

# Suboptimal LULU-estimators in Measurements Containing Outliers

by

Stefan Ludwig Astl

*Thesis presented in partial fulfilment of the requirements for the degree of  
Master of Science at Stellenbosch University*



Department of Physics  
Faculty of Science

Supervisors:

Prof. Hans C. Eggers Dr. Carl H. Rohwer

December 2013

# Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

December 2013

Copyright © 2013 Stellenbosch University

All rights reserved.

# Abstract

## Suboptimal LULU-estimators in Measurements Containing Outliers

S.L. Astl

*Department of Physics  
Faculty of Science*

Thesis: MSc

December 2013

Techniques for estimating a signal in the presence of noise which contains outliers are currently not well developed. In this thesis, we consider a constant signal superimposed by a family of noise distributions structured as a tunable mixture  $f(x) = \alpha g(x) + (1 - \alpha)h(x)$  between finite-support components of “well-behaved” noise with small variance  $g(x)$  and of “impulsive” noise  $h(x)$  with a large amplitude and strongly asymmetric character. When  $\alpha \approx 1$ ,  $h(x)$  can for example model a cosmic ray striking an experimental detector. In the first part of our work, a method for obtaining the expected values of the positive and negative pulses in the first resolution level of a LULU Discrete Pulse Transform (DPT) is established. Subsequent analysis of sequences smoothed by the operators  $L_1U_1$  or  $U_1L_1$  of LULU-theory shows that a robust estimator for the location parameter for  $g$  is achieved in the sense that the contribution by  $h$  to the expected average of the smoothed sequences is suppressed to order  $(1 - \alpha)^2$  or higher. In cases where the specific shape of  $h$  can be difficult to guess due to the assumed lack of data, it is thus also shown to be of lesser importance. Furthermore, upon smoothing a sequence with  $L_1U_1$  or  $U_1L_1$ , estimators for the scale parameters of the model distribution become easily available. In the second part of our work, the same problem and data is approached from a Bayesian inference perspective. The Bayesian estimators are found to be optimal in the sense that they make full use of available information in the data. Heuristic comparison shows, however, that Bayes estimators do not always outperform the LULU estimators. Although the Bayesian perspective provides much insight into the logical connections inherent in the problem, its estimators can be difficult to obtain in analytic form and are slow to compute numerically. Suboptimal LULU-estimators are shown to be reasonable practical compromises in practical problems.

# Uittreksel

## Suboptimale LULU-afskatters in metings wat uitskieters bevat

*(“Suboptimal LULU-estimators in Measurements Containing Outliers”)*

S.L. Astl

*Departement Fisika  
Fakulteit Natuurwetenskappe*

Tesis: MSc

Desember 2013

Tegniese om 'n sein af te skat in die teenwoordigheid van geraas wat uitskieters bevat is tans nie goed ontwikkel nie. In hierdie tesis aanskou ons 'n konstante sein gesuperponeer met 'n familie van geraasverdelings wat as verstelbare mengsel  $f(x) = \alpha g(x) + (1 - \alpha)h(x)$  tussen eindige-uitkomsruimte geraaskomponente  $g(x)$  wat “goeie gedrag” en klein variansie toon, plus “impulsiewe” geraas  $h(x)$  met groot amplitude en sterk asimmetriese karakter. Wanneer  $\alpha \approx 1$  kan  $h(x)$  byvoorbeeld 'n kosmiese straal wat 'n eksperimentele apparaat tref modelleer. In die eerste gedeelte van ons werk word 'n metode om die verwagtingswaardes van die positiewe en negatiewe pulse in die eerste resolusievlak van 'n LULU Diskrete Pulse Transform (DPT) vasgestel. Die analise van rye verkry deur die inwerking van die gladstrykers  $L_1U_1$  en  $U_1L_1$  van die LULU-teorie toon dat hul verwagte gemiddelde waardes as afskatters van die liggingsparameter van  $g$  kan dien wat robuus is in die sin dat die bydrae van  $h$  tot die gemiddeld van orde grootte  $(1 - \alpha)^2$  of hoër is. Die spesifieke vorm van  $h$  word dan ook onbelangrik. Daar word verder gewys dat afskatters vir die relevante skaalparameters van die model maklik verkry kan word na gladstryking met die operatore  $L_1U_1$  of  $U_1L_1$ . In die tweede gedeelte van ons werk word dieselfde probleem en data vanuit 'n Bayesiese inferensie perspektief benader. Die Bayesiese afskatters word as optimaal bevind in die sin dat hulle vol gebruikmaak van die beskikbare inligting in die data. Heuristiese vergelyking wys egter dat Bayesiese afskatters nie altyd beter vaar as die LULU afskatters nie. Alhoewel die Bayesiese sienswyse baie insig in die logiese verbindings van die probleem gee, kan die afskatters moeilik wees om analities af te lei en stadig om numeries te bereken. Suboptimale LULU-beramers word voorgestel as redelike praktiese kompromieë in praktiese probleme.

## Acknowledgements

I would like to express my sincere gratitude to the following people and organisations: Hans Eggers and Carl Rohwer for your friendship, guidance and interesting conversations, Michiel de Kock for being a loyal Bayesian, Hannes Kriel for always making time to help, Frikkie Scholtz for the Group Theory lectures and guidance, Christine Ruperti for all your help. To my parents Ludwig and Adele Astl, and to Janie Swanepoel my dearest friend and life partner, I thank you for your love and belief. Finally I thank the National Research Foundation (NRF) and National Institute of Theoretical Physics (NITheP) for their financial support.

# Dedications

*Diese Arbeit ist meinem Vater Ludwig gewidmet.*

# Contents

<b>Declaration</b>	<b>i</b>
<b>Abstract</b>	<b>ii</b>
<b>Uittreksel</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>Dedications</b>	<b>v</b>
<b>Contents</b>	<b>vi</b>
<b>List of Figures</b>	<b>viii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 LULU Theory and the Discrete Pulse Transform</b>	<b>6</b>
2.1 Introduction . . . . .	6
2.2 LULU basics . . . . .	7
2.3 The highest resolution level of a DPT . . . . .	8
2.4 Theoretical expectation of pulses in $D_1$ and $R_1$ . . . . .	11
2.5 Designing an estimator . . . . .	26
2.6 Analysis of smoothed sequences in the context of outliers . . . . .	33
<b>3 Bayesian Inference</b>	<b>39</b>
3.1 Probability theory as extended logic . . . . .	39
3.2 Parameter estimation . . . . .	41
3.3 Choice of prior probabilities . . . . .	43
<b>4 Asymmetric Uniform Noise: Bayesian solution</b>	<b>46</b>
4.1 Formulating the problem . . . . .	46
4.2 Overview of calculations . . . . .	48
4.3 Moments of the posterior distribution: Exact solution . . . . .	49
4.4 Moments for simplified prior . . . . .	54
4.5 The $M_2$ limit . . . . .	58
4.6 The $n = 1$ case . . . . .	62
4.7 Posteriors for known $X$ . . . . .	67
4.8 Posteriors for known $L$ . . . . .	70
<b>5 Mixture Models: Bayesian Solution</b>	<b>73</b>

<i>CONTENTS</i>	<b>vii</b>
5.1 Mixture models . . . . .	73
5.2 Binomial expansion of likelihood within Bayes' theorem . . . . .	75
5.3 Case 4 . . . . .	81
<b>6 Comparison to LULU based solution</b>	<b>95</b>
6.1 Comparison to LULU based solution . . . . .	95
6.2 Discussion on end-values . . . . .	96
6.3 Results of Comparison . . . . .	99
<b>7 Application to Laser Spectroscopy Data</b>	<b>105</b>
<b>8 Summary and Conclusions</b>	<b>111</b>
<b>Appendices</b>	<b>113</b>
<b>A Window functions</b>	<b>114</b>
<b>B Preparing the window functions for integration</b>	<b>115</b>
<b>List of References</b>	<b>118</b>



# List of Figures

1.1	Histogram of data with outlier . . . . .	2
1.2	Raman Spectroscopy data showing outliers . . . . .	4
2.1	Example of LULU smoothing and the resolution levels of a DPT . . . . .	10
2.2	Drawing of $f(t - z)$ and $g(t)$ for $z < 0$ . . . . .	13
2.3	Triangular distribution . . . . .	17
2.4	The averages $\overline{x - L_1x}$ and $\overline{x - U_1x}$ for the triangular distribution . . . . .	19
2.5	The averages $\overline{U_1x - L_1U_1x}$ and $\overline{L_1x - U_1L_1x}$ for the triangular distribution . . . . .	20
2.6	Mixture model of two uniform distributions . . . . .	20
2.7	The averages $\overline{x - L_1x}$ and $\overline{x - U_1x}$ for the uniform distribution . . . . .	24
2.8	The averages $\overline{U_1x - L_1U_1x}$ and $\overline{L_1x - U_1L_1x}$ for the uniform distribution . . . . .	24
2.9	Expected values and long term averages of $x - L_1x$ and $x - U_1x$ . . . . .	25
2.10	Expected values and long term averages of $x - L_1x$ and $x - U_1x$ . . . . .	26
2.11	Scale parameter estimators for the triangular distribution . . . . .	28
2.12	Scale parameter estimators for the uniform distribution . . . . .	29
2.13	Scale parameter estimators for a mixture model of two uniform distributions with $\alpha = 0.5$ . . . . .	30
2.14	Scale parameter estimators for a mixture model of two uniform distributions with $\alpha = 0.8$ . . . . .	31
2.15	Scale parameter estimators for a mixture model of two uniform distributions with $\alpha = 0.99$ . . . . .	31
2.16	Scale parameter estimators for a mixture model of two uniform distributions with $\alpha = 0.999$ . . . . .	32
2.17	Mixture model with non-overlapping support . . . . .	33
2.18	Long term average of $U_1L_1x$ and $L_1U_1x$ for varying $\alpha$ . . . . .	37
2.19	Data sequences with one-sided impulsive noise smoothed by $U_1L_1$ . . . . .	38
4.1	Asymmetric uniform noise . . . . .	46
4.2	Plot of $\max(M_1, x_{max} - X)$ . . . . .	53
4.3	Ensemble averages of $\langle X \rangle$ and $\langle L \rangle$ for varying $\lambda$ . . . . .	59
	(a) Ensemble averages of $\langle X \rangle$ . . . . .	59
	(b) Ensemble averages of $\langle L \rangle$ . . . . .	59
4.4	Posterior distributions $P(X   DI)$ and $PL   DI$ for the Jeffreys prior . . . . .	61
	(a) $P(X   DI)$ . . . . .	61
	(b) $P(L   DI)$ . . . . .	61
4.5	$\frac{1}{n} \ln(P(X   DI))$ and $\frac{1}{n} \ln(P(L   DI))$ for the Jeffreys prior . . . . .	61
	(a) $\frac{1}{n} \ln(P(X   DI))$ . . . . .	61
	(b) $\frac{1}{n} \ln(P(L   DI))$ . . . . .	61
4.6	Mean estimators and standard deviations of $X$ and $L$ with Jeffreys prior ( $\lambda = 1$ ) . . . . .	62
	(a) $\langle X \rangle \pm \sigma$ . . . . .	62

(b)	$\langle L \rangle \pm \sigma$ . . . . .	62
4.7	$P(L   DXI)$ and $\frac{1}{n} \ln(P(L   DXI))$ with Jeffreys prior . . . . .	69
(a)	$P(L   DXI)$ . . . . .	69
(b)	$\frac{1}{n} \ln(P(L   DXI))$ . . . . .	69
4.8	Ensemble average of $\langle L \rangle$ for varying $\lambda$ . . . . .	69
(a)	Ensemble average of $\langle L \rangle$ . . . . .	69
(b)	$\langle L \rangle$ for one data set. . . . .	69
4.9	Mean estimator and standard deviation of $L$ with Jeffreys prior . . . . .	70
4.10	Posterior distribution and mean estimator for $X$ . . . . .	72
(a)	$P(X   DLI)$ . . . . .	72
(b)	$\langle X \rangle \pm \sigma$ . . . . .	72
5.1	Case 1 . . . . .	78
5.2	Case 2 . . . . .	79
5.3	Case 3 . . . . .	80
5.4	Case 4 . . . . .	81
5.5	Posterior distribution $P(X   DI)$ for the mixture model . . . . .	91
5.6	Means and standard deviations for the parameters of the mixture model . . . . .	92
5.7	Means and standard deviations for the parameters of the mixture model . . . . .	93
5.8	Means and standard deviations for the parameters of the mixture model . . . . .	94
6.1	Comparison of end-values . . . . .	98
6.2	Direct comparison of the LULU and Bayesian estimators for $\alpha = 0.975$ . . . . .	99
6.3	Direct comparison of the LULU and Bayesian estimators for $\alpha = 0.95$ . . . . .	100
6.4	Direct comparison of the LULU and Bayesian estimators for $\alpha = 0.95$ . . . . .	101
6.5	Ensemble averages of the MAEs . . . . .	102
6.6	Ensemble averages of the RMSEs . . . . .	103
7.1	Raman Spectroscopy data with smoothed sequences $F_1$ , $F_2$ and $F_{10}$ . . . . .	105
7.2	Histogram of RAMAN data . . . . .	106
7.3	Section of Raman Spectroscopy data containing outliers . . . . .	107
7.4	Thinned data with smoothed sequences . . . . .	108
7.5	Estimation of location parameter $c$ . . . . .	109
7.6	Estimation of scale parameters $\mu$ and $\epsilon$ . . . . .	110

## Chapter 1

# Introduction

Estimating a signal from noisy measurements is not only common in many physics laboratories, but is a cornerstone of the natural sciences, economics, engineering and everyday life. The aim is to reconstruct a geometric object in space or a signal in time from measured data  $x = \{x_1, x_2, \dots, x_n\}$ . The difficulty is that with every measurement errors appear inevitable. Galileo, looking through his telescope, would have known that in order to determine the position of a stellar body whilst accounting for atmospheric disturbances, calibration errors, or even a wobbly telescope, he would have to rely on making many observations from which the actual values of interest would have to be estimated. Another concern is the choice of a suitable model which is at best an idealized description of the stochastic nature of the errors. The simplest point of departure is to consider a locally constant signal obscured by errors that are independent and identically distributed (i.i.d.). A suitable model for  $x$  is thus  $x_i = c + z_i$  with  $c$  a constant and  $z_i$  coming i.i.d. from an error distribution  $f$ . If  $f$  is symmetric the standard statistical techniques for estimating the model parameter  $c$  is to minimize the error  $x - c$  in some or other norm, where

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \text{and} \quad c = \begin{bmatrix} c \\ c \\ \vdots \\ c \end{bmatrix}.$$

Using the least squares norm, or  $l^2$  norm, we have

$$\frac{\partial}{\partial c} \|x - c\| = \sqrt{\sum_{i=1}^n (x_i - c)^2} = 0$$

$$c = \frac{1}{n} \sum_{i=1}^n x_i,$$

and the average or mean of the data is taken as the estimate for  $c$ . For the  $l^1$  norm we have

$$\frac{\partial}{\partial c} \|x - c\|_1 = \frac{\partial}{\partial c} \sum_{i=1}^n |x_i - c| = \sum_{i=1}^n \frac{x_i - c}{|x_i - c|} = 0$$

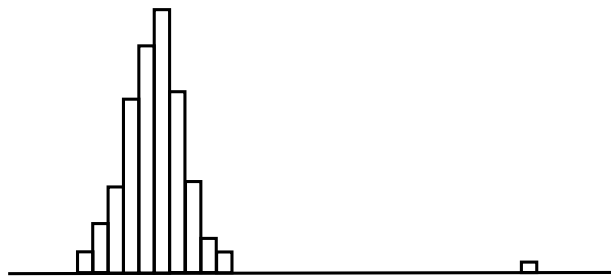
$$c = \text{median}\{x_1, x_2, \dots, x_n\},$$

where the median of the data is taken as the estimate for  $c$ . The  $l^\infty$  norm instead leads one to  $(x_{max} + x_{min})/2$  as the estimate. Another popular approach in the field of statistics is

that of maximum likelihood where the peak of the assumed model distribution is calculated.

However, whereas the above techniques may be merited given the specific problem at hand, our current position on the correct theoretical approach to the problem of data analysis in general is that of *Bayesian inference*. Here one is led to optimal inferences on parameters by a theory that is based on the concept of information in the form of data and prior knowledge, and where only subjective probabilities are used to calculate a probability distribution that describes the degree of belief in a particular value of the parameter. The choice of estimator for this distribution is a choice left for the user to decide. The mean or first moment of the marginal distribution for a parameter is a popular choice, but the median or peak of the distribution may also be used. Often a theoretical justification for the successes of the standard statistical procedures is gained in this framework. For instance, if the noise is assumed Gaussian, the mean estimator in the Bayesian framework is just the average of the data, giving merit to the least squares norm approach.

It is well-known that the average of the data is not a *robust* estimator for the location parameter  $c$ . Qualitatively, it is understood that the term robust means insensitivity to the postulated model and *outliers*, where the occurrence of the latter usually implies the former. Outliers in data sets are data points that on inspection seem to stand apart from the trend followed or cluster made by the majority of data points in the set (see Fig. 1.1 below).



**Figure 1.1:** Histogram of data with outlier.

Take for example the usual case where a location parameter  $c$  of a Gaussian error distribution is to be estimated. A data point that falls far from the mean (say five standard deviations) of the remaining data points will pull the estimate for  $c$  along with it. The median on the other hand is often said to be more robust, since more than one half of the data would have to be corrupt for it to be affected. On encountering outliers, the usual stance is to conclude that something has gone wrong with the means by which the data was accumulated (or apparatus), and then to proceed by throwing the outlier data points away and to draw conclusions from the remaining data which is perceived as good. This may be perfectly acceptable if one knows that the apparatus is unreliable (someone bumps the table on which your experiment is placed), however, it is not always possible to have a person present that can decide about whether or not such an event has occurred given that it were at all possible to identify. Having faith in the apparatus, the experimenter may on the other hand regard the outlier as an important discovery, and may want to postulate a distribution from whence it came and infer on its parameters as well. Furthermore, we are often faced with making use of whatever apparatus we are presented with, no matter how reliable (especially in the fields of Biology and Economics).

In the modern world, with the vast accumulation of high resolution data and given the constraints

of time and space, a need arises for computationally efficient operators that can pre-*smooth* the data. This means to selectively separate what is deemed redundant whilst keeping the signal intact for the purposes of storage, transmission or estimation. This can also be thought of as a form of data reduction which, as frightening as it may sound, is a common practice in fields ranging from Astronomy and Particle Physics to Economics. Examples of linear smoothers<sup>1</sup> or filters are the well-known moving averages. In practical applications, these can help to identify trend amongst noisy measurements but are not robust to outlier events. Examples of nonlinear smoothers are the median smoothers (popularised by Tukey) and the LULU smoothers of Stellenbosch mathematician Carl Rohwer. While the median smoothers can be appropriate for removing outliers, a supporting theory is ‘extremely difficult’ even in the simplest cases [18]. Nonlinear analysis had thus previously been approached by largely heuristic methods [31, 32]. The accompanying theory enabled by the LULU operators promises a deeper understanding.

LULU-theory, is so named due to its constituent nonlinear operators  $L_n$  and  $U_n$  that are usually applied in composition  $L_n U_n$  (or  $U_n L_n$ ). These were initially developed to be applied to one-dimensional sequences in order to remove impulsive noise, but since their appearance a series of articles followed detailing their alluring mathematical properties. A Discrete Pulse Transform (DPT) follows naturally from the theory [24] [26], and is a multiresolutional decomposition of a sequence  $x$  into a sum of positive and negative pulses. Each resolution level  $w$  is then a sequence which contains essentially zeros except for  $w$  consecutive entries of the same constant value (pulses of width  $w$ ) [25]. The DPT is viewed as a competitor to median based decompositions and, although not well-known in the physics community, promises major advantages over currently popular wavelet decompositions in certain applications.

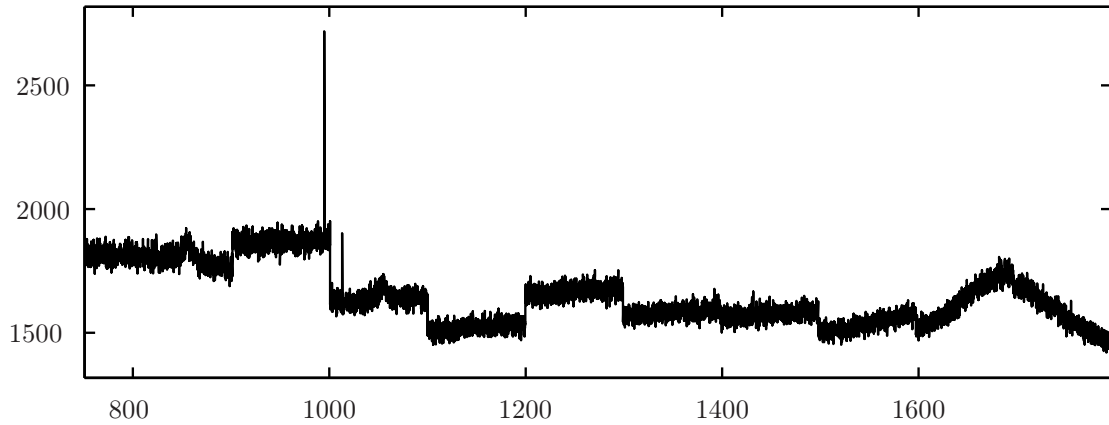
In the context of regular noise in addition to impulsive outlier noise, an appropriate model distribution is

$$f(x) = \alpha g(x) + \beta h(x),$$

where  $\beta = 1 - \alpha \approx 0$ . The signal is considered to be a constant locally but is obscured by noise described by a distribution  $g$  of finite support. In addition to this noise there are occasional one-sided impulses coming from a distribution  $h$ . Thus when  $\beta$  is small outliers taken from  $h$  occur rarely. As an example consider Fig. 1.2 below where a cosmic ray strikes the apparatus during a Laser Spectroscopy experiment, or a noisy transmission line receiving a capacitance discharge from lightning. When  $\alpha = 1$  we are back to the original model and if the average was taken as the estimate for  $c$  then this would be severely affected by an outlier event. When  $\beta > 0$ , an expectation value (denoted by  $\langle \cdot \rangle$ ) calculated over  $f$  is  $\alpha \langle g \rangle + \beta \langle h \rangle$ . We can thus expect that the average of the data points would be severely affected when the average of the outliers points (the points due to  $h$ ) is large. The goal is to reduce this effect to something like  $\beta^2 \times \text{average}(\text{data points due to } h)$  where for the time being  $\text{average}(\cdot)$  means the average of a sequence of data or of a sequence derived from the data. An underlying assumption is that fairly little is known about  $h$  since there is very little data or evidence to estimate it well. Thus, the chosen strategy should accommodate this uncertainty on  $h$ .

---

<sup>1</sup>By linear it is understood that the operations on the data are linear. More precisely, for two sequences  $x_i$  and  $y_i$  it is true that  $S(x_i + y_i) = Sx_i + Sy_i$  and  $S(\alpha x_i) = \alpha(Sx_i)$  if  $S$  is a linear smoother. Clearly the low frequency component of the noise cannot be removed and its effect spreads across to the points neighbouring the outliers.



**Figure 1.2:** Raman Spectroscopy data showing a real signal obscured by regular noise as well as outliers. In Raman Spectroscopy light from a laser is used to illuminate a sample of matter (in this case an Au covered polymer nano and Naphtalene). The scattered light is collected in a detector. The shifts in energy reveal information about the vibrational modes of the system. The outlier can be seen in the more or less constant region between entries 900 and 1000. Its cause was due to a cosmic ray striking the detector.

The objectives of this thesis can be summarized as follows:

1. Of particular interest is the article [25], where Rohwer discusses how the standard deviation of i.i.d. noise on a constant signal is nearly a constant multiplied by the average pulse amplitude of negative pulses in  $x - U_1x$  or positive pulses in  $x - L_1x$  when the distribution is a  $B$ -spline or the limit of such. As will be set out in Section 2.6, a theorem is presented which equates the expected value of positive and negative pulse heights of the first resolution level of a DPT to integrals of the form

$$\int_{-\infty}^{\infty} dt F^{2m}(1 - F), \text{ and } \int_{-\infty}^{\infty} dt F(1 - F)^{2m},$$

with  $m = 1$  or  $m = 2$ , and where  $F$  is the cumulative distribution function (cdf) of  $f$ . These are relatively easy to calculate and thus provide a useful way for estimating scale parameters of the model distribution. In the context of regular noise in addition to outlier noise, the scale parameters can still be estimated assuming  $\alpha$  is known.

2. The above-mentioned theorem provides the means to show that the estimators  $average(U_1L_1x)$  and  $average(L_1U_1x)$  are robust in the sense that they reduce the effect of  $h$  to something of the order  $\beta^2 \times average(\text{data points due to } h)$ . In other words, pre-smoothing with  $U_1L_1$  (or  $L_1U_1$ ) and using the standard technique where the average is taken as the estimate, is a robust procedure in the sense that any contribution to the expected value of the data due to  $h$  is of the order  $\beta^2$  (where  $\beta \approx 0$ ). Making use of the fact that

$$L_1x \leq U_1L_1x \leq M_1x \leq L_1U_1x \leq U_1x,$$

the result supports smoothing with the three-point running median  $M_1$  where before there was little hope for a supporting theory.

3. The corresponding Bayesian estimators to the equivalent problem are explored and tested in comparison to the suboptimal LULU-estimators for certain simple test distributions. In most scenarios, the amount of work and calculation time involved in the Bayesian approach is considerably more than in the LULU approach. Heuristic comparison shows that the Bayes estimators do not always outperform the LULU-estimators as had been expected. Although the Bayesian perspective provides insight into the logic of the problem, the estimators can be difficult to obtain analytically and slow to compute numerically.

In preparing the Bayesian solution for the parameters of a convex combination of  $g$  and  $h$ , the problem of a constant signal obscured by completely asymmetric noise originating from the uniform distribution was studied. The solution is presented in detail and its purpose in the thesis is as a precursor to the two-model problem. It serves as a good preliminary and introductory problem to the subject of Bayesian parameter estimation, and serves as a testament to the power of Bayesian data analysis in the sense that an imagined ad hoc approach is difficult. Furthermore, it shows how prior information enables analysis even when only one data point is available. An example of one-sided noise occurs in the problem of determining the epicentre of an earthquake. Here the times of arrival of seismic waves at numerous sensors are 'late' due to unknown geological features. As an advocate of the Bayesian school of thought, I hope this example stays with the reader as a reminder of the potential of Bayes theorem.

## Chapter 2

# LULU Theory and the Discrete Pulse Transform

### 2.1 Introduction

It was in 1989 when Carl Rohwer, working on practical problems for the Maritime institute in Simonstown, that the founding ideas of LULU-theory and the Discrete Pulse Transform (DPT) emerged [24, 19]. LULU-theory is so named due to its constituent non-linear operators, or smoothers  $L_n$  and  $U_n$  that are usually applied in composition  $L_n U_n$  (or  $U_n L_n$ ). They were initially developed to be applied to one-dimensional sequences in order to remove impulsive noise [24], but since their appearance a series of articles followed by Rohwer [20, 21, 22, 23], Rohwer and Toerien [27] and Rohwer and Wild [28], detailing their alluring mathematical properties which include for example idempotency, co-idempotency, stability, trend preserving and variation decomposing [12]. The work culminated in a self contained monograph on the subject in 2005 [24], and has since continued with numerous ventures. For example, the theory was later extended to higher dimensional arrays [1], knowledge concerning some statistical properties of LULU smoothers has been gained [4] [12] and fast implementation algorithms have been presented [8] [17]. A Discrete Pulse Transform follows naturally from the theory [24] [26], and is a multiresolutional decomposition of a sequence  $x$  into a sum of positive and negative pulses. Each resolution level  $w$  is then a sequence which contains essentially zeros except for  $w$  consecutive entries of the same constant value (pulses of width  $w$ ) [25]. The successes of the DPT are heralded in the field of image processing. Here the acquisition of data can come from digital cameras, long-wave (infrared) and laser capturing devices [7]. The DPT is viewed as a competitor to median based decompositions, and is superior to wavelet decomposition in some applications. The interest in research and applications continues to grow with new publications appearing yearly. Of particular interest for the purpose of this thesis is the article [25], where Rohwer discusses how the variation of i.i.d. noise on a constant signal may be calculated as a near constant multiplied by the average pulse amplitude of downward or upward pulses in the first resolution level of a DPT.

In the following we list some basic definitions and properties of the LULU-theory and introduce the DPT and the pulses in its highest resolution level. A method for simple calculation of the theoretical expectation of the upward and downward pulse heights in the highest resolution level is presented in a theorem and is largely accredited to the article [25] by Rohwer and personal discussions with him. The expectation of the pulse heights are shown to be functions of the parameters of the distribution  $f$  assumed responsible for the data. An avenue opens for proposal of heuristic point estimators of the scale parameters of  $f$  using the averages of pulses in the highest



resolution level. Finally, the theorem is shown to allow for analysis of averages of smoothed sequences leading to interesting results regarding the robustness of the smoothers  $L_1$ ,  $U_1$ ,  $L_1U_1$  and  $U_1L_1$  in the context of removal of outliers. Throughout, simple examples are worked out and tested numerically. The results are argued to be favourable over an optimal solution.

## 2.2 LULU basics

Let  $x$  be a bi-infinite sequence in  $\ell_1$  (the space of absolutely summable sequences)

$$x = \{\dots x_{-1}, x_0, x_1, \dots\}. \quad (2.1)$$

In practice sequences are generally finite, but zeros or end values may be replicated outwards to make them bi-infinite and bounded. This is done so that the  $l^2$  norm may be used. Define the following *rank selectors*

$$(\bigwedge x)_i = \max\{x_i, x_{i+1}\} \text{ and } (\bigvee x)_i = \min\{x_{i-1}, x_i\}. \quad (2.2)$$

Define the *upper* and *lower half smoothers* as

$$U_n(x) = \bigwedge^n \bigvee^n x \text{ and } L_n(x) = \bigvee^n \bigwedge^n x \quad (2.3)$$

respectively. Each entry of  $U_n x$  and  $L_n x$  is thus given by

$$(U_n x)_i = \min(\max(x_{i-n}, \dots, x_i), \dots, \max(x_i, \dots, x_{i+n})) \quad (2.4)$$

and

$$(L_n x)_i = \max(\min(x_{i-n}, \dots, x_i), \dots, \min(x_i, \dots, x_{i+n})). \quad (2.5)$$

Thus, for example,

$$(L_1 x)_i = \max(\min(x_{i-1}, x_i), \min(x_i, x_{i+1}))$$

and

$$(L_2 x)_i = \max(\min(x_{i-2}, x_{i-1}, x_i), \min(x_{i-1}, x_i, x_{i+1}), \min(x_i, x_{i+1}, x_{i+2})).$$

The sequence  $\{x_{i-n}, \dots, x_i, \dots, x_{i+n}\}$  is called a running window. It is also called the *support* of  $(U_n x)_i$  and  $(L_n x)_i$ . An *n-pulse*, or pulse of length  $n$  is a sequence  $x$  such that

$$x = \{\dots, 0, b_1, b_1, \dots, b_n, 0, \dots\} \text{ with } b_1 = b_2 = \dots = b_n = b \quad (2.6)$$

and with infinitely many zeros on both sides. If  $b$  is positive it is called upward, and if  $b$  is negative it is called downward. Applying the smoother  $U_1$  to a sequence will remove the downward pulses of length one whilst  $L_1$  will remove the upward pulses of length one. Similarly,  $L_2$  and  $U_2$  will remove upward and downward 2-pulses. Thus  $U_n$  smooths from below, and  $L_n$  smooths from above. Compositions of the form  $L_n U_n$  and  $U_n L_n$  are called *basic smoothers*. The *composite smoothers*  $C_n$ , called *ceiling*, and  $F_n$ , called *floor*, are defined as

$$C_n = \begin{cases} I, & n = 0 \\ L_1 U_1, & n = 1, \\ L_n U_n C_{n-1}, & n > 1 \end{cases} \quad (2.7)$$

and

$$F_n = \begin{cases} I, & n = 0 \\ U_1 L_1, & n = 1 \\ U_n L_n F_{n-1}, & n > 1 \end{cases} \quad (2.8)$$

respectively.

### 2.3 The highest resolution level of a DPT

A LULU-decomposition of a sequence  $x \in \ell_1$  is defined by

$$x = \sum_{n=1}^{\infty} D_n(x), \quad (2.9)$$

where the different resolution levels are given by the choice

$$D_n = C_{n-1} - C_n, \text{ with } n \geq 1. \quad (2.10)$$

A *dual* decomposition can be defined

$$x = \sum_{n=1}^{\infty} R_n(x), \quad (2.11)$$

with

$$R_n = F_{n-1} - F_n, \text{ } n \geq 1. \quad (2.12)$$

The sequences  $D_n$  and  $R_n$  are essentially made up of zeros except for constant pulses of width  $n$  that are separated sufficiently such that they may be mapped onto the zero sequence by  $L_n U_n$  and  $U_n L_n$  respectively. A DPT is thus the representation of a sequence as a sum of sequences (resolution levels) which each contain positive and negative pulses of width 1, 2, and so on. See Fig. 2.1 for an illustrative example of a DPT. Since  $L_n$  and  $U_n$  are *duals* of each other, meaning  $U_n(-x) = -L_n(x)$ , it follows that  $R_n(-x) = -D_n(x)$ . The first or highest resolution level of a  $C$ -decomposition is given by

$$\begin{aligned} D_1 x &= (I - L_1 U_1) x \\ &= (I - U_1) x + (I - L_1) U_1 x, \end{aligned} \quad (2.13)$$

and of an  $F$ -decomposition by

$$\begin{aligned} R_1 x &= (I - U_1 L_1) x \\ &= (I - L_1) x + (I - L_1) U_1 x. \end{aligned} \quad (2.14)$$

Considering  $D_1$ , since

$$(U_1 x)_i = \min(\max(x_{i-1}, x_i), \max(x_i, x_{i+1})), \quad (2.15)$$

we see that  $(U_1 x)_i \geq x_i$ , and thus

$$x_i - (U_1 x)_i \leq 0. \quad (2.16)$$

Similarly, it is easy to show that

$$(U_1 x)_i - (L_1 U_1 x)_i \geq 0, \quad (2.17)$$

and thus  $D_1$  is the sum of upward (positive) and downward (negative) 1-pulses contained in the sequences  $U_1 x - L_1 U_1 x \geq 0$  and  $x - U_1 x \leq 0$  respectively. Similarly for  $R_1$ , the upward and downward 1-pulses are in  $x - L_1 x \geq 0$  and  $U_1 x - L_1 U_1 x \leq 0$  respectively.

Focussing on  $x - U_1 x$ , the point  $(U_1 x)_i$  differs from  $x_i$  if and only if

$$x_{i-1}, x_{i+1} > x_i, \quad (2.18)$$

or, in other words, when a downward pulse is present in  $x$ . The ordering of the three points  $\{x_{i-1}, x_i, x_{i+1}\}$  in the running window can with equal probability occur in six possible ways, two of which will produce a downward pulse that  $U_1$  will remove. Assuming the sequence  $x$  is i.i.d., the probability (or chance) for the appearance of a downward pulse in the resolution level  $D_1(x)$  is thus  $\frac{1}{3}$ , irrespective of the particular distribution or model! The same is true for the probability of an upward pulse appearing in the first level  $R_1(x)$  of an F-decomposition.

There are 120 ways of arranging the order of 5 points  $\{x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2}\}$ , 24 of which will satisfy

$$x_{i-2}, x_{i-1}, x_{i+1}, x_{i+2} < x_i, \quad (2.19)$$

which is equivalent to the scenario where we have an upward pulse after the effect of  $U_1$  on the sequence. Thus the probability for an upward pulse appearing in  $U_1x - U_1L_1x$  is  $\frac{24}{120} = \frac{1}{5}$ . Because the number of pulses in a decomposition is always smaller than or equal to the number of data points  $N$  [26], we see that more than half ( $\frac{1}{3} + \frac{1}{5} = \frac{8}{15}$ ) of the pulses are expected to be in the first resolution level [25] if the sequence is random (i.i.d.).

An operator  $S$  is defined to be idempotent if  $S^2 = S$ , and co-idempotent if  $(I - S)^2 = I - S$ . An operator is called a separator if it is both idempotent and co-idempotent. It can be shown that the operators  $U_n, L_n, L_nU_n, U_nL_n, C_n$  and  $F_n$  are all separators [24]. As an example, consider the negative pulses in  $x - U_1x$ . They are sufficiently separated (and thus referred to as isolated) in the sense that due to the co-idempotence  $U_1$ , it annihilates the sequence:

$$L_1U_1x - L_1x = -U_1x + U_1U_1x = 0. \quad (2.20)$$

The co-idempotence of  $L_1$  assures that the positive pulses in  $U_1x - L_1U_1x$  are also sufficiently separated:

$$L_1(U_1x - L_1U_1x) = L_1U_1x - L_1L_1U_1x = 0. \quad (2.21)$$

As another example consider the upward and downward pulses in the first resolution level  $D_1(x) = (I - L_1U_1)x$ . Since  $L_1U_1$  is co-idempotent, we have that

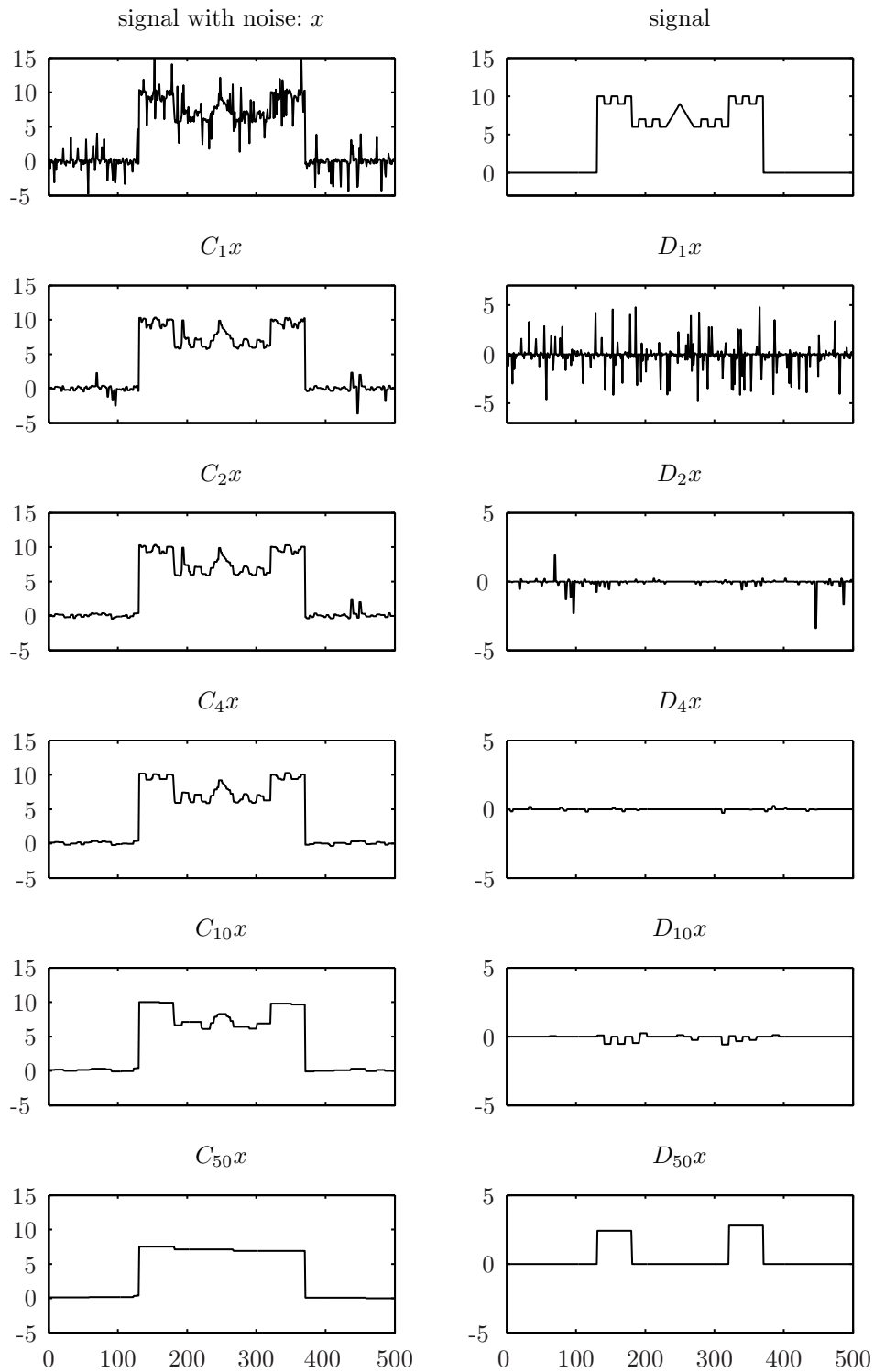
$$L_1U_1D_1 = L_1U_1(I - L_1U_1)x - (L_1U_1 - (L_1U_1)^2)x = 0. \quad (2.22)$$

Thus,

$$x - D_1(x) = \sum_{n=2}^{\infty} D_n(x) - L_1U_1x, \quad (2.23)$$

is a sequence which contains no isolated positive or negative pulses.

Heuristically we can expect to find less than  $\frac{8}{15}$  pulses in the first resolution level if the sequence is correlated (for example a sequence with trend). If the sequence is however assumed i.i.d., and the number of (say) downward pulses is close to  $\frac{1}{3}$  then it can be assumed that most of these are due to the random additive noise. Estimating the parameters of the underlying noise distribution using the pulses in the first resolution level thus seems possible [25].



**Figure 2.1:** Example of LULU smoothing and the resolution levels of a DPT. The image of the castle represents piecewise constant signals as well as trend. The signal is badly distorted and unrecognisable after adding symmetric noise in addition to impulsive noise of relatively large amplitude. The signal is successively smoothed with  $L_nU_n$  up to resolution  $n = 50$ . The smoothed sequences as well as the pulses in the subsequent resolution levels of a  $D$  decomposition are shown.

