turn yields a likelihood of zero. Thus, any estimator $\hat{d}$ is bound below from A . We can truncate the product in (4.1) after $i = d$, assuming $d > A$ as any contribution beyond this index is simply 1. This yields the clearer expression:

$$L(d; y) = \prod_{i=1}^d \frac{i^{y_i} (d-i)^{1-y_i}}{d^d}, A < d$$

This likelihood is complex in the sense that d is not only determining the window length, and thus the upper limit in the product, but it is also a parameter in the factors since it determines the likelihood of each observation. The treatment of the likelihood function is therefore difficult and is not similar to a CP problem, where the change point parameter τ only appears in the limits of the product (or limits of the sum in the log-likelihood). In fact, if $d = 1$, then the whole has no free parameters and is clearly deterministic. Therefore, this example examines the additional phenomena that are only present in, as well as fundamental to, a WoC model.

4.2 Maximum likelihood estimation

Before we turn to the MLE, we will briefly discuss the event A and how we can produce an estimator from it. $A + 1$ is a natural estimator for d , since the observations trailing A are all ones. Note that we must add one to A as $y_A = 0$ by definition, and this in turn requires that $A < d$. In other words,

$$\Pr(A \leq d) = 1.$$

This also shows that this estimator is negatively biased. The only event which makes $A + 1 = d_0$ is if $y_{d-1} = 0$, which has a $1/d$ chance of happening. In all other cases will $A + 1 < d_0$.

Instead, consider the canonical estimator for many statistical problems: the maximum likelihood estimator:

$$\hat{d} = \arg \max_d L(d; y).$$

It is well-defined and unique.

Example 4.2.1 (Bernoulli example). Let $y = (0, 0, 1, 0, 1, 1, 1, 1, 1, 1)$. Then $n = 10$ and $A = 4$. The likelihood of $A + 1$ is:

$$\begin{aligned} L(5; y) &= \prod_{i=1}^{10} \min(i/5, 1)^{y_i} \max(1 - i/5, 1)^{1-y_i} \\ &= \prod_{i=1}^5 (i/5)^{y_i} (1 - i/5)^{1-y_i} \\ &= (1 - 1/5)(1 - 2/5)(3/5)(1 - 4/5) \\ &= 4 \cdot 3 \cdot 3 \cdot 1/5^4 \\ &= 0.0576 \end{aligned}$$

The MLE is $\hat{d} = 6$, with the likelihood $L(6; y) = 0.077$. ▲

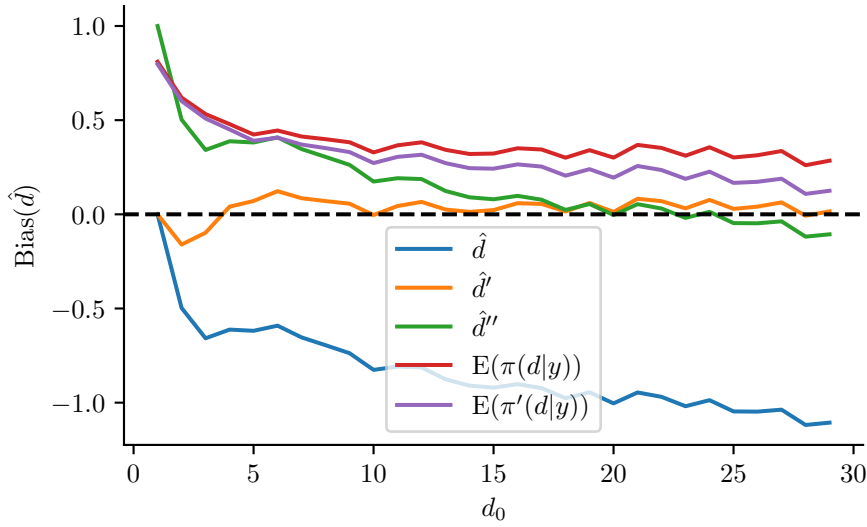


Figure 4.1: Bias for the four estimators. Notice the bias for $\hat{d}$, and that $\hat{d}'$ quite fast returns to zero. $\hat{d}''$ is $\hat{d}$ shifted by one. The posterior means, $\mathbb{E}[\pi(d|y)]$ and $\mathbb{E}[\pi'(d|y)]$ are both positively biased. The latter, which has a decreasing prior for d , is less skewed.

A surprising property of $\hat{d}$ is that it too is negatively biased. Ignoring the case when $d_0 = 1$ (which implies that $y_i = 1$ for all i) is $\text{Bias}(\hat{d}) < 0$. See figure 4.1. It is difficult to find a closed-form expression for the bias as a function of d , but for any particular d , there are only 2^d unique data sets. For each of these data sets, we can compute the MLE and its probability of being observed, thus calculating $\text{Bias}(\hat{d})$ directly. For larger values of d , estimation is necessary but poses no challenge. We can easily sample data sets from the model (4.1) and find the d that maximizes the likelihood.

We now define the correction procedures from Chapter 3, using the correction function $c(d)$. There are no nuisance parameters in this problem, and we can therefore precompute the values for $c(d)$. Since $\hat{d}$ is negatively biased, this function is bounded above by one: $c(d) < 1$. The first correction,

$$\hat{d}' = \hat{d}/c(d),$$

successfully de-biases $\hat{d}$, and it even reduces the mean square error (see figure 4.1). The second correction,

$$\hat{d}'' = \arg \min_d (|\hat{d} - \mathbb{E}_d[\hat{d}]|),$$

also reduces the bias but does not remove it entirely. The reason has to do with the otherwise desirable quality that it only produces integer estimates. Before we see why, notice that it is equivalent to

$$\hat{d}'' = \hat{d} - \text{round}(\text{Bias}(\hat{d})),$$

where $\text{round}(x)$ is the nearest integer to x . For small d_0 , $\text{round}(\text{Bias}(\hat{d})) = -1$, thus $\hat{d}'' = \hat{d} + 1$, which in turn implies that $\text{Bias}(\hat{d}'')$ has the same shape as $\text{Bias}(\hat{d})$, only shifted back towards zero when $\text{Bias}(\hat{d})$ wanders too far away from zero.

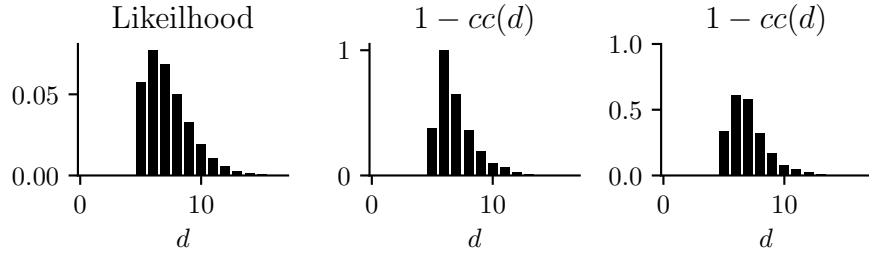


Figure 4.2: Likelihood (left), inverted confidence curve (middle) and inverted confidence curve with half correction (right) for example data (4.3). Note that both are zero for $d < 5$ because $A = 4$. The half corrected cc does not have the prominent peak at the MLE, but is somewhat tighter in the sense that $d = 11$ is excluded from its 95% confidence set.

4.3 Confidence curves

To construct the confidence curves for d , we only need the deviance

$$D(d, Y) = 2[\ell(\hat{d}) - \ell(d)].$$

The confidence curve is then

$$cc(d) = \Pr_d(D(d, Y) < D(d, y)).$$

To calculate the $cc(d)$, we simulate as Cunen, Hermansen and Hjort (2018) suggests:

$$cc(d) \approx \frac{1}{B} \sum_{j=1}^B I\{D(d, Y_j^*) < D(d, y)\}$$

where Y^* are simulated datasets under each d of interest. An example cc using the vector

$$y = 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1 \quad (4.3)$$

is displayed, with the corresponding likelihood in figure 4.2. The 95% confidence set is $\{5, 6, 7, 8, 9, 10, 11\}$, which is quite wide.

We also wish to investigate if the theoretical properties of the confidence curve are achieved by our construction. The necessary property is that for any degree of coverage $1 - \alpha$, the set $\{d : cc(d) < 1 - \alpha\}$ should contain the true d_0 with probability $1 - \alpha$. In other words, let

$$U = cc(d_0, Y)$$

be the random variable defined as the value of the confidence curve at the true parameter d_0 . Here, we let the confidence curve expression explicitly depend on the random variable Y . If the construction is correct, then the distribution of U should be near uniform. However, since it is discrete, there will always be a certain positive probability that $cc(d_0, Y) = 0$, and it cannot therefore be exactly uniform. A common fix for this is *half correction*. It is achieved by adjusting the confidence curve definition slightly, which we call $cc_{1/2}$:

$$cc_{1/2}(d) = \Pr_d(D(d, Y) < D(d, y)) + \frac{1}{2} \Pr_d(D(d, Y) = D(d, y)).$$

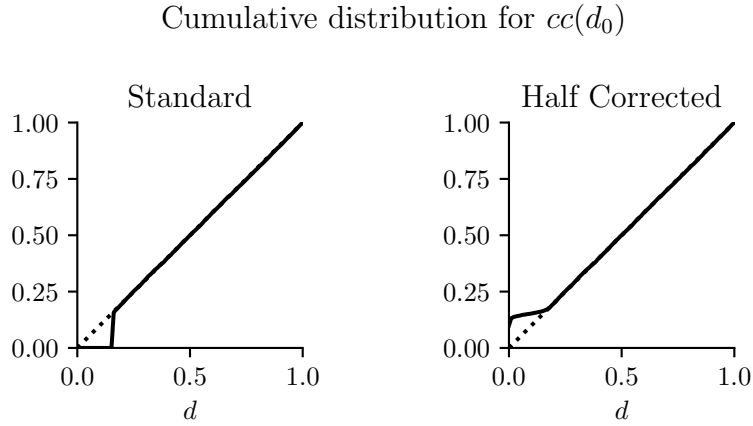


Figure 4.3: Cumulative sample distribution for $cc(d_0, Y)$ (left) and the corrected version (right) for $d_0 = 20$ and $n = 30$. The diagonal is dotted. 10000 samples were used.

In essence, the event that the observed deviance is equal to the random deviance is weighted with a half. In many cases, this adjustment yields a random variable with better coverage properties. However, in our case, it seems not to improve the confidence curve significantly (see figure 4.3). Both the standard and the corrected version of the confidence curve show a large deviation from uniformity around zero, but it quickly conforms to the theoretical distribution for larger values of U . For $0.15 < U$, the observed quantiles are almost identical to the theoretical ones.

The fact that U is very similar to a uniform random variable for large values is very important because it means that the coverage for 95% confidence regions would be correct. Formally, letting $V \sim \text{Uniform}$, this means that

$$\Pr(U < x) \approx \Pr(V < x), \quad 0.15 < x,$$

and, more importantly, that confidence regions at any $1 - \alpha < 0.85$ have the correct coverage. For most practical purposes, we can say that the confidence curve has the correct coverage. On the other hand, turning our attention to estimation once again, the fact that U is not similar to a uniform variable for small values shows that the peak of the log-likelihood function systematically misses the true parameter, which is another way of demonstrating that the MLE is biased.

The point at which U has the same quantiles as a uniform distribution depends on d_0 . When d_0 is smaller, this threshold increases. Without discussing the reason for this in detail, it makes sense that as d_0 increases, it is more accurately approximated by a continuous variable, and the point mass probability on $\Pr(cc(d_0, Y) = 0)$ decreases. Conversely, as d_0 decreases, we expect it to be approximated poorly by a continuous variable.

In summary, the confidence curve has been applied to our very simple problem and shown to have the desired properties regarding confidence sets. Furthermore, it seems that there is no real advantage to half correcting the confidence curve in this example. Half correcting also has the unfortunate effect of removing the very prominent peak of the confidence curve, which points at the MLE. This reduces the visual clarity of the confidence curve as an inference tool, and we will therefore continue to work with the uncorrected one.

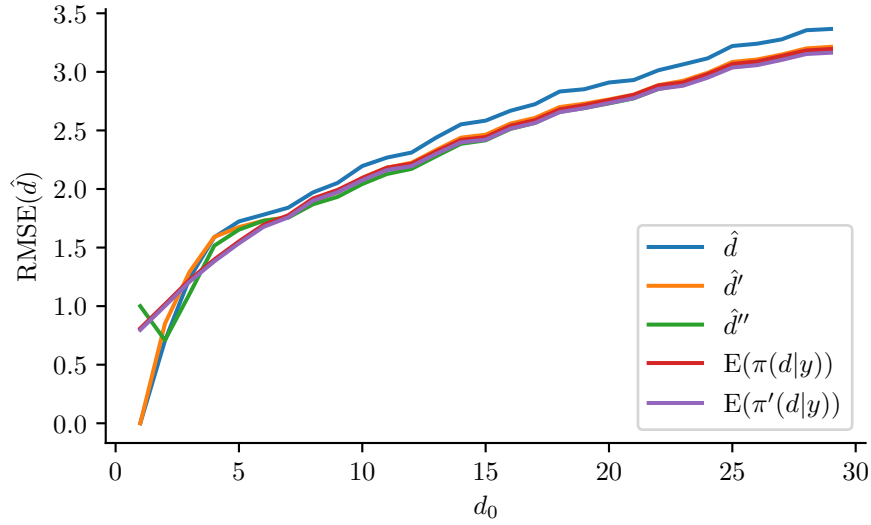


Figure 4.4: Root Mean Square Error (RMSE) for the estimators. It shows that all corrections are comparatively effective at reducing the error.

4.4 Bayesian analysis

The Bayesian analysis requires that we elicit a prior for all unknown parameters, which currently only is d . This means that we are not in the situation discussed in Chapter 3 where the joint prior on a, b induces the prior on d . We might therefore freely choose a flat prior on d . However, we will consider both the constant prior $\pi(d) = \text{const}$ and the would-be-induced prior $\pi'(d) \propto n - d$. For the two priors, define the two posterior distributions

$$\begin{aligned}\pi(d | y) &\propto L(d; y) \\ \pi'(d | y) &\propto L(d; y)(n - d)\end{aligned}$$

using the likelihood (4.1).

With the posterior distribution described, we continue with the posterior estimation. The most common Bayesian estimator is the posterior mean,

$$\hat{d}_B = \mathbb{E}[d | y] = \sum_{d=1}^n d \cdot \pi(d | y),$$

which is the minimizer of the expected square loss. This estimator shares the unfortunate property of (3.3) producing non-integer solutions, and we could therefore consider two other alternatives: the posterior mode and the posterior median. The posterior mode is identical to the MLE in the case of a constant prior and is quite similar. The posterior median requires a more careful definition, since $\pi(d | y)$ is discrete.

Using a uniform prior - since the likelihood is rightly skewed, at least for small values ($d < 100$) of d , the posterior mean is larger than the maximum likelihood estimator. The likelihood is, in fact, so skewed that the Bayes estimator is positively biased, although the magnitude of the bias is smaller than that for the maximum likelihood estimator.

4.5 The unknown window

We continue now with the rather more interesting case when both a and b are unknown. Again, we are concerned with the case where a and $n - b$ are sufficiently large such that edge conditions can be ignored.

The likelihood is

$$L(a, b; y) = \prod_{i=1}^a w_i^{1-y_i} (1 - w_i)^{y_i}, \quad (4.4)$$

where 0^0 is defined to be one. Here, w_i is the weight function (defined at 3.1.1). The reason that w_i is raised to the power of $1 - y_i$ instead of y_i is that the weight function is decreasing, while $\Pr(Y_i = 1)$ is increasing.

While this is a WoC model, we have not discussed what kind of transition we are modelling. The reason for this is that in this very simple example, both the parameter transition and the mixture transition yield the exact same likelihood. First, we can let $Y_i \sim \text{Bernoulli}(1 - w_i)$, thus the model is clearly a transition model where the parameter is the probability of ‘success’. As a mixture model, observe that under f_L , $\Pr(Y_i = 0) = 1$, and under f_R , $\Pr(Y_i = 1) = 1$. The mixture likelihood is then exactly the same as (4.4).

We continue by extending the estimation procedure to two dimensions. The maximum likelihood estimator is readily generalized

$$(\hat{a}, \hat{b}) = \arg \max_{a, b} L(a, b; y)$$

subject to $1 < a, a < b, b < n$. This estimator is, however, also biased, but in a slightly more complicated sense. Consider the following two transformations of a, b :

$$\begin{aligned} \tau &= \frac{a + b}{2} \\ d &= b - a. \end{aligned}$$

The inverse transformations are then

$$\begin{aligned} a &= \tau - d/2 \\ b &= \tau + d/2. \end{aligned}$$

τ is the middle of the window, denoted thus to parallel Change-Point analysis, and d is the window length. $\hat{\tau}$ must be unbiased, which can be seen from the symmetry argument in Chapter 3. We have seen from simulation that $\hat{d}$ is negatively biased when a is known, so the question is whether an unknown a affects this bias. The answer is that it increases the bias almost four times as much (see figure 4.5). Furthermore, if $\hat{\tau}$ is unbiased and $\hat{d}$ is negatively biased, then $\hat{a}$ must be positively biased, and $\hat{b}$ must be negatively biased. The proof that $\hat{\tau}$ is unbiased is not immediately applicable since there are no parameters in this model. However, the same inversion argument works, namely that if, say, $\text{Bias}(\hat{\tau}) < 0$, then $\text{Bias}(\hat{\tau}_x) > 0$ where τ_x is the change point of the reversed sequence. This reversed sequence can be re-expressed with the variable $x = 1 - y$ (hence the subscript x in τ_x), and it would have identical likelihood; thus, we have reached a contradiction.

Another observation of note is that $\hat{d}$ can never be 2. First, if $d_0 = 2$, then y_{a+1} is the only observation from within the window, and therefore the only observation whose value is random. Thus, y must be a contiguous sequence of zeros followed by a contiguous sequence of ones. If $y_{a+1} = 0$, then there are $a + 1$ zeros followed by $n - b$ ones, and if

$y_{a+1} = 0$, then there are a zeros and $n - b + 1$ ones. In both cases $\hat{d} = 1$. If $d_0 > 2$, then either the likelihood when $d = 0$ is zero, or we are in the previously discussed case where y consists of contiguous zeros followed by contiguous ones. Thus, we see that for any given d_0 , there is a certain probability that the observed y could have been produced by some smaller d'_0 .

We extend the correction procedure to two dimensions, and since $\hat{\tau}$ is unbiased, we can correct the window length in isolation. Since we let a and $n - b$ be suitably large, we can fix a and compute the bias for $\hat{d}$ for any given values. Then, when we correct the estimates, we only correct the window length and keep the estimate for τ fixed. In other words, given the estimates $\hat{a}, \hat{b}$, we find $\hat{d}$ and $\hat{\tau}$. Then, let $d' = \hat{d}/c\hat{d}$, and then $a' = \hat{\tau} - d'/2$ and $b' = \hat{\tau} + d'/2$. The procedure is the same for the moment-matching correction, but we take care to let a'' and b'' be integers. This means that if $\hat{d}$ is even, then $\hat{d}''$ must also be even, and vice versa.

Do the corrections succeed? First, all reduce the bias successfully, but only $\hat{d}''$ removes it completely. When a was fixed, they all performed similarly, but $\hat{d}'$ had slightly better RMSE than $\hat{d}''$ (see figure 4.1). Now, $\hat{d}''$ clearly outperforms $\hat{d}'$ in terms of bias correction and is only slightly worse regarding RMSE (see figure 4.5). In all cases, the corrections do not work when d_0 is very small, but as it grows, they all reduce some bias at their own pace.

The Bayesian analysis is also readily extended to two dimensions. We need to elicit a more complicated prior for a, b jointly, but we will keep it constant, as we have already discussed. This induces a linearly decreasing prior in d , and a triangular prior in τ centred around $n/2$. Since a, b are the only unknown parameters in the model, the WoC model is now fully specified:

$$\pi(a, b \mid y) \propto L(a, b; y).$$

The posterior is joint for a, b , so if we want inference on, say, a alone, we must marginalize:

$$\pi(a \mid y) = \sum_{b=a+1}^n \pi(a, b \mid y)$$

and correspondingly for b . Furthermore, if we want inference on d and τ we must re-parametrise accordingly:

$$\pi(d \mid y) = \sum_{a=1}^{n-d} \pi(a, a + d \mid y)$$

and

$$\pi(\tau \mid y) = \begin{cases} \sum_{i=0}^{\min(\tau, n-\tau)} \pi(\tau - i - 0.5, \tau + i + 0.5 \mid y), & \tau \text{ is even} \\ \sum_{i=0}^{\min(\tau, n-\tau)} \pi(\tau - i, \tau + i \mid y), & \tau \text{ is odd} \end{cases}.$$

For the purposes of finding posterior means, however, we can simply manipulate the expectations. Since both d and τ are linear combinations of a and b , can we distribute the expectations:

$$\begin{aligned} \mathbb{E}[d \mid y] &= \mathbb{E}[b \mid y] - \mathbb{E}[a \mid y] \\ \mathbb{E}[\tau \mid y] &= (\mathbb{E}[a \mid y] + \mathbb{E}[b \mid y]) / 2. \end{aligned}$$

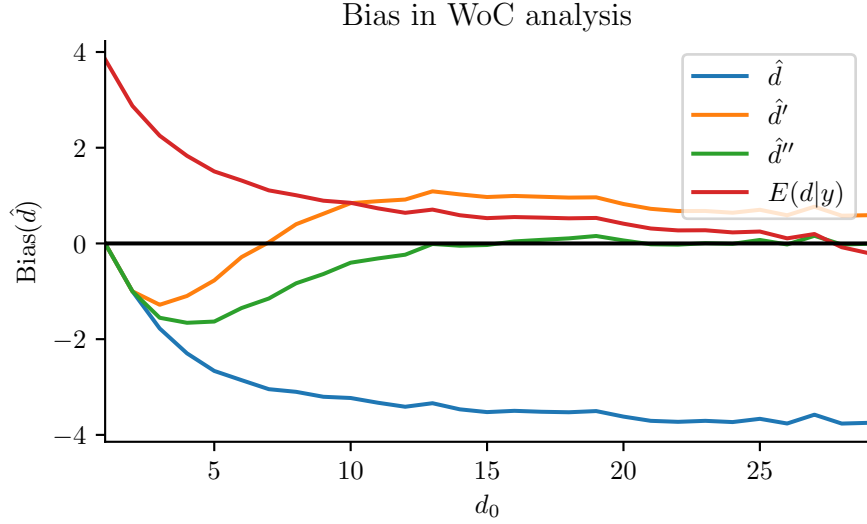


Figure 4.5: Bias for when estimating $d = b - a$ with $a = 10$ and $n = 50$ as a function of d_0 . We see that the MLE is strongly negatively biased, while the posterior mean is slightly positively biased, but recedes towards zero. The first correction, $\hat{d}'$ is an improvement over the MLE, but not as strongly as when a was known. The final correction, $\hat{d}''$, which finds the set $\hat{a}''$, $\hat{b}''$ such that the expected bias matches the MLE, seems to work very well. Calculated using $n_{sim} = 10000$ samples

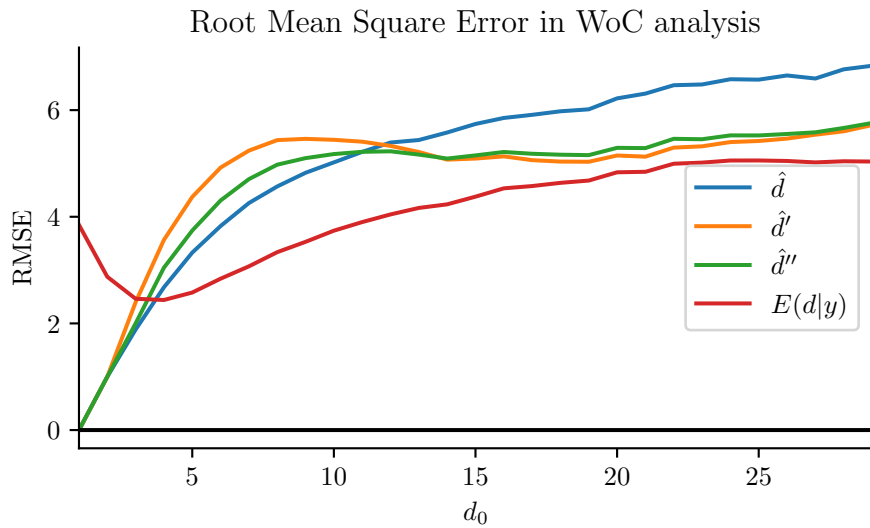


Figure 4.6: RMSE for the four estimators. Contrary to only d unknown does the corrections not universally decrease the error. When $12 < d$, they are all an improvement, but the posterior mean is clearly the estimator with the smallest error.

If we compare the Bayes and the ML estimator for d alone, we see a similar phenomenon, where the posterior mean for d is positively biased but approaches zero, until ultimately it too becomes negatively biased. This effect is even stronger when both a and b are unknown (see figures 4.5 and 4.6).