## 2.4 Theoretical expectation of pulses in $D_1$ and $R_1$

We are looking for a way of understanding the pulses in a DPT in terms of the parameters of the underlying model distribution  $f$  assumed responsible for the data  $x$ . The previous section suggests that since most of the pulses in a DPT are expected to be found in the first resolution level if the signal is assumed near constant, then this may be a good place to begin to look to construct a heuristic method for estimating the parameters of interest. Consider thus the model  $x_i = c + \epsilon_i$  where a constant signal  $c$  is obscured by an error  $\epsilon_i$  which is distributed according to a certain noise distribution. If  $x$  is an i.i.d. random sequence then so are for example the sequences  $x - L_1x$  and  $L_1x - U_1L_1x$  that make up the positive and negative pulses of  $R_1x$ . The following theorem concerns the expected values or theoretical expectation values of the sequences that contain the positive and negative pulses of an  $R_1$  or  $D_1$  decomposition. When we write a sequence in angular brackets, for example  $\langle x - L_1x \rangle$ , it must be interpreted as an expectation value of the variate  $x_i - L_1x_i$  calculated over some distribution (or part thereof). The expectation value  $\langle x - L_1x \rangle$  can thus be thought of as the expected value of the long term average of the pulses in  $x - L_1x$  assuming the model  $f$  is correct. The following theorem is inspired by a proof given in [25] where the expected value of  $x - L_1$  is calculated.

**Theorem 1:** Let  $x$  be a sequence of random numbers generated identically and independently from  $f$ , a piecewise continuous distribution which is positive for  $x \in [a, b]$ . Then the expected values of the four sequences below is given by the integrals involving  $F(t) = \int_{-\infty}^t dx f(x)$  respectively:

- (a)  $\langle x - L_1x \rangle = \int_{-\infty}^{\infty} dt F^2(1 - F)$
- (b)  $\langle U_1x - L_1U_1x \rangle = \int_{-\infty}^{\infty} dt F^4(1 - F)$
- (c)  $\langle x - U_1x \rangle = - \int_{-\infty}^{\infty} dt F(1 - F)^2$
- (d)  $\langle L_1x - U_1L_1x \rangle = - \int_{-\infty}^{\infty} dt F(1 - F)^4$ .

*Proof.*

Assume  $x_i$  are i.i.d. according to the assumed model distribution  $f$ . Consider the first resolution level of the DPT  $D_1x$  and  $R_1x$ . Since

$$(D_1x)_i = \underbrace{(x_i - (U_1x)_i)}_{\leq 0} + \underbrace{((U_1x)_i - (L_1U_1x)_i)}_{\geq 0}, \quad (2.24)$$

$D_1x$  consists of positive and negative 1-pulses which can be found in  $(I - L_1)U_1x$  and  $(I - U_1)x$  respectively. Similarly since

$$(R_1x)_i = \underbrace{(x_i - (L_1x)_i)}_{\geq 0} + \underbrace{((L_1x)_i - (U_1L_1x)_i)}_{\leq 0}, \quad (2.25)$$

$R_1x$ 's constituent positive and negative 1-pulses are found in  $(I - L_1)x$  and  $(I - U_1)L_1x$  respectively. We want to predict the average of these positive and negative pulses by calculating a theoretical expectation value, which we denote by the angular brackets  $\langle \cdot \rangle$ .

Focussing on the positive pulses, the sequence  $(I - L_1)x$  is essentially made up of zeros, except for isolated positive pulses  $(I - L_1)x_i > 0$ . A pulse is only present when the point  $(L_1x)_i$  differs from  $x_i$ , which in turn happens if and only if  $x_{i-1}, x_{i+1} < x_i$ . The sequence  $(I - L_1)U_1x$  also consists mainly of zeros except for positive pulses  $(I - L_1)U_1x_i > 0$ . A pulse is only present when the

point  $(L_1U_1x)_i$  differs from  $U_1x_i$ , which in turn happens if and only if  $x_{i-2}, x_{i-1}, x_{i+1}, x_{i+2} < x_i$ . Construct a new random variable

$$z_i = \max(x_{i-m}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+m}) - x_i. \quad (2.26)$$

Let  $C_m(z)$  be the distribution of  $z_i$ . It is the convolution of  $\max(x_{i-m}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+m})$  and  $-x_i$

$$\begin{aligned} C_m(z) &= p(\max(x_{i-m}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+m}) - x_i) \\ &= p(\max(x_{i-m}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+m}) + (-x_i)) \\ &= \int_{-\infty}^{\infty} dt g(t) f(t - z), \end{aligned} \quad (2.27)$$

where  $g$  is the distribution of  $\max(x_{i-m}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+m})$ . The cumulative distribution function (cdf) of  $g$  is

$$p(\max(x_{i-m}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+m}) < y) = \int_{-\infty}^y dt g(t) = G(y), \text{ with } \frac{d}{dy} G(y) = g(y). \quad (2.28)$$

The following probabilities are equivalent

$$p(\max(x_{i-m}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+m}) < y) \equiv p(x_{i-m} < y, \dots, x_{i-1} < y, x_{i+1} < y, \dots, x_{i+m} < y), \quad (2.29)$$

and since the  $x_i$  are assumed i.i.d. we have that

$$\begin{aligned} p(x_{i-m} < y, \dots, x_{i-1} < y, x_{i+1} < y, \dots, x_{i+m} < y) \\ = p(x_{i-m} < y) \dots p(x_{i-1} < y) p(x_{i+1} < y) \dots p(x_{i+m} < y). \end{aligned} \quad (2.30)$$

The cdf of  $f$  is

$$p(x_i < y) = \int_{-\infty}^y dt f(t) = F(y), \text{ with } \frac{d}{dy} F(y) = f(y). \quad (2.31)$$

Therefore

$$p(\max(x_{i-m}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+m}) < y) = F(y)^{2m}, \quad (2.32)$$

and thus

$$\begin{aligned} g(y) &= \frac{d}{dy} F(y)^{2m} \\ &= 2m F(y)^{2m-1} f(y). \end{aligned} \quad (2.33)$$

Upon substituting  $g$ ,  $C_m(z)$  becomes

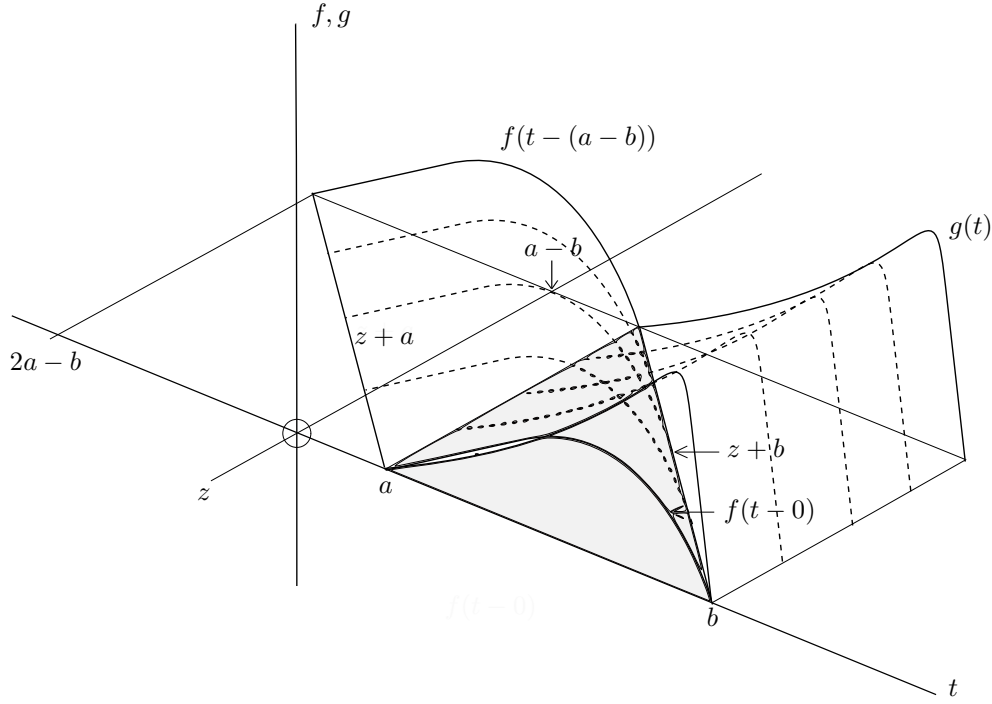
$$C_m(z) = \int_{-\infty}^{\infty} dt 2m F(t)^{2m-1} f(t) f(t - z). \quad (2.34)$$

The expected average pulse height of the positive pulses in the sequences  $x - L_1x$  and  $U_1x - L_1U_1x$ , denoted  $\langle x - L_1x \rangle$  and  $\langle U_1x - L_1U_1x \rangle$  respectively, is an expectation calculated over  $C_m(z)$  where  $z < 0$ , and multiplied by minus one (to get the sign right)

$$-\int_{-\infty}^0 dz z C_m(z) = \begin{cases} \langle x - L_1x \rangle & \text{for } m = 1 \\ \langle U_1x - L_1U_1x \rangle & \text{for } m = 2. \end{cases} \quad (2.35)$$

Substituting  $C_m(z)$  we need

$$- \int_{-\infty}^0 dz z \int_{-\infty}^{\infty} dt 2mF(t)^{2m-1} f(t) f(t-z). \quad (2.36)$$



**Figure 2.2:** Drawing of  $f(t-z)$  and  $g(t)$  for  $z < 0$ . The shaded region indicates where the product of  $f(t-z)$  and  $g(t)$  is non-zero.

As is illustrated in Fig 2.2 above,  $g(t) \geq 0$  for  $t \in [a, b]$  and  $f(t-z) \geq 0$  for  $t \in [a+z, b+z]$ . We can therefore introduce theta functions to write  $g(t)$  and  $f(t-z)$  as follows

$$\begin{aligned} g(t) &= g(t)\theta(b-t)\theta(t-a), \\ f(t-z) &= f(t-z)\theta(z+b-t)\theta(t-(z+a)). \end{aligned} \quad (2.37)$$

Substituting these and rearranging the arguments of the theta functions containing  $z$  we have

$$- \int_{-\infty}^0 dz z \int_{-\infty}^{\infty} dt 2mF(t)^{2m-1} f(t) f(t-z)\theta(b-t)\theta(t-a)\theta(z-(t-b))\theta(t-a-z), \quad (2.38)$$

or

$$- \int_a^b dt 2mF(t)^{2m-1} f(t) \int_{t-b}^0 dz z f(t-z). \quad (2.39)$$

Concentrating on  $\int_{t-b}^0 dz z f(t-z)$  for the moment, make the change of variable  $t-z = -y$  and use integration by parts to get

$$\begin{aligned} \int_{t-b}^0 dz z f(t-z) &= \int_{-b}^{-t} dy (y+t) f(-y) \\ &= \underbrace{(t-t)}_{=0} R(-t) - (t-b) \underbrace{R(-b)}_{=0} - \int_{-b}^{-t} dy R(y) \\ &= -Q(-t), \end{aligned} \quad (2.40)$$

where

$$R(y) = \int_{-\infty}^y dt f(-t) = \int_{-b}^y dt f(-t), \text{ with } f(-y) = \frac{d}{dy} R(y) \quad (2.41)$$

and

$$Q(t) = \int_{-\infty}^t dy R(y) = \int_{-b}^t dy R(y), \text{ with } R(t) = \frac{d}{dt} Q(t). \quad (2.42)$$

Substituting  $-Q(-t)$  and using integration by parts again proves the first half of the theorem

$$\begin{aligned} \int_a^b dt 2mF(t)^{2m-1} f(t) Q(-t) &= \underbrace{F(b)^{2m}}_{=1} \underbrace{Q(-b)}_{=0} - \underbrace{F(a)^{2m}}_{=0} Q(-a) + \int_a^b dt F(t)^{2m} R(-t) \\ &= \int_a^b dt F(t)^{2m} (1 - F(t)), \end{aligned} \quad (2.43)$$

where the final line follows since

$$\begin{aligned} R(-y) &= \int_{-\infty}^{-y} dt f(-y) \\ &= \int_y^{\infty} dt' f(t') \\ &= \int_{-\infty}^{\infty} dt' f(t') - \int_{-\infty}^y dt' f(t') \\ &= 1 - F(y). \end{aligned} \quad (2.44)$$

Focussing on the negative pulses,  $(I-U_1)x$  has a negative pulse at index  $i$  if and only if  $x_{i-1}, x_{i+1} > x_i$  and  $(I-U_1)L_1x$  has a negative pulse at index  $i$  if and only if  $x_{i-2}, x_{i-1}, x_{i+1}, x_{i+2} > x_i$ . Construct a new random variable

$$z_i = \min(x_{i-m}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+m}) - x_i. \quad (2.45)$$

Let  $C_m(z)$  be the distribution of  $z_i$ , it is the convolution of  $\min(x_{i-m}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+m})$  and  $-x_i$

$$\begin{aligned} C_m(z) &= p(\min(x_{i-m}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+m}) - x_i) \\ &= p(\min(x_{i-m}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+m}) + (-x_i)) \\ &= \int_{-\infty}^{\infty} dt g(t) f(t-z), \end{aligned} \quad (2.46)$$



where  $g$  is the distribution of  $\min(x_{i-m}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+m})$ . The cdf of  $g$  is

$$\int_{-\infty}^y dt g(t) = G(y), \text{ with } \frac{d}{dy}G(y) = g(y). \quad (2.47)$$

Therefore

$$\begin{aligned} p(\min(x_{i-m}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+m}) > y) &= \int_y^{\infty} dt g(t) \\ &= \int_{-\infty}^{\infty} dt g(t) - \int_{-\infty}^y dt g(t) \\ &= 1 - G(y). \end{aligned} \quad (2.48)$$

The following probabilities are equivalent

$$p(\min(x_{i-m}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+m}) > y) \equiv p(x_{i-m} > y, \dots, x_{i-1} > y, x_{i+1} > y, \dots, x_{i+m} > y), \quad (2.49)$$

and since the  $x_i$  are assumed i.i.d. we have that

$$\begin{aligned} p(x_{i-m} > y, \dots, x_{i-1} > y, x_{i+1} > y, \dots, x_{i+m} > y) \\ = p(x_{i-m} > y) \dots p(x_{i-1} > y) p(x_{i+1} > y) \dots p(x_{i+m} > y). \end{aligned} \quad (2.50)$$

The cdf of  $f$  is

$$p(x_i < y) = \int_{-\infty}^y dt f(t) = F(y), \text{ with } \frac{d}{dy}F(y) = f(y), \quad (2.51)$$

and since

$$\begin{aligned} p(x_i > y) &= \int_y^{\infty} dt f(t) \\ &= \int_{-\infty}^{\infty} dt f(t) - \int_{-\infty}^y dt f(t) \\ &= 1 - F(y), \end{aligned} \quad (2.52)$$

we have that

$$p(\min(x_{i-m}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+m}) > y) = (1 - F(y))^{2m}. \quad (2.53)$$

Therefore

$$1 - G(y) = (1 - F(y))^{2m}, \quad (2.54)$$

and

$$\begin{aligned} g(y) &= \frac{d}{dy}(1 - (1 - F(y))^{2m}) \\ &= 2m(1 - F(y))^{2m-1} f(y). \end{aligned} \quad (2.55)$$

Upon substituting  $g$ ,  $C_m(z)$  becomes

$$C_m(z) = \int_{-\infty}^{\infty} dt 2m(1 - F(t))^{2m-1} f(t) f(t - z). \quad (2.56)$$

The expected average pulse height of the negative pulses in the sequences  $x - U_1x$  and  $L_1x - U_1L_1x$ , denoted  $\langle x - U_1x \rangle$  and  $\langle L_1x - U_1L_1x \rangle$  respectively, is an expectation calculated over  $C_m(z)$  where  $z > 0$ , and multiplied by minus one (to get the sign right)

$$- \int_0^\infty dz z C_m(z) = \begin{cases} \langle x - U_1x \rangle & \text{for } m = 1 \\ \langle L_1x - U_1L_1x \rangle & \text{for } m = 2. \end{cases} \quad (2.57)$$

Substituting  $C_m(z)$  we need

$$- \int_0^\infty dz z \int_{-\infty}^\infty dt 2m(1 - F(t))^{2m-1} f(t) f(t - z). \quad (2.58)$$

Again, since  $g(t) \geq 0$  for  $t \in [a, b]$ , and  $f(t - z) \geq 0$  for  $t \in [a + z, b + z]$ , one can write  $g(t)$  and  $f(t - z)$  as follows

$$\begin{aligned} g(t) &= g(t)\theta(b - t)\theta(t - a), \\ f(t - z) &= f(t - z)\theta(z + b - t)\theta(t - (z + a)). \end{aligned}$$

Substituting these we have

$$- \int_0^\infty dz z \int_{-\infty}^\infty dt 2m(1 - F(t))^{2m-1} f(t) f(t - z) \theta(b - t) \theta(t - a) \theta(z - (t - b)) \theta(t - a - z), \quad (2.59)$$

or

$$- \int_a^b dt 2m(1 - F(t))^{2m-1} f(t) \int_0^{t-a} dz z f(t - z). \quad (2.60)$$

Concentrating on  $\int_0^{t-a} dz z f(t - z)$  for the moment, make the change of variable  $t - z = y$  and use integration by parts to get

$$\begin{aligned} \int_0^{t-a} dz z f(t - z) &= - \int_t^a dy (t - y) f(y) \\ &= - \left( (t - a) \underbrace{F(a)}_{=0} - \underbrace{(t - t) F(t)}_{=0} + \int_t^a dy F(y) \right) \\ &= \int_a^t dy F(y) \\ &= S(t), \end{aligned} \quad (2.61)$$

where

$$S(t) = \int_{-\infty}^t dy F(y) = \int_a^t dy F(y), \text{ with } F(t) = \frac{d}{dt} S(t) \quad (2.62)$$

Substituting  $S(t)$  and using integration by parts proves the second half of the theorem

$$\begin{aligned} \int_a^b dt (-2m(1 - F(t))^{2m-1} f(t)) S(t) &= (1 - \underbrace{F(b)}_{=1})^{2m} S(b) - (1 - F(a))^{2m} \underbrace{S(a)}_{=0} - \int_a^b dt F(t) (1 - F(t))^{2m} \\ &= - \int_a^b dt F(t) (1 - F(t))^{2m}. \end{aligned} \quad \square \quad (2.63)$$

It is often easier to calculate

$$- \int_{-b}^{-a} dt (1 - R(t))R(t)^{2m}, \quad (2.64)$$

which follows since  $R(-t) = 1 - F(t)$ .  $R(t)$  is simply the cdf of the mirror image of  $f$  about zero. Note further that when  $f$  is symmetric, implying  $f(t) = f(-t)$ , then

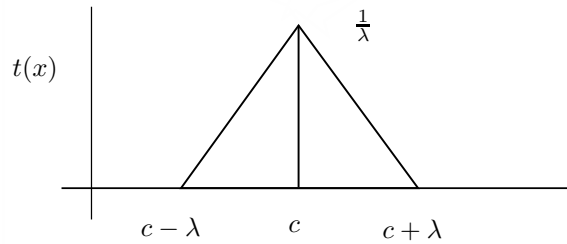
$$\int_a^b dt F(t)^{2m}(1 - F(t)) = \int_{-b}^{-a} dt (1 - R(t))R(t)^{2m}, \quad (2.65)$$

and thus

$$\begin{aligned} \langle x - L_1x \rangle &= -\langle x - U_1x \rangle \\ \langle U_1x - L_1U_1x \rangle &= -\langle L_1x - U_1L_1x \rangle. \end{aligned} \quad (2.66)$$

### 2.4.1 Example 1: Triangular distribution

Consider the model distribution  $t$  where a constant signal  $c$  is obscured by noise coming from a symmetric triangular distribution which is parametrised by  $\lambda$  (see Fig. 2.3 below).



**Figure 2.3:** Triangular distribution.

$$t(y) = \begin{cases} \frac{\lambda - c + y}{\lambda^2}, & y \in [c - \lambda, c] \\ \frac{\lambda + c - y}{\lambda^2}, & y \in (c, c + \lambda] \end{cases}. \quad (2.67)$$

Since  $c$  is irrelevant for estimating the parameter  $\lambda$  let  $c = 0$  (if we keep  $c$  it will cancel out) and  $t$  becomes

$$t(y) = \begin{cases} \frac{\lambda + y}{\lambda^2}, & y \in [-\lambda, 0] \\ \frac{\lambda - y}{\lambda^2}, & y \in (0, \lambda] \end{cases}. \quad (2.68)$$

Since  $t$  is symmetric we need

$$\langle x - L_1x \rangle = -\langle x - U_1x \rangle = \int dx (1 - F(x))F^2(x) \quad (2.69)$$

and

$$\langle U_1x - L_1U_1x \rangle = -\langle L_1x - U_1L_1x \rangle = \int dx(1 - F(x))F^4(x). \quad (2.70)$$

We thus need  $F^q(x)$  for  $q = 2, \dots, 5$ . Observing  $t$  above, the cdf  $F(x) = \int_{-\infty}^x dy t(y)$ , raised to the power  $q$  is

$$F^q(x) = \begin{cases} 0, & x < -\lambda \\ \left[ \int_{-\lambda}^x dy \left( \frac{y}{\lambda^2} + \frac{1}{\lambda} \right) \right]^q, & -\lambda \leq x \leq 0 \\ \left[ \int_{-\lambda}^0 dy \left( \frac{y}{\lambda^2} + \frac{1}{\lambda} \right) + \int_0^x dy \left( \frac{1}{\lambda} - \frac{y}{\lambda^2} \right) \right]^q, & 0 < x \leq \lambda \\ 1, & x > \lambda. \end{cases} \quad (2.71)$$

After performing the integrals this becomes

$$F^q(x) = \begin{cases} 0, & x < -\lambda \\ \left[ \frac{x^2}{2\lambda^2} + \frac{x}{\lambda} + \frac{1}{2} \right]^q, & -\lambda \leq x \leq 0 \\ \left[ -\frac{x^2}{2\lambda^2} + \frac{x}{\lambda} + \frac{1}{2} \right]^q, & 0 < x \leq \lambda \\ 1, & x > \lambda, \end{cases} \quad (2.72)$$

and therefore

$$\int_{-\infty}^{\infty} dx F^q(x) = \int_{-\lambda}^0 dx \left[ \frac{x^2}{2\lambda^2} + \frac{x}{\lambda} + \frac{1}{2} \right]^q + \int_0^{\lambda} dx \left[ -\frac{x^2}{2\lambda^2} + \frac{x}{\lambda} + \frac{1}{2} \right]^q. \quad (2.73)$$

Finally, perform the above integrals to find that

$$\langle x - L_1x \rangle = -\langle x - U_1x \rangle = \int dx F^2(x) - \int dx F^3(x) = \frac{7}{60}\lambda = 0.1167\lambda, \quad (2.74)$$

and

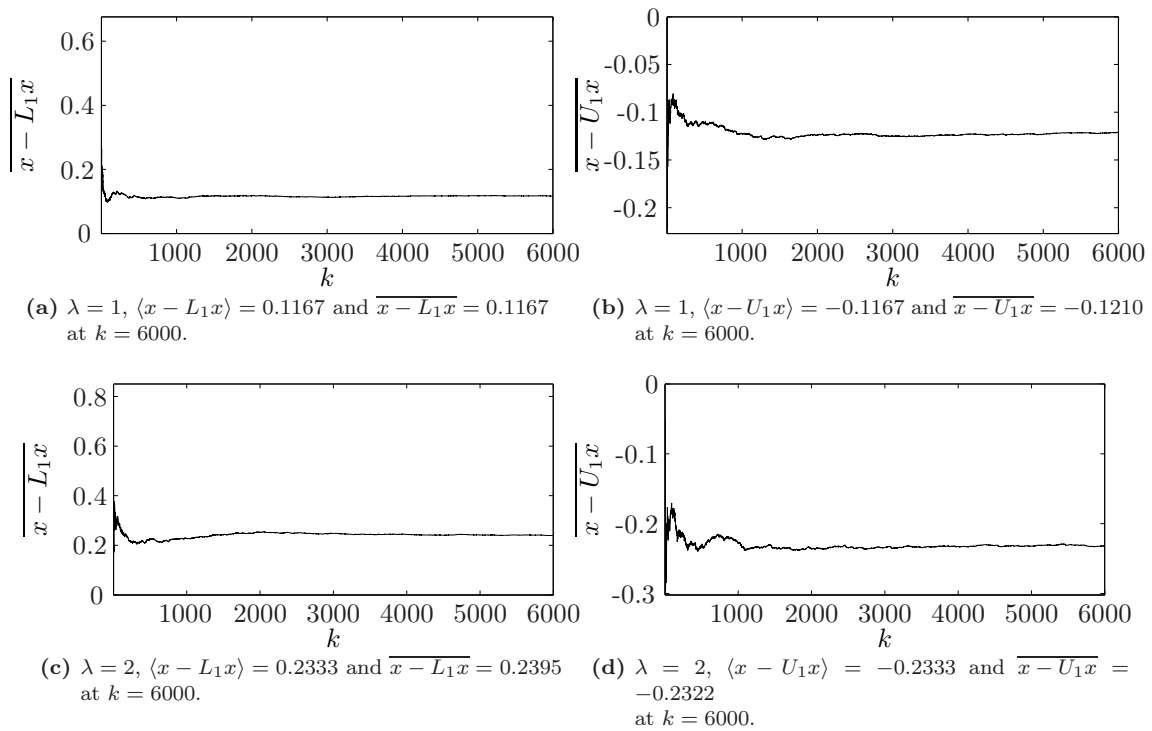
$$\langle U_1x - L_1U_1x \rangle = -\langle L_1x - U_1L_1x \rangle = \int dx F^4(x) - \int dx F^5(x) = \frac{89}{1680}\lambda = 0.05298\lambda. \quad (2.75)$$

One can confirm these results by simulation. Let the average of the pulses in  $x - L_1x$ ,  $x - U_1x$ ,  $U_1x - L_1U_1x$  and  $L_1x - U_1L_1x$  be respectively denoted by

$$\begin{aligned} \overline{x - L_1x}(k) &= \frac{1}{k} \sum_{i=1}^k x_i - (L_1x)_i \geq 0 \\ \overline{x - U_1x}(k) &= \frac{1}{k} \sum_{i=1}^k x_i - (U_1x)_i \leq 0 \\ \overline{U_1x - L_1U_1x}(k) &= \frac{1}{k} \sum_{i=1}^k (U_1x)_i - (L_1U_1x)_i \geq 0 \\ \overline{L_1x - U_1L_1x}(k) &= \frac{1}{k} \sum_{i=1}^k (L_1x)_i - (U_1L_1x)_i \leq 0. \end{aligned} \quad (2.76)$$

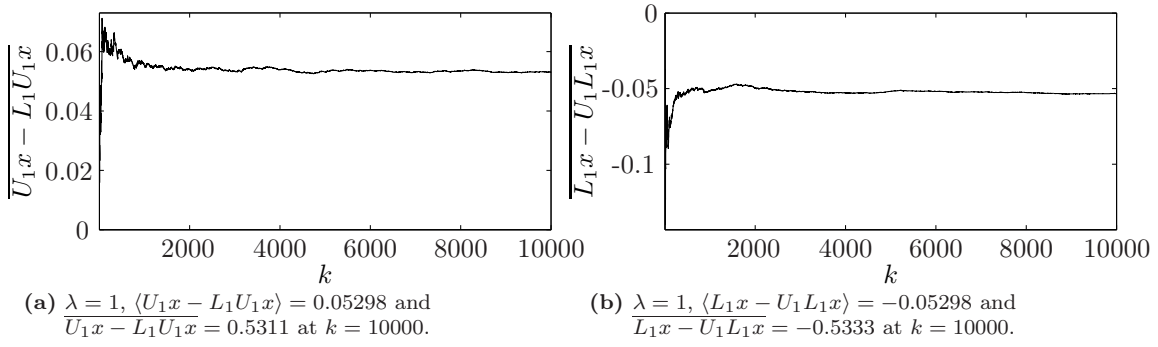
At each trial  $k$  a data point is generated according to the distribution  $t$  and our data sequence grows. Let us denote this sequence by a prime, namely  $x' = \{x_i\}_{i=1}^k$ . For the purpose of the running window of the smoothers  $L_1$  and  $U_1$ , we need a choice of suitable end-values. For example, we can append the median of the data sequence to both sides as follows  $x = \{\text{median}(x'), x_1, x_2, \dots, x_k, \text{median}(x')\}$ . Other methods include for example to replicate the end values as follows  $x = \{x_1, x_1, x_2, \dots, x_k, x_k\}$ , or to add zeros to both sides of the sequence. There is no prescribed way to deal with end values but certain choices seem more justified depending on the particular application. When calculating averages like those in Eq. (2.76) the particular choice of end values becomes less relevant the more data points we have.

The method we shall be using in the remainder of this chapter is called the *omit end-values* rule<sup>1</sup> [12]. At each trial  $k \geq 3$  the sequence is decomposed into the relevant pulses of the first resolution level and an average of these pulses is calculated. The results below (Figs. 2.4 and 2.5) show that the averages gravitate toward the expected values for this particular distribution.



**Figure 2.4:** The averages  $\overline{x - L_1x}$  and  $\overline{x - U_1x}$  for the triangular distribution.

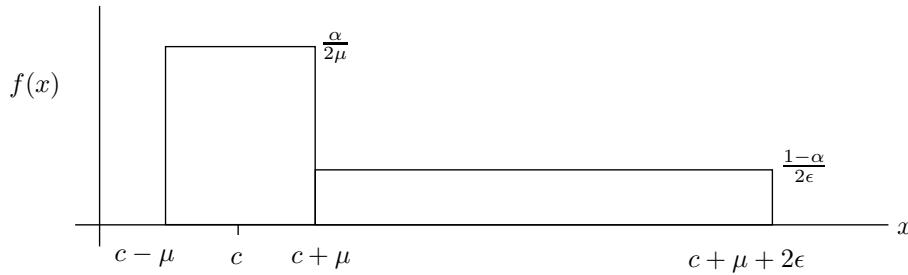
<sup>1</sup>Here one begins smoothing at those original observations for which the full window size applies. In other words, we can begin to smooth at trial  $k = 3$  when the data sequence is big enough to accommodate the size of the running window for  $L_1$  and  $U_1$ . The actual data  $x_1$  and  $x_3$  then performs the role of end-values. At trial  $k = j$  the sequence to be smoothed is  $\{x_1, x_2, \dots, x_{j-1}, x_j\}$  where the points  $x_1$  and  $x_j$  serve as end-values and the pulses (such as  $x - L_1x$ ) are attained from the smoothed sequence and an accordingly reduced data set.



**Figure 2.5:** The averages  $\overline{U_1x - L_1U_1x}$  and  $\overline{L_1x - U_1L_1x}$  for the triangular distribution.

### 2.4.2 Example 2: Mixture model

Consider a model  $f(x) = \alpha g(x) + (1 - \alpha)h(x)$ ,  $\alpha \in [0, 1]$ , which is a convex combination of two uniform distributions  $g$  and  $h$  with scale parameters  $\mu$  and  $\epsilon$  (see Fig. 2.6 below).



**Figure 2.6:** Mixture model of two uniform distributions.

Let

$$g(x) = \begin{cases} \frac{1}{2\mu} & x \in [-\mu, \mu] \\ 0 & \text{elsewhere} \end{cases}, \quad h(x) = \begin{cases} \frac{1}{2\epsilon} & x \in [\mu, \mu + 2\epsilon] \\ 0 & \text{elsewhere} \end{cases}. \quad (2.77)$$

Their respective cdf's  $G(x) = \int_{-\infty}^x dt g(t)$  and  $H(x) = \int_{-\infty}^x dt h(t)$  are

$$G(x) = \begin{cases} 0 & x < -\mu \\ \frac{x+\mu}{2\mu} & x \in [-\mu, \mu] \\ 1 & x > \mu \end{cases}, \quad H(x) = \begin{cases} 0 & x < \mu \\ \frac{x-\mu}{2\epsilon} & x \in [\mu, \mu + 2\epsilon] \\ 1 & x > \mu + 2\epsilon \end{cases}. \quad (2.78)$$

The cdf of  $f$  is  $F = \alpha G + \beta H$ . We want to calculate  $\int_{-\infty}^{\infty} dx (1 - F(x))F^{2m}(x)$  and  $-\int_{-\infty}^{\infty} dx (1 - R(x))R^{2m}(x)$  for  $m = 1$  and  $m = 2$ , as this gives us  $\langle x - L_1x \rangle$ ,  $\langle x - U_1x \rangle$ ,  $\langle U_1x - L_1U_1x \rangle$  and  $\langle L_1x - U_1L_1x \rangle$ . Beginning with  $\int_{-\infty}^{\infty} dx (1 - F(x))F^{2m}(x)$ , the following integrals are calculated

using the definitions of  $G$  and  $H$  and will be of use shortly:

$$\begin{aligned} \int_{-\mu}^{\mu} dx G^q(x) &= \frac{1}{(2\epsilon)^q} \int_{-\mu}^{\mu} dx (x + \mu)^q, \quad I \ni q \leq 0 \\ &= \frac{2\mu}{q+1} \end{aligned} \quad (2.79)$$

$$\begin{aligned} \int_{\mu}^{\mu+2\epsilon} dx H^q(x) &= \frac{1}{(2\epsilon)^q} \int_{\mu}^{\mu+2\epsilon} dx (x - \mu)^q, \quad I \ni q \leq 0 \\ &= \frac{2\epsilon}{q+1} \end{aligned} \quad (2.80)$$

Observing  $G$  and  $H$ , it will also help to write  $F$  in piecewise notation:

$$F^q = \begin{cases} 0 & x < -\mu \\ (\alpha G)^q & x \in [-\mu, \mu) \\ (\alpha + \beta H)^q & x \in [\mu, \mu + 2\epsilon] \\ 1 & x > \mu + 2\epsilon \end{cases}, \quad I \ni q \leq 1. \quad (2.81)$$

Setting  $m = 1$  and  $m = 2$  we need  $\int_{-\infty}^{\infty} dx F^2(x)$ ,  $\int_{-\infty}^{\infty} dx F^3(x)$ ,  $\int_{-\infty}^{\infty} dx F^4(x)$  and  $\int_{-\infty}^{\infty} dx F^5(x)$ . Starting with  $\int_{-\infty}^{\infty} dx F^3(x)$ , use (2.81) to split the integral:

$$\int_{-\infty}^{\infty} dx F^3(x) = \int_{-\mu}^{\mu} dx (\alpha G(x))^3 + \int_{\mu}^{\mu+2\epsilon} dx (\alpha + \beta H(x))^3. \quad (2.82)$$

After expanding we get

$$\alpha^3 \int_{-\mu}^{\mu} dx G^3(x) + \int_{\mu}^{\mu+2\epsilon} dx (\alpha^3 + 3\alpha^2\beta H(x) + 3\alpha\beta^2 H^2(x) + \beta^3 H^3(x)), \quad (2.83)$$

and using (2.97) and (2.80) in the next step and then gathering terms, the result is

$$\int_{-\infty}^{\infty} dx F^3(x) = \left(\frac{\mu}{2} + \frac{\epsilon}{2}\right)\alpha^3 + \frac{\epsilon}{2}\alpha^2 + \frac{\epsilon}{2}\alpha + \frac{\epsilon}{2}. \quad (2.84)$$

Following the same method of calculation we proceed to calculate  $\int_{-\infty}^{\infty} dx F^2(x)$ . The result is

$$\int_{-\infty}^{\infty} dx F^2(x) = \left(\frac{2\mu}{3} + \frac{2\epsilon}{3}\right)\alpha^2 + \frac{2\epsilon}{3}\alpha + \frac{2\epsilon}{3}. \quad (2.85)$$

Subtracting (2.84) from (2.85) and then gathering terms, the expected value of the isolated positive pulses in  $x - L_1x$  is a third degree polynomial in  $\alpha$ , namely

$$\langle x - L_1x \rangle = \int_{-\infty}^{\infty} dx (1 - F(x))F^2(x) = \left(-\frac{\mu}{2} - \frac{\epsilon}{2}\right)\alpha^3 + \left(\frac{2\mu}{3} + \frac{\epsilon}{6}\right)\alpha^2 + \frac{\epsilon}{6}\alpha + \frac{\epsilon}{6}. \quad (2.86)$$

By the same strategy, the expected value of the isolated negative pulses in  $L_1x - U_1L_1x$  is given by

$$\begin{aligned} \langle U_1x - L_1U_1x \rangle &= \int_{-\infty}^{\infty} dx (1 - F(x))F^4(x) \\ &= -\left(\frac{1}{3}\mu + \frac{1}{3}\epsilon\right)\alpha^5 + \left(\frac{2}{5}\mu + \frac{1}{15}\epsilon\right)\alpha^4 + \frac{1}{15}\epsilon\alpha^3 + \frac{1}{15}\epsilon\alpha^2 + \frac{1}{15}\epsilon\alpha + \frac{1}{15}\epsilon \end{aligned} \quad (2.87)$$

To obtain the average of the isolated negative pulses requires a similar calculation, except we consider  $f(-t) = \alpha g(-t) + \beta h(-t)$  or  $f = \alpha g + \beta w$  with  $w = h(-t)$  and  $g = g(-t)$  is symmetric. Therefore

$$w(x) = \begin{cases} \frac{1}{2\epsilon} & x \in [-\mu - 2\epsilon, -\mu] \\ 0 & \text{elsewhere} \end{cases}, \quad (2.88)$$

and its cdf  $W = \int_{-\infty}^x dt w(t)$  is

$$W(x) = \begin{cases} 0 & x < -\mu - 2\epsilon \\ \frac{x + \mu + 2\epsilon}{2\epsilon} & x \in [-\mu - 2\epsilon, -\mu] \\ 1 & x > -\mu \end{cases}. \quad (2.89)$$

The cdf of  $f = \alpha g + \beta w$  is  $R = \alpha G + \beta W$ , and we want to calculate  $-\int_{-\infty}^{\infty} dx (1 - R(x))R^{2m}(x)$  for  $m = 1$  and  $m = 2$ . The following integral is useful:

$$\begin{aligned} \int_{-\mu-2\epsilon}^{-\mu} dx W^q(x) &= \frac{1}{(2\epsilon)^q} \int_{-\mu-2\epsilon}^{-\mu} dx (x + \mu + 2\epsilon)^q, \quad I \ni q \leq 0 \\ &= \frac{2\epsilon}{q+1}. \end{aligned} \quad (2.90)$$

Observing  $G$  and  $W$ , it will help to write  $R$  in piecewise notation:

$$R^q = \begin{cases} 0 & x < -\mu - 2\epsilon \\ (\beta W)^q & x \in [-\mu - 2\epsilon, -\mu] \\ (\beta + \alpha G)^q & x \in [-\mu, \mu] \\ 1 & x > \mu \end{cases}, \quad I \ni q \leq 1. \quad (2.91)$$

Setting  $m = 1$  and  $m = 2$  we need  $\int_{-\infty}^{\infty} dx R^2(x)$ ,  $\int_{-\infty}^{\infty} dx R^3(x)$ ,  $\int_{-\infty}^{\infty} dx R^4(x)$  and  $\int_{-\infty}^{\infty} dx R^5(x)$ . Starting with  $\int_{-\infty}^{\infty} dx R^3(x)$ , use (2.91) to split the integral up:

$$\int_{-\infty}^{\infty} dx R^3(x) = \int_{-\mu-2\epsilon}^{-\mu} dx (\beta W(x))^3 + \int_{-\mu}^{\mu} dx (\beta + \alpha G(x))^3. \quad (2.92)$$

After expanding we get

$$\beta^3 \int_{-\mu-2\epsilon}^{-\mu} dx W^3(x) + \int_{-\mu}^{\mu} dx (\beta^3 + 3\alpha\beta^2 G(x) + 3\alpha^2\beta G^2(x) + \alpha^3 G^3(x)), \quad (2.93)$$

and using (2.79) and (2.90) in the next step and then gathering terms, the result is

$$\int_{-\infty}^{\infty} dx R^3(x) = \left(-\frac{\mu}{2} - \frac{\epsilon}{2}\right)\alpha^3 + \left(2\mu + \frac{3\epsilon}{2}\right)\alpha^2 + \left(-3\mu - \frac{3\epsilon}{2}\right)\alpha + 2\mu + \frac{\epsilon}{2}. \quad (2.94)$$

Following the same method of calculation we proceed to calculate  $\int_{-\infty}^{\infty} dx R^2(x)$ . The result is

$$\int_{-\infty}^{\infty} dx R^2(x) = \left(\frac{2\mu}{3} + \frac{2\epsilon}{3}\right)\alpha^2 + \left(-2\mu - \frac{4\epsilon}{3}\right)\alpha + 2\mu + \frac{2\epsilon}{3}. \quad (2.95)$$

Subtracting (2.95) from (2.94) and then gathering terms, the expectation of the isolated negative pulses in  $x - U_1 x$  is also third degree polynomial in  $\alpha$ , namely

$$\langle x - U_1 x \rangle = - \int_{-\infty}^{\infty} dx (1 - R(x))R^2(x) = \left(-\frac{\mu}{2} - \frac{\epsilon}{2}\right)\alpha^3 + \left(\frac{4\mu}{3} + \frac{5\epsilon}{6}\right)\alpha^2 + \left(-\mu - \frac{\epsilon}{6}\right)\alpha - \frac{\epsilon}{6}. \quad (2.96)$$



By the same strategy the expectation of the negative pulses in  $L_1x - U_1L_1x$  is

$$\begin{aligned}
 & \langle L_1x - U_1L_1x \rangle \\
 &= - \int_{-\infty}^{\infty} dx (1 - R(x))R^4(x) \\
 &= - \left( \frac{1}{3}\mu + \frac{1}{3}\epsilon \right) \alpha^5 + \left( \frac{8}{5}\mu + \frac{19}{15}\epsilon \right) \alpha^4 - \left( 3\mu + \frac{26}{15}\epsilon \right) \alpha^3 + \left( \frac{8}{3}\mu + \frac{14}{15}\epsilon \right) \alpha^2 - \left( \mu + \frac{1}{15}\epsilon \right) \alpha - \frac{1}{15}\epsilon.
 \end{aligned} \tag{2.97}$$

Notice that when we set  $\alpha = 1$  or  $\alpha = 0$ , our current model reduces to a single uniform distribution of length  $2\mu$  or  $2\epsilon$ . Setting  $\alpha = 1$ , our expressions reduce to

$$\langle x - L_1x \rangle = -\langle x - U_1x \rangle = \frac{\mu}{6} = 0.1667\mu, \tag{2.98}$$

and

$$\langle U_1x - L_1U_1x \rangle = -\langle L_1x - U_1L_1x \rangle = \frac{\mu}{15} = 0.06667\mu. \tag{2.99}$$

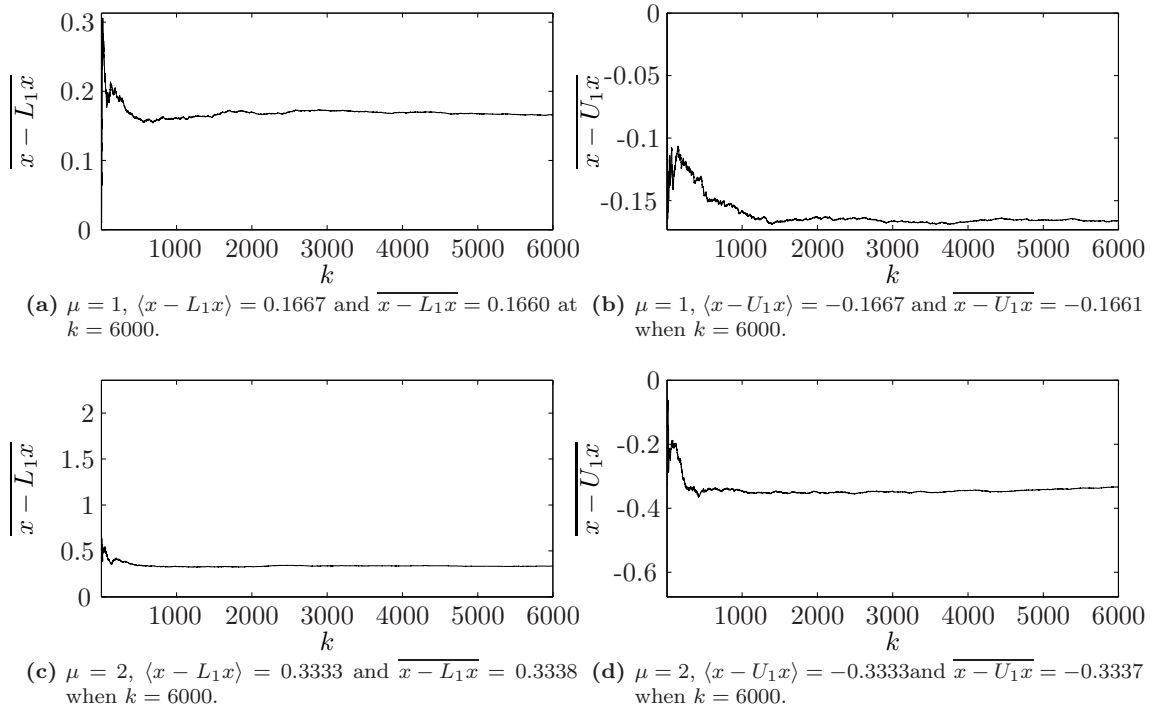
Setting  $\alpha = 0$ , we get

$$\langle x - L_1x \rangle = -\langle x - U_1x \rangle = \frac{\epsilon}{6} = 0.1667\epsilon, \tag{2.100}$$

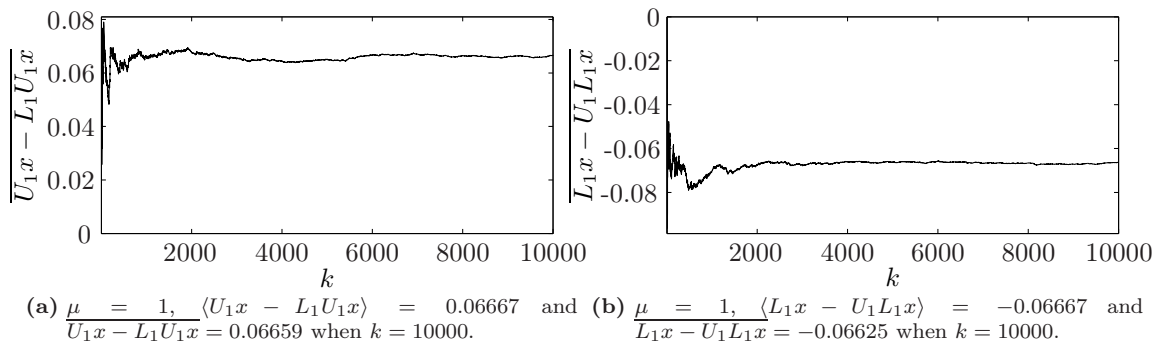
and

$$\langle U_1x - L_1U_1x \rangle = -\langle L_1x - U_1L_1x \rangle = \frac{\epsilon}{15} = 0.06667\epsilon. \tag{2.101}$$

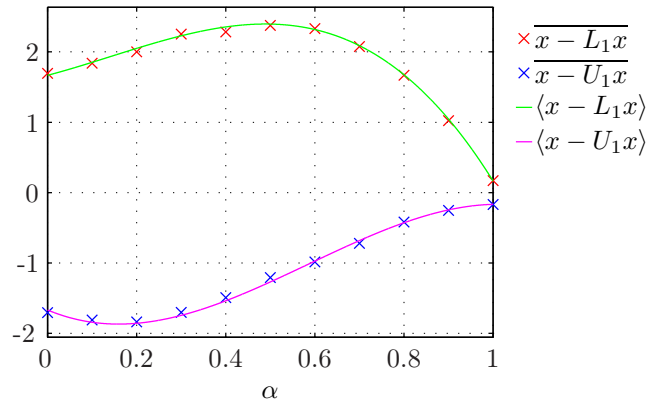
Simulation confirm the above results for  $\alpha = 1$  as well as the polynomials derived for  $0 \leq \alpha \leq 1$  (see Figs. 2.7-2.10 below).