Finally, we wish to construct confidence curves for a and b . We will use the constructions from section 3.7, along with an example data set. The data set was sampled with the true parameters $a_0 = 20$ and $b_0 = 40$, implying that $d_0 = 20$ and $\tau_0 = 30$, and the length of the observation vector was $n = 100$. First, we will consider the joint analysis for a and b , followed by the univariate confidence for τ and d . Figure 4.9 shows the confidence curve for a, b , which is sharply defined along two axes. This is because $\hat{a}$ is bounded from above by the first observed 1, at index $i = 23$, and $\hat{b}$ is bounded below by the last observed 0, the event we called A , at index $i = 34$. The confidence curves for τ and d are more interesting (figure 4.10), which clearly shows that there is more uncertainty when estimating d since the cc is wider.

We now want to compare the confidence curve to its Bayesian counterpart. We cannot easily define the multivariate confidence distribution, and instead, use the highest posterior density (HPD) regions for comparison with the frequentist analysis. An HPD region for a parameter θ and a given coverage $1 - \alpha$ is the smallest region such that $1 - \alpha$ of the posterior mass/density is contained within it. We may define it as

$$HPD^{1-\alpha}(\theta \mid y) = \min \{ \theta : \pi(\theta \mid y) \geq 1 - \alpha \}$$

for some parameter θ and posterior $\pi(\theta \mid y)$. If we want inference on τ or d alone, we can simply marginalize the posterior, and then we can find the posterior credibility curve for τ or d as in the other univariate case. For example, the posterior credibility curve for τ is

$$cc_B(\tau) = \left| 1 - 2 \int_{-\infty}^{\tau} \pi(t \mid y) dt \right|.$$

The results are displayed in figure 4.9 and 4.10.

Comparing the Bayesian and frequentist confidence/credibility sets for the given sample data, we see that they largely concur. The 2-dimensional confidence/credibility curves look almost identical, and the 95% confidence/credibility regions are also almost identical. Thus, at this confidence level, being a Bayesian or a frequentist is a matter of taste. The story is slightly different when looking at the univariate confidence/credibility. We see that the posterior distributions are slightly slimmer for larger α , suggesting that the concurrence depends on the significance level. Regarding estimating τ , the MLE and the posterior mean are almost identical, which is reassuring since the flat prior on a, b actually places larger prior mass on values in the centre of the interval (since $n = 100$ we expect $\mathbb{E}[\tau \mid y]$ to be pushed towards 50). In fact, the posterior mean is slightly smaller. The prior weight on central values for τ is clearly not skewing the posterior too much. The estimation of d tells a similar story, the posterior mean being larger than the MLE, even though the prior places larger mass on small d . As we have seen, the posterior mean for d has a slight positive bias, around $\text{Bias}(\mathbb{E}[d \mid y]) \approx 1$.

4.6 A summary

While the data model in our very simple example might seem like a strange model with few applications, it demonstrates the extra uncertainty introduced solely by the window of change. Remember that estimating τ in the Change-Point version of the very simple example is trivial, as there is no randomness whatsoever. Therefore, all uncertainty

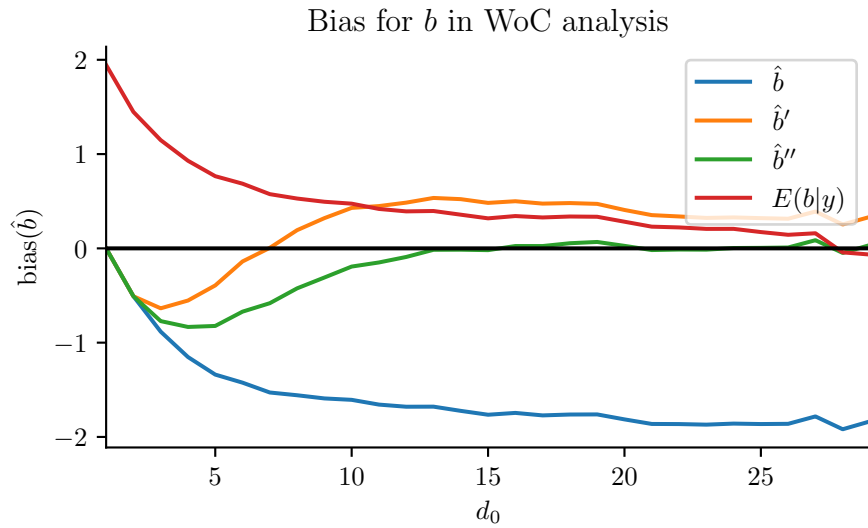


Figure 4.7: Bias for b as a function of d_0 . It is similar to figure 4.1, but the bias is stronger, and

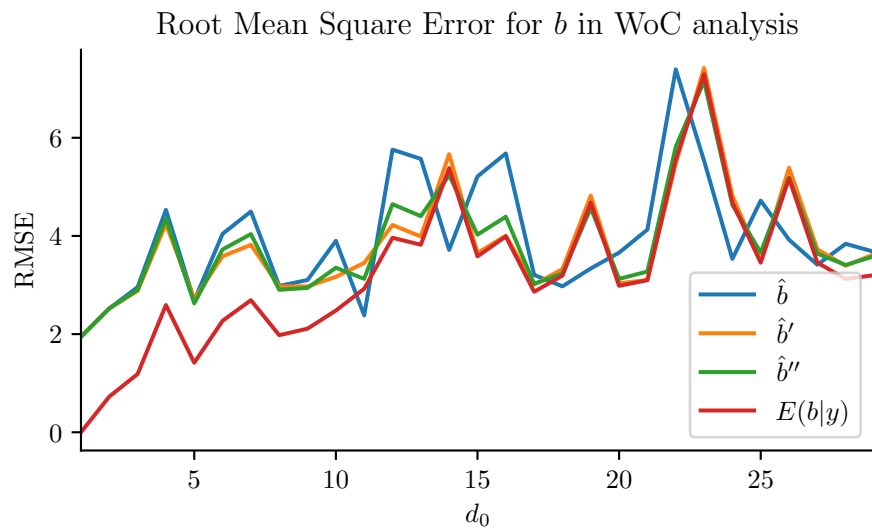


Figure 4.8: RMSE as a function of d . The chaotic nature is not due to using only a few samples, but truly reflect the RMSE.

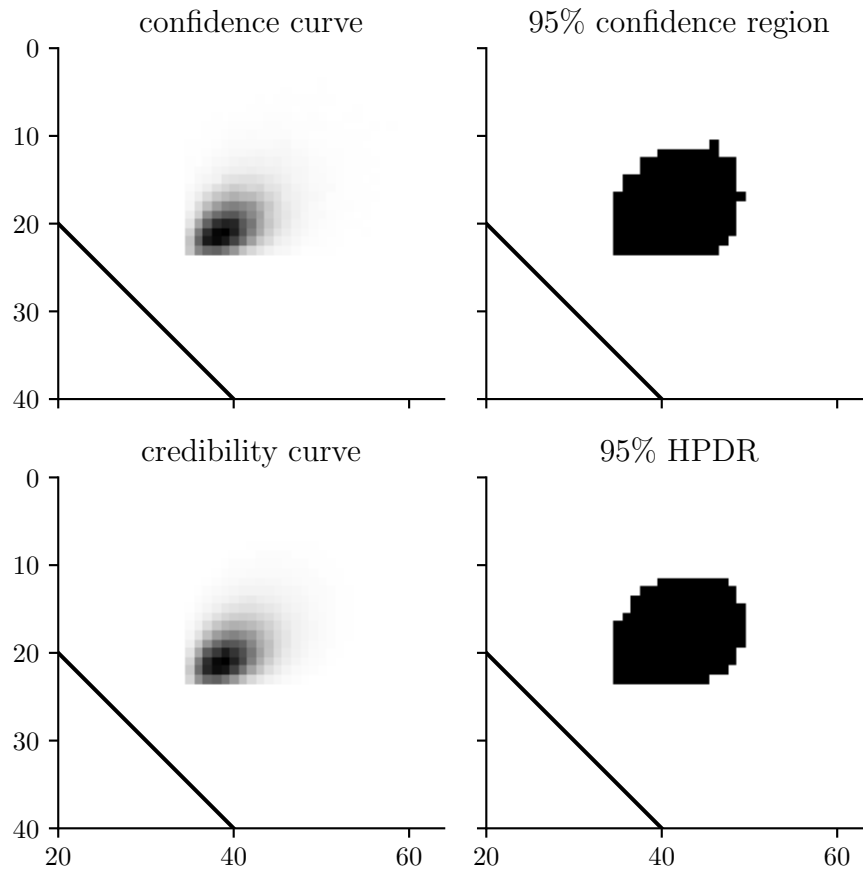


Figure 4.9: Upper left: Confidence curve with $a_0 = 20$, $b_0 = 40$, and $n = 100$. Upper right: 95% confidence region. Lower left: Posterior $\pi(a, b \mid y)$ using flat prior on a, b . Lower right: 95% highest posterior credibility interval.

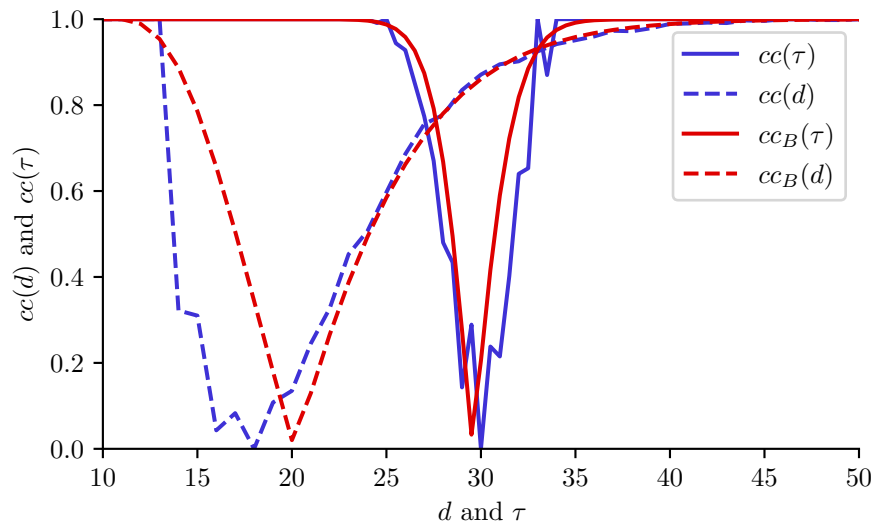


Figure 4.10: Confidence for τ (blue solid), confidence for d (blue dashed), credibility for τ (red solid), and credibility for d (red dashed) for the same data as in figure 4.9. The true values are $d_0 = 20$ and $\tau_0 = 30$.

when estimating τ and d is due to the introduction of the transition window. This does not mean, however, that the total uncertainty in WoC problems is the sum of the CP uncertainty and then the WoC, but it clearly must be larger than in CP problems.

Furthermore, the very simple model is, in fact, the limit of a mixture distribution. The limit here refers to the fact that if two distributions f_L and f_R are sufficiently different, such that we can trivially see which distribution any given sample y_i is from, we may replace the observations y_i with a new variable, say z_i , indicating which distribution it was sampled from. Thus, the results we have shown represent the best case estimation in mixture models, which, frankly, is shockingly bad.

Regarding the parameter distribution, this analysis is not the limit in any reasonable manner. In fact, if f_L and f_R become sufficiently different, we expect to estimate a, b very well, which we shall see in the following chapter.

Chapter 5

Normal simulation

Assuming the observations are deterministic outside the window is, of course, unlikely to be wise in reality. We must therefore extend the model to the case where f_L and f_R are proper probability distributions. As an example of a proper distribution, we let f_L and f_R be normal with known parameters. We are interested in how the estimate for d behaves, if the corrections still work, and if the Bayesian and frequentist inference still resemble each other. Furthermore, now we must choose the transition model, and we are hence interested in if the choice has a considerable effect on the inference. First, we do it for a parameter transition, then a mixture transition.

5.1 Normal data with parameter transition

Starting with the parameter transition, let $f_L = N(0, 1)$ and $f_R = N(\Delta\theta, 1)$. This one-parameter problem is equivalent to any normal model with constant variance. To see this, consider the case where we are interested in $\theta_L = (\mu_L, \sigma^2)$ and $\theta_R = (\mu_R, \sigma^2)$, then $\Delta\theta = (\mu_R - \mu_L)/\sigma^2$ in the one-parameter formulation. $\Delta\theta$ is the relative level difference, and we can consider it the ‘strength’ of the signal. Therefore, if $\Delta\theta$ is small, we expect to struggle to estimate the window, and when it increases we expect to more easily estimate the window. The log-likelihood for an observation y from a WoC model is then:

$$\ell(a, b, \Delta\theta; y) = \sum_{i=1}^n \left[-\frac{1}{2}(y_i - (1 - w_i)\Delta\theta)^2 - \frac{1}{2} \log 2\pi \right] \quad (5.1)$$

$$= -\frac{1}{2} \sum_{i=1}^n (y_i - (1 - w_i)\Delta\theta)^2 + \text{const.} \quad (5.2)$$

Since we assume we know $\Delta\theta$, only a, b are unknown, and we again have a two-parameter WoC model. The MLE is

$$(\hat{a}, \hat{b}) = \arg \max_{a, b} -\frac{1}{2} \sum_{i=1}^n (y_i - (1 - w_i)\Delta\theta)^2$$

subject to $i < a < b < n$. The Bayesian analysis is also trivially extended for the same reason, namely that $\Delta\theta$ is known:

$$\pi(a, b \mid y) \propto \ell(a, b; y),$$

where we have assumed a constant prior on a, b .

Parameter transition

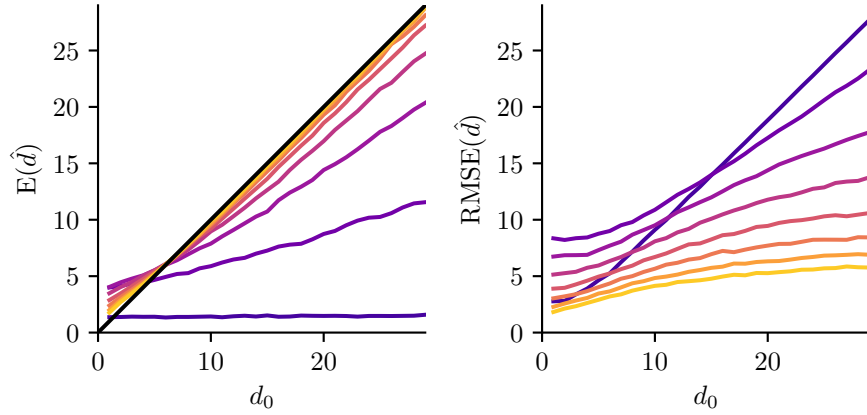


Figure 5.1: Left: Expected value for the ML estimator $\hat{d} = \hat{b} - \hat{a}$ as a function of the true d_0 for a selection of $\Delta\theta \in (\frac{1}{2}, \frac{2}{2}, \frac{3}{2}, \frac{4}{2}, \frac{5}{2}, \frac{6}{2}, \frac{7}{2}, \frac{8}{2})$. The deepest purple is $\Delta\theta = 1/2$, and the brighter colour are higher values with yellow being $\Delta\theta = 4$. The black line is the diagonal $\mathbb{E}[\hat{d}] = d$, which represents an unbiased estimator. Initially, the MLE overestimates d , but eventually all underestimates it. Curiously, for the small $\Delta\theta = 0.5$, seems $\hat{d}$ invariant under d_0 . Right: Expected Root Mean Square Error for the same $\Delta\theta$. As we might expect, the greater the level difference, i.e. $\Delta\theta$, the easier is it to identify it, and thus the smaller the error.

For $\Delta\theta \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$ we simulate 10000 data sets and find the MLE. Since the normal distribution is symmetric, and the random variables before and after the window only differ by a constant, is $\hat{\tau}$ unbiased. We therefore turn our attention again to $\hat{d}$, and compute its expected value given a random data set. The results can be seen in figure 5.1. Indeed, we see that as $\Delta\theta$ is large, we estimate d with little bias. Initially, for small d_0 , the estimates are positively biased, but after around $d_0 = 7$ they all are negatively biased. Notably, even when $\Delta\theta = 4$ the bias is present, although it is only around 0.15. More surprisingly, when $\Delta\theta = 0.5$, $\mathbb{E}[\hat{d}] \approx 2$ regardless of d_0 , and we might say that the estimate utterly fails at its task.

The fact that the bias subsides, but does not disappear is not too surprising considering the difficulties of estimating d in the Bernoulli model, and we could therefore try to correct it similarly. Again, we only correct $\hat{d}$ as $\hat{\tau}$ is unbiased. We can compute the c function

$$c(d, \Delta\theta) = \frac{\mathbb{E}_{d, \Delta\theta}[\hat{d}]}{d},$$

and try the same correction

$$\hat{d}' = \frac{\hat{d}}{c(\hat{d}, \Delta\theta)}.$$

To simplify things, we will only consider the case where $\Delta\theta = 2$, as it seems like a reasonable middle ground where correction seems both necessary and possible. Furthermore, we let $a_0 = 20$ be fixed, $n = 100$, and consider only $d < 30$.

First, we calculate the correction curve $c(d)$ with 10000 samples. Now, there is a considerable probability that $\hat{d}$ is greater than 30, and we must therefore compute $c(\hat{d})$ for all plausible values of $\hat{d}$. To find the maximum value for which $c(d)$ needs to be

calculated, we can simply sample $\hat{d}$ from data generated with $d_0 = 30$, and let the maximum of this sample be the endpoint. If n is very large, this is advisable. Since $n = 100$ is not very large, we simply compute $c(d)$ for all possible values. The results, shown in figure 5.2 and figure 5.3, show a surprising departure from the previous results. None of the corrections seems to provide a unilateral improvement. Firstly, none of the estimators are without bias, which means that they fail at their primary task. Still, $\hat{d}'$ and $\hat{d}''$ reduce it somewhat, approximately halving it, but at the cost of an increase in RMSE. It is well known that introducing bias can improve the RMSE of estimators, and in this regard, the phenomenon is not surprising. While the posterior mean is the most efficient at reducing the RMSE, it introduces a stronger bias than the MLE and the other corrections. It is therefore not trivial to advocate for one estimator over the other, as the optimal one depends on the criterion and what the true value of d_0 is.

If $\Delta\theta$ is unknown, we need the profiled log-likelihood, which in this case can be found analytically. Let $\widehat{\Delta\theta}_{a,b}$ be the MLE for $\Delta\theta$ with a and b fixed. Then,

$$\widehat{\Delta\theta}_{a,b} = \frac{\sum_{i=1}^n (1 - w_i) y_i}{\sum_{i=1}^n (1 - w_i)^2},$$

which gives the profiled log-likelihood

$$\ell_{\text{prof}}(a, b; y) - \frac{1}{2} \sum_{i=1}^n (y_i - (1 - w_i) \widehat{\Delta\theta}_{a,b})^2 + \text{const.}$$

This extension to an unknown level difference does not change the story in any remarkable way. In fact, both the bias and the RMSE seems almost identical, which again suggests that it is estimating the window which is the real challenge, and estimating the remaining parameters is easy. Since the results are so similar, will we not redo the former analysis with it, and rather only define it. For real data sets, we need this full procedure as we do not know all the parameters, but it is very slow to simulate for the selection of $\Delta\theta$ and d_0 we have hitherto concerned ourselves with. Since as $a \rightarrow \infty$ and $n - b \rightarrow \infty$, $\Delta\theta$ can be estimated with arbitrary precision, the reduction in apparent uncertainty becomes less relevant. This correction is thus only necessary when we can not adequately estimate $\Delta\theta$. For example, if $n - b = 100$, which means that there are 100 observations from the right distribution, then the width of the 95% confidence interval for $\Delta\theta$ is approximately $2\Phi(0.95)/\sqrt{100} \approx 0.4$.

Returning to the full correction procedure, we begin by finding the MLE $\hat{a}, \hat{b}$ and the corresponding $\widehat{\Delta\theta}$. With this level difference, simulate the c curve with this as $\Delta\theta$, and finally, produce the correction as before:

$$\hat{d}' = \frac{\hat{d}}{c(\hat{d}, \widehat{\Delta\theta})}. \quad (5.3)$$

Some remarks are in order. Since $\hat{d}$ is negatively biased, $\widehat{\Delta\theta}$ will also be negatively biased. When $n - b \rightarrow \infty$, it is, however, consistent, so the effectiveness of $\hat{d}'$ would depend on the amount of data. Luckily, we have seen from experience that $\widehat{\Delta\theta}$ is well estimated. The full correction procedure is then: find $\widehat{\Delta\theta}$ and compute the bias for $\hat{d}$ as a function of d_0 with this level difference, then compute $\hat{d}$ and apply the formula (5.3). Since this correction procedure requires a unique simulation for each simulated data set, it will not be investigated more closely now.

We construct confidence as we did with the simplest example, which can be seen in figure 5.4. The cc is wider compared to the Bernoulli model, which is to be expected

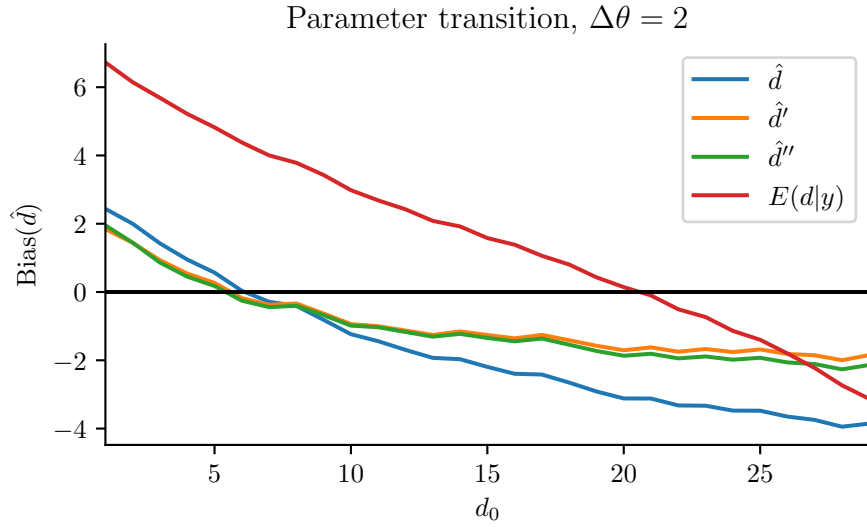


Figure 5.2: Bias for d as a function of d_0 . Here, we see that the corrections are weaker. The posterior mean is initially very strongly positively biased, and decreases almost linearly until it again is negatively biased. Both $\hat{d}'$ and $\hat{d}''$ correct some bias, but not entirely.

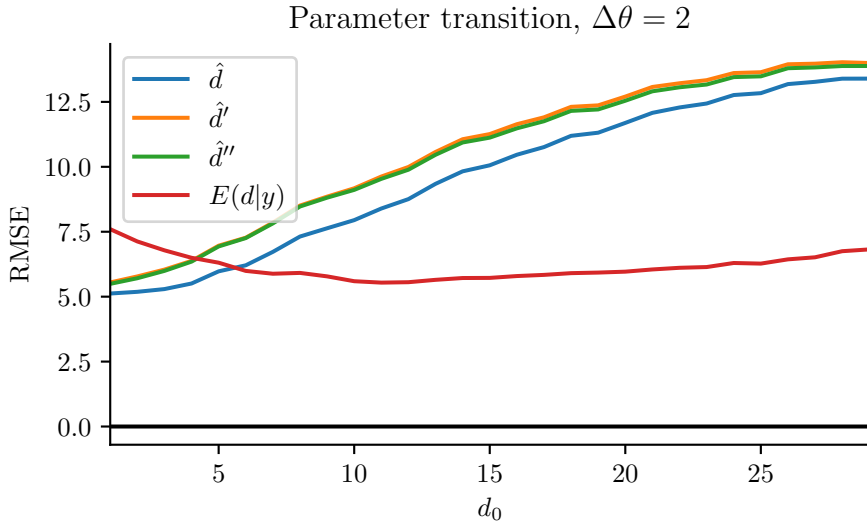


Figure 5.3: RMSE as a function of d . We see that the posterior does indeed give the smallest error for $5 < d_0$, and curiously, $\hat{d}'$ and $\hat{d}''$ gives a larger error than simply the MLE.

as the noise would make estimation more difficult. The cc is still oblong, showing that there is more uncertainty in estimating d . We see that the 95% confidence region for a, b touches the diagonal, which means that there exists some a where $d = 1$ cannot be rejected. However, if we look at the cc for d (figure 5.5), $d = 1$ is in fact rejected. This means that whether to reject $d = 1$ depends on if we are concerned about inference on τ as well or not. We must therefore be careful when interpreting the $cc(a, b)$. The story is mirrored in the Bayesian analysis, where the credibility curve is very similar to the confidence curve, and the corresponding confidence regions too are almost identical. The marginal credibility for d is however slightly different. It is pushed towards larger d , and somewhat slimmer. The 95% credibility interval is similar, but for lower credibility,

the Bayesian interval is slimmer.

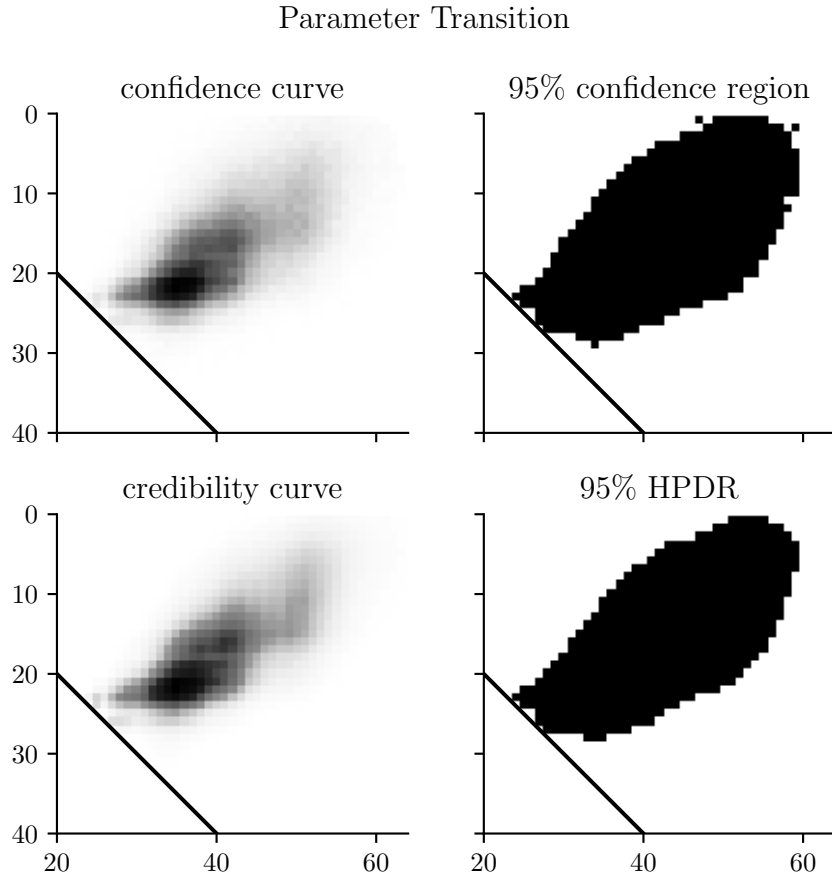


Figure 5.4: Upper left: Confidence curve with $a_0 = 20$, $b_0 = 40$, and $n = 100$. Upper right: 95% confidence region Lower left: Posterior $\pi(a, b \mid y)$ using flat prior on a, b Lower right: 95% highest posterior credibility interval. The black line is the diagonal $a = b$.

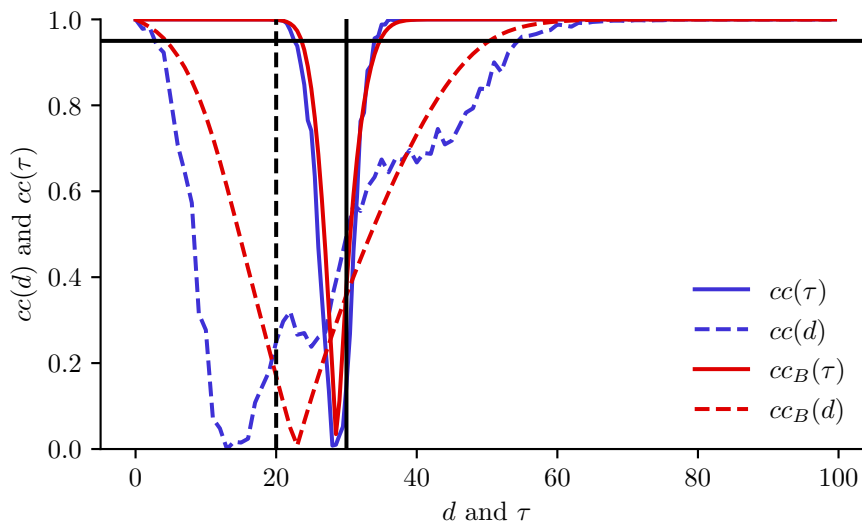


Figure 5.5: Confidence for τ (blue solid), confidence for d (blue dashed), credibility for τ (red solid), and credibility for d (red dashed) for the same data as in figure 5.4. True values are marked in vertical lines.

5.2 Normal data with mixture transition

Continuing with the second transition model, we perform a similar analysis to the parameter transition and note their differences. In the mixture model, the likelihood is the following

$$\ell(a, b, \Delta\theta; y) = \sum_{i=1}^n \log \left[w_i \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} y_i^2 \right\} + (1 - w_i) \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} (y_i - \Delta\theta)^2 \right\} \right],$$

Unfortunately, the inner expression of the logarithm cannot be simplified in a way that allows us to manipulate the sum further. Nevertheless, it is well-defined but more computationally impractical.

The Bayesian analysis is almost identical to the parameter transition and the Bernoulli model, but with a slightly different likelihood.

Again, we simulate 10000 data sets and find the MLE, applying the two corrections in exactly the same manner as in the parameter transition. This time, we see that even as $\Delta\theta$ increases, there is still a strong bias (see figure 5.6), which never seems to go away. For the values of d we have considered, $d - 4$ seems to be the upper limit of $\mathbb{E}_d [\hat{d}]$, which is not surprising considering the analysis of the Bernoulli model, which exhibits the same bias for the MLE. Indeed, when $\Delta\theta \rightarrow \infty$, the normal mixture model is asymptotically equivalent to the Bernoulli model. This is evident when we consider the case where $\Delta\theta$ is very large, which makes it immediately clear from which distribution any given observation is sampled. Thus, the WoC problem reduces to identifying when the window starts and ends, rather than the confounding problem of determining which distribution any given observation is a sample from. At the other extreme, when $\Delta\theta$ is very small, the estimator $\hat{d}$ utterly fails. When $\Delta\theta = 0.5$, $\mathbb{E} [\hat{d}] \approx 2$ regardless of the true d .

Due to the asymptotic similarities to the Bernoulli model, we expect the de-biasing procedure to behave somewhat similarly for large $\Delta\theta$, but it is not as obvious for our middle ground value $\Delta\theta = 2$. Unfortunately, it does not work as well as in the Bernoulli example. It is more reminiscent of the correction results for the parameter transition, in that none of the corrections manage to remove all the bias; however, notice that the scale is different. Particularly, the posterior mean behaves very similarly to the posterior mean in the parameter transition. The bias and RMSE are also greater for the mixture model compared to the parameter model. The bias for $\hat{d}$ is almost twice as large, while RMSE is only slightly larger. However, $\hat{d}'$ is, in fact, less biased in the mixture model but has a larger RMSE. This suggests that it is harder to estimate the window parameters with the mixture transition compared to the parameter transition.

If $\Delta\theta$ is unknown, it is necessary to estimate it. Unfortunately, the profile log-likelihood has no analytic solution for $\widehat{\Delta\theta}$, but as we shall see in Chapter 6, it is possible to use the EM algorithm to efficiently find the maximum through an iterative procedure. Finding the MLE is thus easy, and convergence is typically achieved with fewer than 10 iterations. A full correction procedure can be constructed as in the parameter transition, and for the same asymptotic arguments, we will not perform a simulation of the full procedure.

An example confidence curve and posterior distribution are shown for an example data set in figures 5.9 and 5.10. While we cannot generalize the observations, we can remark that here the posterior credibility curves again give tighter credibility intervals. The Bayesian analysis also favours slightly larger d . The confidence curve and HDPR appear similar to each other (the full confidence curve is cut off as it was far from the true parameters).

Mixture transition

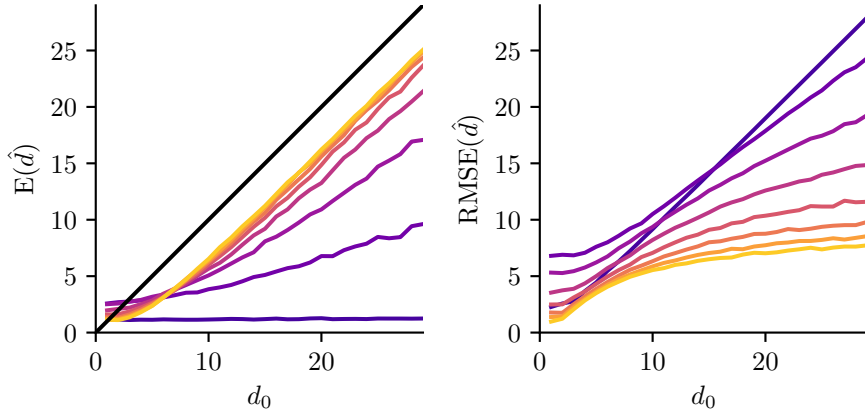


Figure 5.6: Left: Expected value for the ML estimator $\hat{d} = \hat{b} - \hat{a}$ as a function of the true d_0 for a selection of $\Delta\theta$. The black line is the diagonal $\mathbb{E}[\hat{d}] = d$, which represents an unbiased estimator. Right: Expected Root Mean Square Error for the same $\Delta\theta$.

5.3 Discussion

Having studied two different WoC models, from the simplest possible case to the full normal case, we have made many observations. In broad strokes, we can say that WoC problems are more difficult than CP problems in two ways: there is more statistical uncertainty associated with the parameter estimates, and there are more intricate mathematical dependencies. Much of this chapter has been dedicated to one of these phenomena, namely the surprising bias in estimating d . Since the necessity of estimating d is solely a WoC problem, not a CP problem (we might say that it is the defining property of WoC models), it is both new and of extra interest.

The numerous simulations reiterated the empirical observation that $\tau = (a + b)/2$ is consistently and correctly estimated. The algebraic manipulation showed that this is, in fact, due to the fact that τ can be considered the parameter in a true CP model, which we know is unbiasedly estimated if f_L and f_R are symmetric. Studying the bias could therefore be done through the residual random walk representation in equation (3.15).

We have also seen that corrections can be made, but their effectiveness depends on d_0 and is not unilaterally better in the case of normal data. The posterior mean seems to be a good estimator if the RMSE is the primary concern, while $\hat{d}''$ is usually the most unbiased, but it is also the most error-prone.

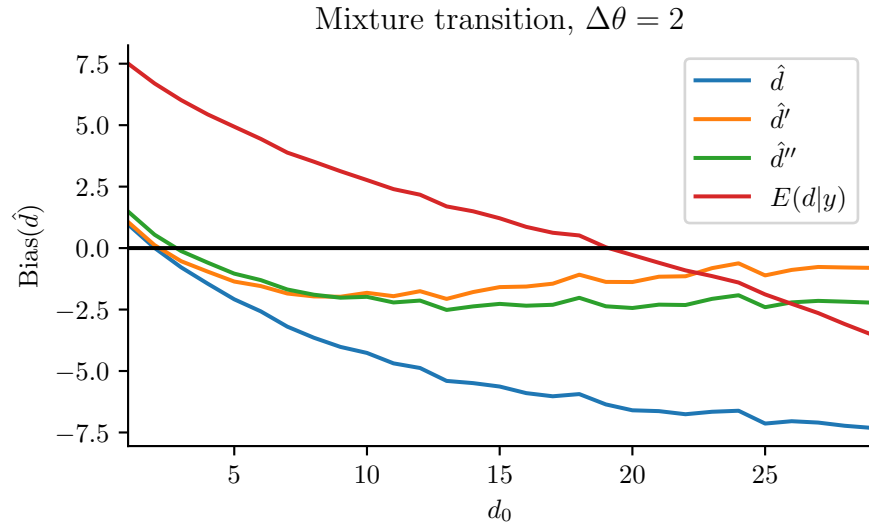


Figure 5.7: Bias for b as a function of d_0 . The de-biasing is slightly more effective, with $\hat{d}''$ getting quite close. The posterior mean is again first quite positively biased, then negatively biased.

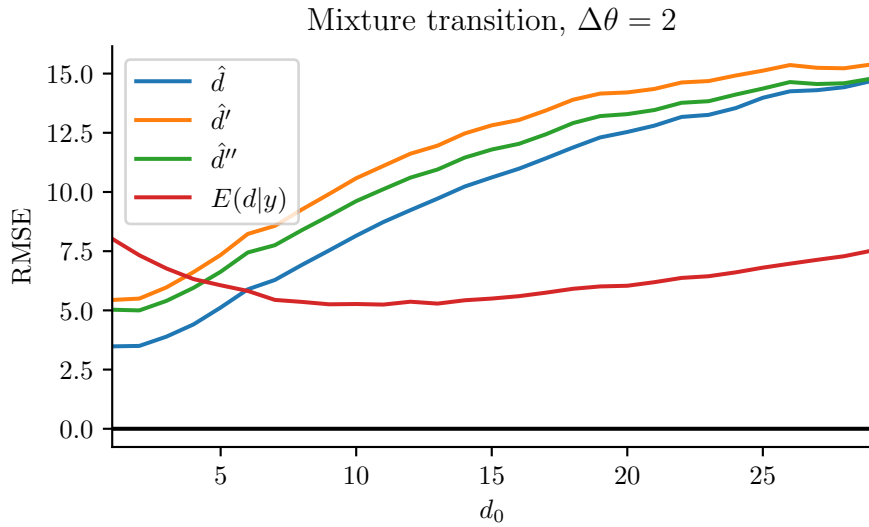


Figure 5.8: RMSE as a function of d . Again, the posterior mean has the lowest RMSE, while the two corrections have larger RMSE than the MLE, but the difference seems to diminish as d_0 increases.

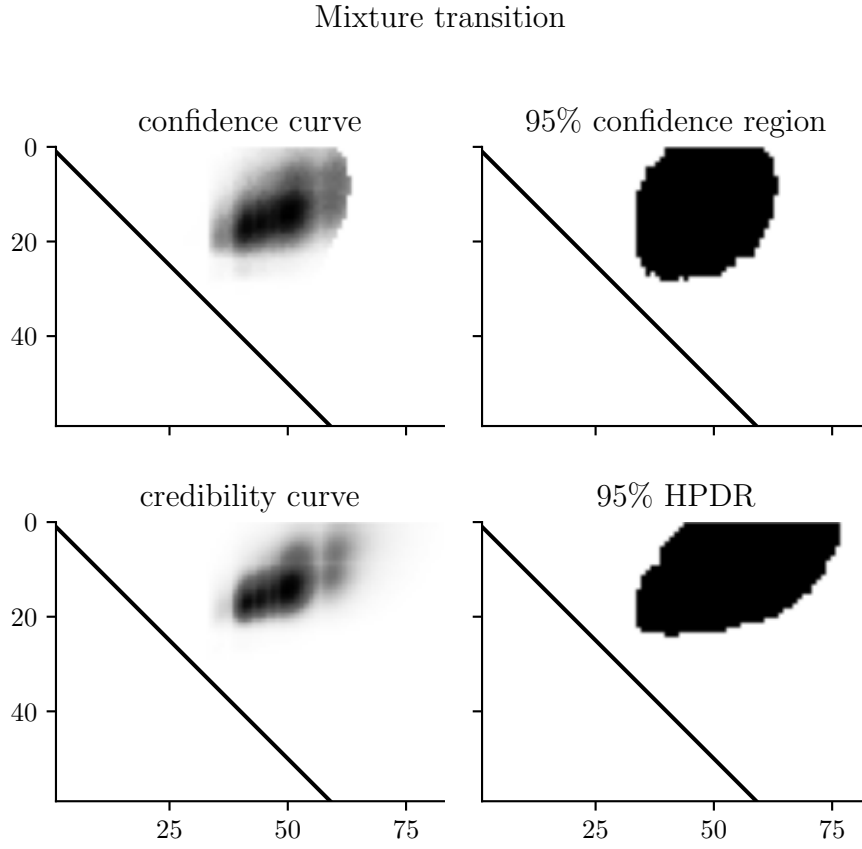


Figure 5.9: Upper left: Confidence curve with $a_0 = 20$, $b_0 = 40$, and $n = 100$. Upper right: The 95% confidence region. Lower left: Posterior $\pi(a, b | y)$ using flat prior on a, b . Lower right: The 95% highest posterior credibility region.

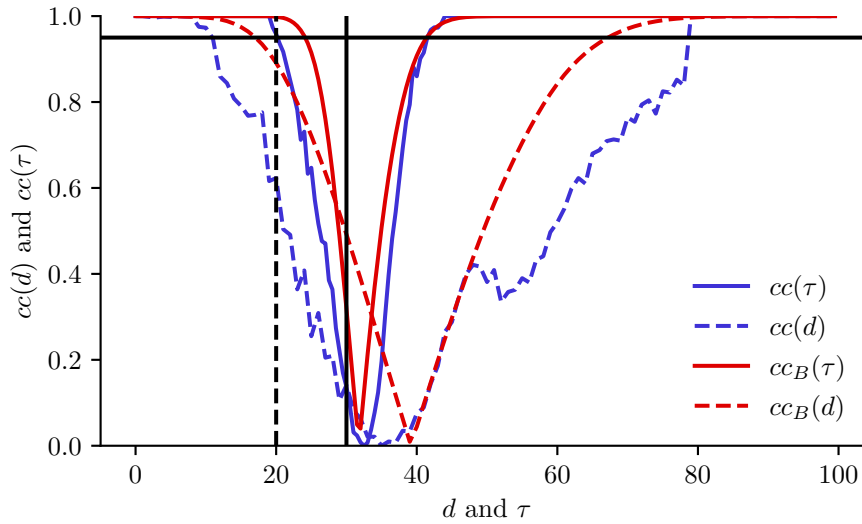


Figure 5.10: Confidence for τ (blue solid), confidence for d (blue dashed), credibility for τ (red solid), and credibility for d (red dashed) for the same data as in figure 5.9. True values are marked in vertical lines.

Chapter 6

Computation

Broadly, this chapter concerns a number of computational methods used to achieve parameter estimates, uncertainty estimates, as well as simulation and sampling algorithms. Many results in this thesis are obtained through numerical work. The complicated nature of the model, in conjunction with the intricate tools of inference, suggests this approach. Additionally, the methods adapted and developed in this chapter allow the WoC model to be easily applied in quite general settings.

As has become common by now, we assume that f_L and f_R are from the same parametric family, say f , such that they only differ in parameter: $f_L(y) = f(y; \theta_L)$ and $f_R(y) = f(y; \theta_R)$. Furthermore, we assume that this family is easy to work with. What this means will become clear. This allows us to study the complexity in the transition window, rather than f itself. As has been shown, even with full knowledge of f_L and f_R , inference about a, b remains difficult. Still, for almost all real-world problems, we don't fully know the distributions f_L and f_R , and must hence estimate them. This chapter develops these tools.

Since the support of a, b is finite, i.e. there exists only a limited number of parameter combinations, we can easily find, say, the MLE or the posterior distribution by simply exhaustively computing all the relevant values. The same is not true for θ_L and θ_R , for which, in general, there are no closed-form solutions for estimators. The problem then remains of obtaining good inference about θ_L and θ_R , which is specifically the goal of this chapter.

6.1 Gradient descent - parameter transition

Let us first consider the parameter transition, where we model the data as follows:

$$Y_i \sim f(\theta_i)$$

where $\theta_i = w_i \theta_L + (1 - w_i) \theta_R$. The log-likelihood is then

$$l = l(a, b, \theta_L, \theta_R; y) = \sum_{i=1}^n \log f(w_i \theta_L + (1 - w_i) \theta_R)$$

We wish to construct an efficient procedure for calculating $\hat{\theta}_L$ and $\hat{\theta}_R$ conditioned on a, b . In Change-Point settings, this problem is as trivial as the i.i.d. case because any one observation y_i is either irrelevant for the estimate of $\hat{\theta}_L$ or $\hat{\theta}_R$. This is clearly not the case here, as all observations within the window, i.e., y_i such that $a < i < b$, contribute

to both estimates. In particular, consider the estimation for θ_L . Using the kernel rule, we may differentiate the log-likelihood:

$$\frac{d}{d\theta_L} l = \sum_{i=1}^n w_i \cdot s(w_i \theta_L + (1 - w_i) \theta_R)$$

where $s(\theta) = \frac{d \log l}{d\theta}$ is the score function. Clearly, some terms dependent on θ_R prevail. Finding the maximum then requires solving a set of non-linear equations, which in general have no simple solutions. For specific functions, however, this is possible, such as for the normal distribution.

We could, of course, employ general optimizing functions, like `nlm` in R or `scipy.optimize.minimize` in Python. General methods do not, however, use the relationship between θ_L and θ_R in the WoC model, and are therefore less efficient than tailored algorithms. Another useful result is if we were to find a closed form estimate for θ_L as a function of θ_R and vice versa. We could then iteratively estimate θ_L and θ_R . It turns out that this is possible in some cases.

If we restrict our view to distributions in the exponential family, things become easier. The exponential family (not to be confused with the exponential *distribution*) is a large and useful family of distributions.

Definition 6.1.1 (Single Parameter Exponential Family). (Based on (Efron, 2022, p. 2)) Let $\theta \in \Theta$ be the *natural parameter*, $S(y) \in R$ be the *sufficient statistic*, $\psi(\theta)$ be the *normalizing function* and $g_0(S(y))$ be *carrying density* or *carrying mass*. Then

$$g_\eta(S(y)) = \exp\{\eta S(y) - \psi(\theta)\} g_0(S(y))$$

is a member of the exponential family. (Note that $g_0(S(y))$ can be replaced with $g_0(S(y))m(dS(y))$, where $m(dS(y))$ is the carrier *measure*. This abstraction is, however, not necessary.)

Definition 6.1.2 (Exponential Family). (Based on (Efron, 2022, p. 2)) Let $S(y) : \mathbb{R}^d \rightarrow \mathbb{R}^k$ be the k -dimensional sufficient statistic of a d -dimensional observation. If y has a distribution of the following form

$$f(y; \theta) = g(S(y)) \exp\{\eta(\theta)^T S(y) - c(\theta)\},$$

Then f is a member of the exponential family, and $\eta(\theta)$ is the *natural parameter*. If $\eta(\theta) = \theta$, then θ is the natural parameter.

Distributions belonging to the exponential family have two properties of particular interest. First is the existence of the sufficient statistic. This is a function $S(\cdot)$ of the observed data vector y which contains *all the relevant information* of the observations. This means that two samples with identical sufficient statistics will have identical properties with respect to the exponential family. (This does not, however, imply that they should be considered identical by the statistician. Erroneous modelling could impose an identical sufficient statistic on wildly dissimilar data, c.f. Anscombe (1973) quartet. It is a set of four data sets with identical mean, variance, and correlation, but are clearly generated by wildly different mechanisms.) The second convenient property relates to maximum likelihood estimation and is accordingly quite relevant. Simply stated, the property is that for i.i.d. data, the MLE for the natural parameter θ is the sufficient statistic itself (Efron, 2022, p. 11):

$$\hat{\theta} = \eta^{-1}(S(y)).$$

Lemma 6.1.3. *If $f_L = f(y; \theta_L)$ and $f_R = f(y; \theta_R)$ are members of the exponential family, then so is the (natural) parameter transition WoC model for a fixed a, b . Furthermore, the sufficient statistic is $S(y_i) = (\sum_{i=1}^n w_i S(y_i), \sum_{i=1}^n (1 - w_i) S(y_i))$.*

Proof. Let $\theta_i = w_i \theta_L + (1 - w_i) \theta_R$. Since f_L and f_R are members of the exponential family, we may write the log-likelihood

$$\begin{aligned} \ell(\theta_L, \theta_R; y, a, b) &= \sum_{i=1}^n \sum_{j=1}^k S(y_i)_j \theta_{i,j} - \sum_{i=1}^n c(\theta_i) + \text{const.} \\ &= \sum_{i=1}^n \sum_{j=1}^k S_j(y_i) (w_i \theta_{L,j} + (1 - w_i) \theta_{R,j}) - \sum_{i=1}^n c(\theta_i). \\ &= \sum_{i=1}^n \sum_{j=1}^k S_j(y_i) w_i \theta_{L,j} + \sum_{i=1}^n \sum_{j=1}^k (1 - w_i) S_j(y_i) \theta_{R,j} - \sum_{i=1}^n c(\theta_i). \\ &= \sum_{j=1}^k \theta_{L,j} \sum_{i=1}^n w_i S_j(y_i) + \sum_{j=1}^k \theta_{R,j} \sum_{i=1}^n (1 - w_i) S_j(y_i) - \sum_{i=1}^n c(\theta_i). \end{aligned}$$

Which shows that $f(\theta_L, \theta_R; y, a, b)$ is in the exponential family with the natural parameters being $\theta' = (\theta_L, \theta_R)$ and sufficient statistic $S(y) = (w^T S(y), (1 - w)^T S(y))$. The new normalizing function c is $c'(\theta_L, \theta_R) = \sum_{i=1}^n c(\theta_i)$. Here, we understand the notation (u, v) , where u and v are vectors as the joint vector beginning with the components of u and followed by the components of v . $\square$

The proof relies on rearranging sums, which is only possible when the log-likelihood is linear in θ . Another useful property of maximum likelihood estimation in exponential families is that

$$\frac{d}{d\theta_j} c(\theta) = \mathbb{E}[S_j(Y)],$$

followed by the observation that finding the MLE amounts to solving the equation:

$$\frac{d}{d\theta_j} \ell(\theta; y) = S_j(y) - \frac{d}{d\theta_j} c(\theta)$$

Returning to WoC models,

$$\begin{aligned} \frac{d}{d\theta_L} c'(\theta_L, \theta_R) &= \sum_{i=1}^n w_i \frac{d}{d\theta_i} c(\theta_i) = \sum_{i=1}^n w_i \mathbb{E}[Y_i] \\ \frac{d}{d\theta_R} c'(\theta_L, \theta_R) &= \sum_{i=1}^n (1 - w_i) \frac{d}{d\theta_i} c(\theta_i) = \sum_{i=1}^n (1 - w_i) \mathbb{E}[Y_i]. \end{aligned}$$

Thus, finding the MLE requires solving this set of equations:

$$w^T S(y) = \sum_{i=1}^n w_i \mathbb{E}[Y_i] \tag{6.1}$$

$$(1 - w)^T S(y) = \sum_{i=1}^n (1 - w_i) \mathbb{E}[Y_i] \tag{6.2}$$

Example 6.1.4. To illustrate the usefulness of this, consider the case with the normal case where $Y_L \sim N(\theta_L, 1)$ and $Y_R \sim N(\theta_R, 1)$. For each Y_i , the expected value is θ_i directly. (6.1) then becomes

$$w^T y = w^T (w\hat{\theta}_L + (1 - w)\theta_R)$$

with the closed form solution

$$\hat{\theta}_L = \frac{w^T y - w^T (1 - w)\theta_R}{w^T w},$$

and conversely,

$$\hat{\theta}_R = \frac{(1 - w)^T y - (1 - w)^T w\theta_L}{(1 - w)^T (1 - w)}.$$

The global maximum can then be found by alternating these two estimates.