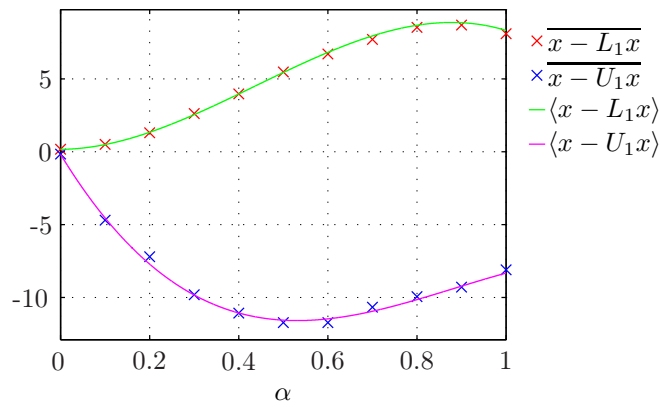
**Figure 2.7:** The averages  $\overline{x - L_1x}$  and  $\overline{x - U_1x}$  for the uniform distribution.



**Figure 2.8:** The averages  $\overline{U_1x - L_1U_1x}$  and  $\overline{L_1x - U_1L_1x}$  for the uniform distribution.

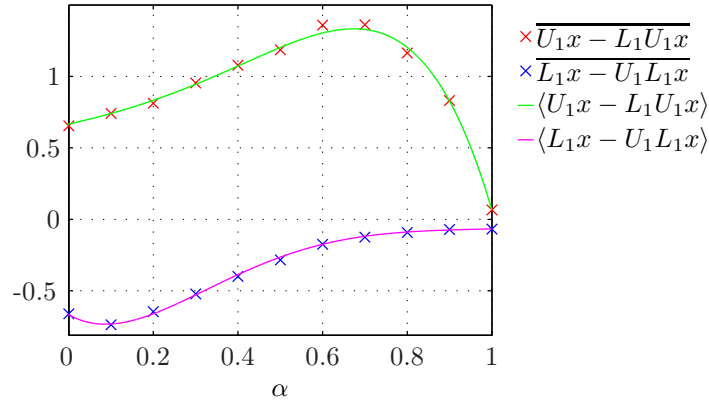
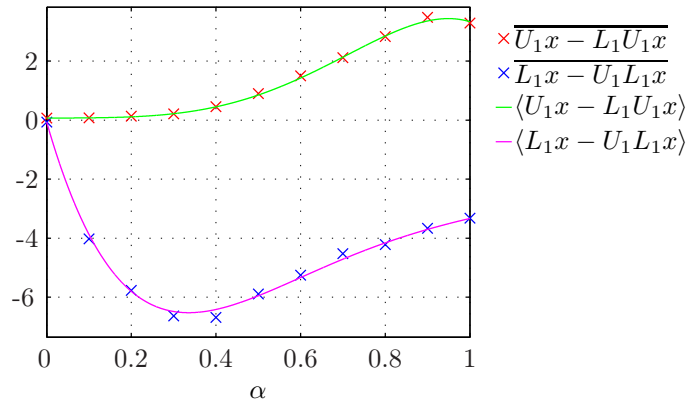


(a)  $\mu = 1, \epsilon = 10$ .



(b)  $\mu = 50, \epsilon = 1$ .

**Figure 2.9:** Expected values and long term averages of  $x - L_1x$  and  $x - U_1x$ . For increasing steps of  $\alpha$ , the averages  $\overline{x - L_1x}$  and  $\overline{x - U_1x}$  are calculated for  $k = 1 \times 10^4$  data points.  $\langle x - L_1x \rangle$  and  $\langle x - U_1x \rangle$  as functions of  $\alpha$  fit the long term averages calculated for different  $\alpha$ .


 (a)  $\mu = 1, \epsilon = 10$ .

 (b)  $\mu = 50, \epsilon = 1$ .

**Figure 2.10:** Expected values and long term averages of  $U_1x - L_1U_1x$  and  $L_1x - U_1L_1x$ . For increasing steps of  $\alpha$ , the averages  $\overline{U_1x - L_1U_1x}$  and  $\overline{L_1x - U_1L_1x}$  are calculated for  $k = 1 \times 10^4$  data points.  $\langle U_1x - L_1U_1x \rangle$  and  $\langle L_1x - U_1L_1x \rangle$  as functions of  $\alpha$  fit the long term averages calculated for different  $\alpha$ .

## 2.5 Designing an estimator

Using numerical simulations, it was shown that the averages of the pulses in  $x - L_1x$  and  $x - U_1x$ , namely  $\overline{x - L_1x}$  and  $\overline{x - U_1x}$ , tend toward their expected values  $\langle x - L_1x \rangle$  and  $\langle x - U_1x \rangle$  as the amount of data points  $k$  was increased. Similarly  $\overline{U_1x - L_1U_1x}$  and  $\overline{L_1x - U_1L_1x}$  was shown to go toward their expected values  $\langle U_1x - L_1U_1x \rangle$  and  $\langle L_1x - U_1L_1x \rangle$ . To summarise, one may write

$$\begin{aligned} \overline{x - L_1x} &\sim \langle x - L_1x \rangle & \overline{x - U_1x} &\sim \langle x - U_1x \rangle \\ \overline{U_1x - L_1U_1x} &\sim \langle U_1x - L_1U_1x \rangle & \overline{L_1x - U_1L_1x} &\sim \langle L_1x - U_1L_1x \rangle. \end{aligned} \quad (2.102)$$

We are now in a position to develop ad hoc estimators for the parameter of the noise distribution given the pulses in the first resolution level of a DPT.

### 2.5.1 Triangular distribution

Drawing on the findings of Example 1 in section 2.4.1 where the triangular distribution was used, it was shown that the expected value of the pulses in  $x - L_1x$  and  $x - U_1x$  is  $\langle x - L_1x \rangle = \frac{7}{60}\lambda$  and  $\langle x - U_1x \rangle = -\frac{7}{60}\lambda$  respectively. Further more the expected value of the pulses in  $U_1x - L_1U_1x$  and  $L_1x - U_1L_1x$  is  $\langle U_1x - L_1U_1x \rangle = \frac{89}{1680}\lambda$  and  $\langle L_1x - U_1L_1x \rangle = \frac{89}{1680}\lambda$  respectively. One can use the positive or negative pulses of a decomposition  $D_1$  or  $R_1$  separately, for example, since  $\langle x - L_1x \rangle = \frac{7}{60}\lambda$ , from (2.102) we have an estimate for  $\lambda$  using only the positive pulses of  $R_1$ :

$$(\lambda)_{est} = \frac{60}{7} \overline{x - L_1x}. \quad (2.103)$$

However, since  $\frac{1}{3}$  of the total number of pulses in  $R_1$  are expected to be in  $(x - L_1x)$ , and  $\frac{1}{5}$  in  $(L_1x - U_1L_1x)$  one can use both  $\overline{x - L_1x}$  and  $\overline{L_1x - U_1L_1x}$  and weigh them as follows:

$$(\lambda)_{est} = w_1 \frac{60}{7} \overline{x - L_1x} - w_2 \frac{1680}{89} \overline{L_1x - U_1L_1x}, \quad (2.104)$$

where

$$w_1 = \frac{\frac{1}{3}}{\frac{1}{3} + \frac{1}{5}} = \frac{5}{8} \text{ and } w_2 = \frac{\frac{1}{5}}{\frac{1}{3} + \frac{1}{5}} = \frac{3}{8}. \quad (2.105)$$

Another appropriate choice is

$$w_1 = \frac{\text{number of pulses in } x - L_1x}{\text{total number of pulses in } R_1} \text{ and } w_2 = \frac{\text{number of pulses in } L_1x - U_1L_1x}{\text{total number of pulses in } R_1}, \quad (2.106)$$

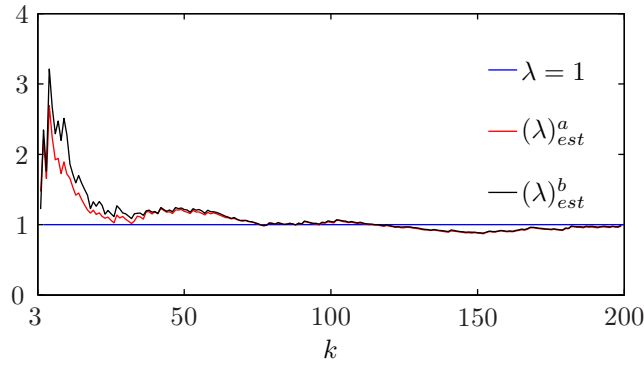
but since these quickly converge to  $\frac{5}{8}$  and  $\frac{3}{8}$  rather quickly, we opt to use the former weighting (2.105) in simulations that follow. If one were to use the pulses in the dual decomposition  $D_1$  the estimator would be

$$(\lambda)_{est} = -w_1 \frac{60}{7} \overline{x - U_1x} + w_2 \frac{1680}{89} \overline{U_1x - L_1U_1x}, \quad (2.107)$$

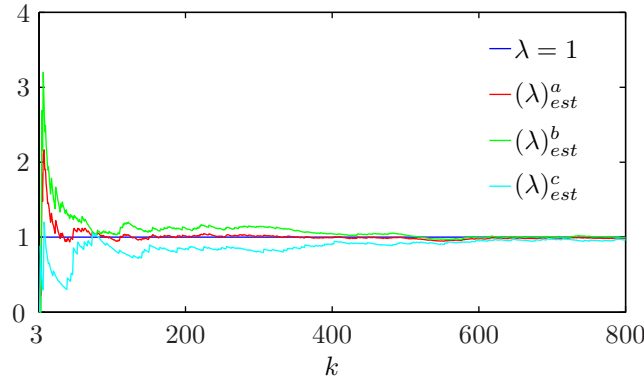
where

$$w_1 = \frac{\text{number of pulses in } x - U_1x}{\text{total number of pulses in } D_1} \text{ and } w_2 = \frac{\text{number of pulses in } U_1x - L_1U_1x}{\text{total number of pulses in } D_1}, \quad (2.108)$$

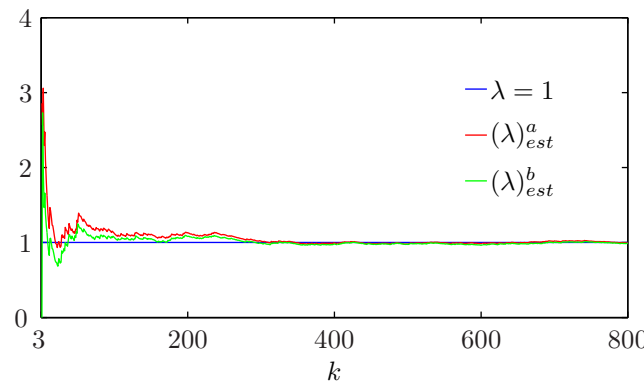
or as chosen in (2.105) above.



(a) Examining the estimator as defined in Eq. (2.107), the weighting coefficients  $w_1$  and  $w_2$  for  $(\lambda)_{est}^a$  and  $(\lambda)_{est}^b$  were chosen as in Eq. (2.105) and Eq. (2.106) respectively. Observing many such plots, both choices display periods of superiority over the other, but as  $w_1$  and  $w_2$  of Eq. (2.106) tend toward  $\frac{5}{8}$  and  $\frac{3}{8}$  respectively,  $(\lambda)_{est}^a$  and  $(\lambda)_{est}^b$  become indistinguishable.



(b) Here  $(\lambda)_{est}^a$  is defined as in Eq. (2.103),  $(\lambda)_{est}^b$  is defined as in Eq. (2.104), and  $(\lambda)_{est}^c = -\frac{1680}{89} L_1 x - U_1 L_1 x$ . The estimator  $(\lambda)_{est}^b$  makes use of the positive pulses in  $R_1$ ,  $(\lambda)_{est}^c$  uses only the negative pulses, and  $(\lambda)_{est}^a$  uses both the positive and negative pulses with the weighting  $w_1 = \frac{5}{8}$  and  $w_2 = \frac{3}{8}$ .



(c)  $(\lambda)_{est}^a$  is defined as in Eq. (2.104) and  $(\lambda)_{est}^b$  is defined as in Eq. (2.107). The estimator  $(\lambda)_{est}^a$  uses the pulses in  $R_1$  while  $(\lambda)_{est}^b$  uses the pulses in  $D_1$ . Depending on exactly which data sequence you draw, each will appear superior over the other.

**Figure 2.11:** Scale parameter estimators for the triangular distribution.

### 2.5.2 Uniform distribution

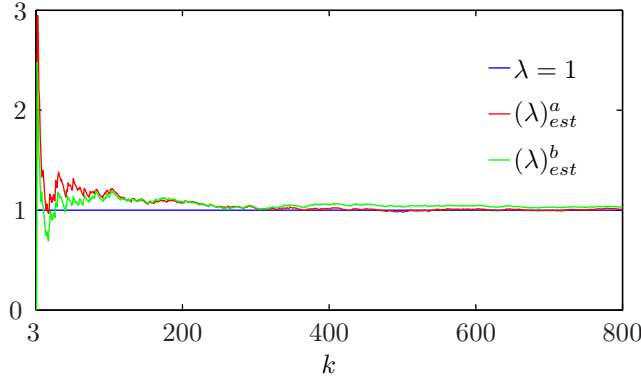
For the uniform distribution the above estimators would be

$$(\mu)_{est} = w_1 \overline{6x - L_1x} - w_2 \overline{15L_1x - U_1L_1x}, \quad (2.109)$$

and

$$(\mu)_{est} = -w_1 \overline{6x - U_1x} + w_2 \overline{15U_1x - L_1U_1x}, \quad (2.110)$$

with  $w_1$  and  $w_2$  selected appropriately.



**Figure 2.12:**  $(\lambda)_{est}^a$  is defined as in Eq. (2.109) and  $(\lambda)_{est}^b$  is defined as in Eq. (2.110). The estimator  $(\lambda)_{est}^a$  uses the pulses in  $R_1$  whilst  $(\lambda)_{est}^b$  uses the pulses in  $D_1$ .

### 2.5.3 Mixture model

For the model  $f = \alpha g + (1 - \alpha)h$  of section 2.4.2 there are two parameters to be estimated. Suppose  $\alpha$  is known and consider the decomposition  $R_1 = (x - L_1x) + (L_1x - U_1L_1x)$ . Consulting Eq. (2.86) and Eq. (2.97), rewrite  $\langle x - L_1x \rangle$  and  $\langle L_1x - U_1L_1x \rangle$  as follows

$$\langle x - L_1x \rangle = \underbrace{\left(-\frac{1}{2}\alpha^3 + \frac{2}{3}\alpha^2\right)}_{=\rho_{11}}\mu + \underbrace{\left(-\frac{1}{2}\alpha^3 + \frac{1}{6}\alpha^2 + \frac{1}{6}\alpha + \frac{1}{6}\right)}_{=\rho_{12}}\epsilon \quad (2.111)$$

$$\langle L_1x - U_1L_1x \rangle = \underbrace{\left(-\frac{1}{3}\alpha^5 + \frac{8}{5}\alpha^4 - 3\alpha^3 + \frac{8}{3}\alpha^2 - \alpha\right)}_{=\eta_{21}}\mu + \underbrace{\left(-\frac{1}{3}\alpha^5 + \frac{19}{15}\alpha^4 - \frac{26}{15}\alpha^3 + \frac{14}{15}\alpha^2 - \frac{1}{15}\alpha - \frac{1}{15}\right)}_{=\eta_{22}}\epsilon, \quad (2.112)$$

and introduce  $\rho_{11}$ ,  $\rho_{12}$ ,  $\eta_{21}$  and  $\eta_{22}$  as indicated. The above in matrix notation is

$$\begin{bmatrix} \langle x - L_1x \rangle \\ \langle L_1x - U_1L_1x \rangle \end{bmatrix} = \begin{bmatrix} \rho_{11} & \rho_{12} \\ \eta_{21} & \eta_{22} \end{bmatrix} \begin{bmatrix} \mu \\ \epsilon \end{bmatrix}. \quad (2.113)$$

The solution for  $\mu$  and  $\epsilon$  is thus

$$\begin{bmatrix} \mu \\ \epsilon \end{bmatrix} = \frac{1}{\rho_{11}\eta_{22} - \rho_{12}\eta_{21}} \begin{bmatrix} \eta_{22} & -\rho_{12} \\ -\eta_{21} & \rho_{11} \end{bmatrix} \begin{bmatrix} \langle x - L_1x \rangle \\ \langle L_1x - U_1L_1x \rangle \end{bmatrix}, \quad (2.114)$$

and an estimate for the two noise parameters can thus be constructed as

$$\begin{bmatrix} (\mu)_{est} \\ (\epsilon)_{est} \end{bmatrix} = \frac{1}{\rho_{11}\eta_{22} - \rho_{12}\eta_{21}} \begin{bmatrix} \eta_{22} & -\rho_{12} \\ -\eta_{21} & \rho_{11} \end{bmatrix} \begin{bmatrix} \overline{x - L_1x} \\ \overline{L_1x - U_1L_1x} \end{bmatrix}. \quad (2.115)$$

For the dual decomposition  $D_1 = (I - U_1)x + (I - L_1)U_1x$ , take Eq. (2.96) and Eq. (2.87) and rewrite  $\langle x - U_1x \rangle$  and  $\langle U_1x - L_1U_1x \rangle$  as follows

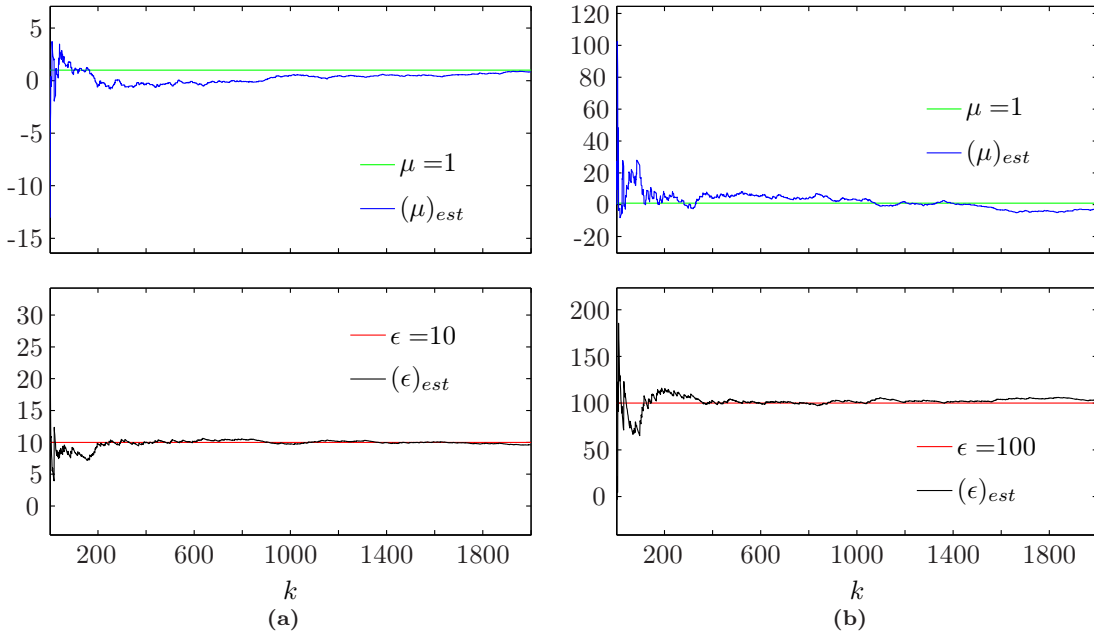
$$\langle x - U_1x \rangle = \underbrace{\left(-\frac{1}{2}\alpha^3 + \frac{4}{3}\alpha^2 - \alpha\right)}_{=\eta_{11}}\mu + \underbrace{\left(-\frac{1}{2}\alpha^3 + \frac{5}{6}\alpha^2 - \frac{1}{6}\alpha - \frac{1}{6}\right)}_{=\eta_{12}}\epsilon \quad (2.116)$$

$$\langle U_1x - L_1U_1x \rangle = \underbrace{\left(-\frac{1}{3}\alpha^5 + \frac{2}{5}\alpha^4\right)}_{=\rho_{21}}\mu + \underbrace{\left(-\frac{1}{3}\alpha^5 + \frac{1}{15}\alpha^4 + \frac{1}{15}\alpha^3 + \frac{1}{15}\alpha^2 + \frac{1}{15}\alpha + \frac{1}{15}\right)}_{=\rho_{22}}\epsilon, \quad (2.117)$$

introducing  $\eta_{11}$ ,  $\eta_{12}$ ,  $\rho_{21}$  and  $\rho_{22}$  as shown. An estimator for the two noise parameters is then

$$\begin{bmatrix} (\mu)_{est} \\ (\epsilon)_{est} \end{bmatrix} = \frac{1}{\eta_{11}\rho_{22} - \eta_{12}\rho_{21}} \begin{bmatrix} \rho_{22} & -\eta_{12} \\ -\rho_{21} & \eta_{11} \end{bmatrix} \begin{bmatrix} \overline{x - U_1x} \\ \overline{U_1x - L_1U_1x} \end{bmatrix}. \quad (2.118)$$

The performance of the estimators for  $R_1$  as given in Eq. (2.114) are tested in numerical simulation for varying values of  $\alpha$  and different choices of the parameters  $\mu$  and  $\lambda$ . The results are displayed in the Figs. 2.13-2.16 below. In each figure the values of the parameters that were used in the simulation and which are to be estimated are indicated by straight lines.



**Figure 2.13:** The estimators  $(\mu)_{est}$  and  $(\lambda)_{est}$  as indicated by Eq. (2.114) for  $\alpha = 0.5$  and different choices of the parameters.



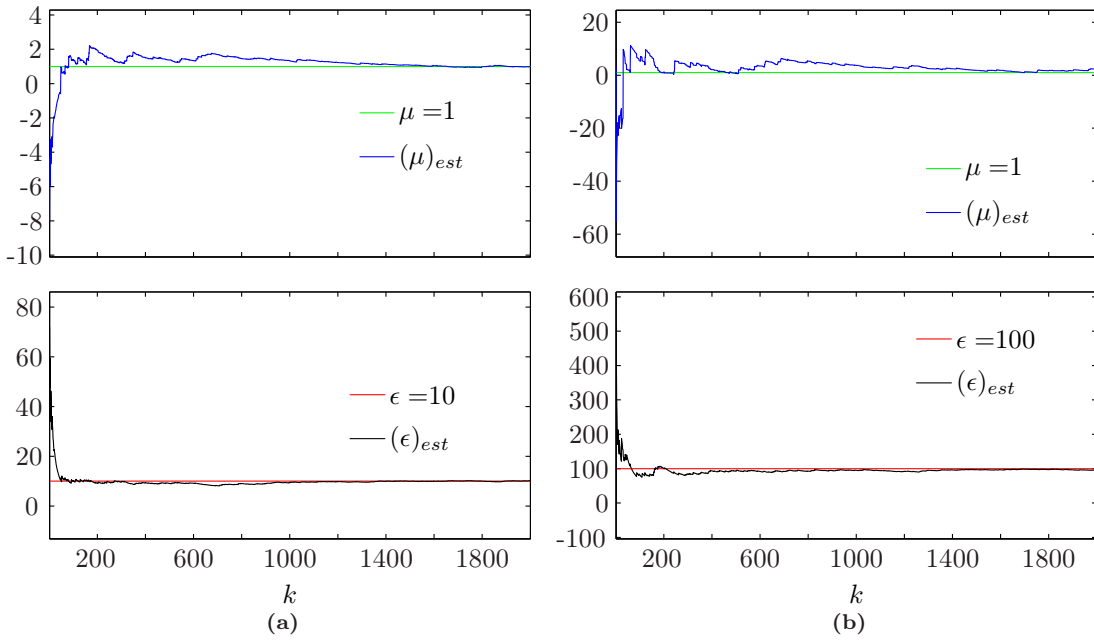


Figure 2.14: The estimators  $(\mu)_{est}$  and  $(\lambda)_{est}$  as indicated by Eq. (2.114) for  $\alpha = 0.8$  and different choices of the parameters.

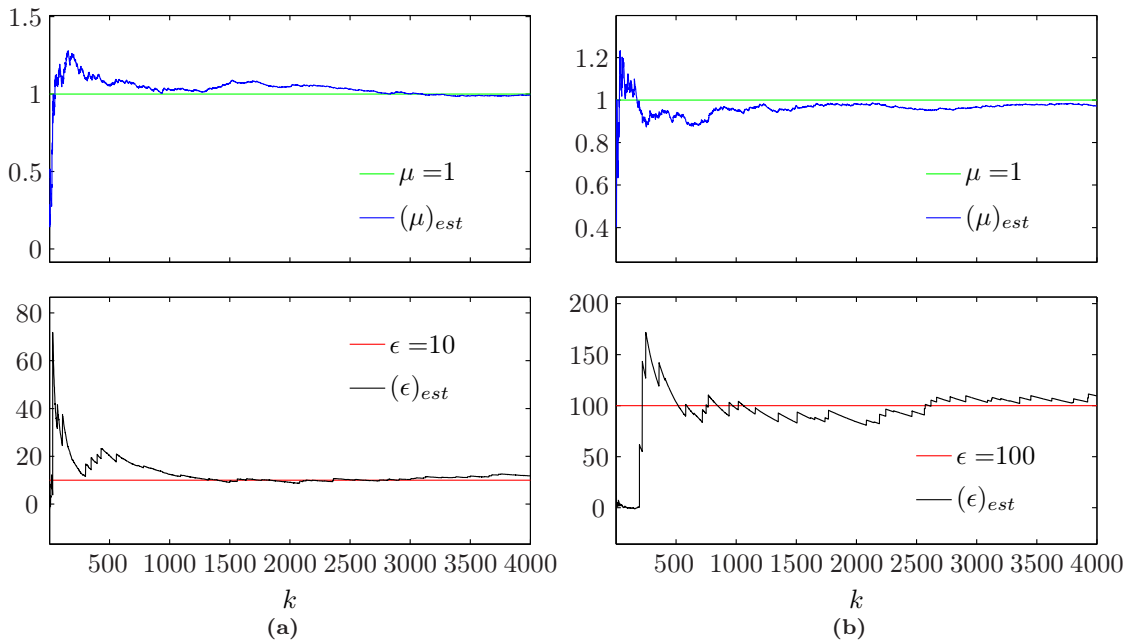
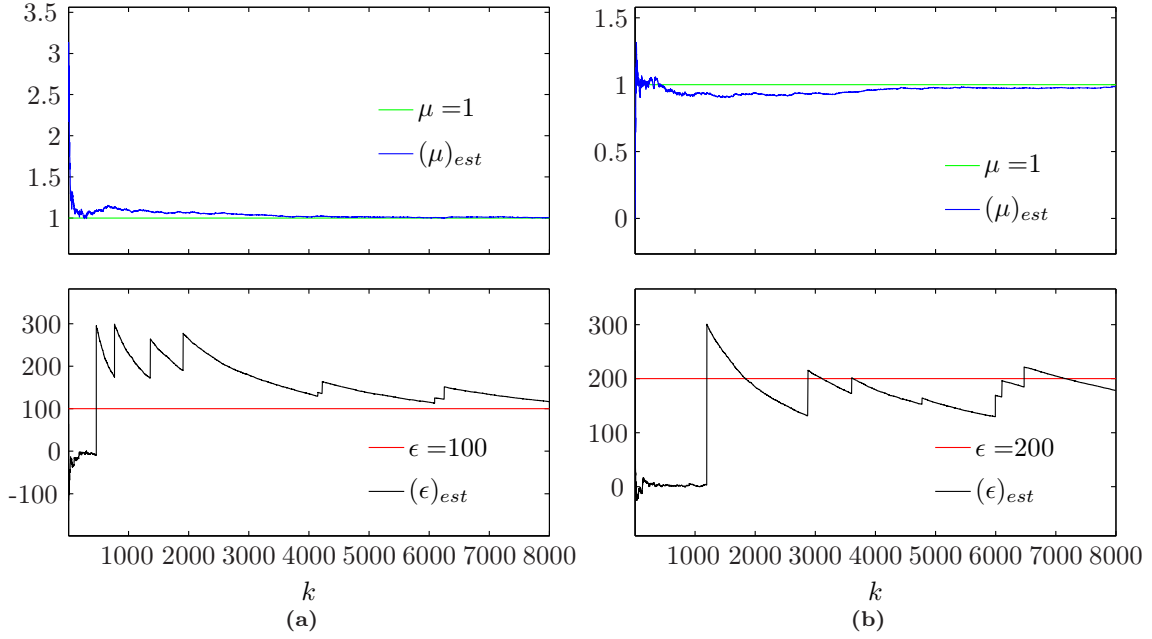


Figure 2.15: The estimators  $(\mu)_{est}$  and  $(\lambda)_{est}$  as indicated by Eq. (2.114) for  $\alpha = 0.99$  and different choices of the parameters.



**Figure 2.16:** The estimators  $(\mu)_{est}$  and  $(\lambda)_{est}$  as indicated by Eq. (2.114) for  $\alpha = 0.999$  and different choices of the parameters.

Notice that although the model does not allow for it, the estimates can be negative. This is an example of the folly of ad hoc devices, and highlights what can happen when one sets an average equal to an expectation value. From (2.113), demanding that  $\mu > 0$  and  $\epsilon > 0$  one finds the inequality

$$\frac{\rho_{12}}{\eta_{22}} < \frac{\langle x - L_1 x \rangle}{\langle L_1 x - U_1 L_1 x \rangle} < \frac{\rho_{11}}{\eta_{21}}. \quad (2.119)$$

A curious result which says that the ratio of the expected pulse heights of positive and negative pulses making up  $R_1$  is bound between two functions of the break down probability  $\alpha$ . Both estimates of  $\mu$  and  $\epsilon$  will thus be positive if

$$\frac{\rho_{12}}{\eta_{22}} < \frac{\overline{x - L_1 x}}{\overline{L_1 x - U_1 L_1 x}} < \frac{\rho_{11}}{\eta_{21}}, \quad (2.120)$$

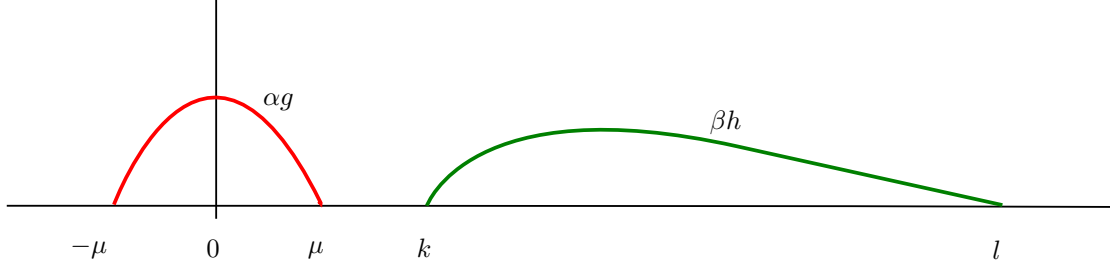
where  $(\mu)_{est} < 0$  when  $\frac{\overline{x - L_1 x}}{\overline{L_1 x - U_1 L_1 x}} < \frac{\rho_{12}}{\eta_{22}}$  and  $(\epsilon)_{est} < 0$  when  $\frac{\rho_{11}}{\eta_{21}} < \frac{\overline{x - L_1 x}}{\overline{L_1 x - U_1 L_1 x}}$ . The reason the estimates can be negative is thus simply because there was not enough data yet for the averages to have converged to their expected values. For the dual decomposition  $D_1$ , the equivalent expression of interest is

$$\frac{\eta_{12}}{\rho_{22}} < \frac{\overline{x - U_1 x}}{\overline{U_1 x - L_1 U_1 x}} < \frac{\eta_{11}}{\rho_{21}}, \quad (2.121)$$

where  $(\mu)_{est} < 0$  when  $\frac{\overline{x - U_1 x}}{\overline{U_1 x - L_1 U_1 x}} < \frac{\eta_{12}}{\rho_{22}}$  and  $(\epsilon)_{est} < 0$  when  $\frac{\eta_{11}}{\rho_{21}} < \frac{\overline{x - U_1 x}}{\overline{U_1 x - L_1 U_1 x}}$ .

## 2.6 Analysis of smoothed sequences in the context of outliers

Consider the example where a sequence is drawn from the mixture of two distributions  $g$  and  $h$  with non-overlapping support as shown in Fig. 2.17 below.



**Figure 2.17:** Mixture model with non-overlapping support  $\mu < k$ . The distribution  $g$  is symmetric about zero and  $h$  can be unsymmetrical but has its left edge at  $k$ .

The combined distribution is

$$f = \alpha g + \beta h, \quad (2.122)$$

where we choose to call  $\beta = (1 - \alpha)$  the break-down probability and

$$F = \alpha G + \beta H \quad (2.123)$$

is the cdf of  $f$ . Let us calculate the expectation value of the negative pulses in  $R_1 x = (I - U_1 L_1)x$ , namely

$$\langle x - L_1 x \rangle = \int_{-\infty}^{\infty} dt F^2 (1 - F) = \int_{-\infty}^{\infty} dt (F^2 - F^3). \quad (2.124)$$

With

$$F^2 = \alpha^2 G^2 + 2\alpha\beta GH + \beta^2 H^2, \quad (2.125)$$

and

$$F^3 = \alpha^3 G^3 + 3\alpha^2\beta G^2 H + 3\alpha\beta^2 GH^2 + \beta^3 H^3, \quad (2.126)$$

we need

$$\int_{-\infty}^{\infty} dt (\alpha^2 G^2 + 2\alpha\beta GH + \beta^2 H^2 - \alpha^3 G^3 - 3\alpha^2\beta G^2 H - 3\alpha\beta^2 GH^2 - \beta^3 H^3). \quad (2.127)$$

Rearrange the terms as follows

$$\int_{-\infty}^{\infty} dt (\alpha^2 G^2 - \alpha^3 G^3 + \alpha\beta(2G - 3\alpha G^2)H + \beta^2(H^2 - 3\alpha GH^2) - \beta^3 H^3). \quad (2.128)$$

Split the integral as follows

$$\langle x - L_1 x \rangle = \underbrace{\int_{-\mu}^{\mu} dt F^2 (1 - F)}_{=I_1} + \underbrace{\int_{\mu}^k dt F^2 (1 - F)}_{=I_2} + \underbrace{\int_k^l dt F^2 (1 - F)}_{=I_3} \quad (2.129)$$

Since  $H = 0$  in the region  $[-\mu, \mu]$ , the first term is

$$I_1 = \int_{-\mu}^{\mu} dt (\alpha^2 G^2 - \alpha^3 G^3). \quad (2.130)$$

Assuming  $k \geq \mu$  we have that  $G = 1$  and  $H = 0$  in the interval  $[\mu, k]$ , and the second term is thus

$$I_2 = \int_{\mu}^k dt (\alpha^2 - \alpha^3) = \alpha^2 \beta (k - \mu). \quad (2.131)$$

Finally since  $G = 1$  in  $[k, l]$ , the last term is

$$I_3 = \int_k^l dt (\alpha^2 - \alpha^3 + \alpha\beta(2 - 3\alpha)H + \beta^2(1 - 3\alpha)H^2 - \beta^3 H^3). \quad (2.132)$$

Concentrating on the above, note that

$$\begin{aligned} \alpha\beta(2 - 3\alpha)H &= \alpha\beta(2 - 2\alpha - \alpha)H \\ &= 2\alpha\beta(1 - \alpha - \frac{\alpha}{2})H \\ &= 2\alpha\beta(1 - \alpha)H - 2\alpha\beta\frac{\alpha}{2}H, \\ &= 2\alpha\beta^2 H - \alpha^2\beta H \end{aligned} \quad (2.133)$$

and

$$\begin{aligned} \beta^2(1 - 3\alpha)H^2 &= \beta^2(1 - \alpha - 2\alpha)H^2 \\ &= \beta^2(1 - \alpha)H^2 - 2\alpha\beta^2 H^2 \\ &= \beta^3 H^2 - 2\alpha\beta^2 H^2 \end{aligned} \quad (2.134)$$

so that

$$I_3 = \int_k^l dt (\alpha^2 - \alpha^3 - \alpha^2\beta H + 2\alpha\beta^2(1 - H)H + \beta^3(1 - H)H^2). \quad (2.135)$$

Since  $H(t) = \int_k^l dt h(t)$ , and thus  $\frac{d}{dt}H(t) = h(t)$  we can use integration by parts on the expectation of  $h$  as follows

$$\langle h \rangle = \int_k^l dt th(t) = tH(t) \Big|_k^l - \int_k^l dt H(t), \quad (2.136)$$

to obtain the result

$$\begin{aligned} \int_k^l dt H(t) &= \underbrace{lH(l)}_{=1} - \underbrace{kH(k)}_{=0} - \langle h \rangle \\ &= l - \langle h \rangle. \end{aligned} \quad (2.137)$$

Letting

$$\sigma = 2\alpha\beta^2 \int_k^l dt (1 - H)H + \beta^3 \int_k^l dt (1 - H)H^2, \quad (2.138)$$

we thus have

$$\begin{aligned} I_3 &= (\alpha^2 - \alpha^3) \int_k^l dt - \alpha^2\beta \int_k^l dt H + \sigma \\ &= \alpha^2\beta(l - k) - \alpha^2\beta(l - \langle h \rangle) + \sigma \\ &= \alpha^2\beta(\langle h \rangle - k) + \sigma. \end{aligned} \quad (2.139)$$

Putting the terms back together again we have that

$$\begin{aligned}
 \langle x - L_1x \rangle &= I_1 + I_2 + I_3 \\
 &= \int_{-\mu}^{\mu} dt (\alpha^2 G^2 - \alpha^3 G^3) + \alpha^2 \beta (k - \mu) + \alpha^2 \beta (\langle h \rangle - k) + \sigma \\
 &= \int_{-\mu}^{\mu} dt (\alpha^2 G^2 - \alpha^3 G^3) + \alpha^2 \beta (\langle h \rangle - \mu) + \sigma.
 \end{aligned} \tag{2.140}$$

Writing  $\langle x - L_1x \rangle$  in this way reveals an interesting truth which we turn to next. Consider a sequence drawn from  $f$  and operated on by  $L_1$ . The expectation of the smoothed sequence can be written

$$\langle L_1x \rangle = \langle x \rangle - \langle x - L_1x \rangle. \tag{2.141}$$

Assuming  $g$  is centered around zero, we have that

$$\langle x \rangle = 0 + \beta \langle h \rangle, \tag{2.142}$$

and inserting our fancily written expression for  $\langle x - L_1x \rangle$ ,  $\langle L_1x \rangle$  becomes

$$\langle L_1x \rangle = \beta \langle h \rangle - \alpha^2 \beta \langle h \rangle + \alpha^2 \beta \mu - \int_{-\mu}^{\mu} dt (\alpha^2 G^2 - \alpha^3 G^3) - \sigma, \tag{2.143}$$

or

$$\begin{aligned}
 \langle L_1x \rangle &= \beta \langle h \rangle (1 - \alpha^2) + \alpha^2 \beta \mu - \int_{-\mu}^{\mu} dt (\alpha^2 G^2 - \alpha^3 G^3) - \sigma \\
 &= (1 + \alpha) \beta^2 \langle h \rangle + \alpha^2 \beta \mu - \int_{-\mu}^{\mu} dt (\alpha^2 G^2 - \alpha^3 G^3) - \sigma.
 \end{aligned} \tag{2.144}$$

Since  $\beta \approx 0$ , the leading order is  $\beta$  and we can thus write

$$\langle L_1x \rangle = \alpha^2 \beta \mu - \int_{-\mu}^{\mu} dt (\alpha^2 G^2 - \alpha^3 G^3) + O(\beta^2) \tag{2.145}$$

where the least-significant terms are summarized by  $O(\beta^2)$ . The above expression reveals an interesting fact. It says that upon smoothing with  $L_1$ , any contribution to the expectation of  $L_1x$  coming from terms dependent on  $h$  decrease like  $\beta^2$  or higher order. In particular we have shown that the dominant term  $\langle h \rangle$  is reduced to order  $\beta^2$ .

Let us now calculate  $\langle L_1x - U_1L_1x \rangle$ . From our theorem this is given by

$$\begin{aligned}
 \langle L_1x - U_1L_1x \rangle &= - \int_{-\infty}^{\infty} dt F(1 - F)^4 = - \int_{-\infty}^{\infty} dt (1 - R)R^4 \\
 &= \int_{-\infty}^{\infty} dt (R^5 - R^4).
 \end{aligned} \tag{2.146}$$

$R(t)$  is the cdf of  $f(-t) = \alpha g(-t) + \beta h(-t)$ . Since  $g$  is symmetric

$$R = \alpha G + \beta W \tag{2.147}$$

where

$$W(x) = \int_{-\infty}^x dt h(-t). \tag{2.148}$$

Thus

$$R^4 = \alpha^4 G^4 + 4\alpha^3 \beta G^3 W + 6\alpha^2 \beta^2 G^2 W^2 + 4\alpha \beta^3 G W^3 + \beta^4 W^4, \tag{2.149}$$

and

$$R^5 = \alpha^5 G^5 + 5\alpha^4 \beta G^4 W + 10\alpha^3 \beta^2 G^3 W^2 + 10\alpha^2 \beta^3 G^2 W^3 + 5\alpha \beta^4 G W^4 + \beta^5 W^5. \quad (2.150)$$

Split the integral as follows

$$\langle L_1 x - U_1 L_1 x \rangle = \underbrace{\int_{-l}^{-k} dt R^5 - R^4}_{=I_4} + \underbrace{\int_{-k}^{-\mu} dt R^5 - R^4}_{=I_5} + \underbrace{\int_{-\mu}^{\mu} dt R^5 - R^4}_{=I_6}. \quad (2.151)$$

Since  $G = 0$  in the region  $[-l, -k]$ , the first term is

$$I_4 = \int_{-l}^{-k} dt (\beta^5 W^5 - \beta^4 W^4). \quad (2.152)$$

Assuming  $k \geq \mu$ , we have that  $W = 1$  and  $G = 0$  in the interval  $[-k, -\mu]$ , and the second term is thus

$$I_5 = \int_{-k}^{-\mu} dt (\beta^5 - \beta^4) = (\beta^4 - \beta^5)\mu + (\beta^5 - \beta^4)k. \quad (2.153)$$

finally since  $W = 1$  in  $[-\mu, \mu]$ , the last term is

$$I_6 = \int_{-\mu}^{\mu} dt (\alpha^5 G^5 + 5\alpha^4 \beta G^4 + 10\alpha^3 \beta^2 G^3 + 10\alpha^2 \beta^3 G^2 + 5\alpha \beta^4 G + \beta^5 - \alpha^4 G^4 - 4\alpha^3 \beta G^3 - 6\alpha^2 \beta^2 G^2 - 4\alpha \beta^3 G - \beta^4). \quad (2.154)$$

Summarizing the terms of order  $\beta^2$  or higher we thus have that

$$\langle L_1 x - U_1 L_1 x \rangle = \int_{-\mu}^{\mu} dt (\alpha^5 G^5 + 5\alpha^4 \beta G^4 - \alpha^4 G^4 - 4\alpha^3 \beta G^3) + O(\beta^2). \quad (2.155)$$

The expectation of the smoothed sequence  $U_1 L_1 x$  can be written

$$\langle U_1 L_1 x \rangle = \langle L_1 x \rangle - \langle L_1 x - U_1 L_1 x \rangle. \quad (2.156)$$

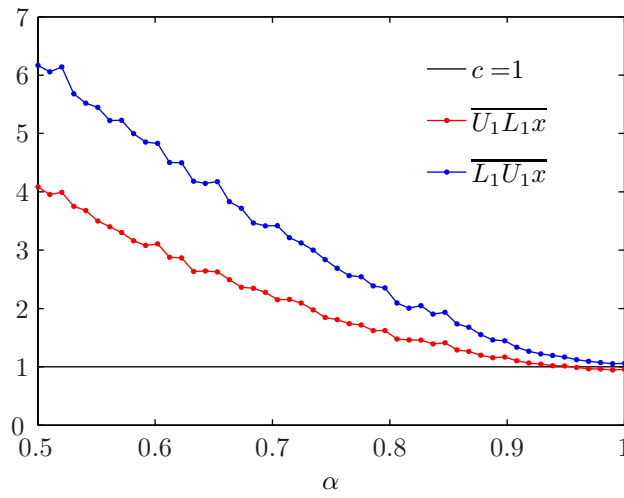
Inserting  $\langle L_1 x \rangle$  and  $\langle L_1 x - U_1 L_1 x \rangle$  we find that

$$\langle U_1 L_1 x \rangle = \alpha^2 \beta \mu - \int_{-\mu}^{\mu} dt (\alpha^2 G^2 - \alpha^3 G^3) - \int_{-\mu}^{\mu} dt (\alpha^5 G^5 + 5\alpha^4 \beta G^4 - \alpha^4 G^4 - 4\alpha^3 \beta G^3) + O(\beta^2). \quad (2.157)$$

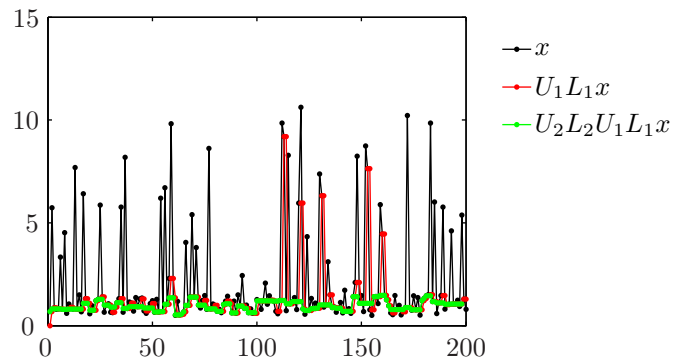
If instead  $\langle g \rangle = c$  and we take the average  $\overline{U_1 L_1 x}$  as our estimate for  $c$  then we can expect that any contribution due to  $h$  is of order  $\beta^2$ . The influence of  $h$  has thus been reduced to order  $\beta^2$  and in this sense  $\overline{U_1 L_1 x}$  can be seen as a robust estimator for the location parameter  $c$  in the context where  $\beta \approx 0$ . Furthermore, under the assumption that the exact  $h$  is relatively unknown due to lack of data, it also becomes less relevant. A similar result can be shown to hold for  $\langle L_1 U_1 x \rangle$ . For the mixture model of section 2.4.2 the above result can be confirmed by simulation (see Fig. 2.18 below). By making use of the fact that

$$L_1 x \leq U_1 L_1 x \leq M_1 x \leq L_1 U_1 x \leq U_1 x, \quad (2.158)$$

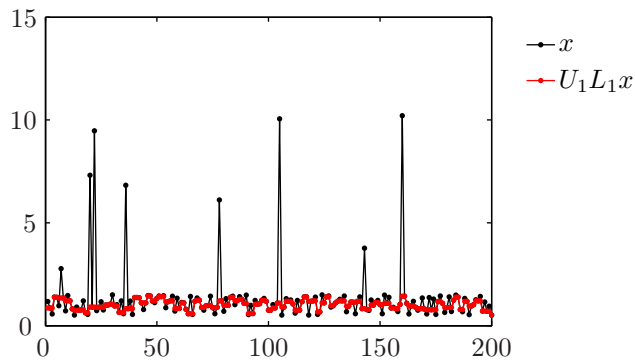
these results are in support of the popular three point running median  $M_1$  where before a supporting theory has been lacking [31, 32, 18].



**Figure 2.18:** For the mixture model of section 2.4.2, the long term averages of  $\overline{U_1 L_1 x}$  and  $\overline{L_1 U_1 x}$  are calculated for varying  $\alpha$ . For each value of the average  $\overline{U_1 L_1 x}$  and  $\overline{L_1 U_1 x}$  a total of  $k = 1 \times 10^4$  data points was used. The parameters were set at  $c = 1$ ,  $\mu = \frac{1}{2}$  and  $\epsilon = 10$ . Observing the averages when  $\alpha$  is close to 1 would suggest that the estimator  $\frac{1}{2}(\overline{U_1 L_1 x} + \overline{L_1 U_1 x})$  would do better than  $\overline{U_1 L_1 x}$  on its own.



(a)  $\alpha = 0.8$



(b)  $\alpha = 0.95$

**Figure 2.19:** Data sequences with one-sided impulsive noise generated according to the mixture model of section 2.4.2 and smoothed by  $U_1L_1$ . Parameters were set at  $c = 1$ ,  $\mu = \frac{1}{2}$  and  $\epsilon = 5$ . In (a) the smoother  $U_1L_1$  was not enough to remove all of the outliers and we had to subsequently smooth with  $U_2L_2$  to remove them.



## Chapter 3

# Bayesian Inference

Probability theory is only common sense reduced to calculus, it enables us to appreciate with exactness that which accurate minds feel with a sort of instinct, often without being able to account for it

Laplace 1814

Bayesian Inference is the mathematical machinery for conducting plausible reasoning or inductive logic. In the Bernoulli urn sense, deductive logic is applied when the contents of an urn is known, and under the assumption of randomness of the contents, the chances of producing certain outcomes upon drawing from the urn are calculated. The real scientist however is concerned with the reverse of this problem: Given the data, what were the probable causes of the data, or contents of the urn? Bernoulli himself pondered this problem [3], but it was Bayes' posthumously published paper [2] that appears to articulate a solution which resembles what we today call Bayes Theorem. Laplace [16], seemingly unaware of Bayes earlier claims, was responsible for writing it down in its general and continuous parameter form as we know it today and, unlike Bayes he motivated the assignment of the *prior probabilities* on a principle of 'insufficient reason' [11]. The potential of their work went undiscovered until Jeffreys [14] inspired thinkers such as Jaynes, de Finetti and others. Aided by Cox's theorems [5] they advocated a new way of interpreting probability theory, in contrast with the predominant classical statistical thinking of their time [30].

### 3.1 Probability theory as extended logic

Cox showed that if degrees of belief of various propositions are represented by real numbers<sup>1</sup> (the larger implying a greater degree of belief), there do exist general quantitative rules for logical and consistent reasoning [5]. These rules, central to probability theory as extended logic, turn out to be nothing more than the standard product and sum rules of probability theory. Their derivation (see for example [5] or [13]), using Boolean algebra and calculus, remarkably follows from only a few verbal statements<sup>2</sup> (qualitative rules, not axioms) describing the desired attributes of a

---

<sup>1</sup>Representation by real numbers automatically introduces a transitive ranking of propositions. This avoids circular argumentation in the sense that if we believe proposition A more than B, and B more than C, then we must necessarily believe A more than C.

<sup>2</sup>The first was representation of degrees of belief by real numbers. The second is that stating ones belief in the truth of a proposition A implicitly specifies how much we believe it is false. The third is that stating first one's

logically consistent reasoning process. Thus, letting  $P$  represent a function that describes our degree of belief in a proposition  $A$  or  $B$ , we have the sum rule

$$P(A|I) + P(\bar{A}|I) = 1 \quad (3.1)$$

and the product rule

$$P(AB|I) = P(A|BI)P(B|I), \quad (3.2)$$

where the *negation*  $\bar{A}$  is the proposition that  $A$  is false, the conditioning symbol ' $|$ ' means 'given' in the sense that all propositions standing to the right of it are taken to be true. Two or more propositions in conjunction, for example  $AB$ , imply 'and', whilst  $A+B$  in Boolean algebra means  $A$  or  $B$ .  $P(\cdot) = 1$  means true with certainty, whilst  $P(\cdot) = 0$  means false with certainty and all real numbers in between measure degree of belief. All probabilities in the Bayesian framework are necessarily conditional on  $I$ , which encompasses *prior* knowledge or information available to the user.

Since conjunction and negation are an adequate set of logical operations to determine the plausibility of any proposition in the Boolean algebra [13], the sum and product rule form the fundamental rules of the theory expounded here. For instance, a rule for the probability of ' $A$  or  $B$ ' can be obtained directly from the above

$$P(A+B|I) = P(A|BI) + P(B|I) - P(AB|I). \quad (3.3)$$

This can be seen as a generalised sum rule which is evident when setting  $B = \bar{A}$ . Another very useful result which is derived from the sum and product rules is the *marginalisation* equation

$$P(A|I) = P(AB|I) + P(A\bar{B}|I). \quad (3.4)$$

If instead we have a set of alternative propositions  $\{B_1, \dots, B_n\}$ , which are *mutually exclusive*, meaning

$$P(B_i B_j | I) = P(B_i | I) \delta_{ij}, \quad (3.5)$$

and which are *exhaustive*, meaning one of  $B_i$  must be true and in which case the rest are false, then the marginalization equation becomes

$$P(A|I) = \sum_{i=1}^n P(AB_i|I). \quad (3.6)$$

By using the product rule on  $P(AB_i|I)$  above, the *normalization* condition

$$\sum_{i=1}^n P(B_i|AI) = 1 \quad (3.7)$$

holds for mutually exclusive and exhaustive propositions  $\{B_1, \dots, B_n\}$ . While these results and their continuous forms are well known, their importance to the theory of data analysis in the current context is as follows. When we replace the propositions  $A$  and  $B$  with  $H$  for *hypothesis* and  $D$  for *data*, and write the product rule twice to get Bayes theorem

$$P(H|DI)P(D|I) = P(D|HI)P(H|I)$$

---

belief in the truth of a proposition  $A$ , and then one's belief in  $B$  given that  $A$  is true, implicitly specifies how much we believe both  $A$  and  $B$  are true. Finally, should different analysis paths exist that use the same information, they should yield the same conclusions.

$$P(H | DI) = \frac{P(D | HI)P(H | I)}{P(D | I)}, \quad (3.9)$$

we are offered a more unified and logical approach to the subject than what conventional statistics has achieved. It gives a direct relationship between  $P(H | DI)$ , the probability that the hypothesis is true given the data, and  $P(D | HI)$ , the probability that we would have observed the data given that a certain hypothesis was true, the latter often being easier to assign. The term  $P(H | I)$  is referred to as the *prior* probability, or simply ‘the prior’, and reflects our state of knowledge or ignorance about the plausibility of a hypothesis before we have any data to analyse. Translating one’s prior information into probabilities is an open-ended problem of logical analysis of which several principles are already well established [13]. The term  $P(H | DI)$  is referred to as the *posterior* or *inverse* probability (sometimes referred to as ‘the posterior’) and is the term we seek when making an inference on the hypothesis given the data. For fixed  $H$ , the term  $P(D | HI)$  is the well-known *sampling* distribution<sup>3</sup> or *forward* probability of sampling theory, where predictions are made about the likely occurrence of certain data sets given that the hypothesis in its explanation of what caused the data is correct. When we consider  $P(D | HI)$  for fixed  $D$  in its dependence on say various hypotheses  $\{H_1, \dots, H_N\}$ , it is referred to as the *likelihood*. The term  $P(D | I)$  is called the *evidence* or *normalizing* constant and can be resolved as follows

$$P(D | I) = \sum_{i=1}^N P(DH_i | I) = \sum_{i=1}^N P(D | H_i I)P(H_i | I). \quad (3.10)$$

Throughout, the emphasis is on logical connections, and not necessarily physical causes. The posterior after one measurement can be used as the prior for the next. In this way, Bayes theorem provides a way for information to be automatically assimilated in its evolution of  $P(H | DI)$ . However, it can be shown that this sequential way of analysing the data is equivalent to considering the data collectively (in one step), if the data is *independent*, meaning that one measurement does not influence the other [30]. If for example the first and second measurements are denoted by  $D_1$  and  $D_2$ , then independence may be expressed mathematically as  $P(D_2 | HD_1 I) = P(D_2 | HI)$ .

### 3.2 Parameter estimation

Introducing a continuous range of hypotheses is straightforward. If we suppose that  $\theta$  is a continuously variable real parameter, the discrete propositions

$$\begin{aligned} F &= (\theta \leq q) \\ \bar{F} &= (\theta > q) \end{aligned} \quad (3.11)$$

are mutually exclusive and exhaustive. Since the proposition  $F$  will generally depend on  $q$ , a function

$$G(q) = P(F | I) \quad (3.12)$$

can be defined. Defining the propositions

$$A = (\theta \leq a), \quad B = (\theta \leq b), \quad C = (a < \theta \leq b), \quad (3.13)$$

the probability that  $\theta$  lies in the interval  $a < \theta \leq b$  is determined uniquely by the sum rule which, since  $A$  and  $C$  are mutually exclusive, reduces to

<sup>3</sup>The popular statistical texts of the latter half of the twentieth century by Feller [9, 10] and volumes 1-2A of Kendall and Stuart [15] concentrate solely on calculating these.

$$P(B|I) = P(A|I) + P(C|I). \quad (3.14)$$

Thus

$$P(a < \theta \leq b|I) = G(b) - G(a), \quad (3.15)$$

or, since  $G(\theta)$  is continuous and differentiable

$$P(a < \theta \leq b|I) = \int_a^b d\theta g(\theta), \quad (3.16)$$

where  $g(\theta) = G'(\theta) \geq 0$  is the *probability density function* for  $\theta$ . Its integral  $G(\theta)$  is called the *cumulative distribution function* for  $\theta$ . Now suppose we want to measure a quantity  $\theta_0$  which is assumed to be in the range  $\theta_1 \leq \theta_0 \leq \theta_N$ . If  $\theta_0$  could take on any finite number of values in this range, we could assign a hypothesis  $H_{\theta_i}$  for each value and thus form a finite set of hypotheses  $\{H_{\theta_1}, \dots, H_{\theta_N}\}$  which are mutually exclusive and exhaustive. When we go to the continuum limit, we are testing an infinite number of hypotheses  $H_{\theta_i}$ . Thus, when we do parameter estimation, we think of  $\theta$  as being continuously varying and represent its possible range by a continuous range of hypotheses. Let  $H_\theta$  represent a continuous range of hypotheses, and let D represent the results of the experiment. If we now let

$$H = \theta_0 \text{ is in the range } (\theta, \theta + d\theta), \quad (3.17)$$

there exists a prior probability

$$P(H|I) = g(\theta|I)d\theta. \quad (3.18)$$

This is the probability that  $\theta_0$  lies in the range  $d\theta$ . Using Bayes theorem, the posterior for  $\theta$  is

$$P(H|DI) = g(\theta|DI)d\theta = P(H|I) \frac{P(D|HI)}{P(D|I)}$$

$$g(\theta|DI) = g(\theta|I) \frac{P(D|HI)}{P(D|I)}. \quad (3.19)$$

Finally, if  $P(D|HI) \rightarrow P(D|H_\theta I)$  as  $d\theta \rightarrow 0$  (which is a subtlety that is not always trivial if there is more than one parameter present), Bayes theorem becomes

$$g(\theta|DI) = g(\theta|I) \frac{P(D|H_\theta I)}{P(D|I)}. \quad (3.20)$$

Using the symbol  $P$  for probability density function, and replacing  $H_\theta$  with  $\theta$ , Bayes theorem for a continuous parameter becomes

$$P(\theta|DI) = \frac{P(D|\theta I)P(\theta|I)}{\int_{-\infty}^{\infty} d\theta P(D|\theta I)P(\theta|I)}. \quad (3.21)$$

Jaynes notes that since we have only applied our product and sum rule to discrete propositions in finite sets, we are protected from the paradoxes of infinite-set theory. Further, since data is always finite, a continuously variable  $\theta$  is only an approximation to the exact discrete theory, while the advantages of using calculus are obvious.

Suppose we want to infer on the plausibility of a number of continuous parameters  $\{\theta_1, \theta_2, \dots, \theta_N\}$ . The correct procedure is to calculate the joint probability of these parameters given all the

available evidence, or data  $D$  at hand, and given any prior information  $I$  we might have regarding these parameters. Using Bayes theorem we have

$$P(\theta_1 \dots \theta_N | DI) = \frac{P(D | \theta_1 \dots \theta_N I) P(\theta_1 \dots \theta_N | I)}{\int d\theta_1 \dots \int d\theta_N P(D | \theta_1 \dots \theta_N I) P(\theta_1 \dots \theta_N | I)}. \quad (3.22)$$

The likelihood  $P(D | \theta_1 \dots \theta_N I)$  describes the proposed mathematical model responsible for producing the data. The joint prior probability distribution is represented by  $P(\theta_1 \dots \theta_N | I)$ . From the posterior  $P(\theta_1 \dots \theta_N | DI)$ , we are particularly interested in the marginal probability distributions for the parameters  $P(\theta_i | DI)$ . The joint posterior represents the full and final answer to the question of what the  $\theta_i$ 's are. From these, we can derive various useful estimators of the parameters. For example, we will consider the mean estimator which minimizes the variance, but the median or peak (mode) may also be used. The marginal probability distribution of a particular parameter  $\theta_i$  is

$$\begin{aligned} P(\theta_i | DI) &= \int d\theta_1 \dots \int d\theta_{i-1} \int d\theta_{i+1} \dots \int d\theta_N P(\theta_1 \dots \theta_N | DI) \\ &= \frac{1}{\mathcal{N}} \int d\theta_1 \dots \int d\theta_{i-1} \int d\theta_{i+1} \dots \int d\theta_N P(D | \theta_1 \dots \theta_N I) P(\theta_1 \dots \theta_N | I) \end{aligned} \quad (3.23)$$

where

$$\mathcal{N} = \int d\theta_1 \dots d\theta_N P(D | \theta_1 \dots \theta_N I) P(\theta_1 \dots \theta_N | I) \quad (3.24)$$

is the normalization constant. The  $m$ -th moment of the marginal distribution for  $\theta_i$  is

$$\begin{aligned} \langle \theta_i^m \rangle &= \int d\theta_i \theta_i^m P(\theta_i | DI) \\ &= \frac{1}{\mathcal{N}} \int d\theta_1 \dots \int d\theta_N \theta_i^m P(D | \theta_1 \dots \theta_N I) P(\theta_1 \dots \theta_N | I) \end{aligned} \quad (3.25)$$

The choice  $m = 1$  gives us the expectation value or mean of the marginal distribution,  $m = 2$  gives the second moment, and so on. For each model we will be investigating, we will proceed to calculate

$$\int d\theta_1 \dots \int d\theta_{i-1} \int d\theta_{i+1} \dots \int d\theta_N \theta_i^m P(D | \theta_1 \dots \theta_N I) P(\theta_1 \dots \theta_N | I) \quad (3.26)$$

for each parameter  $\theta_i$  of interest. Setting  $m = 0$  we then have the numerator of the marginal probability for  $\theta_i$ . Integrating over all  $N$  parameters

$$\int d\theta_1 \dots \int d\theta_N \theta_i^m P(D | \theta_1 \dots \theta_N I) P(\theta_1 \dots \theta_N | I), \quad (3.27)$$

and setting  $m = 0$ , we then immediately have the normalization constant  $\mathcal{N}$ , while the choice  $m \geq 1$  gives us the numerators of the moments. By calculating (3.26) and (3.27) for generic  $m$ , we get all the ingredients needed for the moments of the parameters and their marginal distributions.

### 3.3 Choice of prior probabilities

Suppose we collect the data

$$D = \{x_i\}_i^n \quad (3.28)$$

where each  $x_i \in \mathbb{R}$ , and on prior information  $I$  we know that each data point  $x_i$  contains an unknown constant signal  $X$  contaminated with a random noise part  $y_i$  which is generated i.i.d. according to a certain noise distribution. Each data point thus has the same form (as in Chapter 2), namely

$$x_i = X + y_i, \quad i = 1, \dots, n. \quad (3.29)$$

A typical noise distribution is located somewhere about zero (usually symmetrically) and thus only has a scale parameter, call it  $L$ , which describes its scale or ‘width’. The model distribution is then the noise distribution shifted by  $X$ , where  $X$  is its location parameter (describing its position on the  $x$ -axis) and  $L$  is its scale parameter. For example, if a constant signal  $X$  is obscured by noise from a symmetric uniform noise distribution

$$P(y_i | LI) = \begin{cases} L^{-1} & -\frac{L}{2} \leq y_i \leq \frac{L}{2} \\ 0 & \text{otherwise} \end{cases}, \quad (3.30)$$

then the probability for a single event  $x_i$  is the noise distribution shifted by  $X$

$$P(x_i | XLI) = \begin{cases} L^{-1} & X - \frac{L}{2} \leq x_i \leq X + \frac{L}{2} \\ 0 & \text{otherwise} \end{cases}, \quad (3.31)$$

where  $X$  describes its ‘location’ on the  $x$ -axis and  $L$  describes its ‘scale’. In each case we would like to infer the plausible values of  $X$  and  $L$  given the data. Since we already have the likelihood, namely

$$P(D | XLI) = \prod_{i=1}^n P(x_i | XLI), \quad (3.32)$$

we only need to choose the prior probabilities to proceed in applying Bayes theorem. The prior for a location parameter  $X$  is chosen as

$$P(X | I) = (K_2 - K_1)^{-1} U[X | K_1, K_2], \quad (3.33)$$

where  $K_1$  and  $K_2$  can be negative or positive so long as  $K_2 > K_1$  and the window function  $U$  and its properties are set out in Appendix A.  $K_1$  and  $K_2$  thus give indication of the permissible range of values that  $X$  can have based on our prior information. This choice of uniform prior is based on a *principle of indifference* which was first formulated by Laplace. It reflects our lack of information regarding the possible values of  $X$  (other than it being between  $K_1$  and  $K_2$ ) and is thus called an *ignorance* prior. We will assume a state of ignorance regarding all the parameters we will encounter. The above is a *proper* uniform prior, but we may conveniently change it to *improper* at any stage by taking either the limit  $K_1 \rightarrow -\infty$ , or  $K_2 \rightarrow \infty$ , or both. The term improper simply means that the prior is no longer normalizable. If only the  $K_1$  or  $K_2$  limit is taken, the uniform prior is said to be improper and truncated below or above. Note that in Bayes theorem, any constants like  $(K_2 - K_1)^{-1}$  in the above will cancel out, and thus for simplicity and economy of writing we may omit them already at an early stage. Thus, following this convention, we have

$$P(X | I) = U[X | K_1, K_2]. \quad (3.34)$$

For the scale parameter  $L$ , we choose the prior

$$P(L | I) = L^{-\lambda} U[L | M_1, M_2] \text{ with } \lambda \geq 0, M_2 > M_1 > 0, \quad (3.35)$$

again omitting the constant. Inserting  $L^{-\lambda}$  gives us the freedom to change the prior. Our results, in the form of joint or marginal posterior distributions and estimators will thus be general for any choice of ignorance prior for the scale parameter. If we let  $\lambda = 0$  we have chosen a proper uniform prior for  $L$ , since the constant prefactor  $1/(M_2 - M_1)$  now missing in (3.35) would have cancelled out in Bayes theorem. If we then take the limit  $M_2 \rightarrow \infty$  we change it into an improper uniform prior which is truncated below. When  $\lambda = 1$ , and the limit  $M_2 \rightarrow \infty$  is taken, we have the famous Jeffreys prior. When  $M_2$  is kept, as will sometimes be the case in the calculations to come, we've truncated it above. The choice of Jeffreys priors for scale parameters, and uniform priors for location parameters may seem strange at first, but they have been shown (see transformation group derivation in [13]) to be the correct way to represent ignorance due to their invariance under scale transformation. The different choices of  $\lambda$  correspond with different groups of transformations, the idea being that a change of scale or location should not change our state of knowledge. Consensus is that it is the group of transformations that gives rise to the Jeffreys prior that is appropriate in most imaginable problems [13].

Note that the above priors are ignorance priors, designed to reflect our current supposed state of prior knowledge that we have no reason to favour any particular  $X$  or  $L$  over another. Thus when we build in cut-offs in the form of  $K_1$  and  $K_2$  or  $M_1$  and  $M_2$  we propose to have prior knowledge of the permissible range of the data for our model to be correct. If the data falls outside this range then the joint posterior for the parameters is zero. Also, if you have reason to believe that the data cannot include certain values outside a particular range, then these values will help in the estimation of the parameters when the data does fall within their permissible range. Although heuristics show that the influence of prior wanes as more and more data points are accumulated, we will see that for small samples, the results become dependent on such prior knowledge as is embodied in for example  $M_1$  and  $M_2$ .

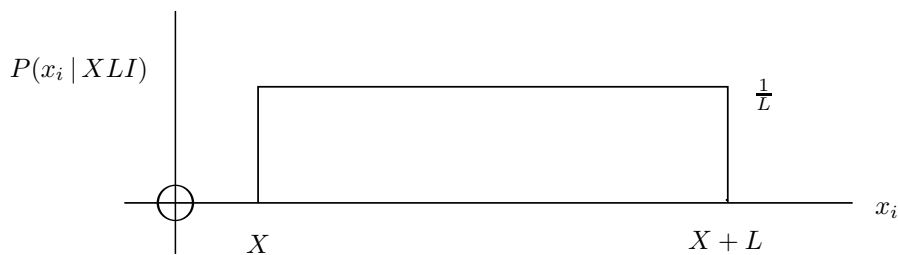
## Chapter 4

# Asymmetric Uniform Noise: Bayesian solution

### 4.1 Formulating the problem

The problem studied here is similar to the popular introductory statistics problem of estimating the parameter of a continuous uniform distribution with the left edge of the distribution starting at zero. The problem is also known as the continuous version of the taxicab problem [13], or the German tank problem, so named because for their discrete versions an estimate is required as to the size of a city inferred only from the observations of numbered taxis or the size of an armies artillery from observed serial numbers on tanks. The continuous version of the problem can be stated as follows: Data is generated according to a uniform distribution between zero and an unknown point, call it  $L$ . What is your guess at the possible length  $L$ , or upper edge of this distribution? The difference in our approach is that our uniform distribution starts at an unknown point, call it  $X$ , and not at zero, and we thus have an extra parameter. But why do we choose to study this problem, and why is it interesting? As is so often the recommended point of departure, we choose this kind of distribution as it is the simplest one imaginable that leads to non-trivial and insightful results.

Let us begin to formulate the problem by writing down the likelihood. Assume that we have a constant but unknown signal  $X$  that is obscured by noise which is distributed according to a continuous and uniform distribution of unknown length  $L$ , as illustrated in Fig. 4.1 below.



**Figure 4.1:** Asymmetric uniform noise of unknown length  $L$  obscuring unknown constant signal  $X$ .



The noise probability distribution is thus

$$P(y_i | LI) = \begin{cases} L^{-1} & 0 \leq y_i \leq L \\ 0 & \text{otherwise} \end{cases}. \quad (4.1)$$

Then, given the signal and the noise distribution, the probability for a single event  $x_i$  is the noise distribution shifted by  $X$

$$P(x_i | XLI) = \begin{cases} L^{-1} & X \leq x_i \leq X + L \\ 0 & \text{otherwise} \end{cases}, \quad (4.2)$$

or in our window function notation (see Appendix A)

$$P(x_i | XLI) = L^{-1}U[x_i | X, X + L]. \quad (4.3)$$

From the product rule and assuming each event  $x_i$  is i.i.d., the likelihood is

$$\begin{aligned} P(D | XLI) &= \prod_{i=1}^n P(x_i | XLI) \\ &= \prod_{i=1}^n L^{-1}U[x_i | X, X + L] \\ &= L^{-n} \prod_{i=1}^n U[x_i | X, X + L]. \end{aligned} \quad (4.4)$$

Now that we have the likelihood, we need a suitable choice of prior probabilities for our parameters  $X$  and  $L$ . Referring to section 3.3 on choice of priors, since  $X$  is a location parameter

$$P(X | I) = U[X | K_1, K_2], \quad K_1, K_2 \in \mathbb{R}, \quad K_2 > K_1, \quad (4.5)$$

and since  $L$  is a scale parameter, we assign a power-law prior

$$P(L | I) = L^{-\lambda}U[L | M_1, M_2], \quad M_1, M_2 \in \mathbb{R}, \quad M_2 > M_1 > 0, \quad \lambda > 0. \quad (4.6)$$

The joint prior  $P(XL | I)$  can be rewritten using the product rule as follows

$$P(XL | I) = P(X | LI)P(L | I) = P(L | XI)P(X | I), \quad (4.7)$$

and since  $X$  and  $L$  are logically independent we know that  $P(X | LI) = P(X | I)$  and thus it is necessarily so from (4.7) that  $P(L | XI) = P(L | I)$ . We could just as well have started by saying that we know that  $P(L | XI) = P(L | I)$ , but either way the joint prior factorises as

$$P(XL | I) = P(X | I)P(L | I), \quad (4.8)$$

and upon substitution of our priors is

$$P(XL | I) = L^{-\lambda}U[L | M_1, M_2]U[X | K_1, K_2]. \quad (4.9)$$

## 4.2 Overview of calculations

In the remainder of this chapter, we shall use the above sampling distribution and priors to find exact solutions to all the posteriors as well as their means and variances. Bayes Theorem for each of the three relevant posteriors yields, respectively,

$$P(XL | DI) = \frac{P(D | XLI)P(XL | I)}{\int dX dL P(D | XLI)P(XL | I)}, \quad (4.10)$$

$$P(L | DI) = \int dX P(XL | DI) = \frac{\int dX P(D | XLI)P(XL | I)}{\int dX dL P(D | XLI)P(XL | I)}, \quad (4.11)$$

$$P(X | DI) = \int dL P(XL | DI) = \frac{\int dL P(D | XLI)P(XL | I)}{\int dX dL P(D | XLI)P(XL | I)}, \quad (4.12)$$

and they represent an exhaustive answer to the question of the values of  $X$  and  $L$ . In Section 4.3 we shall find a complete solution for each of these in terms of the parameters  $K_1, K_2, M_1$  and  $M_2$  and the number of data points  $n$ . Often, however, we are merely interested not so much in the details of the posteriors but only in their means and second moments. With

$$\mathcal{N} = P(D | I) = \int dX dL P(D | XLI)P(XL | I) \quad (4.13)$$

the denominator, these are

$$\langle X \rangle = \int dX X P(X | DI) = \frac{1}{\mathcal{N}} \int dX dL X P(D | XLI)P(XL | I) \quad (4.14)$$

$$\langle X^2 \rangle = \int dX X^2 P(X | DI) = \frac{1}{\mathcal{N}} \int dX dL X^2 P(D | XLI)P(XL | I) \quad (4.15)$$

$$\langle L \rangle = \int dL L P(L | DI) = \frac{1}{\mathcal{N}} \int dX dL L P(D | XLI)P(XL | I) \quad (4.16)$$

$$\langle L^2 \rangle = \int dL L^2 P(L | DI) = \frac{1}{\mathcal{N}} \int dX dL L^2 P(D | XLI)P(XL | I) \quad (4.17)$$

from which the variances  $\langle X^2 \rangle - \langle X \rangle^2$  and  $\langle L^2 \rangle - \langle L \rangle^2$  are easily found. All of the above can be written and calculated succinctly in terms of three generic functions of indices  $m_X, m_L = 0, 1, 2$ , which are integrals over the respective likelihoods and priors,

$$A(L, m_X, m_L) = L^{m_L} \int_{\mathcal{A}(X,L)} dX X^{m_X} P(D | XLI)P(XL | I) \quad (4.18)$$

$$B(X, m_X, m_L) = X^{m_X} \int_{\mathcal{A}(X,L)} dL L^{m_L} P(D | XLI)P(XL | I) \quad (4.19)$$

$$\begin{aligned}
 C(m_X, m_L) &= \int_{\mathcal{A}(X,L)} dX dL X^{m_X} L^{m_L} P(D | XLI) P(XL | I) \\
 &= \int_{\mathcal{A}(X)} dL A(L, m_X, m_L) \\
 &= \int_{\mathcal{A}(L)} dX B(X, m_X, m_L),
 \end{aligned} \tag{4.20}$$

where  $\mathcal{A}(X, L)$  is the outcome space of  $X$  and  $L$  given  $D$ ,  $K_1$ ,  $K_2$ ,  $M_1$  and  $M_2$ . In terms of the above we have  $\mathcal{N} \equiv C(0, 0)$  and

$$P(XL | DI) = \frac{1}{\mathcal{N}} P(D | XLI) P(XL | I) \tag{4.21}$$

$$P(L | DI) = \frac{1}{\mathcal{N}} A(L, 0, 0) \tag{4.22}$$

$$P(X | DI) = \frac{1}{\mathcal{N}} B(X, 0, 0) \tag{4.23}$$

$$\langle X \rangle = \frac{1}{\mathcal{N}} C(1, 0) \tag{4.24}$$

$$\langle L \rangle = \frac{1}{\mathcal{N}} C(0, 1) \tag{4.25}$$

and so on. All of these quantities are, of course, functions of the data  $D$  itself, the number of data points  $n$ , and of prior parameters such as  $\lambda$ .