In addition, the solutions form a linear set of equations, and the MLE can thus be obtained analytically:

$$\begin{aligned}\hat{\theta}_L &= A + B\hat{\theta}_R \\ \hat{\theta}_R &= C + D\hat{\theta}_L\end{aligned}$$

where $A = w^T y / w^T w$, $B = w^T (1 - w) / w^T w$, $C = (1 - w)^T y / (1 - w)^T (1 - w)$, $D = (1 - w)^T w / (1 - w)^T (1 - w)$. The system of equations is then expressed as a matrix equation:

$$\begin{pmatrix} 1 & -B \\ -D & 1 \end{pmatrix} \begin{pmatrix} \hat{\theta}_L \\ \hat{\theta}_R \end{pmatrix} = \begin{pmatrix} A \\ C \end{pmatrix}$$

With solution

$$\begin{pmatrix} \hat{\theta}_L \\ \hat{\theta}_R \end{pmatrix} = \begin{pmatrix} 1 & -B \\ -D & 1 \end{pmatrix}^{-1} \begin{pmatrix} A \\ C \end{pmatrix}$$

▲

6.2 Expectation maximization in mixture WoC models

The mixture transition, as the name suggests, induces a mixture distribution, and we can use the EM algorithm to efficiently estimate θ_L and θ_R . It requires that we fix a, b , but this is just the case when we want to find the profiled log-likelihood for a, b .

The observed likelihood for a mixture model is

$$L(a, b, \theta_L, \theta_R; y) = \prod_{i=1}^n [w_i f(y_i; \theta_L) + (1 - w_i) f(y_i; \theta_R)]. \quad (6.3)$$

But if we introduce the latent variable

$$Z_i = \begin{cases} 0, & \text{if } Y_i \sim f_L \\ 1, & \text{if } Y_i \sim f_R \end{cases},$$

the full likelihood then becomes

$$L(a, b, \theta_L, \theta_R; y, z) = \prod_{i=1}^n w_i^{I(z_i=0)} f(y_i; \theta_L)^{I(z_i=0)} (1 - w_i)^{I(z_i=1)} f(y_i; \theta_r)^{I(z_i=1)}.$$

Note that given a, b , the prior probability that $z_i = 0$ is w_i .

To use the EM algorithm, we need to deduce the updating equations. We begin with the E-step, which estimates the class probabilities. Let $\theta^{(t)}$ be the current estimate for θ_L, θ_R . We use Bayes' rule to calculate the posterior class probabilities:

$$\Pr(Z_i = 0 \mid y_i, \theta^{(t)}) = \frac{\Pr(Y_i \mid Z_i = 0)\Pr(Z_i = 0)}{\sum_{j=0}^1 \Pr(Y_i \mid Z_i = j)\Pr(Z_i = j)} \quad (6.4)$$

$$= \frac{f(y_i; \theta_L)w_i}{f(y_i; \theta_L)w_i + f(y_i; \theta_R)(1 - w_i)}. \quad (6.5)$$

We call this γ_i .

The M-step then maximizes θ_L, θ_R conditioned on knowing z , but now we can observe the decoupling effect of introducing this latent variable. Consider the Q function:

$$Q(\theta \mid \theta^{(t)}) = \mathbb{E} \left[\ell(\theta; Y, Z) \mid y, \theta^{(t)} \right].$$

and we wish to optimize this with respect to θ_L .

$$\frac{d}{d\theta_L} Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^n \Pr(Z_i = 0 \mid y_i, \theta^{(t)}) s(\theta_L; y_i)$$

where s is the score function. Here, we are using the fact that the expectation of an indicator value is the probability of it being 1. Setting it equal to zero yields the score equation, and it is totally free of θ_R . It can therefore be solved in the standard way, but weighted by the leading factors $\Pr(Z_i = 0 \mid y_i, \theta^{(t)})$. Notice that this is simply the standard M-step in a generic mixture model estimation. The only difference for the EM algorithm in a WoC model is thus in the calculation of γ_i , where w_i functions as the prior belief that $Z_i = 0$.

Two comments are in order. First, the procedure must be initiated with some values $\theta_L^{(0)}$ and $\theta_R^{(0)}$, and a good guess ensures faster convergence. A bad guess could possibly lead the procedure to find a local maximum, and while there are conditions that guarantee that the algorithm converges to the global maximum, we will not discuss this. Instead, we propose some good starting guesses. Consider $y_i, \dots, y_a \stackrel{i.i.d}{\sim} f(\theta_L)$, and assuming that f has a well-behaving MLE $h(\cdot)$, let $\theta_L^{(0)} = h(y_i, \dots, y_a)$, and likewise for θ_R . This estimator is asymptotically equivalent to the full $\hat{\theta}_L$, letting $a \rightarrow \infty$ and $n - b \rightarrow \infty$. In practice, this start has ensured fast and stable convergence when a, b are well located in the middle of $1, \dots, n$, usually converging in around 10 steps, but it is less efficient when a, b are located at the extremities.

The second comment addresses the stopping criterion. Every optimization must have a stopping criterion that reflects what is 'good enough'. We must therefore decide what is 'good enough'. The three most common choices are: after some fixed number of iterations, when the parameters change sufficiently little, or when the objective function increases sufficiently little. The first option is convenient to implement and also provides a very predictable runtime, but one must make a trade-off when deciding the number of iterations between accuracy and speed. The second criterion requires a metric on the parameters, say the Euclidean norm, but it has the unfortunate property of being dependent on the parametrization of θ_L and θ_R . The final criterion is clearly what we should use for this task. This is because we primarily care if $\ell_{\text{prof}}(a, b; y)$ has converged, which amounts to the objective function in the EM optimization having converged. To use this final criterion effectively, we need a notion of the scale of the objective function, which in turn determines what we consider 'sufficiently close'. This is no challenge, as some preliminary runs clearly show approximately the scale of the log-likelihood.

Example 6.2.1 (EM in normal model). Let $f(y; \theta_L) = N(\theta_L, 1)$ and $f(y; \theta_R) = N(\theta_R, 1)$. The updating equation for z_i is the same as equation (6.4). The maximizer for θ_L is

$$\hat{\theta}_L = \frac{\sum_{i=1}^n \gamma_i y_i}{\sum_{i=1}^n \gamma_i},$$

where γ_i is the current estimated probability that $z_i = 0$. and conversely for θ_R . ▲

6.3 MCMC for WoC models

The two MCMC algorithms for WoC model we will discuss are Gibbs sampling and the Metropolis-Hastings algorithm. It is common to introduce Metropolis-Hastings first as it is simpler, but it currently makes more sense to start with Gibbs sampling, which we now do.

Carlin, Gelfand and Smith (1992) showed that this is always possible in Change-Point models, thus making it a very effective sampling algorithm for those problems. We will now show that it is also possible in some Window of Change models. The crucial insight, which is really no different from the aforementioned CP case, is that a, b is a bounded (2-dimensional) discrete parameter, and for any bounded discrete parameter with a known PMF up to a constant, it is possible to normalize it. Thus, we can always find its conditional distribution.

$$\begin{aligned} \pi(a, b \mid \theta_L, \theta_R, y) &= \frac{\pi(y, a, b, \theta_L, \theta_R)}{\pi(y, \theta_L, \theta_R)} \\ &= \frac{\pi(y \mid a, b, \theta_L, \theta_R) \pi(a, b)}{\sum_{i=1}^{n-1} \sum_{j=i+1}^n \pi(y \mid \theta_L, \theta_R, i, j) \pi(i, j)} \end{aligned}$$

Note that the prior on θ_L and θ_R cancel, which is a consequence of treating them as known. In CP models, knowing τ decouples the inference on θ_L and θ_R from each other. More precisely, $\pi(\theta_L \mid y, \tau, \theta_R) = \pi(\theta_L \mid y, \tau)$, and we may treat $y_L = (y_1, \dots, y_\tau)$ as i.i.d. observations from f_L . This is not necessarily possible in WoC models, as the observations within the window depend upon both parameters. There are, however, some cases in which decoupling is possible.

The first case is where the algebra simply ‘works out’.

Example 6.3.1 (Conditional distribution for normal data with parameter transition). Let

$$y_i \sim N(w_i \theta_L + (1 - w_i) \theta_R, \sigma^2)$$

with σ known. If we consider a, b to be known, we may write $X_i = (w_i, 1 - w_i)^T$ and $\beta = (\theta_L, \theta_R)^T$. Then,

$$y_i = \beta^T X_i + \epsilon_i,$$

which is a normal linear regression model. Since we are doing a Bayesian analysis, we need a prior on β , and if we choose a binormal, then the posterior is also binormal

$$\pi(\beta) = N(\mu_0, \Sigma_0).$$

The posterior is

$$\pi(\beta \mid y, a, b) = N(\mu_1, \Sigma_1)$$

For some updated parameters μ_1, Σ_1 . ▲

The other case is for mixture transitions, which can be decoupled regardless of likelihood and priors. The trick is introducing latent class variables, which we did for the EM algorithm. As in the frequentist case, if we presume to know which observations were sampled from f_L and from f_R , we can treat the observations separately as i.i.d. observations from their corresponding distributions and treat them as typical. We can also sample a, b , since if we know the class, the likelihood becomes irrelevant. Thus, introducing class variables decouples inference on θ_L, θ_R and a, b . Of course, we do not know the class labels, but if we consider all other parameters known, we can sample them easily.

Let y be a sample from a WoC model with a mixture transition, and let $z_i \in \{0, 1\}$ denote from which distribution y_i is sampled, where $z_i = 0$ means that $y_i \sim f_L$. Then the likelihood is

$$\pi(y \mid z, a, b, \theta_L, \theta_R) = \prod_{i=1}^n f(y_i; \theta_L)^{1-z_i} f(y_i; \theta_R)^{z_i},$$

which means that the posterior distribution for, say θ_L , is

$$\begin{aligned} \pi(\theta_L \mid y, z, a, b, \theta_R) &\propto \prod_{i=1}^n f(y_i; \theta_L)^{1-z_i} \pi(\theta_L) \\ &\propto \pi(\theta_L \mid y, z), \end{aligned}$$

assuming $\pi(\theta_L, \theta_R) = \pi(\theta_L)\pi(\theta_R)$. Note that this posterior is independent of a, b and θ_R , thus decoupling it properly. We continue with deducing the posterior for a, b and z . Clearly,

$$\begin{aligned} \pi(a, b \mid y, z, \theta_L, \theta_R) &\propto \prod_{i=1}^n w_i^{1-z_i} (1 - w_i)^{z_i} \pi(a, b) \\ &\propto \pi(a, b \mid z). \end{aligned}$$

And finally,

$$\pi(z_i \mid y_i, a, b, \theta_L, \theta_R) \propto f(y_i; \theta_R)(1 - w_i)$$

To fully specify the Gibbs sampler, we must also precisely define how the conditional distributions should be sampled. We will use a deterministic scan, that is, for every iteration, sample the parameters in order. The order matters little.

Algorithm 1 Gibbs sampling for WoC mixture models

Input: Target distribution $\pi(\theta \mid y)$ (known up to a constant), proposal distribution $q(\theta' \mid \theta)$, initial state θ_0 , number of iterations N

Initialize: Set $\theta^{(0)} = \theta_0$

for $t = 1$ to N **do**

$\theta_L^{(t)} \sim \pi(\theta_L \mid y, z^{(t-1)})$

$\theta_R^{(t)} \sim \pi(\theta_R \mid y, z^{(t-1)})$

$z^{(t)} \sim \pi(z \mid y, a^{(t-1)}, b^{(t-1)}, \theta_L^{(t)}, \theta_R^{(t)})$

$a^{(t)}, b^{(t)} \sim \pi(a, b \mid z^{(t)})$

end for

Output: Samples $\{\theta^{(t)}\}_{t=1}^N$

	a	b	θ_L	θ_R
mean	37.85	64.44	0.17	2.07
stdev	7.18	9.31	0.16	0.17
M_{eff}	421.87	205.06	1314.12	814.13
Gelman-Rubin	1.00	1.00	1.02	1.02

Table 6.1: Chain diagnostics for the Gibbs sampler using five chains and 1100 samples where the first 100 samples were discarded. Except the Gelman-Rubin statistic where all values calculated using the combined chains sans burn in.

	a	b	θ_L	θ_R
mean	36.25	64.83	0.12	2.05
stdev	9.01	9.88	0.45	0.19
M_{eff}	73.49	50.28	1692.79	693.49
Gelman-Rubin	1.00	1.00	1.01	1.23

Table 6.2: Chain diagnostics for the Metropolis-Hastings sampler using five chains and 10100 samples where the first 100 samples were discarded. That is, 10 times the number of samples compared to the Gibbs sampler.

Example 6.3.2 (Gibbs sampling in normal WoC mixture model). Let $f_L = N(\theta_L, \sigma^2)$ and $f_R = N(\theta_R, \sigma^2)$, and $\pi(\theta_L) = \pi(\theta_R) = N(\mu, \tau^2)$ with σ^2 , μ , and τ^2 known. Given distribution classes $z^{(t)}$, let $y_L = \{y_i : z_i^{(t)} = 0\}$ be the observations from the left distribution with length n_L . From elementary Bayesian analysis, we know that the posterior is itself normal, with the updated parameters

$$\tau_L^{2,(t)} = \left(\frac{1}{\tau^2} + \frac{n_L}{\sigma^2} \right)^{-1}$$

$$\mu_L^{(t)} = \tau_L^{2,(t)} \left(\frac{\mu}{\tau^2} + \frac{n_L \bar{y}_L}{\sigma^2} \right).$$

Then, the conditional distributions are as follows:

$$\pi(\theta_L | y, z^{(t)}) = N(\mu_L^{(t)}, \tau_L^{2,(t)})$$

$$\pi(\theta_R | y, z^{(t)}) = N(\mu_R^{(t)}, \tau_R^{2,(t)})$$

$$\pi(a, b | z^{(t)}) = \frac{\prod_{i=1}^n w_i^{1-z_i^{(t)}} (1-w_i)^{z_i^{(t)}} \pi(a, b)}{\sum_{j=1}^{n-1} \sum_{k=j+1}^n \prod_{i=1}^n w(i, j, k)^{1-z_i^{(t)}} (1-w(i, j, k))^{z_i^{(t)}} \pi(j, k)}$$

$$\pi(z_i | y_i, \theta_R^{(t)}, \theta_L^{(t)}) = \frac{\phi(y_i; \theta_L^{(t)})}{\phi(y_i; \theta_L^{(t)}) + \pi(y_i; \theta_R^{(t)})}.$$

Traces for a, b are shown in figure 6.1, where we see both good convergence and mixing. Note the similarities between this algorithm and the EM algorithm. $\blacktriangle$

The crucial advantage of this algorithm is that sampling a, b and z is entirely independent of the dimensionality of θ_L and θ_R . The algorithm is therefore efficient for very complex models f_L and f_R . The added complexity of introducing the window is therefore fixed; that is, sampling a, b conditioned on knowing all other parameters is really no different from calculating the likelihood of y, z . An additional advantage is

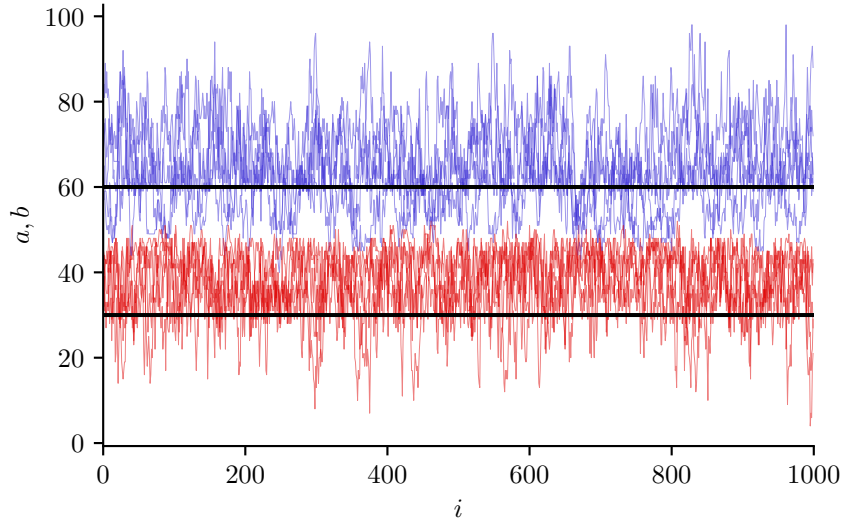


Figure 6.1: Traces of a (red) and b (blue) from 5 chains constructed as in example 6.3.2. The horizontal black lines are the true parameter values. The traces show good mixing, with a total number of samples (across all five runs), of 316 and 189 respectively, however, considering the total number of samples where $5 \cdot 1000$, this mean that we only retrieve 6.3% and 3.8% of the samples. For θ_L, θ_R where the effectiveness 30% and 24% respectively, again illustrating the relative difficulty of estimating a, b . The Gelman-Rubin statistics are almost exactly one, showing that the stationary distribution has been reached. The sampling took 12 seconds on an M1 MacBook Air.

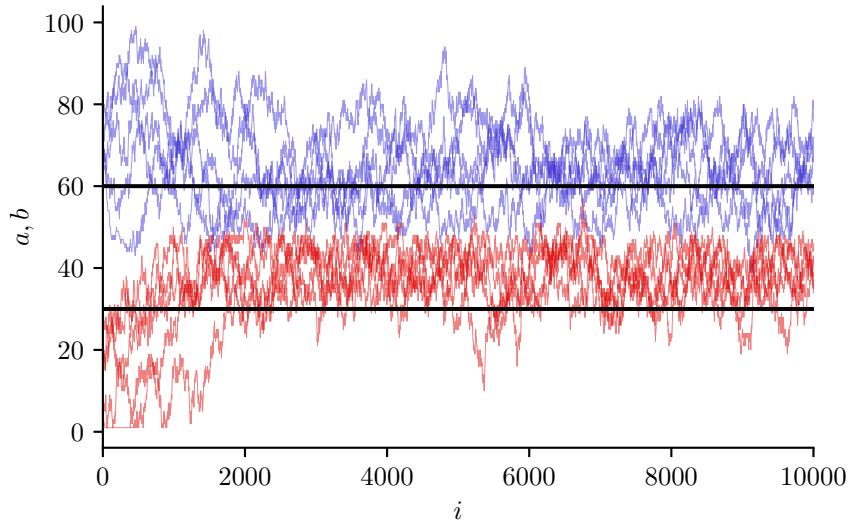


Figure 6.2: Traces for the Metropolis-Hastings algorithm using the same data parameters as above, but 10 times the number of samples. The window parameters a, b were proposed using a symmetric shift, and the distribution parameter θ_L and θ_R were proposed with a symmetric normal shift with standard deviation of 1. The acceptance rate was about 0.7. The traces show stronger autocorrelation and thus worse mixing. Still, the Gelman-Rubin statistic suggest that the stationary distribution has been reached. The sampling took 1.1 seconds on an M1 MacBook Air.

that if finding the posterior for f_L and f_R is a solved problem, it can be used directly, as we did in the example above.

What if finding the conditional distribution is not possible, or perhaps not easy? We continue with the second MCMC algorithm, Metropolis-Hastings, which has no such limitations. The Metropolis-Hastings algorithm requires a proposal distribution $q(\theta'; \theta^{(t)})$, and the proposed θ' is accepted with probability

$$MHR(\theta', \theta^{(t)}) \min \left(1, \frac{\pi(\theta' | y) q(\theta^{(t)} | \theta')}{\pi(\theta^{(t-1)} | y) q(\theta' | \theta^{(t-1)})} \right).$$

We must therefore specify this for WoC models, and we begin by considering the proposal for a, b . The simplest form of proposal is a symmetric shift

$$a' \sim a^{(t)} + U(\{-1, 0, 1\}) \quad (6.6)$$

$$b' \sim b^{(t)} + U(\{-1, 0, 1\}). \quad (6.7)$$

This proposal could potentially propose illegal parameter combinations, like $b < a$, in which cases we must reject the proposal. θ'_L and θ'_R can be sampled as one would otherwise prefer in an i.i.d. setting, for instance with multivariate normal shifts

$$\theta'_L = \theta^{(t)} + \epsilon, \quad \epsilon \sim N(0, \sigma_{\text{prop}} I_k),$$

where I_k is the k -dimensional identity matrix, and k is the dimensionality of θ_L . Often, proposing θ'_L and θ'_R and calculating the (unscaled) posterior poses no greater challenge than in the i.i.d. case. While it is easy to calculate the acceptance probability, the probability of accepting any one proposal decreases as k increases (for a fixed σ_{prop}). Rejecting the proposal too often increases the autocorrelation of the chain, thus reducing the quality of the samples. A solution is to decrease σ_{prop} , but this, too, will eventually increase the autocorrelation.

The advantage of the Metropolis-Hastings algorithm is that both proposing θ' and calculating $MHR(\theta', \theta^{(t)})$ are usually quite fast. Especially in parameter transition models is this easy, as, letting $\theta_i = w_i \theta_L + (1 - w_i) \theta_R$, where the acceptance probability is

$$MHR(\theta', \theta^{(t)}) = \min \left(1, \frac{\prod_{i=1}^n L(y_i; \theta'_i) \pi(a', b') \pi(\theta'_L) \pi(\theta'_R)}{\prod_{i=1}^n L(y_i; \theta_i^{(t)}) \pi(a^{(t)}, b^{(t)}) \pi(\theta_L^{(t)}) \pi(\theta_R^{(t)})} \right).$$

The Metropolis-Hastings algorithm is typically faster if k is small in WoC problems. However, in high-dimensional problems, the Gibbs sampler is preferred, as the autocorrelation of the Metropolis-Hastings chain can be quite large.

Figure 6.1 and 6.2 show traces from a Gibbs sampler and a Metropolis-Hastings sampler sampling the posterior of the same data and model. The Gibbs sampler uses fewer samples, only 1000, and produces higher quality chains. The Metropolis-Hastings algorithm, in turn, uses 10000 samples, and shows more autocorrelation and worse mixing. Table 6.1 and 6.2 shows the chain diagnostics for the two runs. Regarding a, b , we see that the Gibbs sampler had an effective sample size about five times as large. Since it used a tenth of the total number of samples, we can argue that it is 50 times as efficient. However, the runtime of the Gibbs sampler was about 12 seconds, while the Metropolis-Hastings sampler took around 1.1 seconds, making the Metropolis-Hastings algorithm the most time-efficient by a factor of about two. We can also observe that the ratio of effective number of samples is different for the distribution parameter. Consider

θ_L , where both algorithms produced about the same number of samples, or θ_R where the Gibbs sampler only produced the double of the Metropolis-Hastings algorithm. This means that if the dimensionality is small, i.e. $\dim(\theta_L)$ is small, then the Metropolis-Hastings algorithm is preferable.