### 4.3 Moments of the posterior distribution: Exact solution

Beginning with  $A$ , insert the likelihood (4.4) and joint prior (4.9)

$$A = L^{m_L - n - \lambda} \int dX X^{m_X} \prod_{i=1}^n U[x_i | X, X + L] U[L | M_1, M_2] U[X | K_1, K_2]. \tag{4.26}$$

We need to calculate the outcome space or ‘boundary function’ which will enter into  $A$ ,  $B$  and  $C$ :

$$\mathcal{A}(X, L) = \prod_{i=1}^n U[x_i | X, X + L] U[L | M_1, M_2] U[X | K_1, K_2], \tag{4.27}$$

in order to perform an integration over  $X$ . We will soon see how the priors and data will affect the integration boundaries of  $X$  whilst also correctly restricting the range of possible  $L$  values. From result (A.8) we have

$$\prod_{i=1}^n U[x_i | X, X + L] = \theta(x_{min} - X) \theta(X + L - x_{max}), \tag{4.28}$$

where  $x_{min} = \min(x_1, x_2, \dots, x_n)$  and  $x_{max} = \max(x_1, x_2, \dots, x_n)$ . Note that we have now made the assumption that there are at least two data points  $n \geq 2$ . The one data point case is still possible using the Bayesian approach and is presented later. Substitute the above into (4.27) and use definition (A.2) to write out the remaining window functions in terms of Heaviside theta functions to get

$$\mathcal{A}(X, L) = \theta(x_{min} - X)\theta(X + L - x_{max})\theta(L - M_1)\theta(M_2 - L)\theta(X - K_1)\theta(K_2 - X). \quad (4.29)$$

We now rearrange the factors in the above solving successively for  $X$  (as needed to calculate  $A$  in Eq. (4.26)). We want to integrate out  $X$ , so let us make  $X$  the subject of a single window function by using results (A.6) and (A.7) as follows

$$\begin{aligned} & \theta(X - K_1)\theta(X - (x_{max} - L))\theta(x_{min} - X)\theta(K_2 - X)\theta(L - M_1)\theta(M_2 - L) \\ &= \theta(X - \max(K_1, x_{max} - L))\theta(\min(x_{min}, K_2) - X)\theta(L - M_1)\theta(M_2 - L) \\ &= U[X | \max(K_1, x_{max} - L), \min(x_{min}, K_2)]U[L | M_1, M_2]. \end{aligned} \quad (4.30)$$

From (4.29) we have the inequality  $L \geq x_{max} - X$ , but observing (4.30) we see that  $X$  is restricted above as follows  $X \leq \min(x_{min}, K_2)$ . Thus, we know that  $L \geq x_{max} - \min(x_{min}, K_2)$  and (4.30) does not restrict the possibilities for  $L$  correctly. We must build this restriction into (4.30), and we thus use the following instead

$$U[X | \max(K_1, x_{max} - L), \min(x_{min}, K_2)]U[L | M_1, M_2]\theta(L - (x_{max} - \min(x_{min}, K_2))). \quad (4.31)$$

We have done nothing strange by inserting this theta function. All we missed was that we should have written down  $\theta(X + L - x_{max})$  twice from the beginning. Having rewritten one of them as follows  $\theta(L - (x_{max} - X))$ , we would come to the same conclusion that  $L \geq x_{max} - \min(x_{min}, K_2)$  since  $X \leq \min(x_{min}, K_2)$ . It is easy to miss if one has not written something like (4.30) down first.

To continue, absorb the theta function into the window function for  $L$  by writing out (4.31) in terms of theta functions and using result (A.7) as follows

$$\begin{aligned} & U[L | M_1, M_2]\theta(L - (x_{max} - \min(x_{min}, K_2))) \\ &= \theta(L - M_1)\theta(L - (x_{max} - \min(x_{min}, K_2)))\theta(M_2 - L) \\ &= \theta(L - \max(M_1, x_{max} - \min(x_{min}, K_2)))\theta(M_2 - L) \\ &= U[L | \max(M_1, x_{max} - \min(x_{min}, K_2)), M_2]. \end{aligned} \quad (4.32)$$

Finally, the boundary function becomes

$$\mathcal{A}(X, L) = U[X | \max(K_1, x_{max} - L), \min(x_{min}, K_2)]U[L | \max(M_1, x_{max} - \min(x_{min}, K_2)), M_2]. \quad (4.33)$$

Substituting this into  $A$ , we need to integrate

$$\begin{aligned} A &= L^{m_L - n - \lambda} \int dX X^{m_X} U[X | \max(K_1, x_{max} - L), \min(x_{min}, K_2)] \\ &\quad \times U[L | \max(M_1, x_{max} - \min(x_{min}, K_2)), M_2]. \end{aligned} \quad (4.34)$$

The integration boundaries are simply read from the window function, and so the above becomes

$$A = L^{m_L - n - \lambda} \int_{\max(K_1, x_{max} - L)}^{\min(x_{min}, K_2)} dX X^{m_X} U[L | \max(M_1, x_{max} - \min(x_{min}, K_2)), M_2]. \quad (4.35)$$

Performing the integration we get

$$A = \frac{(\min(x_{min}, K_2))^{m_X+1} - (\max(K_1, x_{max} - L))^{m_X+1}}{m_X + 1} L^{m_L - n - \lambda} \times U[L | \max(M_1, x_{max} - \min(x_{min}, K_2)), M_2], \quad (4.36)$$

since  $m_X \geq 0$ .

We turn our focus to calculating  $B$ . We need

$$B = X^{m_X} \int dL L^{m_L - n - \lambda} \prod_{i=1}^n U[x_i | X, X + L] U[L | M_1, M_2] U[X | K_1, K_2]. \quad (4.37)$$

The term (4.27)

$$\mathcal{A}(X, L) = \prod_{i=1}^n U[x_i | X, X + L] U[L | M_1, M_2] U[X | K_1, K_2]$$

must be prepared in order to perform an integration over  $L$  this time. Again, write it down in terms of theta functions but this time rearrange them to make  $L$  the subject of a single theta function as follows

$$\begin{aligned} & \theta(X - K_1) \theta(x_{min} - X) \theta(K_2 - X) \theta(L - M_1) \theta(L - (x_{max} - X)) \theta(M_2 - L) \\ &= \theta(X - K_1) \theta(\min(x_{min}, K_2) - X) \theta(L - \max(M_1, x_{max} - L)) \theta(M_2 - L) \\ &= U[X | K_1, \min(x_{min}, K_2)] U[L | \max(M_1, x_{max} - X), M_2]. \end{aligned} \quad (4.38)$$

Once again, since  $L$  is restricted above by  $M_2$  and  $X \geq x_{max} - L$  we know that  $X \geq x_{max} - M_2$ . Thus, whilst integrating out  $L$ , the possibilities for  $X$  are restricted correctly if we use

$$U[X | K_1, \min(x_{min}, K_2)] U[L | \max(M_1, x_{max} - X), M_2] \theta(X - (x_{max} - M_2)) \quad (4.39)$$

instead. Absorbing this restriction into the window function for  $X$  as follows

$$\begin{aligned} & U[X | K_1, \min(x_{min}, K_2)] \theta(X - (x_{max} - M_2)) \\ &= \theta(X - K_1) \theta(X - (x_{max} - M_2)) \theta(\min(x_{min}, K_2) - X) \\ &= \theta(X - \max(K_1, x_{max} - M_2)) \theta(\min(x_{min}, K_2) - X) \\ &= U[X | \max(K_1, x_{max} - M_2), \min(x_{min}, K_2)] \end{aligned} \quad (4.40)$$

The boundary function (4.27) hence attains its final form

$$\mathcal{A}(X, L) = U[X | \max(K_1, x_{max} - M_2), \min(x_{min}, K_2)] U[L | \max(M_1, x_{max} - X), M_2]. \quad (4.41)$$

Substituting this into  $B$ , and reading the integration boundaries from the window function for  $L$ , we need to integrate

$$B = X^{m_X} \int_{\max(M_1, x_{max} - X)}^{M_2} dL L^{m_L - n - \lambda} U[X | \max(K_1, x_{max} - M_2), \min(x_{min}, K_2)]. \quad (4.42)$$

The result is

$$B = \frac{(M_2)^{m_L - n - \lambda + 1} - (\max(M_1, x_{max} - X))^{m_L - n - \lambda + 1}}{m_L - n - \lambda + 1} X^{m_X} U[X | \max(K_1, x_{max} - M_2), \min(x_{min}, K_2)] \quad (4.43)$$

for  $m_L - n - \lambda \neq -1$ , and

$$B = \ln \left( \frac{M_2}{\max(M_1, x_{max} - X)} \right) X^{m_X} U[X | \max(K_1, x_{max} - M_2), \min(x_{min}, K_2)] \quad (4.44)$$

for  $m_L - n - \lambda = -1$ . Finally we calculate  $C$  by integrating  $X$  out of the above. We need

$$C = \frac{1}{m_L - n - \lambda + 1} \left\{ M_2^{m_L - n - \lambda + 1} \int_{\max(K_1, x_{max} - M_2)}^{\min(x_{min}, K_2)} dX X^{m_X} - \int_{\max(K_1, x_{max} - M_2)}^{\min(x_{min}, K_2)} dX X^{m_X} (\max(M_1, x_{max} - X))^{m_L - n - \lambda + 1} \right\}, \quad (4.45)$$

for  $m_L - n - \lambda \neq -1$ , and

$$C = \ln M_2 \int_{\max(K_1, x_{max} - M_2)}^{\min(x_{min}, K_2)} dX X^{m_X} - \int_{\max(K_1, x_{max} - M_2)}^{\min(x_{min}, K_2)} dX X^{m_X} \ln(\max(M_1, x_{max} - X)) \quad (4.46)$$

for  $m_L - n - \lambda = -1$ . Proceeding with (4.45), we are confronted with the integral

$$\int_{\max(K_1, x_{max} - M_2)}^{\min(x_{min}, K_2)} dX X^{m_X} (\max(M_1, x_{max} - X))^{m_L - n - \lambda + 1}. \quad (4.47)$$

It requires some thought for what are we to do with the integrand  $\max(M_1, x_{max} - X)$ ? It is understood that the criterion for the maximum is checked as we integrate from the lower integration boundary of  $X$  to the upper one, and the integrand will thus change when  $M_1 = x_{max} - X$  or at the point  $X = x_{max} - M_1$  (presuming of course that  $x_{max} - M_1$  lies somewhere in between the integration boundaries  $\max(K_1, x_{max} - M_2) \leq x_{max} - M_1 \leq \min(x_{min}, K_2)$ ) (see Fig. 4.2 below). To proceed, distinguish between the three possible locations of the point  $x_{max} - M_1$  relative to the integration boundaries:

- Case 1:  $x_{max} - M_1 \geq \min(x_{min}, K_2)$

The point  $x_{max} - M_1$  lies above the upper integration boundary. Rewriting this inequality we see that  $M_1 \leq x_{max} - \min(x_{min}, K_2)$  and therefore we know that the integrand  $\max(M_1, x_{max} - X) = x_{max} - X$  for  $\max(K_1, x_{max} - M_2) \leq X \leq \min(x_{min}, K_2)$ .

- Case 2:  $\max(K_1, x_{max} - M_2) \leq x_{max} - M_1 \leq \min(x_{min}, K_2)$

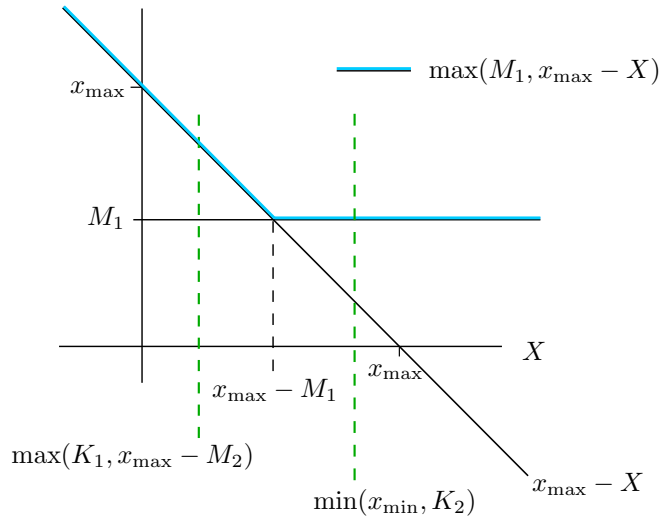
The point  $x_{max} - M_1$  lies in between the integration boundaries of  $X$  (see Fig. 4.2 below). The integral is split into two parts separated by the point  $x_{max} - M_1$ . While  $\max(K_1, x_{max} - M_2) \leq X \leq x_{max} - M_1$  we have the inequality  $M_1 \leq x_{max} - X$  so that the integrand  $\max(M_1, x_{max} - X) = x_{max} - X$ , and similarly the integrand is  $\max(M_1, x_{max} - X) = M_1$  while  $x_{max} - M_1 \leq X \leq \min(x_{min}, K_2)$ .

- Case 3:  $x_{max} - M_1 \leq \max(K_1, x_{max} - M_2)$

The point  $x_{max} - M_1$  lies below the lower integration boundary. Rewriting this inequality we see that  $x_{max} - \max(K_1, x_{max} - M_2) \leq M_1$ , and therefore the integrand is  $\max(M_1, x_{max} - X) = M_1$  for  $\max(K_1, x_{max} - M_2) \leq X \leq \min(x_{min}, K_2)$ .

The three cases are represented as theta functions so that finally, we obtain

$$\begin{aligned}
 & \theta((x_{max} - M_1) - \min(x_{min}, K_2)) \int_{\max(K_1, x_{max} - M_2)}^{\min(x_{min}, K_2)} dX X^{m_X} (x_{max} - X)^{m_L - n - \lambda + 1} \\
 & + \theta((x_{max} - M_1) - \max(K_1, x_{max} - M_2)) \theta(\min(x_{min}, K_2) - (x_{max} - M_1)) \\
 & \times \left\{ \int_{\max(K_1, x_{max} - M_2)}^{x_{max} - M_1} dX X^{m_X} (x_{max} - X)^{m_L - n - \lambda + 1} + M_1^{m_L - n - \lambda + 1} \int_{x_{max} - M_1}^{\min(x_{min}, K_2)} dX X^{m_X} \right\} \\
 & + \theta(\max(K_1, x_{max} - M_2) - (x_{max} - M_1)) M_1^{m_L - n - \lambda + 1} \int_{\max(K_1, x_{max} - M_2)}^{\min(x_{min}, K_2)} dX X^{m_X}.
 \end{aligned} \tag{4.48}$$



**Figure 4.2:** Plot of  $\max(M_1, x_{max} - X)$ . The point  $X = x_{max} - M_1$  lies in between the integration boundaries of  $X$ . This scenario corresponds with the second term in Eq. (4.48).

Writing (4.47) in this way has removed the maximum criterion that stood inside the integral. Unless we choose specific values for the prior bounds  $K_1$ ,  $K_2$ ,  $M_1$  and  $M_2$  then all the terms must be kept because it is not possible to know which term in (4.48) will survive after we have collected new data which may update  $x_{max}$  and  $x_{min}$ . Fortunately however, if we choose to have no bounds on the possibilities for  $X$  then (4.48) is greatly simplified.

Next we turn our attention to the corresponding case (4.46) where we again have a maximum criterion inside the integral

$$\int_{\max(K_1, x_{max} - M_2)}^{\min(x_{min}, K_2)} dX X^{m_X} \ln(\max(M_1, x_{max} - X)). \tag{4.49}$$

This integral is done in the same way as (4.47). The result is just (4.48) after replacing  $M_1^{m_L-n-\lambda+1}$  with  $\ln M_1$  and  $(x_{max} - X)^{m_L-n-\lambda+1}$  with  $\ln(x_{max} - X)$ , namely

$$\begin{aligned}
 & \theta((x_{max} - M_1) - \min(x_{min}, K_2)) \int_{\max(K_1, x_{max} - M_2)}^{\min(x_{min}, K_2)} dX X^{m_X} \ln(x_{max} - X) \\
 & + \theta((x_{max} - M_1) - \max(K_1, x_{max} - M_2)) \theta(\min(x_{min}, K_2) - (x_{max} - M_1)) \\
 & \times \left\{ \int_{\max(K_1, x_{max} - M_2)}^{x_{max} - M_1} dX X^{m_X} \ln(x_{max} - X) + \ln M_1 \int_{x_{max} - M_1}^{\min(x_{min}, K_2)} dX X^{m_X} \right\} \quad (4.50) \\
 & + \theta(\max(K_1, x_{max} - M_2) - (x_{max} - M_1)) \ln M_1 \int_{\max(K_1, x_{max} - M_2)}^{\min(x_{min}, K_2)} dX X^{m_X}.
 \end{aligned}$$

We will rarely need (4.44) and (4.46), but it is important to gather these results now as they will become important later on. For example, if we choose the Jeffreys prior  $\lambda = 1$  and we need the second moment for the  $n = 2$  case, then (4.31) applies.

#### 4.4 Moments for simplified prior

As is already evident from observing (4.48) or (4.50), a more specific choice of priors will greatly simplify the expressions gathered in the previous section. In order to simplify we first take the limits  $K_1 \rightarrow -\infty$  and  $K_2 \rightarrow \infty$ , and keep  $M_1$  and  $M_2$  for the time being. The answer to why we do not just send  $M_1 \rightarrow 0$  will become clear when we study the  $n = 1$  case. Focussing on (4.48), it becomes

$$\begin{aligned}
 & \theta((x_{max} - x_{min}) - M_1) \int_{x_{max} - M_2}^{x_{min}} dX X^{m_X} (x_{max} - X)^{m_L - n - \lambda + 1} \\
 & + \theta(M_2 - M_1) \theta(M_1 - (x_{max} - x_{min})) \\
 & \times \left\{ \int_{x_{max} - M_2}^{x_{max} - M_1} dX X^{m_X} (x_{max} - X)^{m_L - n - \lambda + 1} + M_1^{m_L - n - \lambda + 1} \int_{x_{max} - M_1}^{x_{min}} dX X^{m_X} \right\} \quad (4.51) \\
 & + \theta(M_1 - M_2) M_1^{m_L - n - \lambda + 1} \int_{\max(K_1, x_{max} - M_2)}^{x_{min}} dX X^{m_X}.
 \end{aligned}$$

If we truly know nothing about the length of the noise distribution, and we have collected two unique points of data  $x_{max}$  and  $x_{min}$ , then we imagine that  $M_1$  was chosen so small that it is always smaller than the difference  $x_{max} - x_{min}$ . Assuming that  $x_{max} - x_{min} > M_1$  and since we have already chosen  $M_2 > M_1$ , the only term left in (4.51) is the first one, namely

$$\int_{x_{max} - M_2}^{x_{min}} dX X^{m_X} (x_{max} - X)^{m_L - n - \lambda + 1}. \quad (4.52)$$

In exactly the same way, (4.50) now becomes

$$\int_{x_{max} - M_2}^{x_{min}} dX X^{m_X} \ln(x_{max} - X). \quad (4.53)$$

Taking the same limits in  $C$ , and then substituting the above, we have



$$C = \frac{1}{m_L - n - \lambda + 1} \left\{ M_2^{m_L - n - \lambda + 1} \int_{x_{max} - M_2}^{x_{min}} dX X^{m_X} - \int_{x_{max} - M_2}^{x_{min}} dX X^{m_X} (x_{max} - X)^{m_L - n - \lambda + 1} \right\}, \quad (4.54)$$

for  $m_L - n - \lambda \neq -1$ , and

$$C = \ln M_2 \int_{x_{max} - M_2}^{x_{min}} dX X^{m_X} - \int_{x_{max} - M_2}^{x_{min}} dX X^{m_X} \ln(x_{max} - X), \quad (4.55)$$

for  $m_L - n - \lambda = -1$ . We note that the integration can be written as hypergeometric functions

$$\begin{aligned} & \int dX X^{m_X} (x_{max} - X)^{m_L - n - \lambda + 1} \\ &= \frac{1}{1 + m_X} \left\{ (x_{max} - X)^{m_L - n - \lambda + 1} X^{1 + m_X} \left( 1 - \frac{X}{x_{max}} \right)^{-(m_L - n - \lambda + 1)} \right. \\ & \quad \left. \times {}_2F_1\left(1 + m_X, -(m_L - n - \lambda + 1); 2 + m_X; \frac{X}{x_{max}}\right) \right\} \end{aligned} \quad (4.56)$$

and

$$\begin{aligned} & \int dX X^{m_X} \ln(x_{max} - X) \\ &= \frac{1}{(1 + m_X)^2} X^{1 + m_X} \left( {}_2F_1\left(1, 1 + m_X; 2 + m_X; \frac{X}{x_{max}}\right) + (1 + m_X) \ln(x_{max} - X) - 1 \right), \end{aligned} \quad (4.57)$$

but since we only need solutions for  $m_X \leq 2$  we can still easily evaluate the integrals by hand, and avoid making use of the hypergeometric function  ${}_2F_1$ .

Now that we have simplified  $C$ , let us take the same limits in  $A$  and  $B$ , and set  $m_X = m_L = 0$ , as these expressions will together with the normalization constant give us the marginal posterior distributions of the parameters. Letting  $x_\Delta$  stand for the difference  $x_{max} - x_{min}$ ,  $\frac{1}{\mathcal{N}}B$  and  $\frac{1}{\mathcal{N}}A$  become

$$P(X | DI) = \frac{1}{\mathcal{N}} \frac{M_2^{-n - \lambda + 1} - (x_{max} - X)^{-n - \lambda + 1}}{-n - \lambda + 1} U[X | x_{max} - M_2, x_{min}], \text{ for } n \geq 2 \text{ and } \forall \lambda, \quad (4.58)$$

and

$$P(L | DI) = \frac{1}{\mathcal{N}} (L - x_\Delta) L^{-n - \lambda} U[L | x_\Delta, M_2], \text{ for } n \geq 2 \text{ and } \forall \lambda, \quad (4.59)$$

respectively. The normalization constant is given by  $\mathcal{N} = C(m_X = 0, m_L = 0)$ . Set  $m_X = m_L = 0$  in (4.54) and complete the integration to get

$$\mathcal{N} = \begin{cases} - \left( \ln \frac{x_\Delta}{M_2} - M_2^{-1} x_\Delta + 1 \right), & \text{for } n = 2 - \lambda \\ \frac{1}{-n - \lambda + 1} \left( M_2^{-n - \lambda + 1} (M_2 - x_\Delta) + \frac{x_\Delta^{-n - \lambda + 2} - M_2^{-n - \lambda + 2}}{-n - \lambda + 2} \right), & \text{for } n > 2 - \lambda. \end{cases} \quad (4.60)$$

The reason we don't use (4.55) to get  $\mathcal{N}$  or (4.44) to get  $P(X | DI)$  is because when we set  $m_X = m_L = 0$  we find that they are valid for  $1 - \lambda = n$ . But this is not possible since  $\lambda \geq 0$  and

because we are under the assumption that  $n \geq 2$ . On the other hand (4.54) and (4.43) are both valid for  $1 - \lambda \neq n$  or  $1 - \lambda < n$ , but since  $n \geq 2$  we simply say they are valid for  $n \geq 2$  and  $\forall \lambda$  instead. Substituting the normalization constant in  $P(X | DI)$  and  $P(L | DI)$  above, they become

$$P(X | DI) = \begin{cases} \frac{M_2^{-1} - (x_{max} - X)^{-1}U[X | x_{max} - M_2, x_{min}]}{\ln \frac{x_\Delta}{M_2} - M_2^{-1}x_\Delta + 1}, & \text{for } n = 2 - \lambda \\ \frac{M_2^{-n-\lambda+1} - (x_{max} - X)^{-n-\lambda+1}U[X | x_{max} - M_2, x_{min}]}{M_2^{-n-\lambda+1}(M_2 - x_\Delta) + \frac{x_\Delta^{-n-\lambda+2} - M_2^{-n-\lambda+2}}{-n-\lambda+2}}, & \text{for } n > 2 - \lambda, \end{cases} \quad (4.61)$$

and

$$P(L | DI) = \begin{cases} \frac{(L - x_\Delta)L^{-2}U[L | x_\Delta, M_2]}{M_2^{-1}x_\Delta - \ln \frac{x_\Delta}{M_2} - 1}, & \text{for } n = 2 - \lambda \\ \frac{(-n - \lambda + 1)(L - x_\Delta)L^{-n-\lambda}U[L | x_\Delta, M_2]}{M_2^{-n-\lambda+1}(M_2 - x_\Delta) + \frac{x_\Delta^{-n-\lambda+2} - M_2^{-n-\lambda+2}}{-n-\lambda+2}}, & \text{for } n > 2 - \lambda. \end{cases} \quad (4.62)$$

The first moment, or mean estimator for  $X$  is given by  $\frac{1}{N}C(m_X = 1, m_L = 0)$  or

$$\langle X \rangle = \frac{1}{N} \frac{1}{(-n - \lambda + 1)} \left\{ M_2^{-n-\lambda+1} \int_{x_{max}-M_2}^{x_{min}} dX X - \int_{x_{max}-M_2}^{x_{min}} dX X (x_{max} - X)^{-n-\lambda+1} \right\}. \quad (4.63)$$

The second integral is easily evaluated by making the substitution  $z = x_{max} - X$ . Completing the integration for the three cases that emerge, and dividing by the normalization constant, the results are

$$\langle X \rangle = \begin{cases} \frac{M_2^{-1} \frac{x_{min}^2 - (x_{max} - M_2)^2}{2} + x_{max} \ln \frac{x_\Delta}{M_2} - (x_\Delta - M_2)}{\ln \frac{x_\Delta}{M_2} - \frac{x_\Delta}{M_2} + 1}, & \text{for } n = 2 - \lambda \\ \frac{M_2^{-2} \frac{x_{min}^2 - (x_{max} - M_2)^2}{2} - x_{max}(x_\Delta^{-1} - M_2^{-1}) - \ln \frac{x_\Delta}{M_2}}{2M_2^{-1} - M_2^{-2}x_\Delta - x_\Delta^{-1}}, & \text{for } n = 3 - \lambda \\ \frac{M_2^{-n-\lambda+1} \frac{x_{min}^2 - (x_{max} - M_2)^2}{2} + x_{max} \frac{x_\Delta^{-n-\lambda+2} - M_2^{-n-\lambda+2}}{-n-\lambda+2} - \frac{x_\Delta^{-n-\lambda+3} - M_2^{-n-\lambda+3}}{-n-\lambda+3}}{M_2^{-n-\lambda+1}(M_2 - x_\Delta) + \frac{x_\Delta^{-n-\lambda+2} - M_2^{-n-\lambda+2}}{-n-\lambda+2}}, & \text{for } n > 3 - \lambda. \end{cases} \quad (4.64)$$

The first moment, or mean estimator for  $L$  is  $\frac{1}{N}C(m_X = 0, m_L = 1)$ . The integrals are trivial, and the results are

$$\langle L \rangle = \begin{cases} \frac{M_2 + x_\Delta (\ln \frac{x_\Delta}{M_2} - 1)}{M_2^{-1} x_\Delta - \ln \frac{x_\Delta}{M_2} - 1}, & \text{for } n = 2 - \lambda \\ \frac{2(\ln \frac{x_\Delta}{M_2} - M_2^{-1} x_\Delta + 1)}{2M_2^{-1} - M_2^{-2} x_\Delta - x_\Delta^{-1}}, & \text{for } n = 3 - \lambda \\ \frac{(-n - \lambda + 1) \left( M_2^{-n-\lambda+2} (M_2 - x_\Delta) + \frac{x_\Delta^{-n-\lambda+3} - M_2^{-n-\lambda+3}}{-n-\lambda+3} \right)}{(-n - \lambda + 2) \left( M_2^{-n-\lambda+1} (M_2 - x_\Delta) + \frac{x_\Delta^{-n-\lambda+2} - M_2^{-n-\lambda+2}}{-n-\lambda+2} \right)}, & \text{for } n > 3 - \lambda. \end{cases} \quad (4.65)$$

The second moment of  $X$  is given by  $\frac{1}{N}C(m_X = 2, m_L = 0)$ . The result is

$$\langle X^2 \rangle = \begin{cases} \frac{M_2^{-1} \frac{x_{\min}^3 - (x_{\max} - M_2)^3}{3} + x_{\max}^2 \ln \frac{x_\Delta}{M_2} - 2x_{\max} (x_\Delta - M_2) + \frac{x_\Delta^2 - M_2^2}{2}}{M_2^{-1} x_\Delta - \ln \frac{x_\Delta}{M_2} - 1}, & \text{for } n = 2 - \lambda \\ \frac{M_2^{-2} \frac{x_{\min}^3 - (x_{\max} - M_2)^3}{3} + x_{\max}^2 \frac{x_\Delta^{-1} - M_2^{-1}}{-1} - 2x_{\max} \ln \frac{x_\Delta}{M_2} + x_\Delta - M_2}{2M_2^{-1} - M_2^{-2} x_\Delta - x_\Delta^{-1}}, & \text{for } n = 3 - \lambda \\ \frac{M_2^{-3} \frac{x_{\min}^3 - (x_{\max} - M_2)^3}{3} + x_{\max}^2 \frac{x_\Delta^{-2} - M_2^{-2}}{-2} - 2x_{\max} \frac{x_\Delta^{-1} - M_2^{-1}}{-1} + \ln \frac{x_\Delta}{M_2}}{\frac{3}{2} M_2^{-2} - M_2^{-3} x_\Delta - \frac{1}{2} x_\Delta^{-2}}, & \text{for } n = 4 - \lambda \\ \frac{M_2^{-n-\lambda+1} \frac{x_{\min}^3 - (x_{\max} - M_2)^3}{3} + x_{\max}^2 \frac{x_\Delta^{-n-\lambda+2} - M_2^{-n-\lambda+2}}{-n-\lambda+2} - 2x_{\max} \frac{x_\Delta^{-n-\lambda+3} - M_2^{-n-\lambda+3}}{-n-\lambda+3} + \frac{x_\Delta^{-n-\lambda+4} - M_2^{-n-\lambda+4}}{-n-\lambda+4}}{M_2^{-n-\lambda+1} (M_2 - x_\Delta) + \frac{x_\Delta^{-n-\lambda+2} - M_2^{-n-\lambda+2}}{-n-\lambda+2}}, & \text{for } n > 4 - \lambda \end{cases} \quad (4.66)$$

The second moment of  $L$  is given by  $\frac{1}{N}C(m_X = 0, m_L = 2)$ . The result is

$$\langle L^2 \rangle = \begin{cases} \frac{\frac{1}{2} M_2^2 - M_2 x_\Delta + \frac{1}{2} x_\Delta^2}{M_2^{-1} x_\Delta - \ln \frac{x_\Delta}{M_2} - 1} & \text{for } n = 2 - \lambda \\ \frac{-2 \left( M_2 + x_\Delta (\ln \frac{x_\Delta}{M_2} - 1) \right)}{2M_2^{-1} - M_2^{-2} x_\Delta - x_\Delta^{-1}} & \text{for } n = 3 - \lambda \\ \frac{3 \left( \ln \frac{x_\Delta}{M_2} - M_2^{-1} x_\Delta + 1 \right)}{\frac{3}{2} M_2^{-2} - M_2^{-3} x_\Delta - \frac{1}{2} x_\Delta^{-2}} & \text{for } n = 4 - \lambda \\ \frac{(-n - \lambda + 1) \left( M_2^{-n-\lambda+3} (M_2 - x_\Delta) + \frac{x_\Delta^{-n-\lambda+4} - M_2^{-n-\lambda+4}}{-n-\lambda+4} \right)}{(-n - \lambda + 3) \left( M_2^{-n-\lambda+1} (M_2 - x_\Delta) + \frac{x_\Delta^{-n-\lambda+2} - M_2^{-n-\lambda+2}}{-n-\lambda+2} \right)} & \text{for } n > 4 - \lambda \end{cases} \quad (4.67)$$

## 4.5 The $M_2$ limit

Clearly the above results depend strongly on the choice of  $M_2$ , and if our prior information  $I$  yields its value, the above are the full and final answers. Suppose we do not have any information regarding the size of the parameter  $L$  and we take the limit  $M_2 \rightarrow \infty$ . What effect will this have on the results gathered in the previous section?

Starting with the marginal posterior distributions we take the limit  $M_2 \rightarrow \infty$ . Observing (4.61) and (4.62), we see that taking the limit is only sensible for the  $2 - \lambda < n$  case and thus the marginal distributions become

$$P(X | DI) = \frac{(n + \lambda - 2)(x_{max} - X)^{-n-\lambda+1} U(X | -\infty, x_{min})}{x_{\Delta}^{-n-\lambda+2}}, \text{ for } n > 2 - \lambda, \quad (4.68)$$

and

$$P(L | DI) = \frac{(n + \lambda - 1)(n + \lambda - 2)(L - x_{\Delta}) L^{-n-\lambda} U[L | x_{\Delta}, \infty)}{x_{\Delta}^{-n-\lambda+2}}, \text{ for } n > 2 - \lambda, \quad (4.69)$$

where  $U(z | a, b)$  and  $U[z | a, b)$  are defined in (A.3) and (A.4) of Appendix A. Next, consider the mean estimators of the parameters (4.64) and (4.65), and take the limit  $M_2 \rightarrow \infty$ . The only usable results are for the  $3 - \lambda < n$  case. They are

$$\langle X \rangle = x_{max} - \frac{n + \lambda - 2}{n + \lambda - 3} x_{\Delta}, \text{ for } n > 3 - \lambda, \quad (4.70)$$

and

$$\langle L \rangle = \frac{n + \lambda - 1}{n + \lambda - 3} x_{\Delta}, \text{ for } n > 3 - \lambda. \quad (4.71)$$

Doing the same for the second moment of the parameters, we find from (4.66) and (4.67) that

$$\langle X^2 \rangle = x_{max}^2 - \frac{2(n + \lambda - 2)}{n + \lambda - 3} x_{max} x_{\Delta} + \frac{n + \lambda - 2}{n + \lambda - 4} x_{\Delta}^2, \text{ for } n > 4 - \lambda, \quad (4.72)$$

and

$$\langle L^2 \rangle = \frac{(n + \lambda - 1)(n + \lambda - 2)}{(n + \lambda - 3)(n + \lambda - 4)} x_{\Delta}^2, \text{ for } n > 4 - \lambda. \quad (4.73)$$

The variance is thus

$$\langle L^2 \rangle - \langle L \rangle^2 = \frac{2(n + \lambda - 1)}{(n + \lambda - 3)(n + \lambda - 4)} x_{\Delta}^2, \quad (4.74)$$

which goes like  $\frac{1}{n^2} x_{\Delta}^2$ .

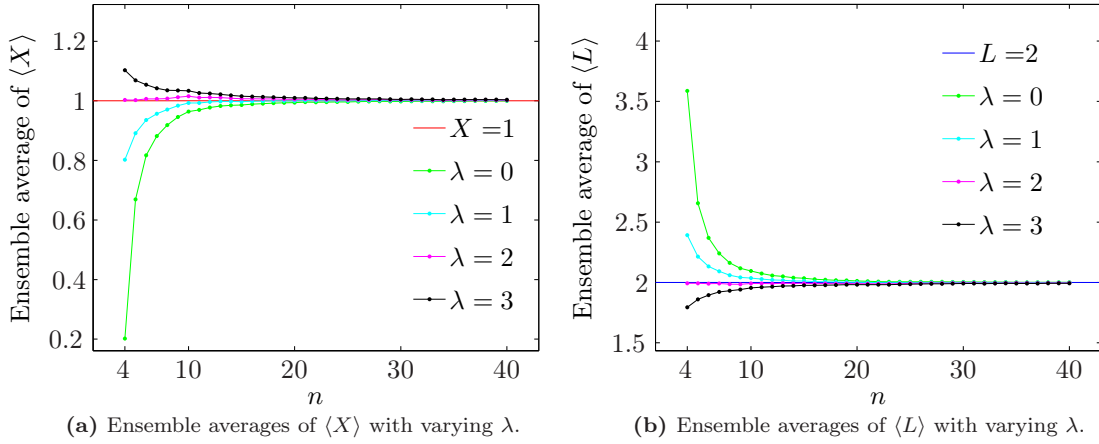
For large  $n$ , we therefore recover the intuitive results

$$\lim_{n \rightarrow \infty} \langle X \rangle = x_{min}, \quad (4.75)$$

and

$$\lim_{n \rightarrow \infty} \langle L \rangle = x_{\Delta}, \quad (4.76)$$

so that the standard deviations do tend to zero even though  $P(X | DI)$  and  $P(L | DI)$  are highly non-gaussian. In Fig. 4.3, we show how these limits are approached in a simulation in which an ensemble average is calculated for different values of  $\lambda$ . Clearly the choice  $\lambda = 1$ ,  $\lambda = 2$  or  $\lambda = 3$  yields faster convergence.



**Figure 4.3:** Ensemble averages of  $\langle X \rangle$  and  $\langle L \rangle$  for varying  $\lambda$ . The parameters were set at  $X = 1$  and  $L = 2$ . 1000 sets of generated data were used.

#### 4.5.1 Uniform prior

Setting  $\lambda = 0$  we now consider the uniform prior. The marginal posterior distributions of the parameters are

$$P(X | DI) = \frac{(n-2)(x_{max} - X)^{-n+1} U(X | -\infty, x_{min})}{x_{\Delta}^{-n+2}}, \quad \text{for } n > 2, \quad (4.77)$$

and

$$P(L | DI) = \frac{(n-1)(n-2)(L - x_{\Delta}) L^{-n} U[L | x_{\Delta}, \infty)}{x_{\Delta}^{-n+2}}, \quad \text{for } n > 2. \quad (4.78)$$

The mean estimators are

$$\langle X \rangle = x_{max} - \frac{n-2}{n-3} x_{\Delta}, \quad \text{for } n > 3, \quad (4.79)$$

and

$$\langle L \rangle = \frac{n-1}{n-3} x_{\Delta}, \quad \text{for } n > 3. \quad (4.80)$$

The second moment is given by

$$\langle X^2 \rangle = x_{max}^2 - \frac{2(n-2)}{n-3} x_{max} x_{\Delta} + \frac{n-2}{n-4} x_{\Delta}^2, \quad \text{for } n > 4, \quad (4.81)$$

and

$$\langle L^2 \rangle = \frac{(n-1)(n-2)}{(n-3)(n-4)} x_{\Delta}^2, \text{ for } n > 4. \quad (4.82)$$

### 4.5.2 Jeffreys prior

By choosing  $\lambda = 1$ , we now consider the Jeffreys prior. The marginal posterior distributions of the parameters shown in Fig. 4.4 are

$$P(X | DI) = \frac{(n-1)(x_{max} - X)^{-n} U(X | -\infty, x_{min})}{x_{\Delta}^{-n+1}}, \text{ for } n > 1, \quad (4.83)$$

and

$$P(L | DI) = \frac{n(n-1)(L - x_{\Delta})L^{-n-1}U[L | x_{\Delta}, \infty)}{x_{\Delta}^{-n+1}}, \text{ for } n > 1. \quad (4.84)$$

They are strongly non-gaussian (asymmetric). Cut-offs determined by  $x_{max}$  can change every time a new data point is collected. The ‘jumps’ (see Fig. 4.6 below) hence accurately reflect the occurrence of a new  $x_{max}$  or  $x_{\Delta}$  value.

The mean estimators are

$$\langle X \rangle = x_{max} - \frac{n-1}{n-2} x_{\Delta}, \text{ for } n > 2, \quad (4.85)$$

and

$$\langle L \rangle = \frac{n}{n-2} x_{\Delta}, \text{ for } n > 2. \quad (4.86)$$

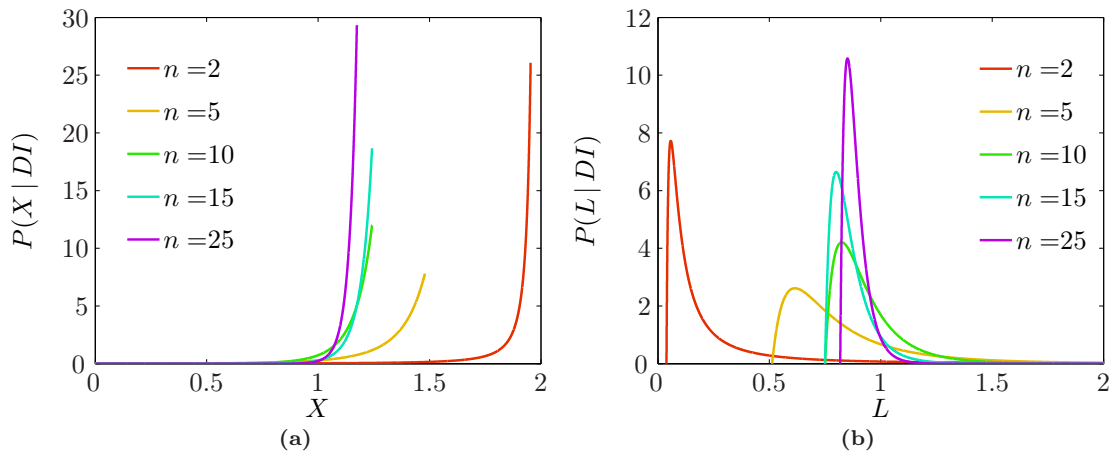
Finally, the second moment is

$$\langle X^2 \rangle = x_{max}^2 - \frac{2(n-1)}{n-2} x_{max} x_{\Delta} + \frac{n-1}{n-3} x_{\Delta}^2, \text{ for } n > 3, \quad (4.87)$$

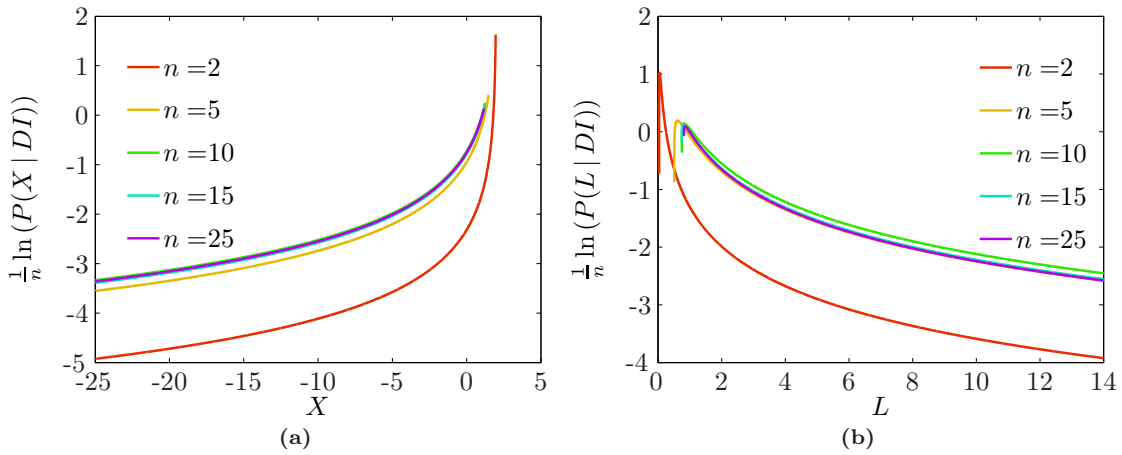
and

$$\langle L^2 \rangle = \frac{n(n-1)}{(n-2)(n-3)} x_{\Delta}^2, \text{ for } n > 3. \quad (4.88)$$

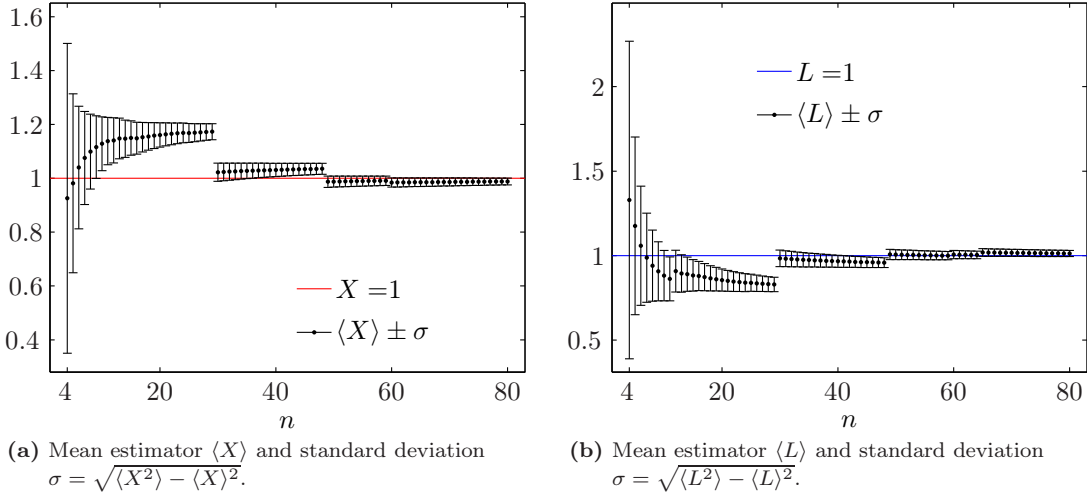
The marginal posteriors for  $X$  and  $L$  are shown in Fig 4.4 and 4.5 below. The mean  $\pm$  standard deviations of  $X$  and  $L$  are shown in Fig 4.6.



**Figure 4.4:** Posterior distributions  $P(X | DI)$  and  $P(L | DI)$  for the Jeffreys prior and increasing  $n$ . The same generated data set is used in both figures with parameters set at  $X = 1$  and  $L = 1$ .



**Figure 4.5:**  $\frac{1}{n} \ln(P(X | DI))$  and  $\frac{1}{n} \ln(P(L | DI))$  for the Jeffreys prior and increasing  $n$ . The same data is used as in the Fig. 4.4 above. The power-law growth is evident.



**Figure 4.6:** Mean estimators and standard deviations of  $X$  and  $L$  with Jeffreys prior. The same generated data set is used in both figures. Evident is the decreasing  $\sigma$  and the ‘jumps’ caused by the updating of  $x_{max}$  or  $x_{min}$ .

#### 4.6 The $n = 1$ case

Since we started the above general derivation we assumed that  $n \geq 2$ . This assumption is first introduced in (4.28) where we let  $x_{min} = \min(x_1, x_2, \dots, x_n)$  and  $x_{max} = \max(x_1, x_2, \dots, x_n)$ . However if  $n = 1$ , and we only have one data point  $x_1$ , then  $\min(x_1, x_2, \dots, x_n) = \min(x_1) = x_1$  and  $\max(x_1, x_2, \dots, x_n) = \max(x_1) = x_1$ , and (4.28) should read

$$\prod_{i=1}^n U[x_i | X, X + L] = \theta(x_1 - X)\theta(X + L - x_1)$$

and we have to do the calculation of  $A$ ,  $B$  and  $C$  over again. Fortunately, going through the above steps carefully, we see that in calculating  $A$  and  $B$ , we are free to use our current results if we replace  $x_{max}$  with  $x_1$  and  $x_{min}$  with  $x_1$  and set  $n = 1$ . Thus, (4.36) becomes

$$A = \frac{(\min(x_1, K_2))^{m_X+1} - (\max(K_1, x_1 - L))^{m_X+1}}{m_X + 1} L^{m_L - \lambda - 1} U[L | \max(M_1, x_1 - \min(x_1, K_2)), M_2], \quad (4.89)$$

since  $m_X \geq 0$ , and (4.43) and (4.44) become

$$B = \frac{(M_2)^{m_L - \lambda} - (\max(M_1, x_1 - X))^{m_L - \lambda}}{m_L - \lambda} X^{m_X} U[X | \max(K_1, x_1 - M_2), \min(x_1, K_2)] \quad (4.90)$$

for  $m_L - \lambda \neq 0$ , and

$$B = \ln \left( \frac{M_2}{\max(M_1, x_1 - X)} \right) X^{m_X} U[X | \max(K_1, x_1 - M_2), \min(x_1, K_2)] \quad (4.91)$$

for  $m_L - \lambda = 0$ . In calculating  $C$  however, we have to make these replacements as early as (4.45) and (4.46), and complete the rest of the calculation with caution. They become



$$C = \frac{1}{m_L - \lambda} \left\{ M_2^{m_L - \lambda} \int_{\max(K_1, x_1 - M_2)}^{\min(x_1, K_2)} dX X^{m_X} - \int_{\max(K_1, x_1 - M_2)}^{\min(x_1, K_2)} dX X^{m_X} (\max(M_1, x_1 - X))^{m_L - \lambda} \right\}, \quad (4.92)$$

for  $m_L - \lambda \neq 0$ , and

$$C = \ln M_2 \int_{\max(K_1, x_1 - M_2)}^{\min(x_1, K_2)} dX X^{m_X} - \int_{\max(K_1, x_1 - M_2)}^{\min(x_1, K_2)} dX X^{m_X} \ln(\max(M_1, x_1 - X)) \quad (4.93)$$

for  $m_L - \lambda = 0$ . Now, similar to what we had before, we have to resolve the terms

$$\int_{\max(K_1, x_1 - M_2)}^{\min(x_1, K_2)} dX X^{m_X} (\max(M_1, x_1 - X))^{m_L - \lambda}, \quad (4.94)$$

and

$$\int_{\max(K_1, x_1 - M_2)}^{\min(x_1, K_2)} dX X^{m_X} \ln(\max(M_1, x_1 - X)) \quad (4.95)$$

where we have a maximum criterion standing inside the integrals. This is done in exactly the same way as illustrated in (4.48) and (4.50), and in fact the results are just (4.48) and (4.50) after setting  $n = 1$  and replacing  $x_{max}$  and  $x_{min}$  with  $x_1$ . Thus (4.48) becomes

$$\begin{aligned} & \theta((x_1 - M_1) - \min(x_1, K_2)) \int_{\max(K_1, x_1 - M_2)}^{\min(x_1, K_2)} dX X^{m_X} (x_1 - X)^{m_L - \lambda} \\ & + \theta((x_1 - M_1) - \max(K_1, x_1 - M_2)) \theta(\min(x_1, K_2) - (x_1 - M_1)) \\ & \times \left\{ \int_{\max(K_1, x_1 - M_2)}^{x_1 - M_1} dX X^{m_X} (x_1 - X)^{m_L - \lambda} + M_1^{m_L - \lambda} \int_{x_1 - M_1}^{\min(x_1, K_2)} dX X^{m_X} \right\} \\ & + \theta(\max(K_1, x_1 - M_2) - (x_1 - M_1)) M_1^{m_L - \lambda} \int_{\max(K_1, x_1 - M_2)}^{\min(x_1, K_2)} dX X^{m_X}, \end{aligned} \quad (4.96)$$

and (4.50) becomes

$$\begin{aligned} & \theta((x_1 - M_1) - \min(x_1, K_2)) \int_{\max(K_1, x_1 - M_2)}^{\min(x_1, K_2)} dX X^{m_X} \ln(x_1 - X) \\ & + \theta((x_1 - M_1) - \max(K_1, x_1 - M_2)) \theta(\min(x_1, K_2) - (x_1 - M_1)) \\ & \times \left\{ \int_{\max(K_1, x_1 - M_2)}^{x_1 - M_1} dX X^{m_X} \ln(x_1 - X) + \ln M_1 \int_{x_1 - M_1}^{\min(x_1, K_2)} dX X^{m_X} \right\} \\ & + \theta(\max(K_1, x_1 - M_2) - (x_1 - M_1)) \ln M_1 \int_{\max(K_1, x_1 - M_2)}^{\min(x_1, K_2)} dX X^{m_X}. \end{aligned} \quad (4.97)$$

Now as before, we make a more specific choice of prior by taking the limits  $K_1 \rightarrow -\infty$  and  $K_2 \rightarrow \infty$ . Concentrating on (4.96) for the moment, it becomes

$$\begin{aligned}
 & \theta(-M_1) \int_{x_1-M_2}^{x_1} dX X^{m_X} (x_1 - X)^{m_L-\lambda} \\
 & + \theta(M_2 - M_1)\theta(M_1) \left\{ \int_{x_1-M_2}^{x_1-M_1} dX X^{m_X} (x_1 - X)^{m_L-\lambda} + M_1^{m_L-\lambda} \int_{x_1-M_1}^{x_1} dX X^{m_X} \right\} \\
 & + \theta(M_1 - M_2)M_1^{m_L-\lambda} \int_{x_1-M_2}^{x_1} dX X^{m_X}.
 \end{aligned} \tag{4.98}$$

However, since we have chosen  $M_2 > M_1 > 0$ , only the second term survives, namely

$$\int_{x_1-M_2}^{x_1-M_1} dX X^{m_X} (x_1 - X)^{m_L-\lambda} + M_1^{m_L-\lambda} \int_{x_1-M_1}^{x_1} dX X^{m_X}. \tag{4.99}$$

In exactly the same way, (4.97) becomes

$$\int_{x_1-M_2}^{x_1-M_1} dX X^{m_X} \ln(x_1 - X) + \ln M_1 \int_{x_1-M_1}^{x_1} dX X^{m_X}. \tag{4.100}$$

Notice that due to the  $\ln M_1$  we do not take the limit  $M_1 \rightarrow 0$  at this stage of the calculation. Taking the limits  $K_1 \rightarrow -\infty$  and  $K_2 \rightarrow \infty$  in  $C$ , and substituting the above we have

$$C = \frac{1}{m_L - \lambda} \left\{ M_2^{m_L-\lambda} \int_{x_1-M_2}^{x_1} dX X^{m_X} - M_1^{m_L-\lambda} \int_{x_1-M_1}^{x_1} dX X^{m_X} - \int_{x_1-M_2}^{x_1-M_1} dX X^{m_X} (x_1 - X)^{m_L-\lambda} \right\}, \tag{4.101}$$

for  $m_L - \lambda \neq 0$ , and

$$C = \ln M_2 \int_{x_1-M_2}^{x_1} dX X^{m_X} - \ln M_1 \int_{x_1-M_1}^{x_1} dX X^{m_X} - \int_{x_1-M_2}^{x_1-M_1} dX X^{m_X} \ln(x_1 - X), \tag{4.102}$$

for  $m_L - \lambda = 0$ . Now that we have simplified  $C$ , let us take the same limits in  $A$  and  $B$ , and then set  $m_X = m_L = 0$ , as these expressions divided by the normalization constant give us the marginal posterior distributions of the parameters. They are

$$P(X | x_1 I) = \begin{cases} \frac{1}{\mathcal{N}} \frac{M_2^{-\lambda} - (\max(M_1, x_1 - X))^{-\lambda}}{-\lambda} U[X | x_1 - M_2, x_1], & \text{for } \lambda > 0 \\ \frac{1}{\mathcal{N}} \ln \left( \frac{M_2}{\max(M_1, x_1 - X)} \right) U[X | x_1 - M_2, x_1], & \text{for } \lambda = 0, \end{cases} \tag{4.103}$$

and

$$P(L | x_1 I) = \frac{1}{\mathcal{N}} L^{-\lambda} U[L | \max(0, M_1), M_2] \forall \lambda. \tag{4.104}$$

Since we have chosen  $M_1 > 0$  however,  $\max(0, M_1) = M_1$  and  $P(L | x_1 I)$  becomes

$$P(L | x_1 I) = \frac{1}{\mathcal{N}} L^{-\lambda} U[L | M_1, M_2] \forall \lambda. \tag{4.105}$$

The normalization constant is given by  $\mathcal{N} = C(m_X = 0, m_L = 0)$ . Completing the integration the result is

$$\mathcal{N} = \begin{cases} M_2 - M_1, & \text{for } \lambda = 0 \\ \ln \frac{M_2}{M_1}, & \text{for } \lambda = 1 \\ \frac{M_2^{1-\lambda} - M_1^{1-\lambda}}{1-\lambda}, & \text{for } \lambda > 1 \end{cases} \quad (4.106)$$

Substituting the normalization constant in  $P(X | x_1 I)$  and  $P(L | x_1 I)$  above, and since  $\max(0, M_1) = M_1$ , the marginal posterior distributions are

$$P(X | x_1 I) = \begin{cases} \frac{\ln \left( \frac{M_2}{\max(M_1, x_1 - X)} \right) U[X | x_1 - M_2, x_1]}{M_2 - M_1}, & \text{for } \lambda = 0 \\ \frac{-(M_2^{-1} - (\max(M_1, x_1 - X))^{-1}) U[X | x_1 - M_2, x_1]}{\ln \frac{M_2}{M_1}}, & \text{for } \lambda = 1 \\ \frac{(1 - \lambda) (M_2^{-\lambda} - (\max(M_1, x_1 - X))^{-\lambda}) U[X | x_1 - M_2, x_1]}{M_2^{1-\lambda} - M_1^{1-\lambda}}, & \text{for } \lambda > 1. \end{cases} \quad (4.107)$$

The first moment of  $X$  is  $\frac{1}{\mathcal{N}} C(m_X = 1, m_L = 0)$ . The result is

$$\langle X \rangle = \begin{cases} x_1 - \frac{M_2^2 (\ln M_2 - \frac{1}{4}) - M_1^2 (\ln M_1 - \frac{1}{4})}{M_2 - M_1}, & \text{for } \lambda = 0 \\ x_1 - \frac{M_2 - M_1}{2 \ln \frac{M_2}{M_1}}, & \text{for } \lambda = 1 \\ x_1 - \frac{\ln \frac{M_2}{M_1}}{2(M_1^{-1} - M_2^{-1})}, & \text{for } \lambda = 2 \\ x_1 - \frac{1 - \lambda}{2(2 - \lambda)} \left( \frac{M_2^{2-\lambda} - M_1^{2-\lambda}}{M_2^{1-\lambda} - M_1^{1-\lambda}} \right), & \text{for } \lambda > 2, \end{cases} \quad (4.108)$$

and

$$P(L | x_1 I) = \begin{cases} \frac{U[L | M_1, M_2]}{M_2 - M_1}, & \text{for } \lambda = 0 \\ \frac{L^{-1} U[L | M_1, M_2]}{\ln \frac{M_2}{M_1}}, & \text{for } \lambda = 1 \\ \frac{(1 - \lambda) L^{-\lambda} U[L | M_1, M_2]}{M_2^{1-\lambda} - M_1^{1-\lambda}}, & \text{for } \lambda > 1. \end{cases} \quad (4.109)$$

The first moment of  $L$  is  $\frac{1}{\mathcal{N}} C(m_X = 0, m_L = 1)$ . The result is

$$\langle L \rangle = \begin{cases} \frac{M_2 + M_1}{2}, & \text{for } \lambda = 0 \\ \frac{M_2 - M_1}{\ln \frac{M_2}{M_1}}, & \text{for } \lambda = 1 \\ \frac{-\ln \frac{M_2}{M_1}}{M_2^{-1} - M_1^{-1}}, & \text{for } \lambda = 2 \\ \frac{1 - \lambda}{2 - \lambda} \left( \frac{M_2^{2-\lambda} - M_1^{2-\lambda}}{M_2^{1-\lambda} - M_1^{1-\lambda}} \right), & \text{for } \lambda > 2 \end{cases} \quad (4.110)$$

The second moment, of  $X$  is  $\frac{1}{\mathcal{N}}C(m_X = 2, m_L = 0)$ . The result is

$$\langle X^2 \rangle = \begin{cases} x_1^2 + \frac{(\frac{2}{3}M_2^3 - 2M_2^2x_1) \ln M_2 - (\frac{2}{3}M_1^3 - 2M_1^2x_1) \ln M_1 + \frac{1}{2}(M_2^2 - M_1^2)x_1 - \frac{1}{9}(M_2^3 - M_1^3)}{M_2 - M_1}, & \text{for } \lambda = 0 \\ x_1^2 - \frac{(M_2 - M_1)x_1 + \frac{1}{6}(M_2^2 - M_1^2)}{\ln \frac{M_2}{M_1}}, & \text{for } \lambda = 1 \\ x_1^2 + \frac{x_1 \ln \frac{M_2}{M_1} - \frac{1}{3}(M_2 - M_1)}{M_2^{-1} - M_1^{-1}}, & \text{for } \lambda = 2 \\ x_1^2 - \frac{2(M_2^{-1} - M_1^{-1})x_1 + \frac{2}{3} \ln \frac{M_2}{M_1}}{M_2^{-2} - M_1^{-2}}, & \text{for } \lambda = 3 \\ x_1^2 - \frac{\frac{1-\lambda}{2-\lambda}(M_2^{2-\lambda} - M_1^{2-\lambda})x_1 - \frac{1-\lambda}{3(3-\lambda)}(M_2^{3-\lambda} - M_1^{3-\lambda})}{M_2^{1-\lambda} - M_1^{1-\lambda}}, & \text{for } \lambda > 3 \end{cases} \quad (4.111)$$

The second moment of  $L$  is  $\frac{1}{\mathcal{N}}C(m_X = 0, m_L = 2)$ . The result is

$$\langle L^2 \rangle = \begin{cases} \frac{1}{3}(M_2^2 + M_2M_1 + M_1^2), & \text{for } \lambda = 0 \\ \frac{M_2^2 - M_1^2}{2 \ln \frac{M_2}{M_1}}, & \text{for } \lambda = 1 \\ M_2M_1, & \text{for } \lambda = 2 \\ \frac{-2 \ln \frac{M_2}{M_1}}{M_2^{-2} - M_1^{-2}}, & \text{for } \lambda = 3 \\ \frac{1 - \lambda}{3 - \lambda} \left( \frac{M_2^{3-\lambda} - M_1^{3-\lambda}}{M_2^{1-\lambda} - M_1^{1-\lambda}} \right), & \text{for } \lambda > 3 \end{cases} \quad (4.112)$$

The results are dominated by the prior parameters as they should be. These are the best possible answers given very limited information.

#### 4.7 Posteriors for known $X$

For completeness, let us consider the cases where  $X$  or  $L$  are assumed to be known as part of our prior information. We begin by assuming  $X$  to be some known constant. When  $X = 0$  this is the popular introductory statistics problem of estimating the length of a uniform distribution also known as the continuous version of the taxicab problem [13] or the German tank problem. The posterior for  $L$  is

$$P(L | DXI) = \frac{P(D | XLI)P(L | XI)}{\int dL P(D | XLI)P(L | XI)}, \quad (4.113)$$

where, as before, the likelihood is

$$P(D | LI) = L^{-n} \prod_{i=1}^n U[x_i | X, X + L], \quad (4.114)$$

and the prior on the scale parameter  $L$  is

$$P(L | XI) = P(L | I) = L^{-\lambda} U[L | M_1, M_2]. \quad (4.115)$$

Substituting the above, we need

$$P(X | DLI) = \frac{L^{-n-\lambda} \prod_{i=1}^n U[x_i | X, X + L] U[L | M_1, M_2]}{\int dX L^{-n-\lambda} \prod_{i=1}^n U[x_i | X, X + L] U[L | M_1, M_2]}. \quad (4.116)$$

Prepare the product of theta functions for an integration over  $L$  as follows

$$\begin{aligned} & \prod_{i=1}^n U[x_i | X, X + L] U[L | M_1, M_2] \\ &= \theta(x_{\min} - X) \theta(X + L - x_{\max}) \theta(L - M_1) \theta(M_2 - L) \\ &= \theta(L - (x_{\max} - X)) \theta(L - M_1) \theta(M_2 - L) \theta(x_{\min} - X) \\ &= \theta(L - \max(x_{\max} - X, M_1)) \theta(M_2 - L) \theta(x_{\min} - X) \\ &= U[L | \max(x_{\max} - X, M_1), M_2] \theta(x_{\min} - X). \end{aligned} \quad (4.117)$$

The following integration is needed to get the moments of the posterior

$$\begin{aligned} & \int dL L^{-n-\lambda+m_L} U[L | \max(x_{\max} - X, M_1), M_2] \theta(x_{\min} - X) \\ &= \begin{cases} \frac{M_2^{-n-\lambda+m_L+1} - (\max(x_{\max} - X, M_1))^{-n-\lambda+m_L+1}}{-n-\lambda+m_L+1}, & \text{for } -n-\lambda+m_L \neq -1 \\ \ln \left( \frac{M_2}{\max(x_{\max} - X, M_1)} \right), & \text{for } -n-\lambda+m_L = -1 \end{cases}. \end{aligned} \quad (4.118)$$

Here we have assumed our model to be consistent with our data and thus set  $\theta(x_{\min} - X) = 1$  from now on. The normalization constant is the above with  $m_L = 0$ , and the first moment (mean)

is the above with  $m_L = 1$  divided by the normalization constant and so on. The normalization constant  $\mathcal{N}$  is thus

$$\mathcal{N} = \frac{M_2^{-n-\lambda+1} - (\max(x_{\max} - X, M_1))^{-n-\lambda+1}}{-n - \lambda + 1}, \text{ for } -n - \lambda \leq -2, \quad (4.119)$$

since  $-n - \lambda = -1$  cannot occur because we are under the assumption that  $n \geq 2$  (and thus  $-n - \lambda \leq -2$ ). The posterior for  $L$  is thus

$$P(L | DXI) = \frac{(-n - \lambda + 1)L^{-n-\lambda}U[L | \max(x_{\max} - X, M_1), M_2]}{M_2^{-n-\lambda+1} - (\max(x_{\max} - X, M_1))^{-n-\lambda+1}}, \text{ for } n \geq 2 - \lambda. \quad (4.120)$$

The mean estimate of  $L$  is

$$\begin{aligned} \langle L \rangle &= \int dL L P(L | DXI) \\ &= \begin{cases} \frac{(-n - \lambda + 1) \ln \left( \frac{M_2}{\max(x_{\max} - X, M_1)} \right)}{M_2^{-n-\lambda+1} - (\max(x_{\max} - X, M_1))^{-n-\lambda+1}}, & \text{for } n = 2 - \lambda \\ \frac{(-n + \lambda - 1) [M_2^{n-\lambda+2} - (\max(x_{\max} - X, M_1))^{-n-\lambda+2}]}{(n + \lambda - 2) [M_2^{-n-\lambda+1} - (\max(x_{\max} - X, M_1))^{-n-\lambda+1}]}, & \text{for } n > 2 - \lambda. \end{cases} \end{aligned} \quad (4.121)$$

The second moment is

$$\begin{aligned} \langle L^2 \rangle &= \int dL L^2 P(L | DXI) \\ &= \begin{cases} \frac{(-n - \lambda + 1) \ln \left( \frac{M_2}{\max(x_{\max} - X, M_1)} \right)}{M_2^{-n-\lambda+1} - (\max(x_{\max} - X, M_1))^{-n-\lambda+1}}, & \text{for } n = 3 - \lambda \\ \frac{(-n - \lambda + 1) [M_2^{n+\lambda-3} - (\max(x_{\max} - X, M_1))^{-n-\lambda+3}]}{(n - \lambda + 3) [M_2^{-n+\lambda-1} - (\max(x_{\max} - X, M_1))^{-n-\lambda+1}]}, & \text{for } n > 3 - \lambda \end{cases} \end{aligned} \quad (4.122)$$

Suppose there is no prior information concerning the size of  $L$ , then we imagine that  $M_1$  was chosen small in the sense that  $\max(x_{\max} - X, M_1) = x_{\max} - X$ . To allow for any possible size  $L$  send  $M_2 \rightarrow \infty$ . The posterior for  $L$  becomes

$$P(L | DXI) = \frac{(n + \lambda - 1)L^{-n-\lambda}U[L | x_{\max} - X, \infty]}{(x_{\max} - X)^{-n-\lambda+1}}, \text{ for } n \geq 2 - \lambda. \quad (4.123)$$

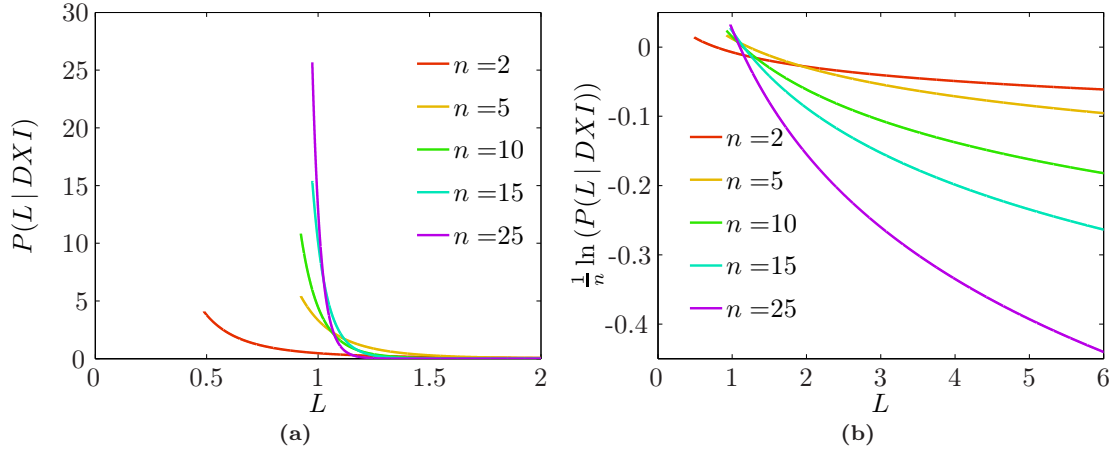
The mean estimate of  $L$  becomes

$$\langle L \rangle = \frac{(n + \lambda - 1)}{(n + \lambda - 2)}(x_{\max} - X), \text{ for } n > 2 - \lambda. \quad (4.124)$$

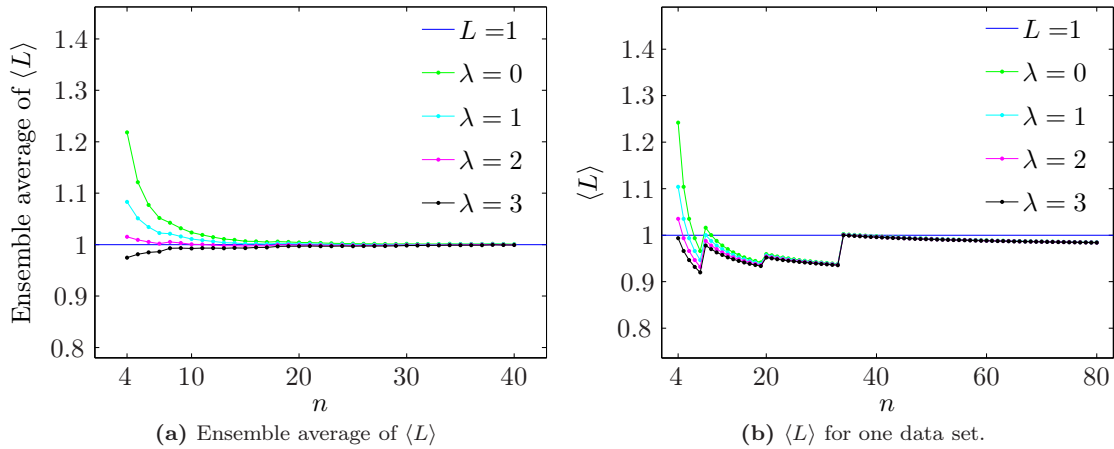
The second moment is

$$\langle L^2 \rangle = \frac{(n + \lambda - 1)}{(n + \lambda - 3)}(x_{\max} - X)^2, \text{ for } n > 3 - \lambda. \quad (4.125)$$

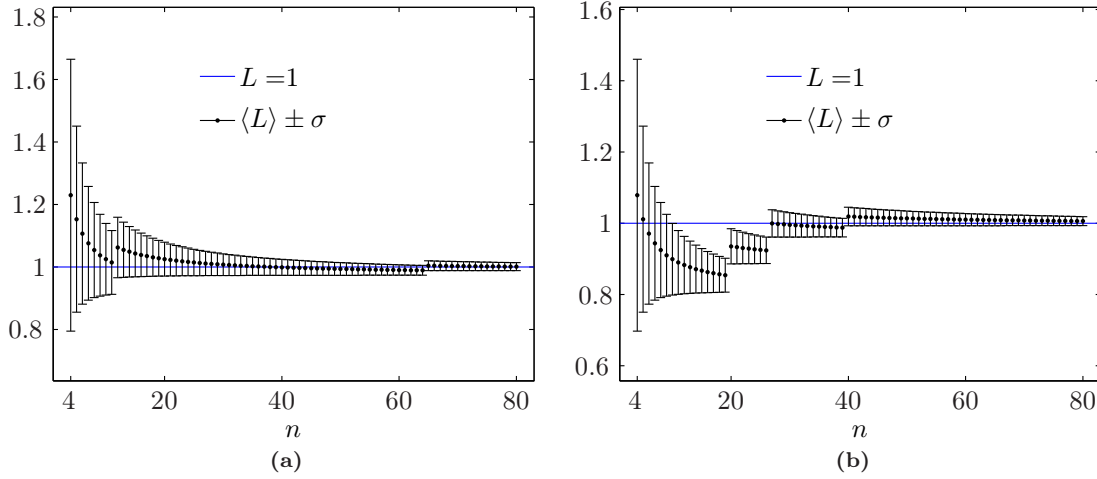
The posterior for  $L$  is shown in Fig 4.7 below. Ensemble average of the mean estimate for  $L$  and for different choices of  $\lambda$  (which describes the particular choice of prior distribution) are shown in Fig 4.8. The mean  $\pm$  standard deviation of  $L$  is shown in Fig 4.9.



**Figure 4.7:**  $P(L|DXI)$  and  $\frac{1}{n} \ln(P(L|DXI))$  with Jeffreys prior and increasing  $n$ . Parameter  $X = 0$  is assumed known. Comparing with Figs. 4.4-4.5, we see that known  $X$  is a pure power-law posterior  $P(L|XDI)$  while unknown  $X$  ‘weakens’ the power law.



**Figure 4.8:** Ensemble average of  $\langle L \rangle$  for varying  $\lambda$ . Parameter  $X = 0$  is assumed known. 1000 sets of generated data was used. While an ensemble average converges quickly to the ‘true’  $L$ , a single data set will do so much more slowly (as one would expect).



**Figure 4.9:** Mean estimator and standard deviation ( $\sigma = \sqrt{\langle L^2 \rangle - \langle L \rangle^2}$ ) of  $L$  with Jeffreys prior ( $\lambda = 1$ ). Parameter  $X = 0$  is assumed known. The results for two different data sets are shown.

As is noted in [29], some of the many classical estimators proposed for this problem can be attained by a certain choice of improper prior. For example, when we set  $X = 0$  and  $\lambda = 2$ , then  $\langle L \rangle$  corresponds with the minimum variance unbiased estimator. When  $\lambda = 3$ , then we get the minimum mean square error estimator. The choice  $\lambda = 0$ , which corresponds to a flat prior, gives the same result had we just calculated the expectation of  $L$  over the likelihood. The choice deemed correct in the Bayesian literature is that of the Jeffreys prior ( $\lambda = 1$ ). The estimator it produces does not correspond with any classical estimator.

#### 4.8 Posteriors for known $L$

We need to calculate

$$P(X | DLI) = \frac{P(D | XLI)P(X | LI)}{\int dX P(D | XLI)P(X | LI)} \quad (4.126)$$

where, as before

$$P(D | XLI) = L^{-n} \prod_{i=1}^n U[x_i | X, X + L], \quad (4.127)$$

and

$$P(X | LI) = P(X | I) = U[X | K_1, K_2]. \quad (4.128)$$

Upon substitution we need

$$P(X | DLI) = \frac{\prod_{i=1}^n U[x_i | X, X + L]U[X | K_1, K_2]}{\int dX \prod_{i=1}^n U[x_i | X, X + L]U[X | K_1, K_2]}. \quad (4.129)$$

Prepare the product of theta functions for an integration over  $X$  as follows



$$\begin{aligned}
 \prod_{i=1}^n U[x_i | X, X + L] U[X | K_1, K_2] &= \\
 &= \theta(x_{\min} - X) \theta(X + L - x_{\max}) \theta(X - K_1) \theta(K_2 - X) \\
 &= \theta(X - (x_{\max} - L)) \theta(X - K_1) \theta(x_{\min} - X) \theta(K_2 - X) \\
 &= \theta(X - \max(x_{\max} - L, K_1)) \theta(\min(x_{\min}, K_2) - X) \\
 &= U[X | \max(x_{\max} - L, K_1), \min(x_{\min}, K_2)].
 \end{aligned} \tag{4.130}$$

The following integration is needed to get the moments of the posterior

$$\begin{aligned}
 \int dX X^{m_X} U[X | \max(x_{\max} - L, K_1), \min(x_{\min}, K_2)] \\
 = \frac{(\min(x_{\min}, K_2)^{m_X+1} - (\max(x_{\max} - L, K_1))^{m_X+1})}{m_X + 1}, \text{ for } n \geq 2, m_X \geq 0
 \end{aligned} \tag{4.131}$$

Setting  $m_X = 0$ , the normalization constant is

$$\mathcal{N} = \min(x_{\min}, K_2) - \max(x_{\max} - L, K_1), \text{ for } n \geq 2. \tag{4.132}$$

The posterior for  $X$  is then

$$P(X | DLI) = \frac{U[X | \max(x_{\max} - L, K_1), \min(x_{\min}, K_2)]}{\min(x_{\min}, K_2) - \max(x_{\max} - L, K_1)}, \text{ for } n \geq 2. \tag{4.133}$$

The mean estimator for  $X$  is

$$\langle X \rangle = \frac{1}{2} (\min(x_{\min}, K_2) + \max(x_{\max} - L, K_1)), \text{ for } n \geq 2. \tag{4.134}$$

The second moment of the posterior is

$$\langle X^2 \rangle = \frac{1}{3} \frac{(\min(x_{\min}, K_2))^3 - (\max(x_{\max} - L, K_1))^3}{\min(x_{\min}, K_2) - \max(x_{\max} - L, K_1)}, \text{ for } n \geq 2. \tag{4.135}$$

In order to allow for any possible signal send  $K_1 \rightarrow -\infty$  and  $K_2 \rightarrow \infty$ . The posterior becomes

$$P(X | DLI) = \frac{U[X | x_{\max} - L, x_{\min}]}{x_{\min} - (x_{\max} - L)}, \text{ for } n \geq 2. \tag{4.136}$$

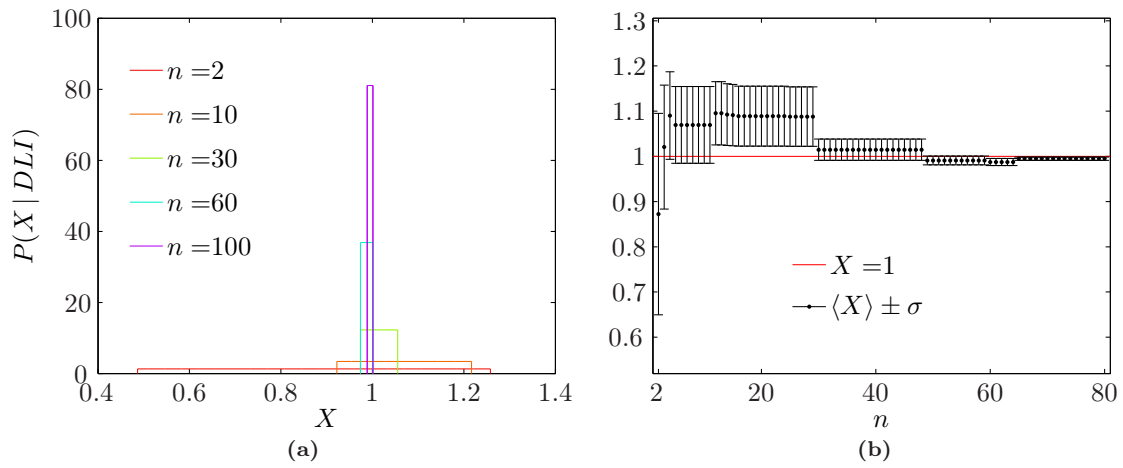
The mean estimator for  $X$  becomes

$$\langle X \rangle = \frac{1}{2} (x_{\min} + x_{\max} - L), \text{ for } n \geq 2. \tag{4.137}$$

The second moment of the posterior becomes

$$\langle X^2 \rangle = \frac{1}{3} \frac{x_{\min}^3 - (x_{\max} - L)^3}{x_{\min} - (x_{\max} - L)}, \text{ for } n \geq 2. \tag{4.138}$$

The posterior for and mean  $\pm$  standard deviation of  $X$  are shown in Fig 4.10 below.



**Figure 4.10:** Posterior distribution and mean estimator for  $X$ . Parameter  $L = 1$  is assumed known. The same generated data set is used in both figures. The location parameter  $X$  has a uniform posterior, the width of which decreases with  $n$ . The more certain we are of the parameter  $X$  the more the posterior goes into a delta function.

## Chapter 5

# Mixture Models: Bayesian Solution

### 5.1 Mixture models

On encountering outliers, the usual stance is to conclude that something has gone wrong with the apparatus, and then to throw the outlier data points away, and to proceed to draw conclusions from the remaining data which is perceived as good. This may be perfectly acceptable if one knows that the apparatus is unreliable. However, what if that outlier happens to be the most significant data point you have? Can one use it to draw conclusions about the parameters of interest or have you perhaps discovered an important physical phenomenon? The solution according to the orthodox statistics is to build estimators that have *robust* or *resistant* qualities, where robust means insensitive to the exact sampling distribution of errors and resistant means that large errors in a small proportion of the data do not greatly affect the conclusions. Jaynes [13] critiques these approaches saying that since one cannot define an optimally robust or resistant estimator one cannot define an optimal inference property. He further critiques that robust or resistant properties are bought at a price since one has poorer performance if the model is correct. From an orthodox view point, Bayesian methods have been criticized for not being robust or resistant for the same reason that the average of the data is not robust or resistant in that it is pulled toward the outlier. But this criticism is unfair and we shall soon see that with the correct choice of model, one which accommodates for the appearance of outliers, that the Bayesian machinery is shown to contain all the robust or resistance qualities desired and even tells us that in some cases it is permissible to throw a data point out altogether. The correct way to describe the situation from a Bayesian perspective is as follows, and is largely inspired from Jaynes' chapter on the subject ([13] Ch. 21). One is trying to measure a quantity  $\theta$ , but there is random noise contaminating the measurements and so the sampling distribution which describes this effect is

$$g(x | \theta, \mu), \quad (5.1)$$

containing possibly one or more uninteresting 'nuisance' parameters  $\mu$  which are to be integrated out of the joint posterior (remember that the conjunction of two propositions or parameters implies 'and'). In addition to this usual random noise, there is sometimes also impulsive noise which is responsible for outliers in the data. Suppose the sampling distribution describing this outlier process is

$$h(x | \theta, \nu) \quad (5.2)$$

containing possibly the nuisance parameter  $\nu$ . Parameters  $\mu$  and  $\nu$  are called nuisance parameters only for now in order to further the discussion generally; later when the present theory is applied, usually all the parameters are interesting and are to be estimated. If we write

$$h(x | \theta, \nu) = h(x | \nu) \quad (5.3)$$

then what we are saying is that the probability for outliers occurring has nothing to do with  $\theta$ , and this data on its own cannot be used to estimate  $\theta$ . In Jaynes' chapter on Outliers he calls such data 'bad data', while the 'good data' is from sampling distribution  $g(x | \theta\mu)$ . If  $h(x | \theta\nu) \neq h(x | \nu)$  then the outlier data must also be relevant to estimating  $\theta$  and it is probably a good idea to keep this data as opposed to throwing it out. The data consists of  $n$  observations

$$D = \{x_1, \dots, x_n\}, \quad (5.4)$$

and although we may speculate, we do not know which distribution,  $g$  or  $h$ , was responsible for each data point if that data point falls within a region where  $g$  and  $h$  overlap. Define the following proposition

$$z_i \equiv \begin{cases} 1 & \text{if the } i\text{-th datum is from } g \\ 0 & \text{if the } i\text{-th datum is not from } g, \text{ but from } h \end{cases}. \quad (5.5)$$

A joint prior probability

$$p(z_1 \dots z_n | I) \quad (5.6)$$

can now be assigned to the  $2^n$  conceivable sequences of data which came from either  $g$  or  $h$ . Assume the probability of any sequence of  $n$  observations of data points coming from either  $g$  or  $h$  depends only on the numbers  $k$  and  $(n - k)$  of observations coming from either  $g$  or  $h$  respectively and not on the particular trials at which they occur. The distribution (5.6) is then invariant under permutations of the  $z_i$ , and is thus what is referred to as an exchangeable prior. There is an important theorem by de Finetti [6] that now becomes relevant to the problem. The de Finetti theorem asserts that any exchangeable probability function is determined by a single generating function  $f(\alpha)$ . Thus there is a function  $f(\alpha)$  such that  $f(\alpha) \geq 0$ ,  $\int_0^1 d\alpha f(\alpha) = 1$ , and the probability that out of a total of  $n$  observations  $k$  of those where from  $g$  ( $k = \sum_{i=1}^n z_i$ ), and the remaining  $(n - k)$  where from  $h$  is given by

$$p(z_1 \dots z_n | I) = \int_0^1 d\alpha \alpha^k (1 - \alpha)^{n-k} f(\alpha). \quad (5.7)$$

There is thus a parameter  $\alpha$ , such that if it were known, then given any data point  $x_i$ , it would with probability  $\alpha$  have come from  $g$ , or with probability  $(1 - \alpha)$  have come from  $h$ . Thus if  $h$  is to represent the sampling distribution of the outliers, then any data point will have  $(1 - \alpha)$  chance of being an outlier and thus it will have been from the distribution  $h$ . The closer  $\alpha$  is to unity, then the less chance there is for the data to be corrupted by impulsive noise. The sampling distribution to describe the the two sources of noise can thus be written as a *mixture* of  $g$  and  $h$ :

$$p(x | \theta\mu\nu I) = \alpha g(x | \theta\mu) + (1 - \alpha)h(x | \theta\nu), \quad 0 \leq \alpha \leq 1. \quad (5.8)$$

Data is thus drawn urn-wise (as in elementary sampling theory where balls of different colours are drawn from an urn) from either  $g$  or  $h$ , and if  $\alpha$  were known we would draw from  $g$  with probability  $\alpha$  and from  $h$  with probability  $1 - \alpha$ . Notice that when  $\alpha = 1$  the possibility for the data to have come from  $h$  is switched off and we are back to a single model as was used previously; similarly when  $\alpha = 0$  then the model consists only of  $h$ . The sampling distribution (5.8) is probably what most of us would write down intuitively when trying to construct a model whose features are made of the joining or superposition of two separate probability functions, for if  $g$  and  $h$  are both normalized and each data point could have either come from the one or the other, then the introduction of the parameter  $\alpha$  normalizes the distribution as a whole. To go via the route of the de Finetti theorem however shows that there is an underlying assumption that the prior on the propositions  $z_i$  is invariant under permutation.