Finally, the hybrid Gibbs sampler can also be used, and it is employed throughout the analyses. Now, we sample a, b as in the Gibbs sampler, but sample θ_L and θ_R with a Metropolis-Hastings step. This can always be done since $\pi(a, b \mid y, \theta_L, \theta_R)$ is discrete and bounded. In fact, it is the sampler used for the Bayesian analysis of the British mining disasters. The advantage of the hybrid approach is that a, b has lower autocorrelation and more efficiently explores the parameter space. Using the symmetric shift from equation 6.6 limits the amount a, b can move from one step to another, while the conditional distribution for a, b has no such restrictions. We see this phenomenon when we consider the ratio of the effective number of samples in 6.1 and 6.2, which suggests that the Metropolis-Hastings algorithm disproportionately underperforms when sampling a, b and is otherwise comparable to the Gibbs sampler.

Chapter 7

Hypothesis testing

Having defined, analysed and applied this new proposed model, some might interject: ‘when should we use it?’ Such questions naturally arise and need answers. In this chapter, we develop tools to test when a WoC model should be selected over a CP model. We will concern ourselves less with testing which distribution f fits the data best.

It is natural to test the hypothesis $H_0 : d = 1$ against the alternate $H_a : d > 1$ for two reasons: the conceptual roots of WoC analysis are in CP analysis, which is equivalent to $d = 1$. Therefore, we wish to know if we should consider this extension for a given CP problem. The second reason is that CP models are more restrictive, and, in general, more flexible models require more justification. Although problems of ‘over-fitting’ seem somewhat far when jumping from one parameter to two, the WoC is more complicated. Without good reason, the model choice should be deferred to the simpler model. First, we consider frequentist analysis.

7.1 Defining a test statistic

To perform hypothesis testing, we need a test statistic, T , which can be any function that maps the observed data to a scalar. Two natural choices for the test statistic are the MLE $\hat{d}$ itself,

$$T_{ML}(y) = \arg \max_d \max_a \ell_{\text{prof}}(a, a + d; y) \quad (7.1)$$

and twice the log-likelihood ratio (also known as deviance), which is

$$T_{LR}(y) = 2(\max_{a,b} \ell_{\text{prof}}(a, b; y) - \max_a \ell_{\text{prof}}(a, a + 1; y)). \quad (7.2)$$

In equation 7.2, the first term is the maximum likelihood of the unconstrained WoC model, and the second term is the maximum likelihood of the CP model. We may use the same likelihood function, as forcing $b = a + 1$ collapses the WoC model to a CP model.

The choice of T_{ML} is intuitive, as under H_0 , we expect $\hat{d}$ to be small, while for large d_0 , we expect $\hat{d}$ to also be large. The second statistic is interesting as the famous Neyman-Pearson lemma shows that under certain conditions, the likelihood ratio test is universally the most powerful if H_a is *simple*, that is, a single hypothesis. Since we are testing for a family of distributions, parametrized by $d > 1$, the guarantee does not hold. Furthermore, it is equivalent to the deviance, and in smooth parametric models, it will have a χ^2 distribution, and, crucially, the distribution will be independent of

nuisance parameters. For this reason alone, it is quite useful, but we are not in the smooth parametric case, so such guarantees do not hold.

7.2 The null distribution

To perform the tests, we need to calculate their p -values of the null distribution, which we cannot do analytically, and must therefore simulate. Simulating this is comparatively easy since the null hypothesis is singular, i.e. we only need to generate data from a CP model. Thus, the search space is constrained, and sampling is comparatively easier than constructing confidence curves. We cannot, however, simulate for all possible distributions f , and we will therefore limit ourselves to normal distributions with different means.

Let $f_L = N(0, 1)$ and $f_R = N(\Delta\theta, 1)$, and let $N = 100$ and $\tau = 40$. For $\Delta\theta \in \{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4\}$, we estimate the empirical cumulative distribution F_T , and given that, we calculate the probability of accepting or rejecting the null hypothesis at any given α . We then simulate $n_{sim} = 10000$ data sets, and compute the test statistics T_{ML} and T_{LR} , as shown in figures 7.1 and 7.2.

Figure 7.1 shows that T_{LR} has a clear dependence on $\Delta\theta$, which shows that T_{LR} is not pivotal. This is unfortunate, as with real data, $\Delta\theta$ must be estimated, and this estimation would introduce uncertainty and possible biases. For all $\Delta\theta$, we see that there is a positive (and frankly quite large) probability that $T_{LR} = 0$, that is, a Change-Point model is the maximum likely model. Interestingly, an increase in $\Delta\theta$ also increases the 95% quantile, which in turn affects our critical value when testing with $\alpha = 0.05$. At the same time, an increase in $\Delta\theta$ increases the probability that $T_{LR} = 0$. This is not surprising; with a larger level difference, it is easier to identify that the change is immediate, rather than gradually increasing.

Comparing the log-likelihood ratio statistic to the maximum likelihood statistic, we see something different. There is still a dependency on $\Delta\theta$, but the dependency is not so easy to identify. When $\Delta\theta = 0.5$, $P(\hat{d} = 1) = 0.86$, $\hat{d} = 2$ are already larger than the critical value. Initially, it might seem strange that $\hat{d} > 1$ is so unlikely, but if we consider the bias from Chapter 5, we should not be surprised. If $\hat{d}$ fails as an estimator, then we should not expect it to work as a test statistic.

With this in mind, it is clear that for a given data set and model, the critical value must be determined through simulation, which depends on the nuisance parameters θ_L and θ_R , which must be estimated from the data. This dependency on estimated values introduces more uncertainty and potential biases in the estimation of the test statistic's distribution. We also need to decide how we are to estimate θ_L and θ_R . The suggestion is to find the best CP model, and use the MLE estimates from this. The reason for this is that if one actually believes that a CP model is correct, then that model should be the foundation for estimating the nuisance parameters. If, on the other hand, we were to estimate the nuisance parameters from a WoC model, our reasoning would be unsound. (There are still cases where not strictly adhering to the null hypothesis is optimal.)

Example 7.2.1 (Test statistic for British Mining Disasters). In the introduction, we tested for $H_0 : d = 1$ in the analysis of British Mining disasters. The procedure is as follows, still using a Poisson data model: Under H_0 , we estimate the change-point $\hat{\tau} = 41$ in addition to the parameters $\hat{\theta}_L = \bar{y}_{1:\tau} = 3.10$ and $\hat{\theta}_R = \bar{y}_{\tau+1:n} = 0.90$. Using these values, simulate $n_{sim} = 1000$ data sets y^* and calculate their test statistic

$$T_{LR}^*(y^*) = 2(\max_{a,b} l_{\text{prof}}(a, b; y^*) - \max_a l_{\text{prof}}(a, a + 1; y^*))$$

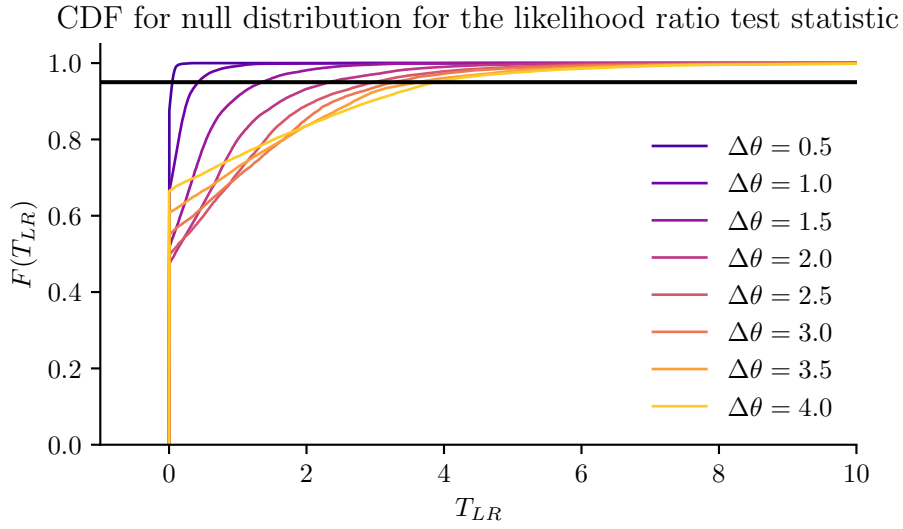


Figure 7.1: CDF of null distribution for T_{LR} for a normal parameter transition WoC model given different level differences. The curve was constructed from 10000 samples, and the 0.95 level is marked by the black line.

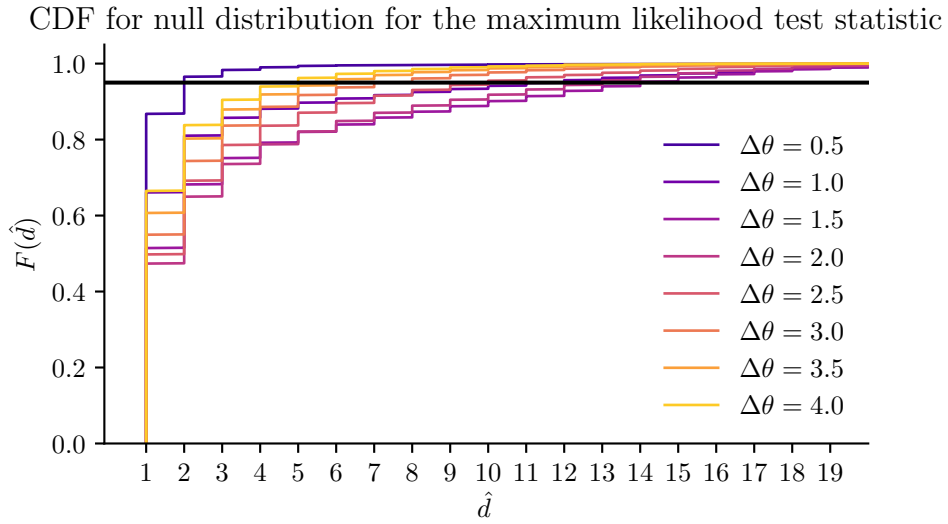


Figure 7.2: CDF for the MLE test statistic for a normal parameter transition WoC model given different level differences. The curve was constructed from 10000 samples, and the 0.95 level is marked by the black line.

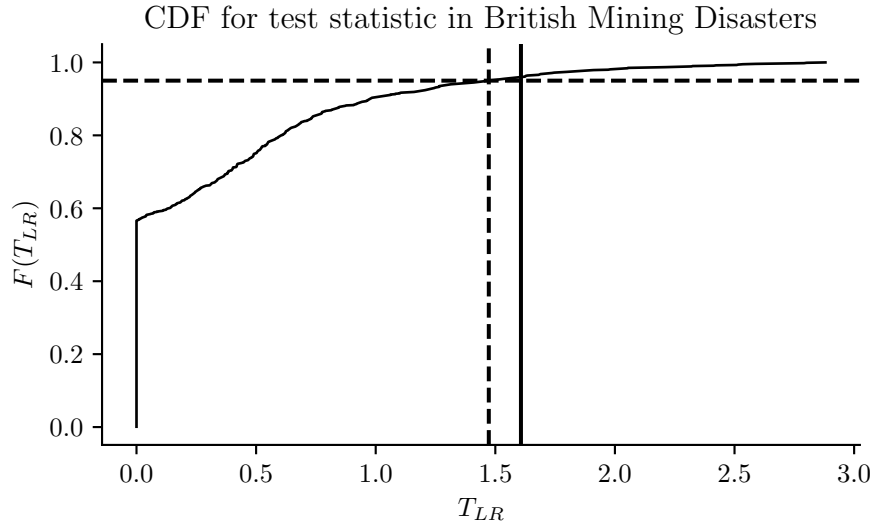


Figure 7.3: Empirical CDF for the Likelihood ratio test statistic. The significance level, alongside its critical value, is marked by the dashed line. The observed value is the full line.

We now have a sample of the random test statistic and can compute its properties empirically. For example, we can compute the 0.95% quantile, which is the critical value for rejecting H_0 . It happens to be 1.47. The observed test statistic is $T_{LR} = 1.61$, and we can therefore reject the null hypothesis. Furthermore, we can estimate its p -value $p = \sum_{i=1}^{n_{sim}} I(T < T_i^*) = 0.041$. The full distribution is shown in figure 7.3 ▲

7.3 Power

The power of a statistic is the probability of rejecting the null hypothesis as a function of the true parameter value. It is therefore a function of the true parameter value, and we are here only concerned with the window length d . Given two test statistics with correct coverage, we prefer the one with the most power, as it will more often correctly reject the null hypothesis. By construction, T_{LR} has the correct coverage, and simulations show that T_{ML} is somewhat too conservative. However, it cannot be helped since $\hat{d}$ is discrete, and the critical value is almost certainly not exactly an integer.

To find the power, we first calculate the critical value T_c for our statistic at the desired level α . We have decided to let $\alpha = 0.05$, but the procedure will work for any desired level of significance. Given a sample $\{T^*\}$ of simulated test statistics under $H_0 : d = 1$, the critical value is the $1 - \alpha$ quantile of $\{T^*\}$. Having found T_c , we then simulate 10000 data sets for each $d \in \{1, \dots, 50\}$, and calculate their respective T^{**} . The power is then the total number of times $T_c < T^{**}$ divided by the number of sample data sets.

We continue this for the case of the normal parameter transition WoC model, and the critical values can be found in table 7.1. For the likelihood ratio statistic, we see that the critical value increases with $\Delta\theta$, which means that simulation is indeed necessary to determine T_c for any given problem. Interestingly, for the maximum likelihood estimate statistic, the critical value initially increases with $\Delta\theta$ before shrinking again. From the selected $\Delta\theta$, it suggests that $13 < \hat{d}$ would always reject H_0 , but using this critical value is far too conservative.

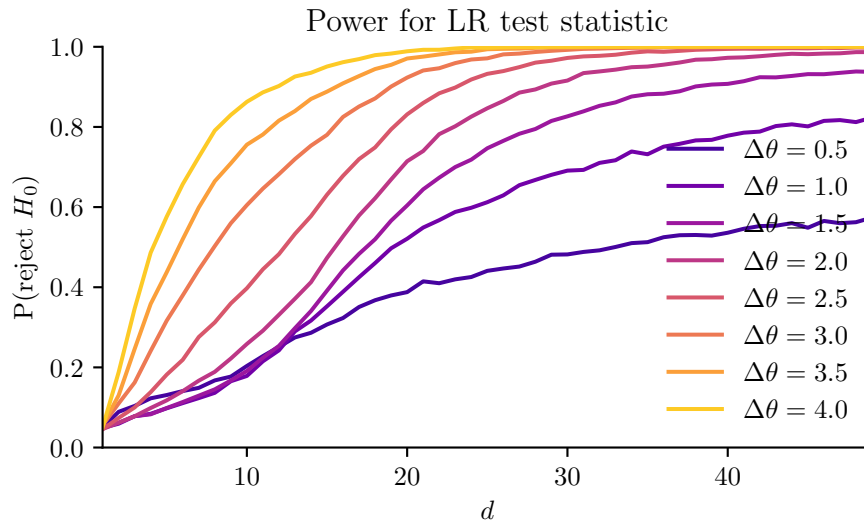


Figure 7.4: The power of the statistics in figure 7.1

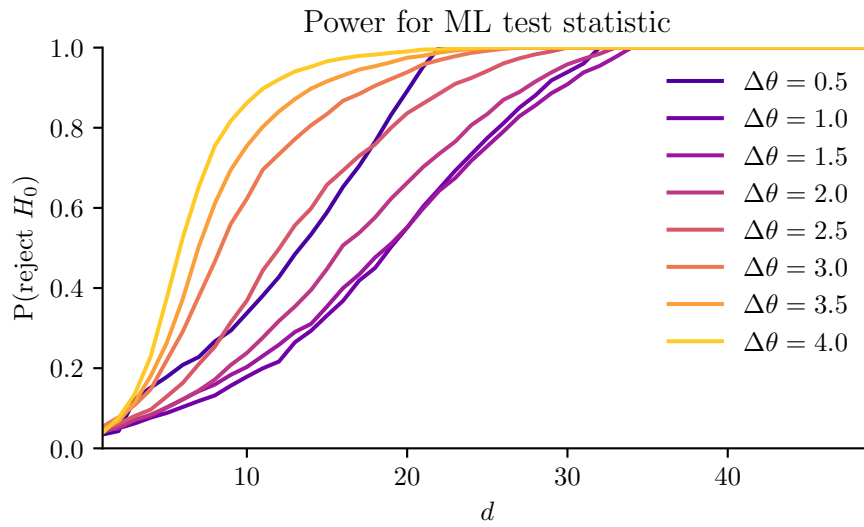


Figure 7.5: The power of the statistics in figure 7.2

	T_{LR}	T_{ML}
$\Delta\theta$		
0.5	0.05	2
1.0	0.43	12
1.5	1.35	14
2.0	2.33	13
2.5	2.99	10
3.0	3.28	7
3.5	3.62	6
4.0	3.88	5

Table 7.1: Critical values for $\alpha = 0.05$.

It is surprising to see how little power these tests have. Even when $\Delta\theta = 4$ and $d_0 = 10$ are the probabilities of rejecting H_0 only about 89%. The results also show that the power of T_{LR} and T_{ML} depends on d_0 and $\Delta\theta$ very differently. In general, T_{LR} initially performs slightly better than T_{ML} , until the former ‘catches up’, but this cannot be claimed universally. For example, when $\Delta\theta = 0.5$ must $33 < d_0$, there is not even a 50% probability of rejecting H_0 using T_{LR} . On the contrary, $22 < \hat{d}$ gives an estimated 100% chance of rejection, making T_{ML} not only preferred, but T_{LR} almost useless. In summary, we can say that the likelihood ratio statistic has better performance when the probability of rejecting H_0 is low, while the MLE statistic is better when the probability of rejecting H_0 is large. That is, T_{LR} is preferred in difficult cases, and T_{ML} is preferred in easy cases.

7.4 The model selection perspective

A commonly used alternative to hypothesis testing is *model selection*. In this field, we are presented with two competing models ‘explaining’ the data, and we wish to select which one explains it the best. Model selection can be considered a broader concept than hypothesis testing. The models being compared need not be nested, nor similar in any regard, except that they model the same data. This is useful if we want to contrast our WoC analysis not only with CP analysis, but, say, stationary models, regression models, or any other kind of statistical or machine learning model.

Formally, model selection consists of a finite family of models $\mathcal{G} = g_1, g_2, \dots$ and a goodness-of-fit function $Q(g, y) \rightarrow \mathbb{R}$. Then, the model choice is the minimizer:

$$g^* = \arg \min_{g \in \mathcal{G}} Q(g, y)$$

The most common choice of criterion is the Akaike information criterion (AIC)

$$\text{AIC} = 2k - 2l(\hat{\theta}_{ML}; y),$$

where k is the number of free parameters in the model. The criterion was developed to find the model that minimizes the expected prediction error, and it is thus a measure of generalizability (Akaike, 1974). This criterion is very similar to the likelihood ratio test, which is no coincidence, as it is similarly derived. However, the criterion is only asymptotically justified and relies on large sample theory, which, in our particular case, is not universally valid. It relies on the asymptotic normality of the MLE, which we have

established does not hold for $\hat{a}, \hat{b}$. We therefore expect the criterion to underperform, but its flexibility makes it still interesting where other means of comparison are difficult. However, if we are to choose between data models, i.e. selecting the best fitting f_L and f_R , AIC would often be appropriate.

An alternate criterion is the Bayesian information criterion (BIC)

$$\text{BIC} = k \log n - 2l(\hat{\theta}_{ML}; y),$$

which is based on the asymptotic normality of the posterior distribution (Schwarz, 1978). Since it is not confined to the Bayesian paradigm (merely inspired by it), it is applicable for comparison with the AIC.

With our particular problem of selecting between a CP model and a WoC model, k is fixed at 1 in the former case, and at 2 in the latter case. This means that the AIC would prefer the larger model, the WoC, if it is more than e times likely than the smaller model. Furthermore, when $\log n > 2$, which it will for all our examples, BIC selects sparser models. While in general this could be appropriate, in our case it is not. Since observations far from either the CP or the WoC contribute vanishingly little to the variance of the estimates, the $\log n$ term would penalize model complexity due to observations that do not meaningfully contribute to the estimation. In other words, as $n \rightarrow \infty$, $k \log n$ would dominate the difference in BIC between the CP and WoC model, and the CP model would always be selected.

The BIC is derived from Bayesian model selection, by means of a normal approximation of the posterior distribution. It therefore relies on $\pi(a, b | y)$ being well approximated by a (in this case bi-variate) normal distribution. This approximation is similar to the one in the AIC and fails for similar reasons. Bayesian model selection can, however, still be done directly, as there is no need for approximation. In a general model selection problem, we must elicit a prior on the competing models and calculate the posterior probability for the null model. Then we can reject the null model if this probability is smaller than α . This is, however, not entirely necessary, since the null model is nested within the full model. Instead, we can use the posterior probability $\Pr(d = 1 | y)$. This means that the prior on d serves a double purpose, namely both reflecting our (educated) prior belief on d and controlling for the probability of erroneously rejecting the null hypothesis. One could perhaps construct the following prior:

$$\pi_0(d) = (1 - \alpha)I(d = 1) + \alpha\pi(d | d > 1) \quad (7.3)$$

The augmented prior (7.3) certainly has the advantage of being correct in the sense that it reflects our prior preference for a CP model. Unfortunately, it is too conservative, and under $H_0 : d = 1$, it never rejects the null hypothesis. The null hypothesis is, however, rejected too often using the original prior, i.e. $\pi(a, b) = \text{const}$, but as $\Delta\theta$ increases, the probability of a type I error approaches $\alpha = 0.05$, as shown in figure 7.6. The posteriors were calculated using knowledge of $\Delta\theta$, but using the profiled log-likelihood, i.e. using $\hat{\theta}_L$ and $\hat{\theta}_R$, yielded indistinguishable results.

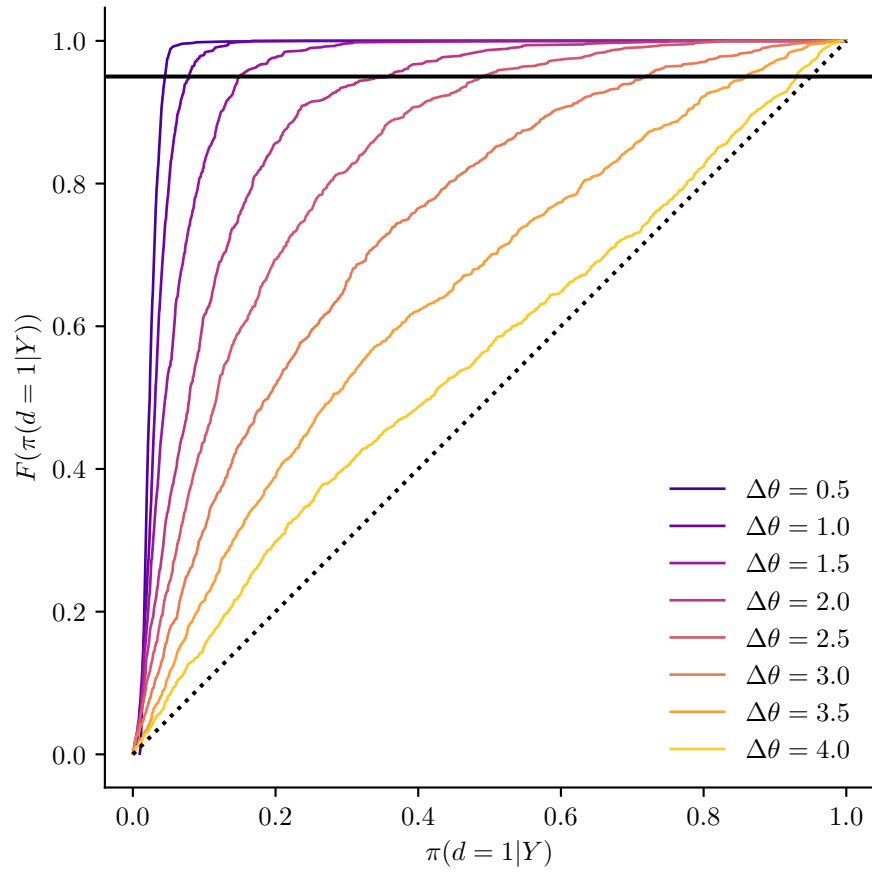


Figure 7.6: CDF for $\pi(d = 1 | Y)$. If $\pi(d = 1 | Y)$ were to have correct coverage for any levels, the CFDs follow the diagonal, which they seem to approach as $\Delta\theta$ increases.

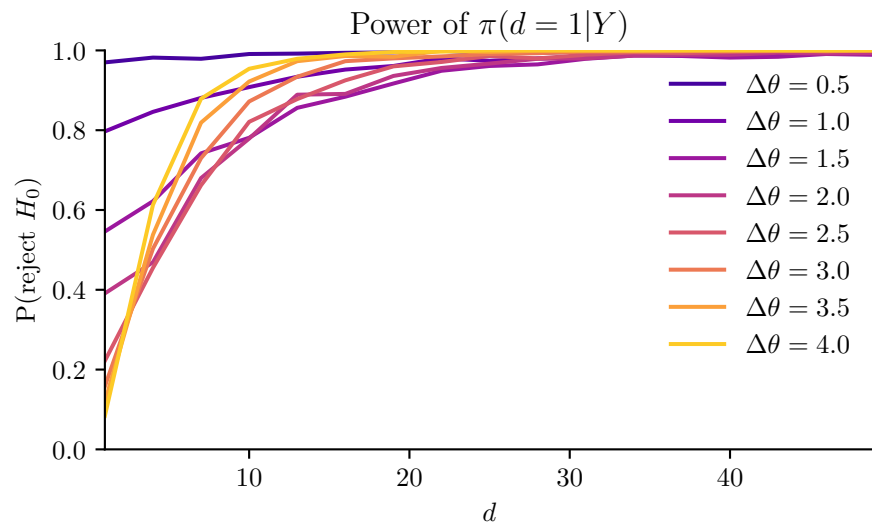


Figure 7.7: Power for the statistics. Note that none of them start at 0.05, which means that they do not have correct coverage, and we can therefore hardly compare them to the other test statistics. When $\Delta\theta = 4$, $\Pr(\text{Reject } H_0 | H_0) = 0.084$ thus still not correct.

Chapter 8

Application: Bjørnholt

Here we analyse the number of skiing days at Bjørnholt, a DNT cabin in Nordmarka. The data represent the number of days it was possible to ski at Bjørnholt per season and tell the foreboding story of warmer winters.

This analysis serves the purpose of showing how WoC models can be applied to models with dependent data, which we have not done hitherto. The model we will choose is a simple linear regression with i.i.d. deviates:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \\ &= \beta^T \mathbf{X}_i + \epsilon_i \end{aligned}$$

Where Y_i is the number of skiing days, and X_i is the year. Since $y_i \in (1, \dots, 366)$ (stopping at 366 due to leap years), one could argue that Bernoulli regression should be used. However, since a Bernoulli model would be approximately normal for the magnitude of y in relation to the number of days in a year, we will ignore this and continue with the assumption of normality. The assumption of normality is not entirely justified, as previous analyses suggest that the ϵ_i are autocorrelated (Cunen, Hermansen & Hjort, 2018), but we will not model this here. Finally, there are some observations missing from the data sets, namely during the Second World War, but this poses no challenge for a regression model. There exist imputed values, but we will keep the data as it is, highlighting the flexibility a regression model enables.

8.1 The WoC model

How exactly should we model this as a Window of Change model? The change now needs to occur in the dimension of one of the covariates x . Furthermore, we can exclude a mixture transition since we are mainly interested in the trend in y with respect to x . Using a mixture transition would instead model two different trends, and in the window, y_i would either follow one or the other. In other words, for any observation within the window, the mixture transition would model that it is either $Y_i = \beta_{0,L} + \beta_{1,L}x_i$ or $Y_i = \beta_{0,R} + \beta_{1,R}x_i$.

Therefore, we use the mixture transitions:

$$Y_i = w_x(\beta_{0,L} + \beta_{1,L}x_i) + (1 - w_x)(\beta_{0,R} + \beta_{1,R}x_i) + \epsilon_i$$

letting the weights $w_x = w(x, a, b)$ now depend on x rather than the index i . We also want the trends to intersect, i.e. $\mathbb{E}_x[Y]$ is a continuous function of x , and we are currently

not concerned about the intersection. This suggests using the following model.

$$Y_i = \beta_0 + \beta_L x_i + (\beta_R - \beta_L)(1 - w_x)(x_i - a) + \epsilon_i.$$

The final term, $(\beta_R - \beta_L)(1 - w_x)(X - a)$, partially adds the effect of $\delta\beta = \beta_R - \beta_L$, only when $x = a$, and it fully adds the effect when $x = b$. We have to subtract a so that it starts at zero. We may therefore interpret $x - a$ as ‘anno fenestra’ (the year of the window), as compared to x : ‘anno domini’ (the year of the lord). Letting $\beta = (\beta_0, \beta_L, \beta_R - \beta_L)^T$ and $\mathbf{X}_{design} = (1, X, (1 - w)(X - a))$, we have reduced the WoC model to an ordinary multiple linear regression model. $\beta_R - \beta_L$ is the familiar change in parameter. Note that if $b - a = 1$, then $(1 - w_x)(x - a) = 0$ if $x < a$ and $x - a$ for $a < x$.

Indeed, the CP model is a special case of the WoC model, but interestingly, having fixed a, b , the CP and WoC regression models have equal degrees of freedom. This is of interest as normally, when comparing regression models, we penalize the complexity of the models, for instance, with the AIC. If we were to naively compare the AIC of the CP and WoC regression models, we would only compare log-likelihoods, which necessarily cannot be lower for the WoC model. Of course, the WoC model has an additional parameter b , but as we have seen in Chapter 7, we cannot use normal complexity penalization for comparing CP and WoC models. Instead, we must find the critical value for the deviance through simulation.

The term $(1 - w(x))(x - a)$ warrants a comment. Within the window, it is a product of two linear terms: both $(1 - w(x))$ and $(x - a)$ increasing. The contribution from this term is therefore not linear, but quadratic within the window, and as the figure 8.1 shows, perhaps slightly more smooth when transitioning. There are other ways of constructing smooth, piecewise polynomials, like with splines,

We have currently only concerned ourselves with estimating β , but the full model requires that we estimate the variance of the deviates as well. In normal regression settings, estimation of β can be done without regard to σ^2 , but since we are, in effect, comparing many regression models — one for each combination of a, b — it will affect model choice. How we model the variance will also affect the model selection procedure when comparing the WoC model with the CP model. We will therefore consider both cases: where the variance is constant, and when the variance changes linearly from σ_L^2 to σ_R^2 . Unfortunately, even though the normal distribution with an unknown mean and variance is part of the exponential family, it is not possible to solve for the MLE for σ^2 , or expressed in the natural parameter $\eta = \frac{1}{2\sigma^2}$. The sufficient statistic for η is y^2 , thus finding the MLE requires solving the following set of equations

$$\begin{aligned} w^T y^2 &= \sum_{i=1}^n \frac{w(x_i)}{w(x_i)\eta_L + (1 - w(x_i))\eta_R} \\ (1 - w)^T y^2 &= \sum_{i=1}^n \frac{1 - w(x_i)}{w(x_i)\eta_L + (1 - w(x_i))\eta_R}, \end{aligned}$$

which is equivalent to an n degree polynomial. We must therefore use generic numerical methods instead.

Finally, while it is outside the scope of this thesis to detect the presence of a change point, we will also compare the two alternative models: no change, i.e. that the skiing days are stationary, and a normal linear regression.

	$\hat{a}$	$\hat{b}$	l_{max}	$\hat{\beta}$	$\hat{\sigma}$	dof
Stationary	1897	1897	-512.49	[123.68 0. 0.]	37.69	2
Linear	1897	2022	-500.86	[151.32 -0.43 0.]	34.19	3
CP _{CV}	1976	1977	-497.52	[142.08 -0.14 -0.97]	33.31	4
CP	1979	1980	-495.36	[142.69 -0.16 -0.99]	[28.65 38.3]	5
WoC _{CV}	1969	1991	-496.73	[140.57 -0.08 -0.94]	33.07	4
WoC	1975	1990	-494.60	[141.56 -0.12 -0.99]	[28.51 38.49]	6

Table 8.1: Table of summary statistic of Bjørnholt data set for different models. All assume normal deviates. The Stationary assumes no change, the linear assumes a single linear trend, CP_{CV} assumes two linear trends which intersect at the change point with constant variance, CP assumes the same, but allow the variance to change as well. WoC_{CV} assumes the two linear trends with a WoC and constant variance, while WoC allows the variance to change as well.

8.2 Results

The results of all five models are summarized in table 8.1. Not surprisingly, the stationary model is vastly unlikely compared to the alternative models, $l_{max} = -512.49$, and we therefore reject the stationary model and accept that the number of skiing days is shrinking. According to the linear regression model, the number of skiing days is reducing by just under half a day per year. This model is also quite unlikely, given its low $l_{max} = -500.86$.

Turning to the CP and WoC models, they all seem to agree that initially, before around 1975, the change was slower at approximately -0.1 per year. After around 1980, it shrinks by nearly a full day per year. Comparing the models with constant variance against those with changing variance, we see that both the WoC and the CP models increase their likelihoods by more than two. They are thus preferred over their constant counterparts using the AIC. We may use AIC to select the models here, since we are comparing the distributions f_L and f_R , not a CP model against a WoC model where the AIC does not apply. The preferred WoC model estimates that the change in the trend started in 1975, and that the change had been followed though in 1990 - a span of fifteen years.

Exactly how does the variance change? It increases after the change, which is in accordance with climate change not only increasing the temperature of the Earth, but also making the climate more ‘extreme’ or in this case, more variable.

It is interesting to observe that the relative relationship between the CP and WoC models with constant variance is almost exactly the same as that with variable variance. Furthermore, their estimates are also very similar to each other. This is perhaps not very surprising when we inspect the regression curve of the WoC model: it follows the CP curve extremely closely. Thus, the residuals $y - \beta^T X_{design}$, the sum of squares of which are the sufficient statistics for σ^2 , are also almost identical.

With a log-likelihood ratio of only 0.76, we would be inclined to reject the WoC model in favour of the CP model. However, as we have seen, it is necessary to calculate the critical value for this test statistic. We test both the case where the variance is constant and unique for the two distributions. It is not necessary to compute the test statistic for the models with constant variance, as we have already observed that changing variance is the preferred model. We only do it to inspect the effect that the model choice has on T_{LR} . In both cases, we simulate data under $H_0 : d = 1$, using the MLE for τ, β, σ^2 of the corresponding models. The resulting CDF is shown in figure 8.2.

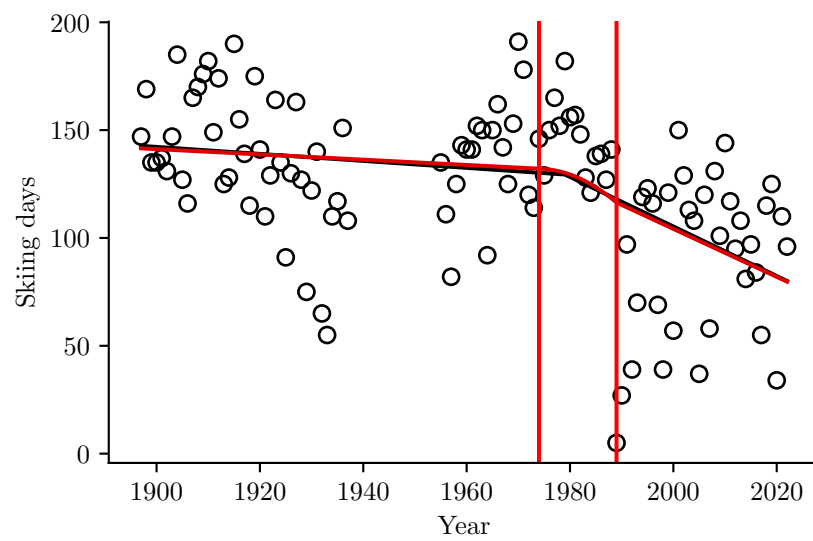


Figure 8.1: Skiing days at Bjørnholt (circles) with best fit CP model (black line) and best fit WoC model (red line). The two regression lines almost overlap.

For both tests, the observed test statistic is larger than the critical value, and we reject the null hypothesis. Still, the p values are just barely smaller than 0.05. Considering that the current model is probably not preferred over a model with autocorrelation, the result is not as convincing as it could be.

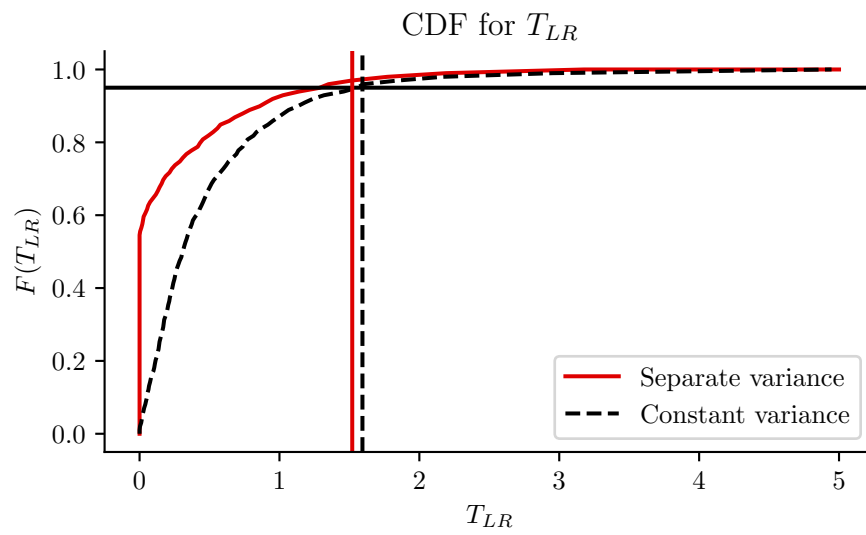


Figure 8.2: Null distribution for $H_0 : d = 1$. H_0 is rejected with a p -value of 0.03 if f_L and f_R has unique variance, and with a p -value of 0.045 with constant variance. Computed using 1000 samples.

Chapter 9

Application: Tirant lo Blanch

Parts of this analysis are built upon a course in Change-Point analysis, where the final project was a Window of Change extension. All text and figures are, however, written anew.

9.1 A tale of two authors

Tirant lo Blanch is a medieval chivalry novel written in Catalan. Outside its historic significance as an early and very firm example of a distinct Catalan culture, it is of interest in change point analysis, as it is widely known that the book had two authors. Joanot Martorell began the book but did not complete it before his death. Martí Joan de Galba continued his work and completed the book, claiming to have written the final quarter of the manuscript. Our goal is then to discover, through the data alone, if this claim has merit, by means of fitting a WoC model.

If one author picked up the pen where the former one placed it, what sort of transition do we expect? A CP model assumes that once Martorell wrote his last chapter, Joan de Galba continued without regard to the previous work, in the sense that his writing style was already developed and did not adapt further throughout the work. If we reject this belief for the purposes of this example, then we must construct a model that allows Joan de Galba to transition into his own writing style. One possible way is to assume that Martorell had left drafts for future chapters, and Joan de Galba patched them together. Then, we should model the transition as a mixture transition, which we now do.

Having decided on the transition model, we need to decide on the data model, and before that, discuss the observations we are working with. While there are many ways of conducting natural language and linguistic analysis, we will use a common and fruitful representation. Consider the 10 vector $y_i^T = (y_{i,1}, y_{i,2}, \dots, y_{i,10})$ where $y_{i,j}$ is the count of words of length j in chapter i . Words longer than 10 are aggregated in the final component, i.e. $y_{i,10}$ is the count of words of length 10 and longer. Then we can represent the writing style as the relative preference for long, short, or medium words. Furthermore, we only consider chapters with at least 200 words, of which there are $N = 425$. Thus, our observation matrix y is the 10 aggregates of word count by length for 425 chapters.

9.2 Two WoC models

The simplest possible way to model the data is with a multinomial distribution. The multinomial model is parametrized by a probability vector. Given a $k = 10$ dimensional

probability vector p , subject to $\sum_{j=1}^k p_j = 1$ and $p_j > 0$, the likelihood of a single observation y_i is

$$L(y_i, p) = m_i! \prod_{j=1}^k \frac{1}{y_{i,j}!} p_j^{y_{i,j}} \quad (9.1)$$

Where $\sum_{j=1}^k y_{i,j} = m_i$. The multinomial model is of interest because of its clear interpretation: Any given word has a fixed probability of being of length p_j . It would be as if the author cast a 10-sided die for every word he wrote, and that Martorell had a slightly different die than Joan de Galba.

To construct the multinomial WoC model, let $f_L = \text{Multinomial}(\theta_L)$ and $f_R = \text{Multinomial}(\theta_R)$. The likelihood is then

$$L(a, b, \theta_L, \theta_R; y, m) = \prod_{i=1}^n \left[w_i m_i! \prod_{j=1}^k \frac{1}{y_{i,j}!} \theta_{L,j}^{y_{i,j}} + (1 - w_i) m_i! \prod_{j=1}^k \frac{1}{y_{i,j}!} \theta_{R,j}^{y_{i,j}} \right] \quad (9.2)$$

Since we treat m_i and y_i as fixed we can simplify the expression as

$$\propto \prod_{i=1}^n \left[w_i \prod_{j=1}^k \theta_{L,j}^{y_{i,j}} + (1 - w_i) \prod_{j=1}^k \theta_{R,j}^{y_{i,j}} \right] \quad (9.3)$$

We now want to find the profiled log-likelihood, but this has no closed form solution, and we therefore use the EM algorithm for optimization. Then we need to compute the two steps required for the EM algorithm. The E step calculates the probability that chapter i is written by Martorell, and we can use equation (6.4) directly, using the likelihood from equation (9.2). The M step maximizes the estimated nuisance parameters given the estimates from the E step. They are simple to deduce since the multinomial distribution is part of the exponential family

$$\hat{\theta}_{L,j} = \frac{\sum_{i=1}^n \gamma_i y_{i,j}}{\sum_{i=1}^n \gamma_i m_i}$$

and

$$\hat{\theta}_{R,j} = \frac{\sum_{i=1}^n (1 - \gamma_i) y_{i,j}}{\sum_{i=1}^n (1 - \gamma_i) m_i}.$$

The EM algorithm can then be iterated until convergence, which is typically quite fast. Especially in the current case, since the number of chapters is so large.

How does this model do? Very poorly, in fact: $\hat{a} = 1$ and $\hat{b} = n - 1$, i.e. it pushes the window to the far corners of the parameter space. The reason for this is very simple; a multinomial model is inadequate as the data is overdispersed. Alas, our authors did not throw a die for each word they wrote. This revelation is not new, as Cunen, Hermansen and Hjort (2018) already remarked. In their analysis, they find that a simple multivariate normal is the best candidate model for the CP problem.

What is interesting, compared with the CP model, is that the estimated change point, $\hat{\tau} = 320$, was identical for the multinomial model and the multivariate normal model. Thus, for the purposes of estimation, the more complex model offered limited benefit. When constructing confidence curves, however, the model fit is more important. The estimated change point with the naive multinomial WoC model is, of course, $n/2 = 212.5$.

We continue with their preferred model: a heteroscedastic multivariate normal distribution. Let $x_i = y_i/m_i$, and we ignore the final word count, i.e. $y_{i,10}$. The reason

for ignoring the last word count is that $\sum_{j=1}^k y_{i,j}/m_i = 1$, and we do not want our mean vector μ_L to be bound by this condition. Then, we model the x_i as they suggest:

$$\begin{aligned} f_{L,i}(x_i) &= N_9(\mu_L, \Sigma_L/m_i) \\ f_{R,i}(x_i) &= N_9(\mu_R, \Sigma_R/m_i). \end{aligned}$$

The data is heteroscedastic because every chapter i has its own covariance matrix, but this matrix is just Σ_R scaled by the number of words m_i . The consequences of this modelling choice are that μ_L is confined to the unit cube, and that chapters with more words have a smaller variance. The likelihood is the familiar

$$L(a, b, \theta_L, \theta_R; x, m) = \prod_{i=1}^n [w_i f_L(x_i) + (1 - w_i) f_R(x_i)].$$

Where

$$f_L(x_i) = (2\pi)^{-9/2} |\Sigma_L/m_i|^{-1/2} \exp\left\{-\frac{1}{2}(x_i - \mu_L)^T (\Sigma_L/m_i)^{-1} (x_i - \mu_L)\right\}$$

and conversely for f_R .

To find the profiled log-likelihood, we again use the EM algorithm. The E-step is still (6.4), and we deduce the M step now. We need to find the maximizer for both the mean vector and the covariance matrix. Beginning with the mean vector, the maximizer is

$$\mu_{\hat{L},j} = \frac{\sum_{i=1}^n \gamma_i m_i x_{i,j}}{\sum_{i=1}^n \gamma_i m_i}.$$

We see that the contribution for any one chapter of $\theta_{L,j}$ is weighted by both the number of words, m_i , in the chapter, and the current posterior belief that $z_i = 0$. The procedure is equivalent for the covariance matrix, using the observed covariance as the base estimator:

$$\hat{\Sigma}_L = \frac{\sum_{i=1}^n \gamma_i m_i \{x_i - \hat{\mu}_L\} \{x_i - \hat{\mu}_L\}^T}{\sum_{i=1}^n \gamma_i}$$

How does this perform? It results in $\hat{a} = 320$ and $\hat{b} = 321$, that is, a Change-Point model is the most likely (see figure 9.1). It is reassuring that it concurs with the CP analysis.

9.3 The Bayesian analysis

We now extend the Gibbs sampler to the multivariate case with heteroscedastic variance. First, we define the priors for Σ_L, Σ_R and μ_L, μ_R .

$$\begin{aligned} \pi(\Sigma) &= \text{Inverse-Wishart}(\nu, \Lambda) \\ \pi(\mu \mid \Sigma) &= N_9(\xi, \Sigma/\kappa) \end{aligned}$$

We let $\nu = 1 + d$ and $\lambda = I_d$, where $d = 9$, which results in a suitably flexible prior in Σ . Regarding the prior on μ , we can be more informative since we know that x_i is contained within the unit hypercube of dimension 9. Therefore, let $\xi_j = 1/9$ for all components, and $\kappa = 100$. This results in $\mathbb{E}[\mu \mid \Sigma] = 0.12$ and $\text{Var}[\mu \mid \Sigma] = 0.18^2$. Given knowledge of which class the sample is from, i.e. knowledge of z , the data model is

$$\pi(x \mid \mu) = N(\mu, \Sigma/m).$$

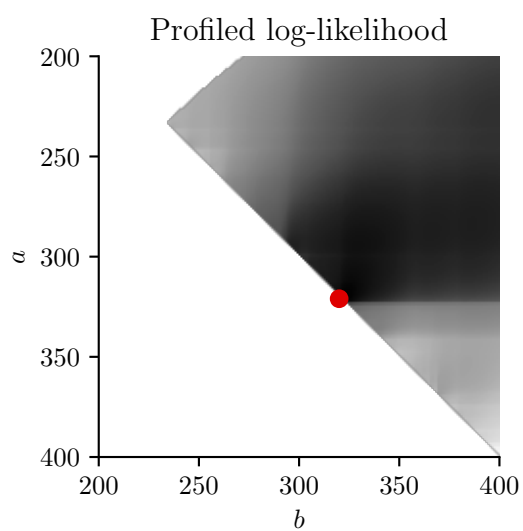


Figure 9.1: Profiled log-likelihood for the multivariate normal model. The MLE is marked in red. The edge in the upper left corner is the edge of the search space, and should not be interpreted that the likelihood is zero outside it.

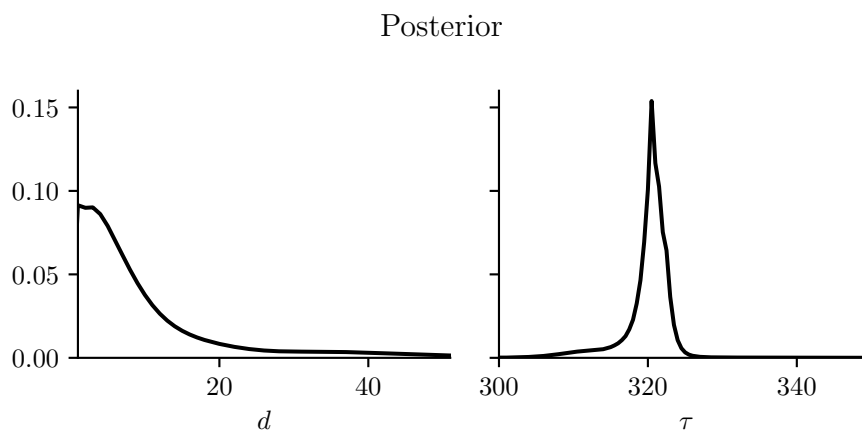


Figure 9.2: Posterior distributions for d and τ calculated by normalizing the profiled log likelihood then marginalizing.

	a	b
mean	297.13	412.08
stdev	9.75	10.29
M_{eff}	878.63	309.00
Gelman-Rubin	1.00	1.00

Table 9.1: Chain diagnostics for the MCMC.