## 5.2 Binomial expansion of likelihood within Bayes' theorem

We can now proceed with the calculation. The likelihood function is

$$P(D | \theta\mu\nu\alpha I) = \prod_{i=1}^n [\alpha g(x_i | \theta\mu) + (1 - \alpha)h(x_i | \theta\nu)], \quad (5.9)$$

and by Bayes' theorem the joint posterior is

$$P(\theta\mu\nu\alpha | DI) = \frac{P(\theta\mu\nu\alpha | I)P(D | \theta\mu\nu\alpha)}{\iiint d\theta d\mu d\nu d\alpha P(\theta\mu\nu\alpha | I)P(D | \theta\mu\nu\alpha)}. \quad (5.10)$$

In particular, we are after the marginal posterior distribution of  $\theta$ . It is the joint posterior with the nuisance parameters integrated out, namely

$$P(\theta | DI) = \frac{\iiint d\mu d\nu d\alpha P(\theta\mu\nu\alpha | I)P(D | \theta\mu\nu\alpha)}{\iiint d\theta d\mu d\nu d\alpha P(\theta\mu\nu\alpha | I)P(D | \theta\mu\nu\alpha)}. \quad (5.11)$$

We thus need

$$\iiint d\mu d\nu d\alpha P(\theta\mu\nu\alpha | I)P(D | \theta\mu\nu\alpha) \quad (5.12)$$

followed by another integration over  $\theta$  to get the normalization constant. Substituting in the likelihood (5.9) into the above, the joint posterior for  $\theta$  is proportional to

$$P(\theta | DI) \propto \iiint d\mu d\nu d\alpha P(\theta\mu\nu\alpha | I) \prod_{i=1}^n [\alpha g(x_i | \theta\mu) + (1 - \alpha)h(x_i | \theta\nu)]. \quad (5.13)$$

The likelihood can be written in a more revealing form as follows

$$\begin{aligned} & \prod_{i=1}^n [\alpha g(x_i | \theta\mu) + (1 - \alpha)h(x_i | \theta\nu)] \\ &= \alpha^n \prod_{i=1}^n g(x_i | \theta\mu) \\ &+ \alpha^{n-1}(1 - \alpha) \sum_{j=1}^n h(x_j | \theta\nu) \prod_{i \neq j} g(x_i | \theta\mu) \\ &+ \alpha^{n-2}(1 - \alpha)^2 \sum_{j < k} h(x_j | \theta\nu)h(x_k | \theta\nu) \prod_{i \neq j, k} g(x_i | \theta\mu) \\ &+ \dots \\ &+ \alpha^2(1 - \alpha)^{n-2} \sum_{j < k} g(x_j | \theta\mu)g(x_k | \theta\mu) \prod_{i \neq j, k} h(x_i | \theta\nu) \\ &+ \alpha(1 - \alpha)^{n-1} \sum_{j=1}^n g(x_j | \theta\mu) \prod_{i \neq j} h(x_i | \theta\nu) \\ &+ (1 - \alpha)^n \prod_{i=1}^n h(x_i | \theta\nu). \end{aligned} \quad (5.14)$$

Before trying to understand under what conditions the  $2^n$  terms it produces are reduced, let us first show what they mean and at the same time shed light upon how Bayes' theorem organizes the

problem. In order to gain transparency to what is happening, make the simplifying assumption that ‘normal’ and ‘outlier’ noise are independent (each does not depend on the parameters of the other)

$$g(x_i | \theta \mu) = g(x_i | \theta) \quad (5.15)$$

and

$$h(x_i | \theta \nu) = h(x_i | \nu). \quad (5.16)$$

In other words,  $g$  contains only the interesting parameter  $\theta$ , and the distribution  $h$  is independent of  $\theta$  and contains only the uninteresting parameter  $\nu$ . The marginal posterior for  $\theta$  will, instead of (5.13), be proportional to

$$\iint d\nu d\alpha P(\theta \nu \alpha | I) \left[ \alpha^n \prod_{i=1}^n g(x_i | \theta) + \alpha^{n-1} (1 - \alpha) \sum_{j=1}^n h(x_j | \nu) \prod_{i \neq j} g(x_i | \theta) + \dots \right]. \quad (5.17)$$

Using the product rule, the prior density may be factored as follows

$$P(\theta \nu \alpha | I) = P(\theta | I) P(\nu \alpha | \theta I). \quad (5.18)$$

Now (5.17) becomes

$$P(\theta | I) \iint d\nu d\alpha P(\nu \alpha | \theta I) \left[ \alpha^n \prod_{i=1}^n g(x_i | \theta) + \alpha^{n-1} (1 - \alpha) \sum_{j=1}^n h(x_j | \nu) \prod_{i \neq j} g(x_i | \theta) + \dots \right]. \quad (5.19)$$

The first term in (5.19) simplifies to

$$\begin{aligned} & \prod_{i=1}^n g(x_i | \theta) \iint d\nu d\alpha P(\nu \alpha | \theta I) \alpha^n \\ &= \prod_{i=1}^n g(x_i | \theta) \int d\alpha \int d\nu P(\nu \alpha | \theta I) \alpha^n \\ &= \prod_{i=1}^n g(x_i | \theta) \int_0^1 d\alpha \alpha^n P(\alpha | \theta I). \end{aligned} \quad (5.20)$$

Here  $\prod_{i=1}^n g(x_i | \theta)$  is a sequence of likelihood functions of  $g$  using all of the data, and as can be seen in the Laplace-de Finetti form (5.7),  $\int_0^1 \alpha^n d\alpha P(\alpha | \theta I)$  is the probability, conditional on  $\theta$  and  $I$ , that all of the data are from  $g$ . The generating function  $f(\alpha)$  is then the prior density  $P(\alpha | \theta I)$  conditional on  $\theta$ . In most cases  $P(\alpha | \theta I) = P(\alpha | I)$  since  $\alpha$  refers to something entirely different than  $\theta$ . The second term, for any particular  $j$  in the summation, is

$$\begin{aligned} & \prod_{i \neq j} g(x_i | \theta) \iint d\nu d\alpha P(\nu \alpha | \theta I) \alpha^{n-1} (1 - \alpha) h(x_j | \nu) \\ &= \prod_{i \neq j} g(x_i | \theta) \int_0^1 d\alpha \alpha^{n-1} (1 - \alpha) \int d\nu P(\nu \alpha | \theta I) h(x_j | \nu). \end{aligned} \quad (5.21)$$

Here  $\prod_{i \neq j} g(x_i | \theta)$  is a sequence of likelihood functions of  $g$  using all the data except  $x_j$ . The factor

$$d\nu \int_0^1 d\alpha \alpha^{n-1} (1 - \alpha) P(\nu \alpha | \theta I), \quad (5.22)$$

is the joint probability density, given  $\theta$  and  $I$ , that any specific data point  $x_j$  comes from  $h$ , that the other  $(n-1)$  data points come from  $g$ , and that  $\nu$  lies in  $(\nu, \nu + d\nu)$ . Therefore

$$\int_0^1 d\alpha \alpha^{n-1} (1-\alpha) \int d\nu P(\nu\alpha | \theta I) h(x_j | \nu) \quad (5.23)$$

is the probability, given  $\theta$  and  $I$ , that the  $j$ -th data point comes from  $h$  and has the value  $x_j$ , and the other data points come from  $g$ . To put it in words then, (5.13) says

$$P(\theta | DI) \propto$$

$$\begin{aligned} & P(\theta | I) [\text{prob}(\text{all the data come from } g) \times (\text{sequence of } n \text{ likelihood functions } g \text{ using all the data}) \\ & + \sum_{j=1}^n \text{prob}(\text{only the } j\text{-th data point comes from } h \text{ and has the value } x_j) \\ & \quad \times (\text{sequence of } (n-1) \text{ likelihood functions } g \text{ using all data except } x_j) \\ & + \sum_{j=1}^n \sum_{k=1}^n \text{prob}(\text{the } j\text{-th and } k\text{-th data points come from } h \text{ and have the values } x_j \text{ and } x_k) \\ & \quad \times (\text{sequence of } n-2 \text{ likelihood functions } g \text{ using all the data except } x_j \text{ and } x_k) \\ & + \dots \\ & + \sum_{j=1}^n \text{prob}(\text{only the } j\text{-th data point comes from } g \text{ and has the value } x_j) \\ & \quad \times (\text{sequence of likelihood functions } h \text{ using all the data except } x_j) \\ & + \text{prob}(\text{all the data come from } h) \times (\text{sequence of } n \text{ likelihood functions } h \text{ using all the data})]. \end{aligned} \quad (5.24)$$

Going back to the more general problem where  $g(x_i | \theta\mu)$  and  $h(x_i | \theta\nu)$  are considered, we want to comment on the meaning of (5.13):

$$\iiint d\mu d\nu d\alpha P(\theta\mu\nu\alpha | I) \prod_{i=1}^n [\alpha g(x_i | \theta\mu) + (1-\alpha)h(x_i | \theta\nu)].$$

As our short detour might suggest, the revealing form (5.24) will not change except for the definition of  $g$  and  $h$ . Going through the reasoning steps again for clarity, factor the joint prior and write (5.13) as follows

$$P(\theta | I) \iiint d\mu d\nu d\alpha P(\mu\nu\alpha | \theta I) \left[ \alpha^n \prod_{i=1}^n g(x_i | \theta\mu) + \alpha^{n-1} (1-\alpha) \sum_{j=1}^n h(x_j | \theta\nu) \prod_{i \neq j} g(x_i | \theta\mu) + \dots \right] \quad (5.25)$$

The first term is

$$\begin{aligned} & \iiint d\mu d\nu d\alpha P(\mu\nu\alpha | \theta I) \alpha^n \prod_{i=1}^n g(x_i | \theta\mu) \\ & = \int_0^1 d\alpha \alpha^n \int d\mu \prod_{i=1}^n g(x_i | \theta\mu) \int d\nu P(\mu\nu\alpha | \theta I) \\ & = \int_0^1 d\alpha \alpha^n \int d\mu P(\mu\alpha | \theta I) \prod_{i=1}^n g(x_i | \theta\mu). \end{aligned} \quad (5.26)$$

The factor

$$d\mu \int_0^1 d\alpha \alpha^n P(\mu\alpha | \theta I) \quad (5.27)$$

is the joint probability, conditional on  $\theta$  and  $I$ , that all the data are from  $g$  and that  $\mu$  lies in  $(\mu, \mu + d\mu)$ . Therefore the first term (5.26) is the probability given  $\theta$  and  $I$ , that all the data are from  $g$ , multiplied by a sequence of likelihood functions  $g$  using all of the data. The second term, for any particular  $j$  in the summation, is

$$\begin{aligned} & \iiint d\mu d\nu d\alpha P(\mu\nu\alpha | \theta I) \alpha^{n-1} (1-\alpha) h(x_j | \theta\nu) \prod_{i \neq j} g(x_i | \theta\mu) \\ &= \int_0^1 d\alpha \alpha^{n-1} (1-\alpha) \iint d\mu d\nu \prod_{i \neq j} g(x_i | \theta\mu) h(x_j | \theta\nu) P(\mu\nu\alpha | \theta I). \end{aligned} \quad (5.28)$$

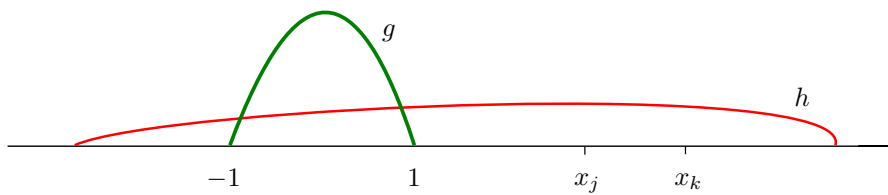
The factor

$$d\mu d\nu \int_0^1 d\alpha \alpha^{n-1} (1-\alpha) P(\mu\nu\alpha | \theta I) \quad (5.29)$$

is the probability, conditional on  $\theta$  and  $I$  that any specific data point  $x_j$  comes from  $h$  and that the other  $(n-1)$  data points come from  $g$ , and that  $\mu$  and  $\nu$  lie in  $(\mu, \mu + d\mu)$  and  $(\nu, \nu + d\nu)$  respectively. Therefore the second term (5.28) is the probability, given  $\theta$  and  $I$ , that the  $j$ -th data point comes from  $h$  and has the value  $x_j$ , and the other  $(n-1)$  data points come from  $g$ , multiplied by a sequence of likelihood function  $g$  using all of the data except  $x_j$ . Thus, to put it in words we would again write down something like (5.24).

A full nontrivial solution of the problem becomes intricate as Bayes' theorem considers every small contributing detail to the problem. But under what circumstances do the  $2^n$  terms reduce? The question depends on how one defines the relationship between the outcome spaces of  $g$  and  $h$ , as well as on the relevant priors on the parameters, for when a datum falls in a region where  $g$  and  $h$  overlap one cannot tell from which distribution it came and as a result a certain amount of terms will continue to remain relevant. To illustrate these ideas, we distinguish between four cases concerning the outcome spaces of  $g$  and  $h$ .

### 5.2.1 Case 1



**Figure 5.1:** The outcome spaces of  $g$  and  $h$  overlap, but certain points can be identified to have come from  $h$ .

Suppose that

$$g(x | \theta) \begin{cases} > 0 & \text{where } |x| < 1 \\ = 0 & \text{elsewhere,} \end{cases} \quad (5.30)$$



and that

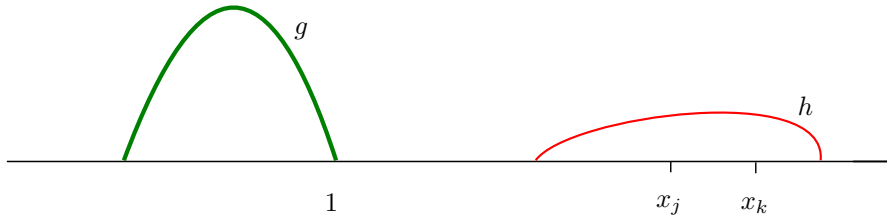
$$h(x|\nu) > 0 \text{ for } x \in [a, b] \text{ where } a < -1 \text{ and } b > 1, \quad (5.31)$$

where the prior information about the scale and location of  $h$  is encoded in  $a$  and  $b$ . Suppose  $n$  data points are collected, two of which,  $x_j$  and  $x_k$  are larger than one, then according to the chosen model they are with certainty from  $h$ . The problem is we do not know which distribution was responsible for the other  $n - 2$  data points, and after examining (5.13), (5.14) and the revealing form (5.24), the joint posterior for  $\theta$  is proportional to

$$\begin{aligned} \iint d\nu d\alpha P(\theta\nu\alpha | I) & \left[ \alpha^{n-2}(1-\alpha)^2 h(x_j|\nu)h(x_k|\nu) \prod_{i \neq j,k} g(x_i|\theta) \right. \\ & + \alpha^{n-3}(1-\alpha)^3 h(x_j|\nu)h(x_k|\nu) \sum_{l \neq j,k} h(x_l|\nu) \prod_{i \neq j,k,l} g(x_i|\theta) \\ & + \cdots + \alpha(1-\alpha)^{n-1} h(x_j|\nu)h(x_k|\nu) \sum_{i \neq j,k} g(x_i|\theta) \prod_{l \neq j,k,i} h(x_l|\nu) \\ & \left. (1-\alpha)^n h(x_1|\nu) \dots h(x_n|\nu) \right]. \end{aligned} \quad (5.32)$$

The  $2^n$  terms in the likelihood expansion are reduced to  $n^2 - 5n + 8$  terms, still quite a few to work with.

### 5.2.2 Case 2



**Figure 5.2:** The origin of all the data points (either from  $g$  or  $h$ ) can be distinguished.

Suppose that given the way  $g(x|\theta\mu)$  and  $h(x|\theta\nu)$  are defined, and given the way in which we have stated our prior information regarding the parameter at hand,  $g$  and  $h$  have disjoint outcome spaces. Let  $\mathcal{A}_g(x)$  and  $\mathcal{A}_h(x)$  denote the outcome spaces of  $g$  and  $h$  respectively. For given  $x_j, x_k \in \mathcal{A}_h(x)$ , and all other  $x_i \in \mathcal{A}_g(x)$ , the marginal posterior distribution for  $\theta$  is proportional to

$$\iiint d\mu d\nu d\alpha P(\theta\mu\nu\alpha | I) \alpha^{n-2}(1-\alpha)^2 h(x_j|\theta\nu)h(x_k|\theta\nu) \prod_{i \neq j,k} g(x_i|\theta\mu). \quad (5.33)$$

Using the product rule consecutively, factor the joint prior as follows

$$\begin{aligned} P(\theta\mu\nu\alpha | I) & = P(\theta\nu\alpha | \mu I)P(\mu | I) \\ & = P(\nu\alpha | \theta\mu I)P(\theta | \mu I)P(\mu | I) \\ & = P(\nu\alpha | \theta\mu I)P(\theta\mu | I). \end{aligned} \quad (5.34)$$

In most cases  $\alpha$  refers to something entirely different than all the parameters belonging to the distributions  $g$  and  $h$ , and if we further assume that  $\nu$  is in no way related to  $\mu$ , which is also usually the case, then we can write

$$P(\nu\alpha | \theta\mu I) = P(\nu\alpha | I) \quad (5.35)$$

and thus

$$P(\theta\mu\nu\alpha | I) = P(\nu\alpha | I)P(\theta\mu | I). \quad (5.36)$$

Inserting this into the above we have

$$C_{j,k}(\theta) \int d\mu P(\theta\mu | I) \prod_{i \neq j,k} g(x_i | \theta\mu), \quad (5.37)$$

where

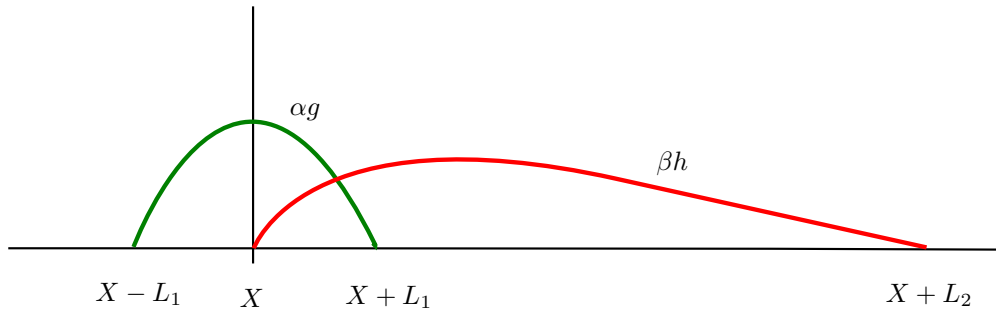
$$C_{j,k}(\theta) = \iint d\nu d\alpha P(\nu\alpha | I) \alpha^{n-2} (1-\alpha)^2 h(x_j | \theta\nu) h(x_k | \theta\nu) \quad (5.38)$$

is a function of  $\theta$  only as long as  $h$  is. Thus if  $h(x | \theta\nu) = h(x | \nu)$  then  $C_{j,k}(\theta) = C_{j,k}$  is a constant which will cancel out upon normalization. The result is the same had only the data from  $g$  been used, namely

$$P(\theta | DI) \propto \int d\mu P(\theta\mu | I) \prod_{i \neq j,k} g(x_i | \theta\mu). \quad (5.39)$$

This is the only case where it is permissible to throw out data when inferring on  $\theta$ . The next example is considered in greater detail.

### 5.2.3 Case 3



**Figure 5.3:** Partially overlapping outcome spaces implying that a given data point cannot uniquely be identified as having come from  $g$  or  $h$ . Whether the data points are from  $g$  or from  $h$  is not certain.

The sampling distribution is

$$P(x_i | \alpha X L_1 L_2 I) = \alpha g(x_i | X L_1) U[x_i | X - L_1, X + L_1] + (1 - \alpha) h(x_i | X L_2) U[x_i | X, X + L_2], \quad (5.40)$$

which can be rewritten as

$$\begin{aligned} & \alpha g(U[x_i | X - L_1, X] + U[x_i | X, X + L_1]) + (1 - \alpha)h(U[x_i | X, X + L_1] + U[x_i | X + L_1, X + L_2]) \\ & = \alpha gU[x_i | X - L_1, X] + (\alpha g + (1 - \alpha)h)U[x_i | X, X + L_1] + (1 - \alpha)hU[x_i | X + L_1, X + L_2]. \end{aligned} \quad (5.41)$$

The likelihood is given by

$$P(D | \alpha XL_1L_2I) = \prod_{i=1}^n P(x_i | \alpha XL_1L_2I), \quad (5.42)$$

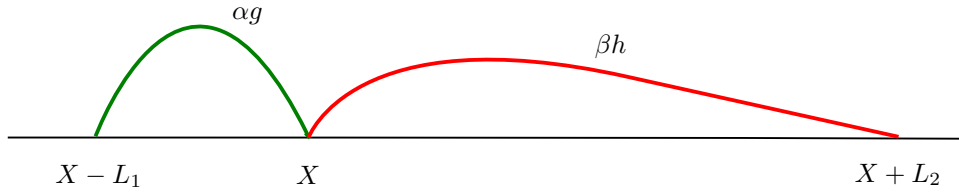
which upon ordering the data ( $x_1 < \dots < x_n$ ) gives a multi-binomial

$$\begin{aligned} & P(D | \alpha XL_1L_2I) \\ & = \sum_{j=0}^n \sum_{k=0}^n (\alpha g)^{n-k-j} (\alpha g + (1 - \alpha)h)^k ((1 - \alpha)h)^j \\ & \quad \times \prod_{i=1}^{n-k-j} U[x_i | X - L_1, X] \prod_{\gamma=n-k-j+1}^{n-j+1} U[x_i | X, X + L_1] \prod_{\sigma=n-j+2}^n U[x_i | X + L_1, X + L_2]. \end{aligned} \quad (5.43)$$

The amount of terms in the likelihood is greatly reduced from  $2^n$  to  $\binom{n+2}{2}$ . The strings of theta functions contain information in the form min max functions of the data and prior bounds. Similarly to what was done in Appendix B, when these are carefully prepared, the moments of such a model can be obtained by numerical integration.

### 5.3 Case 4

#### 5.3.1 Preparing the calculation



**Figure 5.4:** A mixture model of two distributions sharing the same location parameter. If  $h$  is a long and flat tail then it can model the occurrence of outliers in the data, for example a cosmic ray striking a measuring apparatus in a laboratory.

We will now calculate in detail the fourth case. Consider a model that is a convex combination of two distributions  $g$  and  $h$

$$f = \alpha g + (1 - \alpha)h. \quad (5.44)$$

As illustrated in the Fig. 5.4 above,  $g$  and  $h$  have non-overlapping support and share the location parameter  $X$ , while their respective scale parameters are  $L_1$  and  $L_2$ . The sampling distribution is

$$P(x_i | \alpha X L_1 L_2 I) = \alpha g(x_i | X L_1) U[x_i | X - L_1, X] + (1 - \alpha) h(x_i | X L_2) U[x_i | X, X + L_2]. \quad (5.45)$$

The likelihood for iid generated data is

$$P(D | \alpha X L_1 L_2 I) = \prod_{i=1}^n [\alpha g(x_i | X L_1) U[x_i | X - L_1, X] + (1 - \alpha) h(x_i | X L_2) U[x_i | X, X + L_2]]. \quad (5.46)$$

Observing the above, there is no reason why we can not relabel the data as follows  $x_1 < x_2 < \dots < x_n$ , so that the indices now represent the ordering of the data. With this relabelling it is clear that  $\min(D) = x_{\min} = x_1$  and  $\max(D) = x_{\max} = x_n$ . By ordering the data we realize that in its current form, many of the  $2^n$  terms in the binomial expansion of the likelihood produces are zero. To illustrate this, take the  $n = 2$  case as an example. Writing out  $P(D | \alpha X L_1 L_2 I)$  we have

$$\begin{aligned} P(D | \alpha X L_1 L_2 I) &= \prod_{i=1}^2 [\alpha g(x_i | X L_1) U[x_i | X - L_1, X] + (1 - \alpha) h(x_i | X L_2) U[x_i | X, X + L_2]] \\ &= \alpha^2 g(x_1 | X L_1) g(x_2 | X L_1) U[x_1 | X - L_1, X] U[x_2 | X - L_1, X] \\ &\quad + \alpha(1 - \alpha) g(x_1 | X L_1) h(x_2 | X L_2) U[x_1 | X - L_1, X] U[x_2 | X, X + L_2] \\ &\quad + \alpha(1 - \alpha) g(x_2 | X L_1) h(x_1 | X L_2) U[x_2 | X - L_1, X] U[x_1 | X, X + L_2] \\ &\quad + (1 - \alpha)^2 h(x_1 | X L_2) h(x_2 | X L_2) U[x_1 | X, X + L_2] U[x_2 | X, X + L_2]. \end{aligned} \quad (5.47)$$

The term  $\alpha(1 - \alpha) g(x_2 | X L_1) h(x_1 | X L_2) U[x_2 | X - L_1, X] U[x_1 | X, X + L_2]$  should be zero because it says that the data point  $x_1 \in [X, X + L_2]$  and  $x_2 \in [X - L_1, X]$ , which is not possible since  $x_1 < x_2$ . To see this, write out the term's window functions in terms of Heaviside theta functions:

$$U[x_2 | X - L_1, X] U[x_1 | X, X + L_2] = \theta(x_2 - X + L_1) \theta(X - x_2) \theta(x_1 - X) \theta(X + L_2 - x_1). \quad (5.48)$$

Here  $\theta(x_1 - X) \theta(X - x_2)$  is only nonzero if  $x_2 \leq X < x_1$  which is not possible. Thus the current form in which the likelihood is written produces unnecessary terms. The following truncated series produces all  $n + 1$  relevant terms:

$$\begin{aligned} P(D | \alpha X L_1 L_2 I) &= \sum_{k=0}^n \alpha^{n-k} (1 - \alpha)^k \prod_{i=1}^{n-k} g(x_i | X L_1) U[x_i | X - L_1, X] \prod_{j=n-k+1}^n h(x_j | X L_2) U[x_j | X, X + L_2], \end{aligned} \quad (5.49)$$

since  $U[x_l | X - L_1, X] U[x_m | X, X + L_2] = 0$  for  $m < l$ . In the above the convention  $\prod_{i=l}^m U[x_i | a, b] = 1$  for  $l > m$  is used. The likelihood permeates the data points between the two distributions. The  $k$ -th term in the sum corresponds with the scenario where  $k$  data points are possibly due to  $h$  and  $n - k$  are possibly due to  $g$ . If one has reason to believe that at most, out of the total number of data points accumulated, only a certain number are from either one of the distributions, then one may be justified to adjust the amount of terms in the sum accordingly. For instance, if we believe at most two data points in a set could have been due to  $h$  then we would adjust the sum above to run between  $0 \leq k \leq 2$ , effectively saying that the other probabilities in the sum are zero.

Next we choose priors for the parameters. For the location parameters  $\alpha$  and  $X$ , uniform priors are chosen:

$$P(\alpha | I) = U[\alpha | A_1, A_2], \quad 0 \leq A_1 < A_2 \leq 1, \quad (5.50)$$

$$P(X | I) = U[X | K_1, K_2], \quad K_1 < K_2. \quad (5.51)$$

For the scale parameters  $L_1$  and  $L_2$  we choose Jeffreys Priors

$$P(L_1 | I) = L_1^{-1} U[L_1 | M_1, M_2], \quad 0 < M_1 < M_2 \quad (5.52)$$

and

$$P(L_2 | I) = L_2^{-1} U[L_2 | Q_1, Q_2], \quad 0 < M_1 < M_2, \quad (5.53)$$

where we have omitted the constants since they will cancel in Bayes' theorem. As will emerge in the calculation below, it is not permissible to take certain 'obvious' limits such as  $M_1, M_2 = 0$  or  $M_2, Q_2 \rightarrow \infty$ , and we must necessarily do the full calculation for general values of prior parameters. Notice that the condition  $\max(K_1, x_n - Q_2) < \min(K_2, x_1 + M_2)$  must hold, else the data contradicts the priors and the joint posterior is identically zero. The parameters are logically independent, and therefore

$$\begin{aligned} P(\alpha X L_1 L_2 | I) &= P(\alpha | X L_1 L_2 I) P(X | L_1 L_2 I) P(L_1 | L_2 I) P(L_2 | I) \\ &= P(\alpha | I) P(X | I) P(L_1 | I) P(L_2 | I). \end{aligned} \quad (5.54)$$

Of interest are the marginal probability distributions and moments of the parameters  $\alpha$ ,  $X$ ,  $L_1$  and  $L_2$ . We will see that an estimator for  $\alpha$  can easily be calculated once we have done the calculations for the other parameters first. We may want to assume  $\alpha$  known, and this choice can easily be implemented at any stage. We want to calculate

$$A = A(X, m_\alpha, m_X, m_{L_1}, m_{L_2}) = \iiint d\alpha dL_1 dL_2 T, \quad (5.55)$$

$$B = B(L_2, m_\alpha, m_X, m_{L_1}, m_{L_2}) = \iiint d\alpha dX dL_1 T, \quad (5.56)$$

$$C = C(L_1, m_\alpha, m_X, m_{L_1}, m_{L_2}) = \iiint d\alpha dX dL_2 T, \quad (5.57)$$

$$D = D(m_\alpha, m_X, m_{L_1}, m_{L_2}) = \iiint d\alpha dX dL_1 dL_2 T, \quad (5.58)$$

where the integrand is

$$T = \alpha^{m_\alpha} X^{m_X} L_1^{m_{L_1}} L_2^{m_{L_2}} P(D | \alpha X L_1 L_2 I) P(\alpha X L_1 L_2 | I). \quad (5.59)$$

The marginal probability distributions and moments of all the parameters can be attained from  $A$ ,  $B$ ,  $C$  and  $D$ . We will calculate the following distributions and moments

$$\begin{aligned} P(X | DI) &= \frac{1}{N} A(X, 0, 0, 0, 0) & \langle X \rangle &= \frac{1}{N} D(0, 1, 0, 0) & \langle X^2 \rangle &= \frac{1}{N} D(0, 2, 0, 0) \\ P(L_1 | DI) &= \frac{1}{N} C(L_1, 0, 0, 0, 0) & \langle L_1 \rangle &= \frac{1}{N} D(0, 0, 1, 0) & \langle L_1^2 \rangle &= \frac{1}{N} D(0, 0, 2, 0) \\ P(L_2 | DI) &= \frac{1}{N} B(L_2, 0, 0, 0, 0) & \langle L_2 \rangle &= \frac{1}{N} D(0, 0, 0, 1) & \langle L_2^2 \rangle &= \frac{1}{N} D(0, 0, 0, 2) \end{aligned} \quad (5.60)$$

where

$$N = D(0, 0, 0, 0) \quad (5.61)$$

is the normalization constant. The calculations are tedious, so we will opt to simplify matters where we can. Of particular interest is the estimators for  $L_1$  and  $L_2$ , so let us begin by calculating B,C and D. Begin by preparing  $T$  as follows. Substitute the likelihood and joint prior:

$$\begin{aligned} T &= \sum_{k=0}^n \alpha^{m_\alpha+n-k} (1-\alpha)^k X^{m_X} L_1^{m_{L_1}-1} L_2^{m_{L_2}-1} \prod_{i=1}^{n-k} g(x_i | XL_1) U[x_i | X - L_1, X] \\ &\times \prod_{j=n-k+1}^n h(x_j | XL_2) U[x_j | X, X + L_2] U[\alpha | A_1, A_2] U[X | K_1, K_2] U[L_1 | M_1, M_2] U[L_2 | Q_1, Q_2]. \end{aligned} \quad (5.62)$$

Integrating out  $\alpha$  we have

$$\int d\alpha T = \sum_{k=0}^n \int_{A_1}^{A_2} d\alpha \alpha^{m_\alpha+n-k} (1-\alpha)^k \times \text{the rest} \quad (5.63)$$

where

$$\int_{A_1}^{A_2} d\alpha \alpha^{m_\alpha+n-k} (1-\alpha)^k = \beta(A_2, m_\alpha + n - k + 1, k + 1) - \beta(A_1, m_\alpha + n - k + 1, k + 1), \quad (5.64)$$

or if  $A_1 = 0$  and  $A_2 = 1$  then

$$\begin{aligned} \int_0^1 d\alpha \alpha^{m_\alpha+n-k} (1-\alpha)^k &= \frac{\Gamma(k+1)\Gamma(m_\alpha+n-k+1)}{\Gamma(m_\alpha+n+2)} \\ &= \binom{m_\alpha+n}{k}^{-1} (m_\alpha+n+1)^{-1}, \end{aligned} \quad (5.65)$$

where  $\beta$  is the incomplete beta function and  $\Gamma$  is the Euler gamma function. Let

$$\Lambda(m_\alpha, n, k) = \frac{\Gamma(k+1)\Gamma(m_\alpha+n-k+1)}{\Gamma(m_\alpha+n+2)}, \quad (5.66)$$

and note that while the moments of  $\alpha$  are given by  $\frac{1}{N}D(m_\alpha, 0, 0, 0)$ , the marginal distribution of  $\alpha$  can also be obtained if one replaces  $\Lambda$  with  $\alpha^{n-k}(1-\alpha)^k$  in  $D(m_\alpha, 0, 0, 0)$ . Similarly if we want to assume  $\alpha$  to be known for any part of the calculation, we can simply replace  $\Lambda$  by  $\alpha^{n-k}(1-\alpha)^k$  in  $A$ ,  $B$  and  $C$ . For economy of writing let  $U[X, L_1, L_2]$  represent the string of window functions which refer to the boundaries of the priors on the parameters  $X$ ,  $L_1$  and  $L_2$ :

$$U[X, L_1, L_2] = U[X | K_1, K_2] U[L_1 | M_1, M_2] U[L_2 | Q_1, Q_2]. \quad (5.67)$$

Awaiting integration over  $X$ ,  $L_1$  and  $L_2$  we have

$$\begin{aligned} \int d\alpha T &= \sum_{k=0}^n \Lambda(m_\alpha, n, k) X^{m_X} L_1^{m_{L_1}-1} L_2^{m_{L_2}-1} \prod_{i=1}^{n-k} g(x_i | XL_1) U[x_i | X - L_1, X] \\ &\times \prod_{j=n-k+1}^n h(x_j | XL_2) U[x_j | X, X + L_2] U[X, L_1, L_2]. \end{aligned} \quad (5.68)$$

Next split the sum in three parts as follows

$$\begin{aligned}
 \int d\alpha T &= T_1 + T_2 + T_3 \\
 &= \Lambda(m_\alpha, n, 0) X^{m_X} L_1^{m_{L_1}-1} L_2^{m_{L_2}-1} \prod_{i=1}^n g(x_i | X L_1) U_1 \\
 &+ \sum_{k=1}^{n-1} \Lambda(m_\alpha, n, k) X^{m_X} L_1^{m_{L_1}-1} L_2^{m_{L_2}-1} \prod_{i=1}^{n-k} g(x_i | X L_1) \prod_{j=n-k+1}^n h(x_j | X L_2) U_2 \\
 &+ \Lambda(m_\alpha, n, n) X^{m_X} L_1^{m_{L_1}-1} L_2^{m_{L_2}-1} \prod_{j=n-k+1}^n h(x_j | X L_2) U_3,
 \end{aligned} \tag{5.69}$$

where we have introduced  $U_1$ ,  $U_2$  and  $U_3$  to represent the window functions in each term:

$$\begin{aligned}
 U_1 &= \prod_{i=1}^n U[x_i | X - L_1, X] U[X, L_1, L_2] \\
 U_2 &= \prod_{i=1}^{n-k} U[x_i | X - L_1, X] \prod_{j=n-k+1}^n U[x_j | X, X + L_2] U[X, L_1, L_2] \\
 U_3 &= \prod_{j=n-k+1}^n U[x_j | X, X + L_2] U[X, L_1, L_2].
 \end{aligned} \tag{5.70}$$

In order to proceed with an integration over  $X$ ,  $L_1$  or  $L_2$ , the above strings of theta functions must be prepared in such away to give us the correct boundaries of integration. To summarize the results of Appendix B, if we want to calculate  $B$ ,  $C$  or  $D$ , the terms  $U_1$ ,  $U_2$  and  $U_3$  are chosen as shown in (B.14):

$$\begin{aligned}
 U_1 &= U[X | \max(K_1, x_n), \min(K_2, x_1 + L_1)] U[L_1 | \max(M_1, \max(K_1, x_n) - x_1), M_2] U[L_2 | Q_1, Q_2] \\
 U_2 &= U[X | \max(K_1, x_{n-k}, x_n - L_2), \min(K_2, x_{n-k+1}, x_1 + L_1)] \\
 &\quad \times U[L_1 | \max(M_1, \max(K_1, x_{n-k}) - x_1), M_2] U[L_2 | \max(Q_1, x_n - \min(K_2, x_{n-k+1})), Q_2] \\
 U_3 &= U[X | \max(K_1, x_n - L_2), \min(K_2, x_1)] U[L_1 | M_1, M_2] U[L_2 | \max(Q_1, x_n - \min(x_1, K_2)), Q_2]
 \end{aligned}$$

If we want to calculate  $A$ , the terms  $U_1$ ,  $U_2$  and  $U_3$  are chosen as shown in (B.15):

$$\begin{aligned}
 U_1 &= U[X | \max(K_1, x_n), \min(K_2, x_1 + M_2)] U[L_1 | \max(M_1, X - x_1), M_2] U[L_2 | Q_1, Q_2] \\
 U_2 &= U[X | \max(K_1, x_{n-k}, x_n - Q_2), \min(K_2, x_{n-k+1}, x_1 + M_2)] \\
 &\quad \times U[L_1 | \max(M_1, X - x_1), M_2] U[L_2 | \max(Q_1, x_n - X), Q_2]. \\
 U_3 &= U[X | \max(K_1, x_n - Q_2), \min(K_2, x_1)] U[L_1 | M_1, M_2] U[L_2 | \max(Q_1, x_n - X), Q_2]
 \end{aligned}$$

### 5.3.2 Calculation of Moments

The moments can be obtained from

$$D(m_\alpha, m_X, m_{L_1}, m_{L_2}) = \iiint dX dL_1 dL_2 (T_1 + T_2 + T_3). \tag{5.71}$$

Here we want to integrate out  $L_1$  and  $L_2$  followed by an integration over  $X$ . Each of the terms  $T_1$ ,  $T_2$  and  $T_3$  are integrated separately. The window functions  $U_1$ ,  $U_2$  and  $U_3$  are chosen as shown

in (B.15).