Let $\bar{x}_m = \sum_{i=1}^n \frac{m_i x_i}{n\bar{m}}$ be the weighted mean of x and $S = \sum_{i=1}^n m_i (\bar{x}_m - x_i)(\bar{x}_m - x_i)^T$ be the weighted observed covariance. Then, the updated parameters conditioned on the class sample x (cf. (Gelman, 2013, p. 72)):

$$\begin{aligned}\kappa_n &= \kappa + n\bar{m} \\ \xi_n &= \frac{(\kappa\xi + n\bar{y})}{\kappa_n} \\ \nu_n &= \nu + n \\ \Lambda_n &= \Lambda + S \frac{\kappa n \bar{m}}{\kappa_n} (\bar{x}_m - \mu)(\bar{x}_m - \mu)^T\end{aligned}$$

And the conditional distributions for Σ and μ are

$$\begin{aligned}\Sigma &| y \sim \text{Inverse-Wishart}(\nu_n, \Lambda_n) \\ \mu &| \Sigma = N_9(\xi_n, \Sigma/\kappa_n)\end{aligned}$$

To use the WoC Gibbs sampler, replace x and m in the above expressions with $x_L = \{x_i : z_i = 0\}$, $m_L = \{m_i : z_i = 0\}$ and re-index them such that $x_L = x_{L,1}, x_{L,2}, \dots, x_{L,n_L}$ where $n_L = |z_i : z_i = 0|$ (and conversely for x_R etc.)

Before we apply the Gibbs sampler, we need a prior on the window parameters a, b . There are a couple of considerations to be made. Since the number of observations is somewhat large, we might wish to restrict d from above by some d_{max} . This could potentially be justified by a belief that there are only so many notes left by the first author for the second author to use, or perhaps by the belief that after a certain number of chapters, the second author has fully developed their style. Another consideration is that since we know that $\hat{d} = 1$, we might wish to bias $\pi(d)$ towards larger values, but this is, of course, not the appropriate Bayesian method. As can be seen in figure 9.3, the Bayesian estimate for b is pushed towards the edge, and for this reason, biasing the prior on d is ill-advised.

Now, we can use the WoC Gibbs sampler as described, conducting five runs with $n_{sim} = 1000$ samples each. The traces can be seen in figure 9.3, along with their posterior means. The posterior probability that Joan de Galba wrote chapter i is shown in figure 9.4.

The Bayesian analysis tells a very different story, namely that of a very long window of change, starting at $\mathbb{E}[a | x] = 298$ and ending at $\mathbb{E}[b | x] = 410$. The mean window length is $\mathbb{E}[d | x] = 113$ and $\mathbb{E}[\tau | x] = 354$. In fact, b towards the edge, suggesting that the window of change never really concluded within the book. Because of this, another model was tried, which biased d by a factor of $1/d$, i.e. $\pi'(a, b) = \pi(a, b)/(b - a)$. However, this had little effect, and the posterior distributions were largely identical.

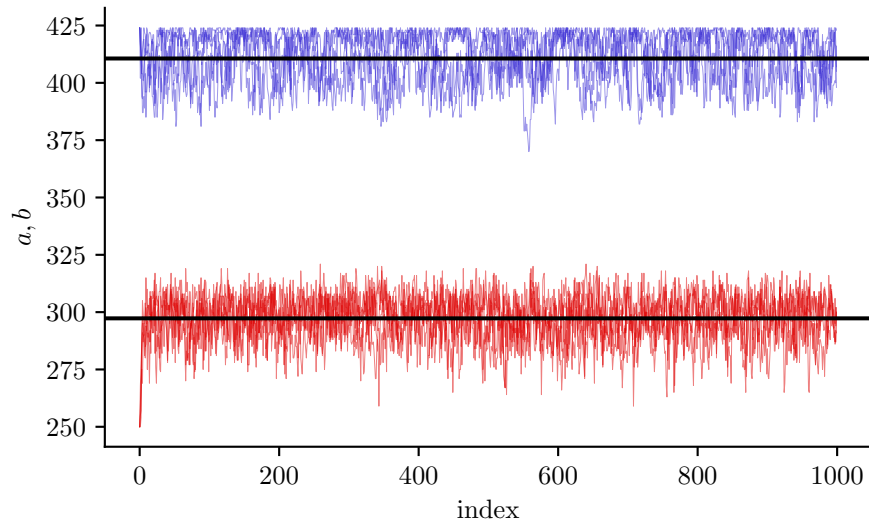


Figure 9.3: Traces from the five MCMC runs for the multivariate normal model. The red traces are a , and the blue traces are b . Starts where randomized but close to their mean value. Starting further away, i.e. letting say $a < 200$ only resulted in slower convergence, while having the same stationary distribution. Using $n_{sim} = 1000$ samples.

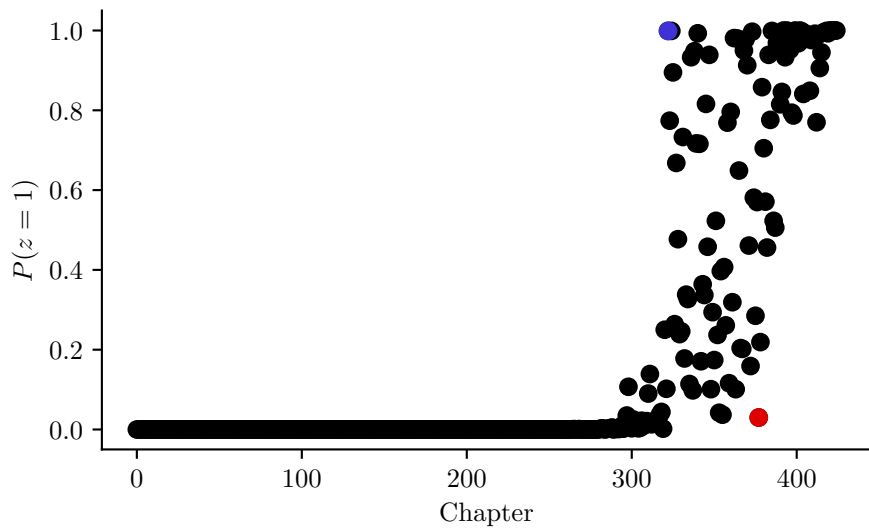


Figure 9.4: Posterior probability that chapter i was written by Joan de Galba. It shows clearly that Martorell was solely writing the book until about $i = 320$, then more and more chapters were written by Joan de Galba, but Martorell was influential until late into the book. At chapter 378 was he still overwhelmingly the most likely author (red dot). The first chapter almost certainly written by Joan de Galba was chapter 322 (blue dot)

9.4 Discussion

This analysis highlights a caveat with WoC models, namely that $\hat{d}$ seems to be more sensitive to misspecification than $\hat{t}u$. While in the CP analysis, the model choice had little impact on the estimation of the change point, overdispersed data in the WoC model completely dominated the analysis. In particular, the multinomial WoC model was entirely overshadowed by the overdispersion, making the estimates meaningless. They are meaningless in the sense that the estimates provide no actionable information about the location of the window. Instead, the estimates simply suggest that the data is overdispersed.

Regarding the better-fitting multivariate normal model, we are now in a situation where the frequentist and Bayesian approaches disagree considerably, and we must reconcile their conclusions. This discrepancy could be explained by two different phenomena that we have observed throughout the thesis. First, the MLE can be very biased, particularly for mixture transitions. While we only considered the univariate case in Chapter 5, it is hard to believe that the bias subsides in higher dimensions; this must be investigated further. The profiled log-likelihood function was also very flat around $\hat{a}, \hat{b}$, suggesting large uncertainty. Interestingly, there is a sharp decline in log-likelihood after $a = 322$ (a vertical edge in figure 9.1), corresponding to the first chapter almost certainly written by Joan de Galba.

The second phenomenon is the inherent bias in the Bayesian posterior for pushing a, b to the edges. The degree to which it increases with the dimensionality of θ_L and θ_R , which in the current model is $9 + 9 \cdot 9 = 90$ for each, has not been determined. We have not calculated the exact dependency, and it is unclear how the estimation of a covariance matrix affects this bias. There is a stark difference between the ‘posterior’ in figure 9.2, which was calculated by simply normalizing the posterior, and the full Bayesian procedure. If shortcuts are to be made, correction is a necessity.

Nevertheless, the conclusion from the WoC analysis is almost identical to the CP analysis if we only consider the data before $i = 320$. This suggests that there is little doubt that the beginning of the book was written by Martorell. However, after this index, the WoC analysis suggests that the influence of Martorell never ceased, and while Joan de Galba certainly wrote some chapters, he most likely did not write them all. This extra insight is facilitated by the WoC model and is therefore a good example of the type of insight that this model can contribute to and enrich the literature.

Chapter 10

Concluding remarks

10.1 Summary

The Window of Change model offers a fundamentally different perspective from the traditional Change-Point model. This distinction is evident in the random walk decomposition, which separates into a CP walk and a ‘window’ walk. Our analysis shows that proper treatment of the latter is essential, as it introduces additional complexity not present in CP models.

Using both confidence curves and Bayesian methods, we found that uncertainty about the window length d is generally greater than that about the midpoint τ . Estimating d often leads to bias, and we developed several correction techniques with varying effectiveness. These approaches rely on efficient estimation algorithms, which we have implemented across a range of models.

In the simplest case (Chapter 4), Bayesian credibility intervals and frequentist confidence intervals yielded comparable results. However, in more complex settings — such as the multivariate normal model in Chapter 9 — the two frameworks diverged considerably.

Testing WoC versus CP models using likelihood-ratio statistics is complicated by dependence on nuisance parameters. As a result, critical values must be determined for each specific dataset and model. It is therefore inappropriate to reject the WoC model in favor of the CP model based solely on the log-likelihood ratio without accounting for model and its parameters.

Overall, our findings suggest that applying the WoC model requires careful validation for the data and context at hand. In particular, if inference on the window length is the main goal, it is necessary to address the biases associated with both estimation, whether frequentist or not, and the model choice. Although the Bayesian analysis yielded the smallest root mean square error in the univariate normal case, it did not convincingly remove the bias. More investigation is necessary to determine if the bias becomes increasingly negative or if it flattens out.

In conclusion, the Window of Change models work, provide unique inference, but require more careful treatment than a first glance would suggest.

10.2 Future work

Our investigations have led to many still unanswered questions, which this thesis does not cover.

If we are concerned with when a change happens, we are most likely also interested in the change itself, which amounts to studying the relationship between θ_L and θ_R . We have been mainly concerned with inference on the window parameters a, b , and not on the nuisance parameters.

For the purposes of this thesis, we have assumed that f_L and f_R are simple, and in particular that $y_L \sim f_L(y)$ are i.i.d. (except for the regression model for Bjørnholt in Chapter 8). We could instead model f_L and f_R as time series, which would complicate our calculations slightly. While in the most general form, the WoC we have considered does encompass these models, we have only worked with likelihoods that factor nicely. Extending f_L and f_R to time series models is often readily justifiable, as many of our example usages involve time-dependent data. This extension would be very interesting for the regression model in Chapter 8, where modelling the residuals as autoregressive is advantageous in an CP analysis (Cunen, Hermansen & Hjort, 2018). Another interesting case arises if the observations stem from a hidden Markov model. Carlin, Gelfand and Smith (1992) briefly considers this as an example of the generality of their estimation procedure.

The problem of finding matching priors, or approximately matching priors, could also be investigated further. In Chapter 3, we deduce the matching prior for a univariate normal model with a parameter transition. This could be extended to the multidimensional case and applied to *Tirant lo Blanch* in Chapter 9. Since most of the degrees of freedom in the multivariate normal model are allocated to the covariance matrix, an approximation of its posterior distribution should be considered in particular.

Only single change points or, alternatively, single windows have been considered. The methods should therefore be tested in the cases where we do not know the number of changes, or if there are any changes at all.

Finally, we have assumed throughout this thesis that observations occur at a steady pace, i.e. that the intervals between i are fixed. In the case of Bjørnholt, we had some missing values but still assumed that the sampling rate was constant. If we instead assume that the observations come with random time intervals (this might reflect some data sets more appropriately), the model becomes more complicated.

References

- Akaike, H. (1974). A new look at the statistical model identification [Conference Name: IEEE Transactions on Automatic Control]. *IEEE Transactions on Automatic Control*, 19(6), 716–723.
- Anscombe, F. J. (1973). Graphs in statistical analysis [Publisher: [American Statistical Association, Taylor & Francis, Ltd.]]. *The American Statistician*, 27(1), 17–21.
- Berger, J. O., Bernardo, J. M., & Sun, D. (2012). Objective priors for discrete parameter spaces [Publisher: ASA Website _eprint: <https://doi.org/10.1080/01621459.2012.682538>]. *Journal of the American Statistical Association*, 107(498), 636–648.
- Carlin, B. P., Gelfand, A. E., & Smith, A. F. M. (1992). Hierarchical bayesian analysis of changepoint problems [Publisher: [Wiley, Royal Statistical Society]]. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2), 389–405.
- Cobb, G. W. (1978). The problem of the Nile: Conditional solution to a changepoint problem [Publisher: [Oxford University Press, Biometrika Trust]]. *Biometrika*, 65(2), 243–251.
- Cunen, C., Hermansen, G., & Hjort, N. L. (2018). Confidence distributions for change-points and regime shifts. *Journal of Statistical Planning and Inference*, 195, 14–34.
- Efron, B. (2022). *Exponential families in theory and practice*. Cambridge University Press.
- Fisher, R. A., & Russell, E. J. (1997). On the mathematical foundations of theoretical statistics [Publisher: Royal Society]. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594), 309–368.
- Gelman, A. (2013). *Bayesian data analysis, third edition, 3rd edition* (3rd edition.). CRC Press.
- Givens, G. H., & Hoeting, J. A. (2013). *Computational statistics* (2nd ed). Wiley.
- Hagen, E. N. (2020, January 9). *Tests of homogeneity and change-point inference* [Master Thesis]. University of Oslo. <http://urn.nb.no/URN:NBN:no-87124>
- Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables [Publisher: [Oxford University Press, Biometrika Trust]]. *Biometrika*, 57(1), 1–17.
- Jarrett, R. G. (1979). A note on the intervals between coal-mining disasters [Publisher: [Oxford University Press, Biometrika Trust]]. *Biometrika*, 66(1), 191–193.
- Neyman, J., Pearson, E. S., & Pearson, K. (1997). IX. on the problem of the most efficient tests of statistical hypotheses [Publisher: Royal Society]. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694), 289–337.

References

- Raftery, A. E., & Akman, V. E. (1986). Bayesian analysis of a poisson process with a change-point [Publisher: [Oxford University Press, Biometrika Trust]]. *Biometrika*, 73(1), 85–89.
- Schwarz, G. (1978). Estimating the dimension of a model [Publisher: Institute of Mathematical Statistics]. *The Annals of Statistics*, 6(2), 461–464.
- Taylor, A. J. (1968). The coal industry. In D. H. Aldcroft (Ed.), *The development of british industry and foreign competition 1875–1914* (pp. 37–70). University of Toronto Press. Retrieved January 27, 2025, from <https://www.jstor.org/stable/10.3138/j.ctvcj2kg9.5>

Appendix A

Sample code

A.1 Gibbs sampler for mixture of normal observations

```
import numpy as np
from numpy.typing import NDArray
import matplotlib.pyplot as plt
from numba import njit

@njit(cache=True)
def weight(a: int, b: int, n: int) -> NDArray[np.float64]:
    """
    Generate a linear transition from 1 to 0 between indices 'a' and 'b'
    over 'n' elements.

    This function returns an array of length 'n' where each value is:
    - 1 for positions 'i' such that 'i <= a - 1'
    - 0 for positions 'i' such that 'i > b - 1'
    - Linearly decreasing from 1 to 0 for positions between 'a' and 'b
    - 1'

    Parameters
    -----
    a : int
        Index at which the values begin to decrease from 1.
    b : int
        Index at which the values reach 0.
    n : int
        Total number of values in the output array.

    Returns
    -----
    res : ndarray of shape (n,)
        Array of float weights transitioning from 1 to 0 between indices
        'a' and 'b'.
    """
    res = []
    assert int(a) == a
    for i in range(1, n + 1):
        if i <= a - 1:
            val = 1
        elif b - 1 < i:
            val = 0
        else:
            val = (b - i) / (b - a)
        res.append(val)
```

Appendix A. Sample code

```
    return np.array(res, dtype=np.float64)

@njit
def sample_ab(z, a_c, b_c, r=10):
    """
    Compute the posterior distribution over (a, b) gating parameters
    given class assignments.

    This function evaluates the log-likelihood of observed class labels '
    z' for all
    valid pairs of (a, b) parameters defining the gating function. It
    returns a
    probability distribution over these parameter pairs based on their
    likelihoods.

    Parameters
    -----
    z : ndarray
        Binary class assignments (0 or 1) for each data point, indicating
        membership in one of two components.
    a_c : int
        Current sample of a, used to narrow sample space
    b_c : int
        Current sample of b, used to narrow sample space
    r : int
        Radius of the search space

    Returns
    -----
    probs : ndarray
        Normalized posterior probabilities for each (a, b) parameter pair
    .
    params : list of [int, int]
        Corresponding list of (a, b) parameter pairs considered.
    """
    liks = []
    params = []
    n = len(z)
    for a in range(max(1, a_c-r), min(a_c+r, n-1)):
        for b in range(max(a+1, b_c-r), min(b_c + r, n)):
            params.append([a,b])
            w = weight(a,b,n)
            l = np.log(w[z==0]).sum() + np.log((1-w)[z==1]).sum()
            liks.append(l)

    liks = np.array(liks)
    # shifting log likelihood decreases floating point error
    liks -= liks.max()
    probs = np.exp(liks)
    probs /= probs.sum()
    return probs, params

@njit
def sample_z(y, a, b, theta_L, theta_R):
    """
    Compute the posterior probability that each data point belongs to the
    left distribution.

    This function calculates the probability of each data point in 'y'
    belonging
```

A.1. Gibbs sampler for mixture of normal observations

```

to the "left" distribution (with mean 'theta_L') of a two-component
Gaussian mixture model.

Parameters
-----
y : ndarray
    Observed data points (1D array).
a : float
    Start of the window
b : float
    End of the window
theta_L : float
    Mean of the left Gaussian distribution.
theta_R : float
    Mean of the right Gaussian distribution.

Returns
-----
p_class_L : ndarray
    Posterior probability of each data point belonging to the left
distribution.
"""

w = weight(a,b,len(y))
liks_L = -0.5*(y-theta_L)**2
liks_R = -0.5*(y-theta_R)**2
# shifting log likelihood decreases floating point error
shift = max(liks_L.max(), liks_R.max())
p_class_L = np.exp(liks_L-shift)*w
p_class_R = np.exp(liks_R-shift)*(1-w)
p_class_L/=(p_class_L + p_class_R)

return p_class_L

def gibbs_sampler_normal(y, n_samples, theta_0, rng, tau2=100, mu=0):
    """
    Perform Gibbs sampling for a mixture of two univariate normal
distributions.

    This function uses Gibbs sampling to draw samples from the posterior
distribution
of the parameters in a mixture model of two normal distributions.
Each data point
is assumed to be generated from one of two components, with class
assignments
inferred via latent variables 'z'.

Parameters
-----
y : array-like
    Observed data, assumed to be 1-dimensional.
n_samples : int
    Number of Gibbs samples to draw.
theta_0 : array-like of shape (4,)
    Initial values for the parameters [a, b, theta_L, theta_R], where
:
    - a, b window parameters
    - theta_L and theta_R are means of the two normal components.
rng : numpy.random.Generator
    NumPy random number generator for reproducibility.
tau2 : float, optional

```

Appendix A. Sample code

```
Prior variance for the component means (default is 100).
mu : float, optional
    Prior mean for the component means (default is 0).

Returns
-----
theta : ndarray of shape (n_samples, 4)
    Samples of the model parameters [a, b, theta_L, theta_R] at each
iteration.
zs : ndarray of shape (n_samples, len(y))
    Latent class sample for each data point at each iteration.
"""

theta = np.zeros([n_samples, len(theta_0)])
theta[0] = theta_0
n = len(y)
a, b, theta_L, theta_R = theta_0
w = weight(a,b,n)
z = (w<rng.uniform(size=n)).astype(int)
zs = np.zeros([n_samples, len(z)], dtype=int)
zs[0] = z

for i in range(1, n_samples):
    # Sample theta_L and theta_R
    y_L = y[z==0]
    y_R = y[z==1]
    tau2_L = 1/(1/tau2 + len(y_L)/1)
    mu_L = tau2_L*(mu/tau2 + y_L.sum()/1)
    theta_L = rng.normal(loc = mu_L, scale=np.sqrt(tau2_L))

    tau2_R = 1/(1/tau2 + len(y_R)/1)
    mu_R = tau2_R*(mu/tau2 + y_R.sum()/1)
    theta_R = rng.normal(loc = mu_R, scale=np.sqrt(tau2_R))

    # sample a, b
    probs, params = sample_ab(z, a, b, r=20)
    index = np.random.choice(len(params), p=probs)
    a, b = params[index]

    # sample z
    p_class_L = sample_z(y,a,b,theta_L, theta_R)
    z = (np.random.uniform(size=n)>p_class_L).astype(int)

    # Storing samples
    zs[i] = z
    theta[i] = [a,b, theta_L, theta_R]
return theta, zs
```

The following are some notes on the implementation. First, the sampler uses the preferred `numpy.random.Generator` abstraction for random number generator (RNG). This allows more flexible RNG algorithm choices, but requires a reference to the object through the algorithm. Second, this algorithm benefits very much from just-in-time (JIT) compilation, but the compiler of choice, `numba`, is incompatible with the `numpy.random.Generator` class. Therefore, we must JIT the computation intensive operations, like calculating the likelihoods for all a, b in a simple function, and returning the probabilities as well as the corresponding parameters such that sampling can be done from outside the compiled function.

Appendix B

Data sets

B.1 British mining disaster

year	accidents
1851	4
1852	5
1853	4
1854	1
1855	0
1856	4
1857	3
1858	4
1859	0
1860	6
1861	3
1862	3
1863	4
1864	0
1865	2
1866	6
1867	3
1868	3
1869	5
1870	4
1871	5
1872	3
1873	1
1874	4
1875	4
1876	1
1877	5
1878	5
1879	3
1880	4

Continued on next page

Appendix B. Data sets

year	accidents
1881	2
1882	5
1883	2
1884	2
1885	3
1886	4
1887	2
1888	1
1889	3
1890	2
1891	2
1892	1
1893	1
1894	1
1895	1
1896	3
1897	0
1898	0
1899	1
1900	0
1901	1
1902	1
1903	0
1904	0
1905	3
1906	1
1907	0
1908	3
1909	2
1910	2
1911	0
1912	1
1913	1
1914	1
1915	0
1916	1
1917	0
1918	1
1919	0
1920	0
1921	0
1922	2
1923	1
1924	0
1925	0
1926	0
Continued on next page	

year	accidents
1927	1
1928	1
1929	0
1930	2
1931	3
1932	3
1933	1
1934	1
1935	2
1936	1
1937	1
1938	1
1939	1
1940	2
1941	4
1942	2
1943	0
1944	0
1945	0
1946	1
1947	4
1948	0
1949	0
1950	0
1951	1
1952	0
1953	0
1954	0
1955	0
1956	0
1957	1
1958	0
1959	0
1960	1
1961	0
1962	1

B.2 Skiing days at Bjørnholt

year	days
1897	147
1898	169
1899	135
Continued on next page	

Appendix B. Data sets

year	days
1900	135
1901	137
1902	131
1903	147
1904	185
1905	127
1906	116
1907	165
1908	170
1909	176
1910	182
1911	149
1912	174
1913	125
1914	128
1915	190
1916	155
1917	139
1918	115
1919	175
1920	141
1921	110
1922	129
1923	164
1924	135
1925	91
1926	130
1927	163
1928	127
1929	75
1930	122
1931	140
1932	65
1933	55
1934	110
1935	117
1936	151
1937	108
1955	135
1956	111
1957	82
1958	125
1959	143
1960	141
1961	141
1962	152
Continued on next page	

B.2. Skiing days at Bjørnholt

year	days
1963	150
1964	92
1965	150
1966	162
1967	142
1968	125
1969	153
1970	191
1971	178
1972	120
1973	114
1974	146
1975	129
1976	150
1977	165
1978	152
1979	182
1980	156
1981	157
1982	148
1983	128
1984	121
1985	138
1986	139
1987	127
1988	141
1989	5
1990	27
1991	97
1992	39
1993	70
1994	119
1995	123
1996	116
1997	69
1998	39
1999	121
2000	57
2001	150
2002	129
2003	113
2004	108
2005	37
2006	120
2007	58
2008	131
Continued on next page	

year	days
2009	101
2010	144
2011	117
2012	95
2013	108
2014	81
2015	97
2016	84
2017	55
2018	115
2019	125
2020	34
2021	110
2022	96

B.3 Tirant lo Blanch

Note that the 'chapter' index is not the original chapters, but the index once the chapters with fewer than 200 words have been removed

chapter	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	m
1	21	59	44	19	33	20	16	17	9	17	255
2	53	113	80	49	52	33	28	36	16	16	476
3	109	274	239	128	112	110	76	51	43	32	1174
4	69	150	126	71	60	71	47	32	23	21	670
5	119	207	231	123	128	102	61	55	29	34	1089
6	69	136	126	69	60	61	37	27	15	15	615
7	32	63	51	18	29	28	15	15	19	13	283
8	26	52	41	19	27	29	11	16	5	11	237
9	23	42	48	16	15	28	12	15	14	10	223
10	92	191	190	93	84	72	47	47	27	24	867
11	60	132	144	62	55	50	26	27	9	14	579
12	43	66	94	66	44	29	27	13	9	13	404
14	125	217	245	92	101	131	59	54	28	33	1085
15	41	118	129	38	40	44	27	25	16	7	485
16	27	64	59	42	16	28	16	17	14	6	289
17	68	151	135	73	57	82	33	46	20	15	680
18	57	134	119	48	65	74	25	25	10	9	566
19	130	302	328	150	128	132	75	59	26	37	1367
20	53	117	98	70	51	60	28	26	14	22	539
21	56	142	136	67	45	64	25	28	14	13	590
22	69	179	137	85	58	66	32	47	20	26	719
23	85	163	217	99	108	79	53	27	37	24	892
24	66	139	171	77	78	59	35	33	13	11	682
25	116	174	205	128	134	98	68	29	22	13	987

Continued on next page

chapter	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	m
26	158	316	274	136	157	150	80	61	37	32	1401
27	222	442	452	188	221	218	94	89	49	36	2011
28	52	72	78	48	43	54	18	15	21	20	421
29	47	128	111	49	50	47	26	19	13	17	507
30	33	54	48	30	16	26	6	16	5	10	244
31	38	67	60	27	31	45	13	10	4	9	304
32	84	120	152	77	85	66	42	42	18	18	704
33	76	133	162	75	80	66	31	50	17	19	709
34	72	143	134	76	56	43	23	35	16	27	625
35	73	184	174	79	96	79	37	40	23	23	808
36	43	92	111	47	54	34	28	31	11	17	468
37	38	53	55	22	35	37	15	15	18	16	304
38	79	154	150	66	76	77	33	35	9	13	692
39	132	212	211	116	82	134	51	36	25	33	1032
40	24	51	55	28	22	33	14	10	9	6	252
41	83	194	208	100	112	100	42	27	22	34	922
42	75	115	106	75	91	94	57	22	14	14	663
43	28	62	72	36	25	23	17	11	9	13	296
44	90	160	125	86	78	82	51	43	24	16	755
52	111	128	139	73	107	62	34	39	30	22	745
53	41	91	83	52	48	44	13	17	9	8	406
54	27	65	59	22	29	37	12	24	12	10	297
55	115	220	195	135	128	105	49	59	23	32	1061
56	109	207	220	95	104	96	34	44	28	23	960
57	129	304	282	144	146	121	70	57	28	32	1313
58	41	61	74	27	42	31	15	13	8	14	326
59	181	297	282	157	162	140	66	76	48	39	1448
60	127	272	238	132	126	121	71	45	24	25	1181
61	23	45	60	20	29	27	12	7	6	4	233
63	44	77	79	25	39	46	11	9	9	4	343
64	115	235	221	80	117	90	46	33	34	26	997
65	164	320	335	121	181	144	67	54	45	31	1462
66	27	75	92	25	37	34	8	12	10	7	327
67	273	532	523	273	301	240	115	86	66	45	2454
68	467	798	781	420	452	380	191	156	98	86	3829
71	291	490	516	268	282	258	113	103	103	63	2487
72	128	192	191	111	111	120	34	37	14	20	958
73	151	226	224	103	138	134	72	58	26	13	1145
74	152	302	295	128	134	137	84	76	38	33	1379
75	34	87	96	33	38	36	16	17	3	6	366
76	87	196	162	94	66	81	36	45	20	18	805
78	79	143	151	80	94	68	22	31	22	19	709
80	134	235	263	131	141	114	47	50	27	36	1178
81	329	652	561	256	313	270	143	146	72	53	2795
82	84	130	127	74	59	66	30	38	12	8	628
83	29	63	57	26	23	32	15	17	6	10	278

Continued on next page

Appendix B. Data sets

chapter	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	m
84	133	222	179	105	133	124	59	64	23	28	1070
85	182	354	310	122	152	188	80	75	38	47	1548
92	160	210	161	108	115	91	46	37	20	29	977
96	40	71	72	26	36	31	16	17	8	10	327
97	79	105	102	60	49	76	33	24	18	11	557
98	246	633	529	321	241	262	157	115	73	52	2629
99	161	309	317	176	160	170	91	78	62	41	1565
100	383	768	779	354	387	374	169	168	73	76	3531
101	240	507	418	198	208	228	110	73	46	36	2064
102	29	65	59	30	28	29	18	10	3	7	278
103	39	116	108	50	48	42	21	11	13	8	456
104	204	466	415	197	190	217	118	73	77	41	1998
105	176	324	330	149	175	209	95	56	36	35	1585
106	240	438	438	227	260	205	112	65	32	21	2038
107	92	172	144	91	77	96	48	31	37	23	811
107	58	111	96	43	56	64	33	36	31	15	543
108	145	284	309	138	142	173	88	49	37	23	1388
109	272	559	533	245	297	262	156	85	72	62	2543
110	374	873	775	338	386	371	221	136	45	38	3557
111	108	226	208	89	103	119	58	33	30	22	996
112	45	70	74	32	26	47	19	14	8	5	340
113	156	306	302	138	170	150	74	57	36	30	1419
114	202	440	376	172	210	207	96	80	41	34	1858
116	154	264	261	122	124	124	70	64	49	35	1267
117	141	311	251	139	120	157	77	49	43	30	1318
118	73	147	123	76	83	65	33	29	18	12	659
119	422	813	766	373	417	440	237	196	130	86	3880
120	25	59	55	31	19	24	9	9	4	3	238
121	129	238	255	121	144	138	68	66	39	28	1226
122	23	55	50	28	27	38	15	19	13	18	286
123	78	157	146	66	78	82	29	37	20	17	710
124	213	384	390	195	175	198	108	97	77	32	1869
125	257	461	440	198	216	246	106	112	63	52	2151
126	136	258	272	122	111	138	54	49	35	27	1202
127	231	477	380	209	216	212	117	119	51	32	2044
128	68	102	105	53	64	40	30	21	12	15	510
129	184	397	395	160	182	197	82	106	50	51	1804
130	46	112	95	48	44	58	30	23	16	15	487
131	146	211	273	142	127	134	59	44	41	31	1208
132	215	506	392	184	235	195	112	104	57	29	2029
133	465	872	970	587	517	471	242	189	130	93	4536
134	238	382	387	204	182	219	114	83	69	46	1924
136	26	59	45	28	18	30	13	10	12	5	246
137	159	297	352	152	173	139	84	64	46	33	1499
138	261	563	482	271	252	297	133	142	94	44	2539
139	49	65	75	62	51	59	22	23	9	6	421

Continued on next page

chapter	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	m
140	153	259	293	178	185	153	64	66	37	34	1422
141	422	792	808	404	407	393	208	148	140	92	3814
142	73	164	183	111	81	101	35	31	33	9	821
143	475	897	789	394	466	427	302	219	131	106	4206
144	54	84	98	59	52	42	19	24	22	19	473
145	245	444	449	214	229	214	104	91	96	57	2143
146	366	800	705	341	352	318	152	175	136	80	3425
148	146	257	290	128	129	157	70	57	39	21	1294
149	199	428	335	209	165	202	90	63	44	18	1753
151	78	224	163	76	83	83	44	46	22	10	829
153	150	309	329	141	146	163	84	55	39	30	1446
154	398	1033	932	414	414	466	228	170	165	75	4295
155	236	461	450	255	178	247	110	74	75	43	2129
156	66	131	126	57	73	75	46	37	32	12	655
157	406	782	753	383	403	455	232	153	109	66	3742
159	199	499	459	205	218	263	110	84	67	27	2131
161	502	1125	971	453	477	475	239	213	153	69	4677
162	178	358	307	186	161	155	96	64	53	39	1597
163	329	746	667	344	345	372	156	119	99	69	3246
164	319	634	665	380	387	351	180	116	60	59	3151
165	22	44	41	21	25	38	16	25	15	7	254
166	179	380	342	189	167	165	68	81	50	32	1653
167	125	220	200	99	108	118	51	54	34	31	1040
169	59	124	108	49	47	63	23	32	16	15	536
170	23	61	61	32	29	24	7	11	9	8	265
171	37	92	71	41	32	29	22	15	7	10	356
172	115	271	299	124	116	129	63	51	43	19	1230
173	109	196	204	90	107	110	41	40	23	36	956
174	29	86	52	22	39	39	17	14	5	5	308
175	39	74	79	40	32	29	9	10	9	6	327
176	55	115	103	54	45	53	24	17	13	15	494
177	58	121	114	52	36	60	25	25	16	21	528
178	109	233	190	115	94	104	42	41	29	21	978
179	57	156	121	59	63	36	33	27	13	14	579
180	61	148	139	68	68	52	34	39	14	14	637
181	68	80	148	58	61	53	41	41	18	12	580
182	81	204	205	86	75	83	48	66	23	21	892
183	76	170	182	64	61	90	41	39	20	22	765
184	30	100	105	30	29	26	21	18	12	9	380
185	47	69	50	21	26	32	31	16	25	9	326
189	754	1462	1246	642	739	717	315	307	200	139	6521
190	36	71	63	35	47	34	35	14	20	11	366
191	39	67	53	37	29	28	18	11	15	4	301
192	37	99	111	63	47	53	18	27	19	12	486
194	29	61	50	31	26	23	19	21	9	6	275
195	43	89	93	33	44	49	17	7	16	18	409

Continued on next page

Appendix B. Data sets

chapter	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	m
201	69	129	103	60	71	52	42	27	18	11	582
202	62	138	119	61	77	54	28	21	19	15	594
207	30	55	68	26	31	25	13	11	9	7	275
208	54	170	143	64	72	63	30	34	18	22	670
209	53	162	134	61	65	67	37	34	12	13	638
210	78	239	191	90	102	106	59	56	31	31	983
211	39	111	82	37	42	42	23	20	26	10	432
212	29	67	59	34	30	19	20	18	12	10	298
214	78	198	178	63	72	85	34	34	20	19	781
215	92	216	220	80	95	102	39	33	29	20	926
216	80	211	208	110	99	75	38	50	23	13	907
217	24	107	82	38	40	44	26	24	6	13	404
218	172	370	296	178	144	149	84	73	60	41	1567
219	30	68	66	29	30	33	18	18	19	15	326
220	139	249	217	93	98	97	54	28	40	38	1053
221	110	238	187	107	109	112	69	66	50	47	1095
222	229	346	371	196	223	158	80	80	80	46	1809
223	28	71	73	24	34	44	16	13	9	6	318
224	97	205	181	88	74	111	40	47	30	18	891
225	72	223	194	95	78	85	31	31	10	19	838
226	86	239	229	105	94	106	43	59	15	20	996
227	71	176	168	81	90	86	30	46	21	13	782
228	123	342	310	125	122	132	71	71	32	17	1345
229	147	307	264	135	145	154	62	67	30	24	1335
230	66	208	143	49	77	55	23	33	16	15	685
231	125	351	293	135	124	158	45	65	35	31	1362
232	48	84	100	36	39	53	21	18	12	6	417
233	248	495	429	195	247	201	103	86	38	37	2079
234	192	447	393	175	205	181	109	86	39	36	1863
235	19	57	54	20	22	25	11	15	4	9	236
236	209	493	446	206	209	216	83	83	54	34	2033
237	30	88	86	42	39	43	22	16	15	18	399
238	109	225	197	105	78	112	39	41	31	15	952
239	144	305	301	149	128	167	66	47	46	30	1383
240	33	132	122	31	57	50	23	30	8	12	498
241	42	110	70	33	51	46	25	17	9	7	410
242	24	106	76	38	38	32	14	24	5	10	367
244	47	136	129	57	57	54	39	21	15	16	571
245	16	60	39	14	16	25	17	15	1	8	211
248	43	103	70	42	46	42	23	16	12	15	412
249	54	146	120	50	63	56	29	21	12	20	571
250	35	116	57	32	42	57	21	25	14	12	411
251	56	148	145	59	63	54	46	29	11	15	626
252	33	101	91	44	32	68	23	26	9	8	435
253	41	87	81	50	39	58	14	15	7	7	399
254	100	157	159	82	78	75	42	53	20	20	786

Continued on next page

chapter	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	m
255	33	109	101	29	43	49	28	18	12	9	431
256	48	80	79	44	37	26	23	20	11	6	374
257	38	88	63	47	50	51	15	17	8	14	391
258	129	263	263	113	142	123	53	45	35	36	1202
259	24	68	54	42	23	25	16	10	5	10	277
260	139	303	238	130	131	121	76	59	29	36	1262
262	352	844	701	328	361	353	205	128	79	75	3426
263	141	286	256	141	129	126	61	42	15	29	1226
264	211	418	375	196	176	196	92	70	38	53	1825
265	79	153	132	68	62	67	51	21	15	14	662
266	41	133	119	42	44	45	27	17	27	12	507
267	21	67	63	23	34	25	18	19	12	3	285
268	93	270	226	97	118	108	69	35	24	24	1064
269	179	413	369	179	185	168	82	83	41	40	1739
270	18	67	55	17	26	20	23	21	3	6	256
271	58	131	124	65	56	35	24	43	13	10	559
272	37	69	57	23	39	28	20	24	11	8	316
273	30	81	82	35	42	30	23	16	12	12	363
274	20	78	58	29	30	39	19	22	4	6	305
275	74	118	117	58	54	61	39	31	21	14	587
276	33	107	81	62	47	45	21	25	14	12	447
277	105	202	164	108	91	91	53	48	23	20	905
278	57	105	76	38	53	62	36	32	11	17	487
279	65	128	106	47	52	43	22	17	12	17	509
280	42	110	86	39	41	39	22	17	16	12	424
281	165	305	272	161	141	157	62	73	34	41	1411
282	117	252	236	102	122	122	56	62	26	25	1120
283	187	387	351	169	197	165	84	67	26	55	1688
284	25	62	47	29	35	35	23	20	8	18	302
285	31	99	66	39	35	26	20	13	4	18	351
286	141	264	284	148	162	122	72	51	38	26	1308
288	177	357	339	168	168	164	90	70	44	36	1613
289	23	43	45	24	25	16	11	9	16	3	215
290	83	222	183	112	87	82	43	27	35	23	897
291	220	472	457	220	198	214	100	94	76	43	2094
292	205	399	384	200	175	214	84	81	50	38	1830
293	91	153	143	77	67	84	38	40	26	20	739
294	32	99	72	43	44	37	32	30	13	9	411
295	74	242	201	112	122	77	75	45	41	35	1024
296	82	169	153	76	78	65	51	42	13	24	753
297	34	109	107	41	44	47	25	17	11	15	450
299	232	583	520	300	288	233	136	93	45	56	2486
300	24	89	85	43	31	32	22	12	6	7	351
301	173	374	367	180	165	159	80	91	44	30	1663
302	48	95	74	35	50	32	35	16	12	12	409
303	34	114	112	36	39	32	40	16	9	7	439

Continued on next page

Appendix B. Data sets

chapter	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	m
304	207	361	373	205	207	224	94	76	43	36	1826
305	17	57	56	29	22	14	16	10	6	11	238
306	29	51	46	9	27	37	22	18	11	5	255
307	133	293	282	140	127	133	78	43	38	22	1289
308	29	119	93	40	45	28	19	19	14	9	415
309	235	499	495	235	230	229	131	106	63	55	2278
310	265	616	569	294	286	231	158	111	59	51	2640
311	28	84	67	40	26	29	20	10	10	7	321
312	105	233	203	90	85	126	38	52	30	13	975
313	59	102	107	60	39	51	24	15	9	12	478
314	32	74	50	26	31	28	16	12	8	7	284
315	254	419	422	240	226	204	112	65	48	51	2041
316	24	91	90	39	43	24	18	23	11	9	372
317	82	227	211	98	90	79	42	51	24	28	932
318	42	86	91	45	28	38	19	23	10	8	390
319	145	365	380	160	143	160	88	68	52	27	1588
320	22	49	52	21	16	22	10	14	3	8	217
321	146	315	340	135	159	144	81	60	41	33	1454
322	40	107	104	41	45	41	30	18	10	12	448
323	46	136	133	47	60	55	35	31	11	15	569
324	57	196	135	88	60	59	39	34	17	20	705
325	48	146	106	57	51	58	37	25	23	19	570
326	24	67	59	29	34	19	12	23	14	8	289
327	51	156	138	59	56	51	31	30	26	17	615
328	94	128	188	66	82	76	53	44	25	19	775
329	67	139	130	53	66	56	36	35	20	11	613
330	121	228	232	110	110	106	72	56	37	20	1092
331	46	118	87	61	67	38	23	19	9	10	478
332	17	72	68	16	34	37	13	11	8	3	279
333	61	153	155	78	79	66	30	32	26	13	693
334	182	320	335	187	180	154	98	76	54	33	1619
335	77	135	158	90	82	64	35	27	24	9	701
336	12	63	51	17	21	21	5	12	3	11	216
337	111	220	228	106	117	119	46	28	30	28	1033
338	24	54	59	22	24	28	22	18	8	10	269
339	100	207	225	99	94	111	71	32	15	16	970
340	241	353	476	235	237	225	113	85	79	39	2083
341	22	61	61	25	17	29	10	14	10	3	252
343	147	241	308	142	157	140	81	58	35	29	1338
344	109	203	185	108	107	94	61	43	28	16	954
345	84	163	186	86	72	80	38	29	28	19	785
346	48	107	107	51	33	41	29	34	9	20	479
347	65	145	167	76	74	62	41	30	26	24	710
348	26	57	36	22	28	18	20	15	8	5	235
349	234	446	446	246	253	254	126	104	84	76	2269
350	121	232	240	104	122	120	54	49	38	37	1117

Continued on next page

chapter	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	m
351	42	96	96	30	49	33	27	33	10	12	428
352	23	43	40	22	27	24	17	8	6	7	217
353	71	156	156	86	60	58	42	55	31	23	738
354	77	198	188	92	73	71	39	46	44	24	852
355	145	361	313	154	140	163	102	81	47	50	1556
356	29	106	98	35	41	40	26	31	12	10	428
357	87	231	217	102	96	105	50	42	31	41	1002
358	21	67	58	38	19	26	18	16	9	7	279
359	38	67	69	27	32	27	21	19	5	6	311
360	37	91	89	45	34	45	26	24	9	16	416
362	44	87	100	43	35	58	24	23	15	17	446
363	21	53	46	40	31	20	11	6	10	6	244
364	24	84	80	34	38	30	11	26	10	12	349
365	18	69	54	27	29	22	12	17	5	8	261
366	68	134	105	69	52	66	31	20	14	20	579
367	68	134	105	69	52	66	31	20	14	20	579
368	37	82	71	42	23	27	19	11	6	13	331
369	35	119	88	42	44	51	36	25	8	15	463
370	53	116	92	52	48	62	31	22	16	13	505
371	71	197	156	78	80	59	26	45	17	21	750
372	91	184	172	86	93	96	49	58	27	36	892
373	42	79	80	40	46	38	19	18	14	12	388
374	147	296	241	120	138	130	142	78	73	76	1441
375	42	77	53	37	29	31	24	26	16	22	357
376	86	140	108	74	64	77	57	48	52	46	752
377	34	74	45	21	33	23	27	13	16	17	303
378	46	118	92	71	48	34	45	30	18	21	523
379	12	51	41	15	34	26	15	18	3	15	230
380	31	97	66	44	51	38	19	27	20	21	414
381	27	66	43	24	21	30	14	22	12	13	272
382	41	72	61	30	32	44	12	20	8	10	330
383	42	60	52	30	30	30	16	17	16	16	309
384	57	101	110	58	43	50	32	24	19	18	512
385	30	53	58	37	16	27	21	16	13	16	287
386	64	117	113	55	53	79	39	34	20	14	588
387	163	321	350	170	180	165	103	80	63	44	1639
388	56	76	53	39	32	29	19	17	16	20	357
390	43	99	109	40	38	52	15	29	17	12	454
391	54	71	60	27	30	30	13	22	9	17	333
392	42	51	59	25	24	39	9	10	14	14	287
393	55	70	58	19	28	36	10	18	14	19	327
394	106	190	198	121	117	131	71	50	43	32	1059
395	33	46	43	38	26	29	14	9	9	10	257
397	41	76	76	30	36	47	9	17	10	14	356
399	26	42	42	15	29	21	6	5	8	9	203
400	32	88	57	27	30	38	28	24	17	21	362

Continued on next page

Appendix B. Data sets

chapter	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	m
401	109	175	202	95	96	92	43	28	33	34	907
402	29	62	46	31	40	36	26	27	17	13	327
403	102	258	225	117	101	98	69	54	45	39	1108
404	63	91	92	55	58	49	17	16	19	11	471
405	33	65	53	33	19	26	11	4	4	13	261
406	71	128	111	58	55	51	30	28	20	21	573
407	36	79	72	52	37	38	15	16	8	17	370
408	151	291	252	133	139	125	59	40	32	32	1254
409	137	241	281	134	127	114	53	47	52	37	1223
410	229	382	373	219	194	150	86	73	34	37	1777
411	29	69	81	39	38	38	26	16	14	9	359
412	29	68	45	28	28	27	15	17	7	7	271
413	49	68	49	34	54	39	13	16	15	20	357
414	137	256	243	133	106	135	63	39	42	34	1188
415	117	206	165	75	107	101	40	45	31	35	922
416	85	164	134	67	94	58	39	31	19	23	714
417	29	42	26	20	24	26	26	17	10	13	233
418	147	241	276	160	171	147	48	48	34	34	1306
420	141	192	204	78	89	115	36	44	39	22	960
422	50	111	129	52	53	66	34	22	22	24	563
423	62	76	93	51	48	60	22	18	8	20	458
424	85	180	177	64	96	103	29	33	19	23	809
425	43	88	101	69	69	59	26	26	20	31	532
426	39	55	47	35	39	31	18	23	23	24	334
427	57	130	127	74	73	63	33	23	31	30	641
428	31	71	91	22	33	43	25	15	17	9	357
429	29	61	77	25	27	24	18	24	8	14	307
430	129	205	152	90	108	77	67	53	40	30	951
431	76	166	159	76	68	72	44	52	33	35	781
432	92	196	202	69	86	110	57	53	17	31	913
433	27	63	54	33	34	31	14	16	8	0	280
434	120	291	235	102	117	113	62	54	32	28	1154
435	25	101	70	35	40	45	18	16	14	10	374
436	30	66	53	26	25	48	24	23	11	6	312
437	17	61	55	20	28	31	16	18	9	13	268
438	47	97	92	42	48	52	30	25	7	21	461
439	43	108	58	34	37	52	21	28	11	12	404
440	54	74	60	29	32	40	23	12	16	5	345
442	86	163	155	61	75	89	41	44	28	25	767
443	26	47	21	9	20	36	17	20	13	18	227
444	80	114	133	60	79	61	46	36	24	26	659
445	96	225	168	75	75	112	53	58	37	23	922
446	115	233	240	142	117	121	68	52	47	57	1192
447	58	74	64	45	33	43	19	21	17	18	392
448	147	213	226	129	127	133	72	62	49	28	1186
449	56	100	112	48	46	42	31	22	27	13	497

Continued on next page

chapter	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	m
450	178	276	208	119	159	135	96	68	52	39	1330
451	57	126	89	46	41	77	41	45	18	16	556
452	176	254	239	107	121	170	85	67	70	54	1343
454	37	77	53	22	35	22	27	21	12	16	322
456	169	303	275	155	153	163	95	68	55	51	1487
457	31	45	45	16	27	28	13	12	5	9	231
458	34	78	55	23	39	47	21	11	12	10	330
459	135	302	220	138	158	146	109	68	57	29	1362
460	30	72	66	27	36	55	27	16	16	15	360
462	28	72	68	39	40	32	24	11	15	14	343
463	211	335	219	148	189	149	120	91	63	54	1579
464	46	59	53	30	24	36	18	17	21	17	321
465	58	95	71	40	45	66	32	23	8	15	453
466	73	111	93	40	68	56	30	33	24	23	551
467	56	128	114	49	48	73	36	22	28	28	582
468	45	66	62	51	40	51	26	32	19	19	411
471	96	196	183	80	93	96	42	36	33	26	881
472	100	182	171	92	99	103	52	49	41	24	913
473	53	126	106	72	75	68	29	29	30	15	603
474	99	260	252	126	157	118	84	37	50	41	1224
475	34	98	76	36	38	39	19	23	15	6	384
476	128	198	199	93	89	79	53	52	31	29	951
477	140	233	179	130	105	88	42	38	33	24	1012
478	165	298	283	131	127	152	62	72	43	38	1371
479	107	183	175	77	89	84	51	35	34	34	869
480	78	123	150	57	54	65	42	25	34	13	641
481	159	282	262	137	124	122	63	71	56	46	1322
482	50	47	61	18	32	47	23	32	14	11	335
483	158	220	207	80	120	93	65	54	62	50	1109
484	59	67	68	37	26	32	15	14	17	6	341
485	96	174	106	57	77	86	42	54	24	25	741
486	45	88	91	46	40	28	13	30	11	10	402
487	48	49	62	53	41	36	21	9	16	13	348