Starting with  $T_2$ , it is given by

$$T_2 = \sum_{k=1}^{n-1} \Lambda(m_\alpha, n, k) X^{m_X} L_1^{m_{L_1} - n + k - 1} L_2^{m_{L_2} - k - 1} U[X \mid \max(K_1, x_{n-k}, x_n - Q_2), \min(K_2, x_{n-k+1}, x_1 + M_2)] \\ \times U[L_1 \mid \max(M_1, X - x_1), M_2] U[L_2 \mid \max(Q_1, x_n - X), Q_2]. \quad (5.72)$$

Integrating out  $L_1$  and  $L_2$  we get

$$\iint dL_1 dL_2 T_2 = \sum_{k=1}^{n-1} \Lambda(m_\alpha, n, k) X^{m_X} I_1 I_2 U[X \mid \max(K_1, x_{n-k}, x_n - Q_2), \min(K_2, x_{n-k+1}, x_1 + M_2)], \quad (5.73)$$

where

$$I_1(\max(M_1, X - x_1)) = \int_{\max(M_1, X - x_1)}^{M_2} dL_1 L_1^{m_{L_1} - n + k - 1} \\ = \begin{cases} \ln\left(\frac{M_2}{\max(M_1, X - x_1)}\right) & , \text{ for } m_{L_1} = n - k \\ \frac{M_2^{m_{L_1} - n + k} - (\max(M_1, X - x_1))^{m_{L_1} - n + k}}{m_{L_1} - n + k} & , \text{ for } m_{L_1} \neq n - k. \end{cases} \quad (5.74)$$

and

$$I_2(\max(Q_1, x_n - X)) = \int_{\max(Q_1, x_n - X)}^{Q_2} dL_2 L_2^{m_{L_2} - k - 1} \\ = \begin{cases} \ln\left(\frac{Q_2}{\max(Q_1, x_n - X)}\right) & , \text{ for } m_{L_2} = k \\ \frac{Q_2^{m_{L_2} - k} - (\max(Q_1, x_n - X))^{m_{L_2} - k}}{m_{L_2} - k} & , \text{ for } m_{L_2} \neq k. \end{cases} \quad (5.75)$$

Finally integrate out  $X$ :

$$\iiint dX dL_1 dL_2 T_2 = \sum_{k=1}^{n-1} \Lambda(m_\alpha, n, k) \int_{\max(K_1, x_{n-k}, x_n - Q_2)}^{\min(K_2, x_{n-k+1}, x_1 + M_2)} dX X^{m_X} I_1 I_2. \quad (5.76)$$

Letting  $a = \max(K_1, x_{n-k}, x_n - Q_2)$  and  $b = \min(K_2, x_{n-k+1}, x_1 + M_2)$ , and noting that

$$\max(M_1, X - x_1) = \begin{cases} M_1 & , \text{ if } X \leq M_1 + x_1 \\ X - x_1 & , \text{ if } X \geq M_1 + x_1 \end{cases} \quad (5.77)$$

and

$$\max(Q_1, x_n - X) = \begin{cases} Q_1 & , \text{ if } X \geq x_n - Q_1 \\ x_n - X & , \text{ if } X \leq x_n - Q_1, \end{cases} \quad (5.78)$$



we can rid the integral over  $X$  of the max functions if the numerous cases they produce are represented by theta functions indicating the positions of the points  $a$  and  $b$  relative to the points  $x_n - Q_1$  and  $M_1 + x_1$ . If we let  $Q_1 = M_1$ , then the following six possibilities arise

$$\begin{aligned}
 & \int_a^b dX X^{m_X} I_1 I_2 \\
 &= \theta(M_1 + x_1 - a)\theta(M_1 + x_1 - b) \int_a^b dX I_1(M_1)I_2(x_n - X) \\
 &+ \theta(a - (x_n - Q_1))\theta(b - (x_n - Q_1)) \int_a^b dX I_1(X - x_1)I_2(Q_1) \\
 &+ \theta(a - (M_1 + x_1))\theta((x_n - Q_1) - b) \int_a^b dX I_1(X - x_1)I_2(x_n - X) \\
 &\theta((M_1 + x_1) - a)\theta(b - (x_n - Q_1)) \left[ \int_a^{M_1+x_1} dX I_1(M_1)I_2(x_n - X) \right. \\
 &\quad \left. + \int_{M_1+x_1}^{x_n-Q_1} dX I_1(X - x_1)I_2(x_n - X) \right. \\
 &\quad \left. + \int_{x_n-Q_1}^b dX I_1(X - x_1)I_2(Q_1) \right] \\
 &+ \theta((M_1 + x_1) - a)\theta((x_n - Q_1) - b)\theta(b - (M_1 + x_1)) \\
 &\quad \times \left[ \int_a^{M_1+x_1} dX I_1(M_1)I_2(x_1 - X) + \int_{M_1+x_1}^b dX I_1(X - x_1)I_2(x_n - X) \right] \\
 &+ \theta(a - (M_1 + x_1))\theta((x_n - Q_1) - a)\theta(b - (x_n - Q_1)) \\
 &\quad \times \left[ \int_a^{x_n-Q_1} dX I_1(X - x_1)I_2(x_1 - X) + \int_{x_n-Q_1}^b dX I_1(X - x_1)I_2(Q_1) \right].
 \end{aligned} \tag{5.79}$$

To simplify we can take the limits  $K_1 \rightarrow -\infty$  and  $K_2 \rightarrow \infty$ , then  $a = \max(x_{n-k}, x_n - Q_2)$  and  $b = \min(x_{n-k+1}, x_1 + M_2)$ . If  $M_2$  and  $Q_2$  was chosen large enough so that  $M_2 \geq x_{n-k+1} - x_1$  and  $Q_2 \geq x_n - x_{n-k}$ , then  $a = x_{n-k}$  and  $b = x_{n-k+1}$ , and we are thus integrating over successive data points. If our prior knowledge is limited, we may want to choose  $M_1$  and  $Q_1$  small, but we cannot send their limits to zero. If we did then only the third term in the above expression would survive, and this leads to dividing by zero or integrating  $\ln$  at zero when  $k = 1$  and  $k = n - 1$ . The last two terms thus protect us from doing just that so long as  $M_1, Q_1 > 0$ . The user can also not take the limits  $M_2 \rightarrow \infty$  and  $Q_2 \rightarrow \infty$  in an attempt to simplify matters, as doing so causes  $I_1$  and  $I_2$  to diverge when  $m_{L_1} = n - k$  and when  $m_{L_2} = k$ . Thus after appropriate values for the prior bounds have been chosen, the integrals can be performed exactly, and are also uncomplicated enough to be performed numerically.

Now we turn our attention to the first and last terms of (5.69). These are the scenarios where Bayes' theorem enumerates the possibilities that all the data points have come from either  $g$  or all have come from  $h$ . The first term of (5.69) is

$$\begin{aligned}
 T_1 = & \Lambda(m_\alpha, n, 0) X^{m_X} L_1^{m_{L_1} - n - 1} L_2^{m_{L_2} - 1} U[X | \max(K_1, x_n), \min(K_2, x_1 + M_2)] \\
 & \times U[L_1 | \max(M_1, X - x_1), M_2] U[L_2 | Q_1, Q_2].
 \end{aligned} \tag{5.80}$$

Integrating out  $L_1$  and  $L_2$  we get

$$\iint dL_1 dL_2 T_1 = \Lambda(m_\alpha, n, 0) X^{m_X} I_3 I_4 U[X | \max(K_1, x_n), \min(K_2, x_1 + M_2)], \quad (5.81)$$

where

$$\begin{aligned} I_3(\max(M_1, X - x_1)) &= \int_{\max(M_1, X - x_1)}^{M_2} dL_1 L_1^{m_{L_1} - n - 1} \\ &= \begin{cases} \ln\left(\frac{M_2}{\max(M_1, X - x_1)}\right) & , \text{ for } m_{L_1} - n = 0 \\ \frac{M_2^{m_{L_1} - n} - (\max(M_1, X - x_1))^{m_{L_1} - n}}{m_{L_1} - n} & \text{ for } m_{L_1} - n \neq 0 \end{cases} \end{aligned} \quad (5.82)$$

and

$$I_4 = \int_{Q_1}^{Q_2} dL_2 L_2^{m_{L_2} - 1} = \begin{cases} \ln\left(\frac{Q_2}{Q_1}\right) & , \text{ for } m_{L_2} = 0 \\ \frac{Q_2^{m_{L_2}} - Q_1^{m_{L_2}}}{m_{L_2}} & \text{ for } m_{L_2} \neq 0. \end{cases} \quad (5.83)$$

Integrating out  $X$  we get

$$\iiint dX dL_1 dL_2 T_1 = \Lambda(m_\alpha, n, 0) I_4 \int_{\max(K_1, x_n)}^{\min(K_2, x_1 + M_2)} dX X^{m_X} I_3. \quad (5.84)$$

Letting  $c = \max(K_1, x_n)$  and  $d = \min(K_2, x_1 + M_2)$ , the  $X$  integral is

$$\begin{aligned} &\int_c^d dX X^{m_X} I_3 \\ &= \theta(x_1 + M_1 - c)\theta(x_1 + M_1 - d) \int_c^d dX X^{m_X} I_3(M_1) \\ &+ \theta(x_1 + M_1 - c)\theta(d - (x_1 + M_1)) \\ &\quad \times \left[ \int_c^{x_1 + M_1} dX X^{m_X} I_3(M_1) + \int_{x_1 + M_1}^d dX X^{m_X} I_3(X - x_1) \right] \\ &+ \theta(c - (x_1 + M_1))\theta(d - (x_1 + M_1)) \int_c^d dX X^{m_X} I_3(X - x_1), \end{aligned} \quad (5.85)$$

where we have once again rid the integral of the max function by introducing theta functions. After taking the limits  $K_1 \rightarrow -\infty$  and  $K_2 \rightarrow \infty$  we have that  $c = x_n$  and  $d = x_1 + M_2$ .

The third term of (5.69) is

$$\begin{aligned} T_3 &= \Lambda(m_\alpha, n, n) X^{m_X} L_1^{m_{L_1} - 1} L_2^{m_{L_2} - n - 1} U[X | \max(K_1, x_n - Q_2), \min(K_2, x_1)] \\ &\quad \times U[L_1 | M_1, M_2] U[L_2 | \max(Q_1, x_n - X), Q_2]. \end{aligned} \quad (5.86)$$

Integrating out  $L_1$  and  $L_2$  we get

$$\iint dL_1 dL_2 T_3 = \Lambda(m_\alpha, n, n) X^{m_X} I_5 I_6 U[X | \max(K_1, x_n - Q_2), \min(K_2, x_1)], \quad (5.87)$$

where

$$\begin{aligned}
 I_5(\max(Q_1, x_n - X)) &= \int_{\max(Q_1, x_n - X)}^{Q_2} dL_2 L_2^{m_{L_2} - n - 1} \\
 &= \begin{cases} \ln\left(\frac{Q_2}{\max(Q_1, x_n - X)}\right) & , \text{ for } m_{L_2} - n = 0 \\ \frac{Q_2^{m_{L_2} - n} - (\max(Q_1, x_n - X))^{m_{L_2} - n}}{m_{L_2} - n} & \text{ for } m_{L_2} - n \neq 0. \end{cases} \quad (5.88)
 \end{aligned}$$

and

$$I_6 = \int_{M_1}^{M_2} dL_1 L_1^{m_{L_1} - 1} = \begin{cases} \ln\left(\frac{M_2}{M_1}\right) & , \text{ for } m_{L_1} = 0 \\ \frac{M_2^{m_{L_1}} - M_1^{m_{L_1}}}{m_{L_1}} & \text{ for } m_{L_1} \neq 0. \end{cases} \quad (5.89)$$

Integrating out  $X$  we get

$$\iiint dX dL_1 dL_2 T_3 = \Lambda(m_\alpha, n, n) I_6 \int_{\max(K_1, x_n - Q_2)}^{\min(K_2, x_1)} dX X^{m_X} I_5. \quad (5.90)$$

Letting  $u = \max(K_1, x_n - Q_2)$  and  $v = \min(K_2, x_1)$  the  $X$  integral is

$$\begin{aligned}
 &\int_u^v dX X^{m_X} I_5 \\
 &= \theta(x_n - Q_1 - u)\theta(x_n - Q_1 - v) \int_u^v dX X^{m_X} I_5(x_n - X) \\
 &+ \theta(x_n - Q_1 - u)\theta(v - (x_n - Q_1)) \\
 &\quad \times \left[ \int_u^{x_n - Q_1} dX X^{m_X} I_5(x_n - X) + \int_{x_n - Q_1}^v dX X^{m_X} I_5(Q_1) \right] \\
 &+ \theta(u - (x_n - Q_1))\theta(v - (x_n - Q_1)) \int_u^v dX X^{m_X} I_5(Q_1). \quad (5.91)
 \end{aligned}$$

After taking the limits  $K_1 \rightarrow -\infty$  and  $K_2 \rightarrow \infty$  we have that  $u = x_n - Q_2$  and  $v = x_1$ .

### 5.3.3 Marginal posterior distribution for $X$

From (5.60), the marginal posterior distribution for  $X$  is given by

$$P(X | DI) = \frac{1}{N} A(0, 0, 0, 0). \quad (5.92)$$

Putting together the results of the previous section we find that

$$\begin{aligned}
 A(0, 0, 0, 0) &= \iint dL_1 dL_2 (T_1 + T_2 + T_3) \\
 &= \Lambda(0, n, 0) I_3 I_4 U[X | \max(K_1, x_n), \min(K_2, x_1 + M_2)] \\
 &+ \sum_{k=1}^{n-1} \Lambda(0, n, k) I_1 I_2 U[X | \max(K_1, x_{n-k}, x_n - Q_2), \min(K_2, x_{n-k+1}, x_1 + M_2)] \\
 &+ \Lambda(0, n, n) I_5 I_6 U[X | \max(K_1, x_n - Q_2), \min(K_2, x_1)], \quad (5.93)
 \end{aligned}$$

where

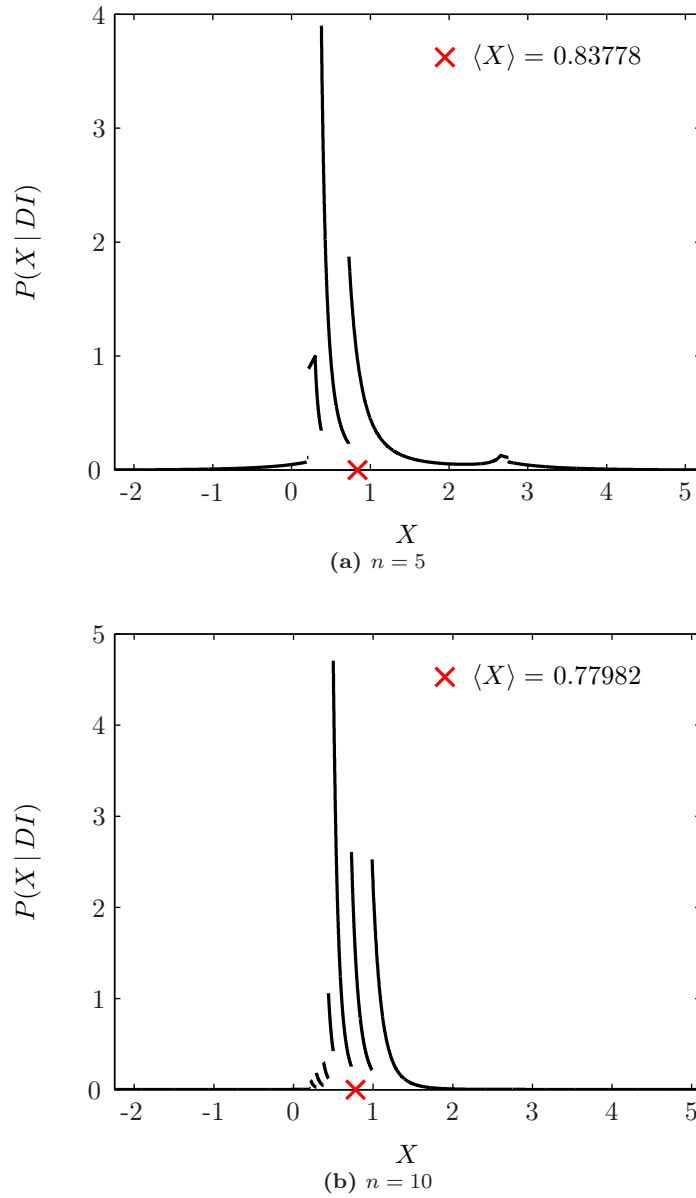
$$\begin{aligned}
 I_1(m_{L_1} = 0) &= \frac{M_2^{-n+k} - (\max(M_1, X - x_1))^{-n+k}}{-n+k} \\
 I_2(m_{L_2} = 0) &= \frac{Q_2^{-k} - (\max(Q_1, x_n - X))^{-k}}{-k} \\
 I_3(m_{L_1} = 0) &= \frac{M_2^{-n} - (\max(M_1, X - x_1))^{-n}}{-n} \\
 I_4(m_{L_2} = 0) &= \ln\left(\frac{Q_2}{Q_1}\right) \\
 I_5(m_{L_2} = 0) &= \frac{Q_2^{-n} - (\max(Q_1, x_n - X))^{-n}}{-n} \\
 I_6(m_{L_1} = 0) &= \ln\left(\frac{M_2}{M_1}\right).
 \end{aligned} \tag{5.94}$$

After taking the limits  $K_1 \rightarrow -\infty$  and  $K_2 \rightarrow \infty$  we have that

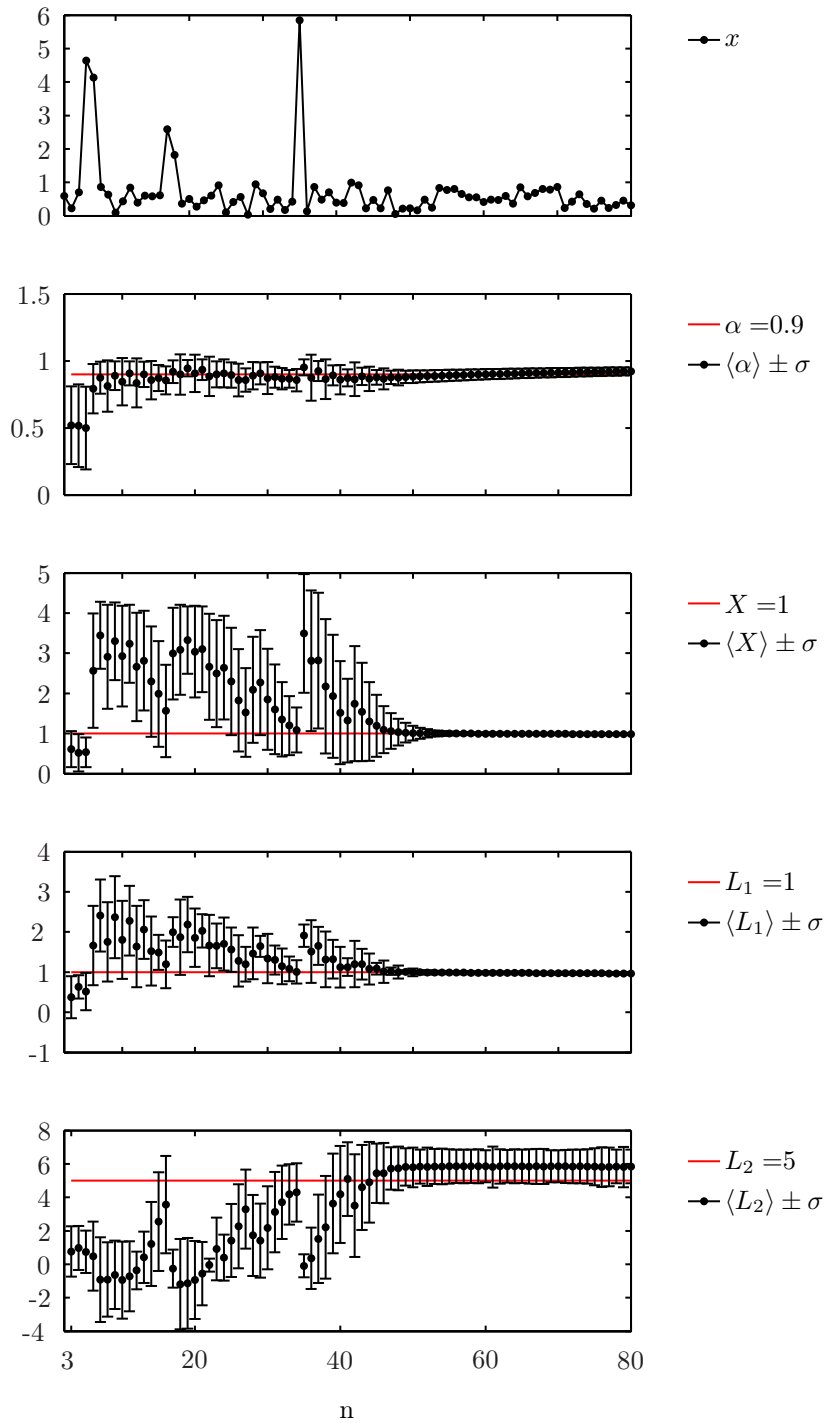
$$\begin{aligned}
 P(X | DI) &\propto A(0, 0, 0, 0) \\
 &= \Lambda(0, n, 0) I_3 I_4 U[X | x_n, x_1 + M_2] \\
 &\quad + \sum_{k=1}^{n-1} \Lambda(0, n, k) I_1 I_2 U[X | \max(x_{n-k}, x_n - Q_2), \min(x_{n-k+1}, x_1 + M_2)] \\
 &\quad + \Lambda(0, n, n) I_5 I_6 U[X | x_n - Q_2, x_1].
 \end{aligned} \tag{5.95}$$

### 5.3.4 Results

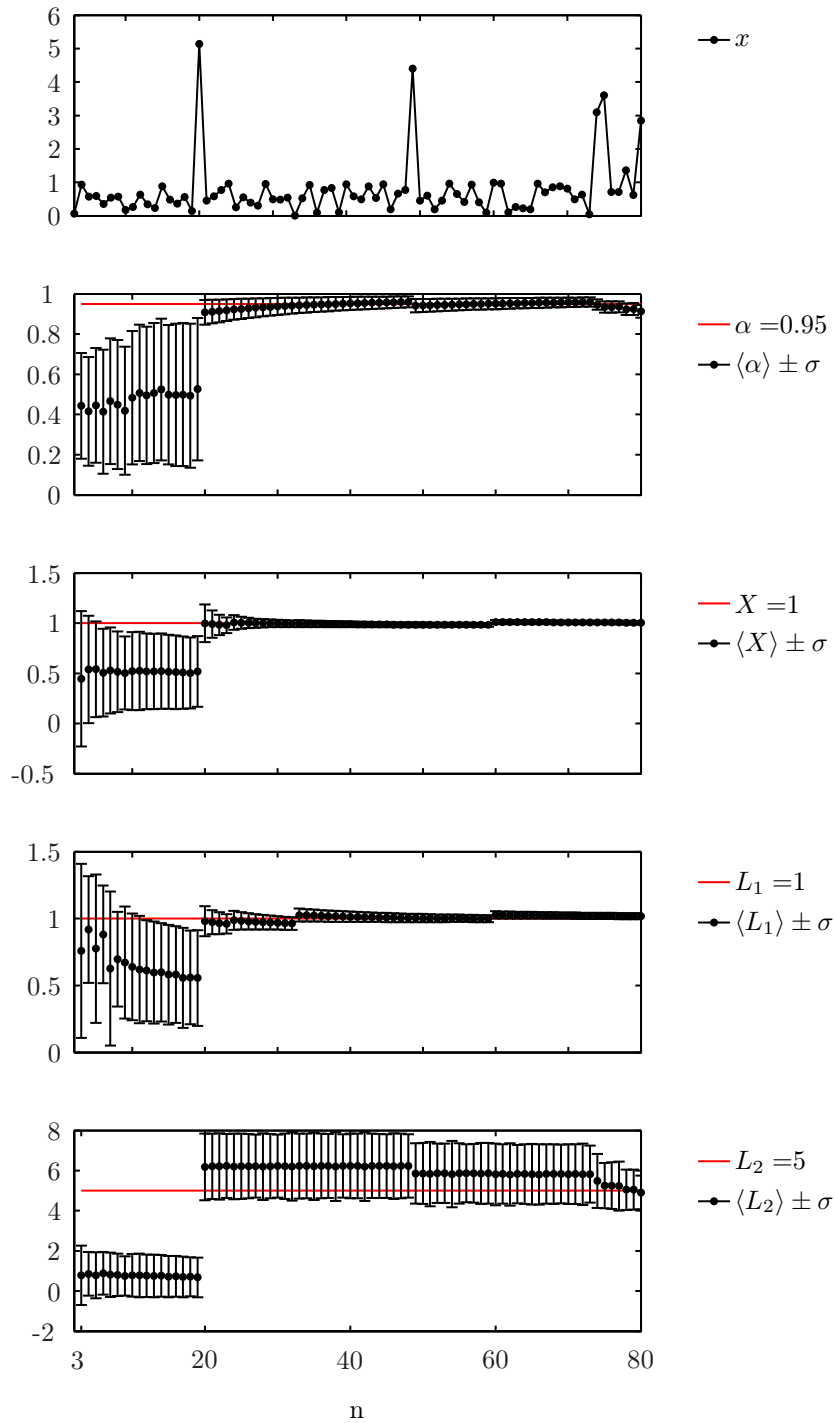
The results of sections 5.3.2 and 5.3.3 are now implemented for simulated data (see Figs. 5.5-5.8 below).



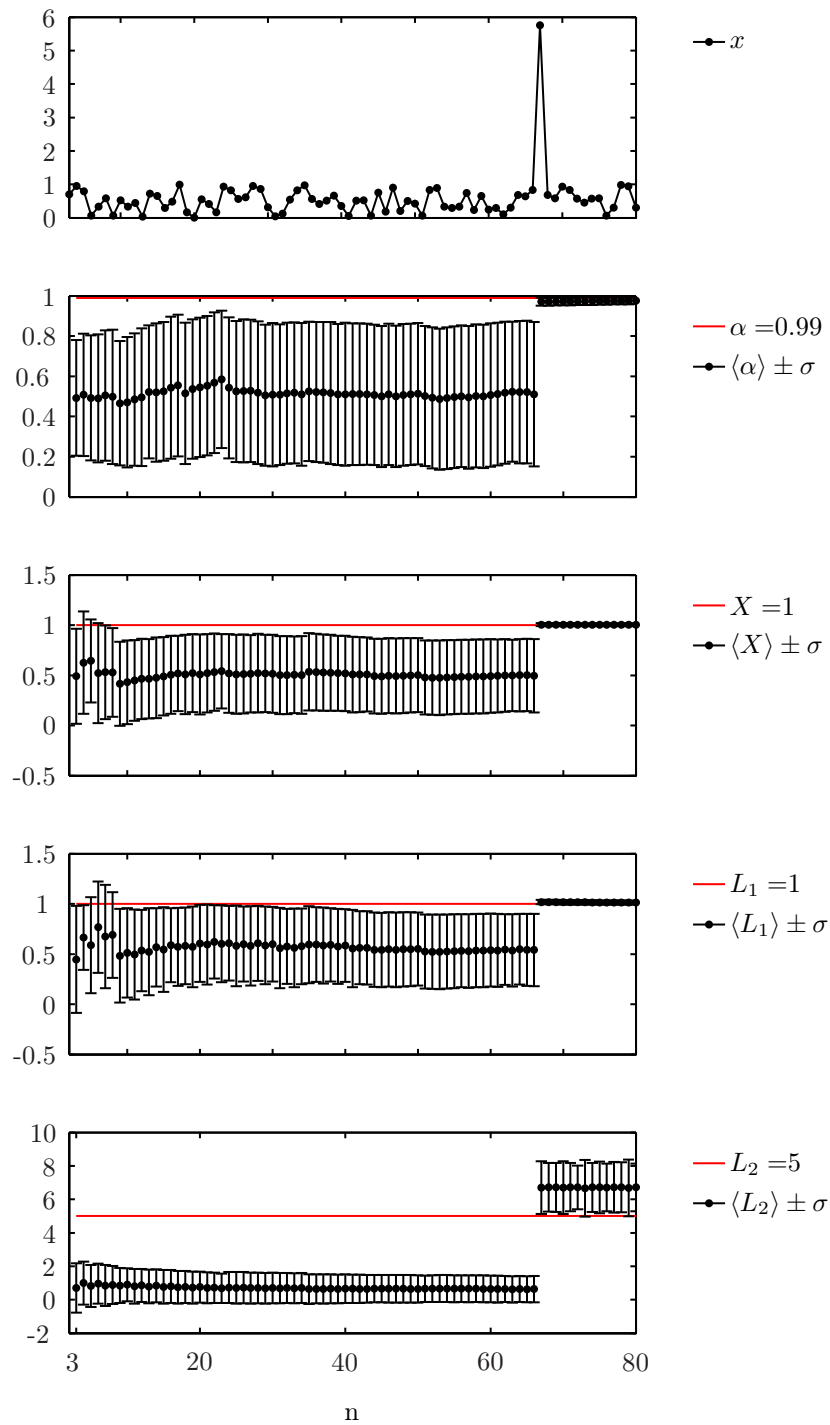
**Figure 5.5:** Posterior distribution  $P(X | DI)$  for the mixture model (Eq. 5.91). Each of the spikes represents one of the terms in the binomial expansion.



**Figure 5.6:** At each trial  $n$  a datum is generated according to the mixture model of two uniform distributions and the means and standard deviations of all ( $\alpha$  is considered unknown) the parameters are calculated. All figures share the same horizontal axis. The top figure shows the original data sequence  $x$ . Each figure thereafter has its own legend. Indicated by red lines are the model parameters used during the simulation. The prior bounds were set at  $M_1 = Q_1 = 0.01$ ,  $M_2 = 2$ ,  $Q_2 = 10$ ,  $A_1 = 0$  and  $A_2 = 1$ .



**Figure 5.7:** At each trial  $n$  a datum is generated according to the mixture model of two uniform distributions and the means and standard deviations of all ( $\alpha$  is considered unknown) the parameters are calculated. All figures share the same horizontal axis. The top figure shows the original data sequence  $x$ . Each figure thereafter has its own legend. Indicated by red lines are the model parameters used during the simulation. The prior bounds were set at  $M_1 = Q_1 = 0.01$ ,  $M_2 = 2$ ,  $Q_2 = 10$ ,  $A_1 = 0$  and  $A_2 = 1$ .



**Figure 5.8:** At each trial  $n$  a datum is generated according to the mixture model of two uniform distributions and the means and standard deviations of all ( $\alpha$  is considered unknown) the parameters are calculated. All figures share the same horizontal axis. The top figure shows the original data sequence  $x$ . Each figure thereafter has its own legend. Indicated by red lines are the model parameters used during the simulation. The prior bounds were set at  $M_1 = Q_1 = 0.01$ ,  $M_2 = 2$ ,  $Q_2 = 10$ ,  $A_1 = 0$  and  $A_2 = 1$ . A particularly interesting example. The moment one outlier appears, the parameters are more or less fixed, but not before.



## Chapter 6

# Comparison to LULU based solution

### 6.1 Comparison to LULU based solution

In this Chapter the estimators of the LULU and Bayesian approach to the mixture model of two uniform distributions are compared directly. Although one expects the optimal Bayesian estimators to outperform the LULU-estimators, qualitatively the latter are shown to perform admirably. Consider the mixture model of Chapter 2 section 2.4.2. The LULU-estimators of the scale parameters are given by Eq. (2.114). They are

$$(\mu)_{est}^{UL} = \frac{\overline{\eta_{22}x - L_1x} - \rho_{12}\overline{L_1x - U_1L_1x}}{\rho_{11}\eta_{22} - \rho_{12}\eta_{21}} \quad (6.1)$$

and

$$(\epsilon)_{est}^{UL} = \frac{-\overline{\eta_{21}x - L_1x} + \rho_{11}\overline{L_1x - U_1L_1x}}{\rho_{11}\eta_{22} - \rho_{12}\eta_{21}}, \quad (6.2)$$

where  $\rho_{11}$ ,  $\rho_{12}$ ,  $\eta_{21}$  and  $\eta_{22}$  are polynomials in  $\alpha$  indicated by equations (2.110) and (2.111). The LULU-estimate of the location parameter is chosen as

$$(c)_{est}^{UL} = \overline{U_1L_1x}, \quad (6.3)$$

the average of the smoothed sequence  $U_1L_1x$ . The mixture model studied in section 5.3.1 is parametrised differently where the parameters are related by  $L_1 = 2\mu$ ,  $L_2 = 2\epsilon$  and therefore  $c = X - \mu = X - \frac{1}{2}L_1$ . Correspondingly, Bayesian estimators which can be compared to the LULU-estimators are thus

$$(\mu)_{est}^B = \frac{1}{2}\langle L_1 \rangle, \quad (6.4)$$

and

$$(\epsilon)_{est}^B = \frac{1}{2}\langle L_2 \rangle. \quad (6.5)$$

Although the models studied for the LULU and Bayesian approach have their location parameters positioned differently, we can compare

$$(c)_{est}^B = \langle X \rangle - \frac{1}{2}\langle L_1 \rangle \quad (6.6)$$

to the LULU-estimate  $\overline{U_1 L_1 x}$  since the expectation value operator is linear:

$$\begin{aligned} \langle X - \frac{L_1}{2} \rangle &= \iint dX dL_1 \left( X - \frac{L_1}{2} \right) P(X L_1 | \alpha_2 DI) \\ &= \iint dX dL_1 X P(X L_1 | \alpha L_2 DI) + \frac{1}{2} \iint dX dL_1 L_1 P(X L_1 | \alpha L_2 DI) \\ &= \langle X \rangle - \frac{1}{2} \langle L_1 \rangle. \end{aligned} \quad (6.7)$$

One way of comparing estimators is to consider for example the mean absolute error (MAE) of an estimate with the known numerical experimental value of the parameter being estimated. Suppose  $\theta$  is the known parameter and  $(\theta)_{est}$  is an estimate thereof, then the MAE is given by

$$\text{MAE}(\theta, (\theta)_{est}, k) = \sum_{i=1}^k \frac{|(\theta)_{est}(k) - \theta|}{k}, \quad (6.8)$$

where  $k$  is the amount of data points collected after  $k$  trials. In this chapter  $k$  is not to be confused with the summation index  $k$  of Chapter 5<sup>1</sup>. A quantity like the MAE for a single experiment is of no more use than a direct comparison, so quantities such as the MAE will be calculated over an ensemble of data sets. This may indicate which estimator performs better on average for data sets drawn from the same distribution. In the simulations that follow the ensemble average of the MAE is written as

$$\|\text{MAE}(\theta, (\theta)_{est}, k)\|_{ens}.$$

Define also the root mean square error

$$\text{RMSE}(\theta, (\theta)_{est}, k) = \sqrt{\sum_{i=1}^k \frac{((\theta)_{est}(k) - \theta)^2}{k}}, \quad (6.9)$$

and its ensemble average

$$\|\text{RMSE}(\theta, (\theta)_{est}, k)\|_{ens}.$$

Before a comparison is made, a discussion on end-values is appropriate.

## 6.2 Discussion on end-values

Every time we collect a new datum at a trial  $k$ , the question of end-values for the purpose of the running window of the smoothers  $L_1$  and  $U_1$  comes into question. Since for our purposes we will only be smoothing up to the first resolution level, the running window is of size three and we only need each data point to have at least two neighbours. The question is thus what should we do with the first and  $k$ -th data point which have no neighbours to the left or the right of the sequence respectively? There is no prescribed way of dealing with end-values except some approaches may be favoured depending on the particular application [12].

In the previous simulations we opted not to add end-values. This is the so-called *omit end-value rule*. Smoothing can only begin after the third data point has been accumulated ( $k \geq 3$ ). The smoothed sequence is thus shorter than the original and, in order to calculate the correct and

---

<sup>1</sup>In Chapter 5,  $n$  represents the amount of accumulated data points whereas in Chapter 2  $k$  is used for this purpose whilst  $n$  is for resolution level.

corresponding pulses, the original sequence is also shortened accordingly.

Other methods include for example the popular *replicate end-value* rule. Here the first and the  $k$ -th datum is appended to both ends of the sequence that is to be smoothed. In other words if at trial  $k$  our data sequence is  $x' = \{x_i\}_{i=1}^k$ , then the sequence that is to be smoothed is  $x = \{x_1, x_1, x_2, \dots, x_{k-1}, x_k, x_k\}$ . Although the sequence will be constant, smoothing can begin after one datum has been collected. The smoothed sequence is the same length as the original data sequence and no data points are ‘lost’ to the ends [12].

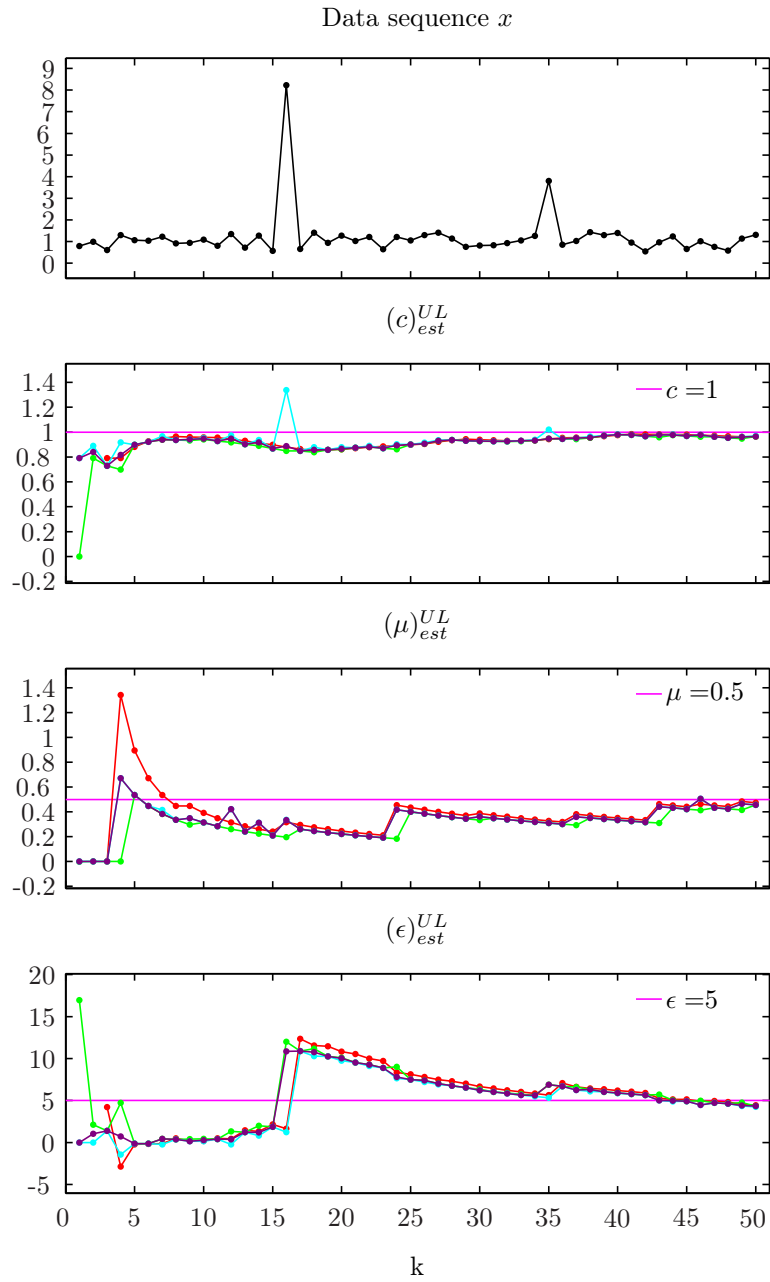
We shall refer to it as the *zero end-value* rule when we place zeros at the ends of the data sequence as follows  $x = \{0, x_1, x_2, \dots, x_{k-1}, x_k, 0\}$ . This can clearly have a devastating effect on the average pulse height when we only have one data point. For later trials however any 1-pulse at the ends can only be of magnitude  $|x_2 - x_1|$  or  $|x_k - x_{k-1}|$  unless  $x_1 > 0 > x_2$  or  $x_k > 0 > x_{k-1}$  in which case the pulse magnitude will be  $|x_1|$  or  $|x_k|$  respectively. It is important to note that since at every trial the average of the positive and negative pulses of the first resolution level is updated, no effect like that just described can permeate through to the averages of the pulses at later trials  $k \geq 2$ .

Finally, we shall refer to it as the *median end-value* rule when we place the median of the data sequence at the ends as follows  $x = \{\text{median}(x'), x_1, x_2, \dots, x_{k-1}, x_k, \text{median}(x')\}$ .

Where we had chosen the omit-end value rule before, this was of little consequence especially when long term averages were calculated. However, since we are about to compare the performance of our LULU-estimators with those of the Bayesian method for a simulated time series, a discussion on end-values becomes relevant. Furthermore, when a comparison is made using only a few data points (say  $k \leq 50$ ), then the choice of end-values does have a noticeable effect on our LULU-estimators.

Studying Fig. 6.1 below, estimators using the omit end-value rule (red) are only possible for  $k \geq 3$ . Considering  $(\epsilon)_{est}^{UL}$ , where estimators using the zero (green) and median (purple) end-value rules display an obvious ‘updating’ when an outlier (at  $k = 16$  and  $k = 35$ ) occurs, estimators using the omit (red) and replicate (blue) end-value rules lag two trials in their response to the outliers. The estimator  $(c)_{est}^{UL} = \overline{U_1 L_1 x}$  for the replicate end-value rule (blue) is seriously affected at the points where the outliers occur. The zero end-value rule (green) works well, except when only one datum is available. The estimator  $(\mu)_{est}^{UL}$  using the median for end values (purple) displays ‘bumps’ in comparison to the other estimators, an effect due to the particular extrapolated end value.

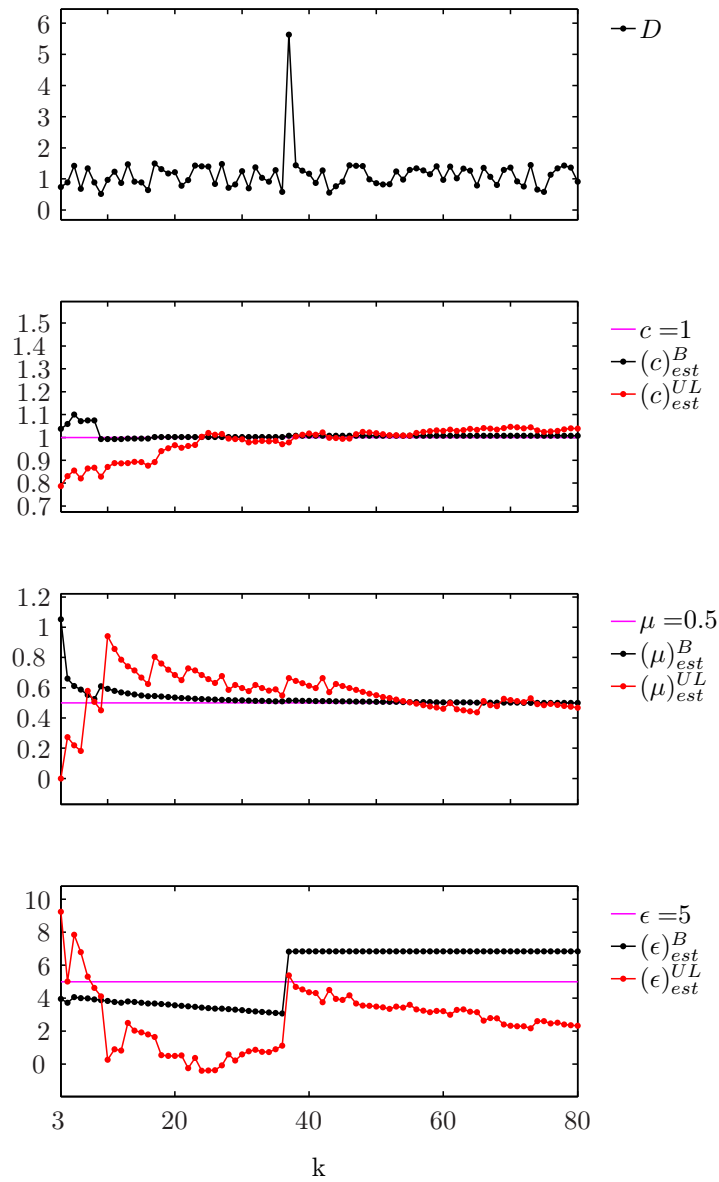
In conclusion, omitting end values works well if the application does not demand real-time estimates. The replicate end-value rule also gives lagged estimates and users should be cautious when estimating  $c$ , as  $\overline{U_1 L_1 x}$  is no longer robust (although the effect is not spread to future estimates). The zero end-value rule works well if we have more than one data point. The median end-value rule is a good choice for real-time estimates and will be our choice for comparison with the Bayesian estimates.



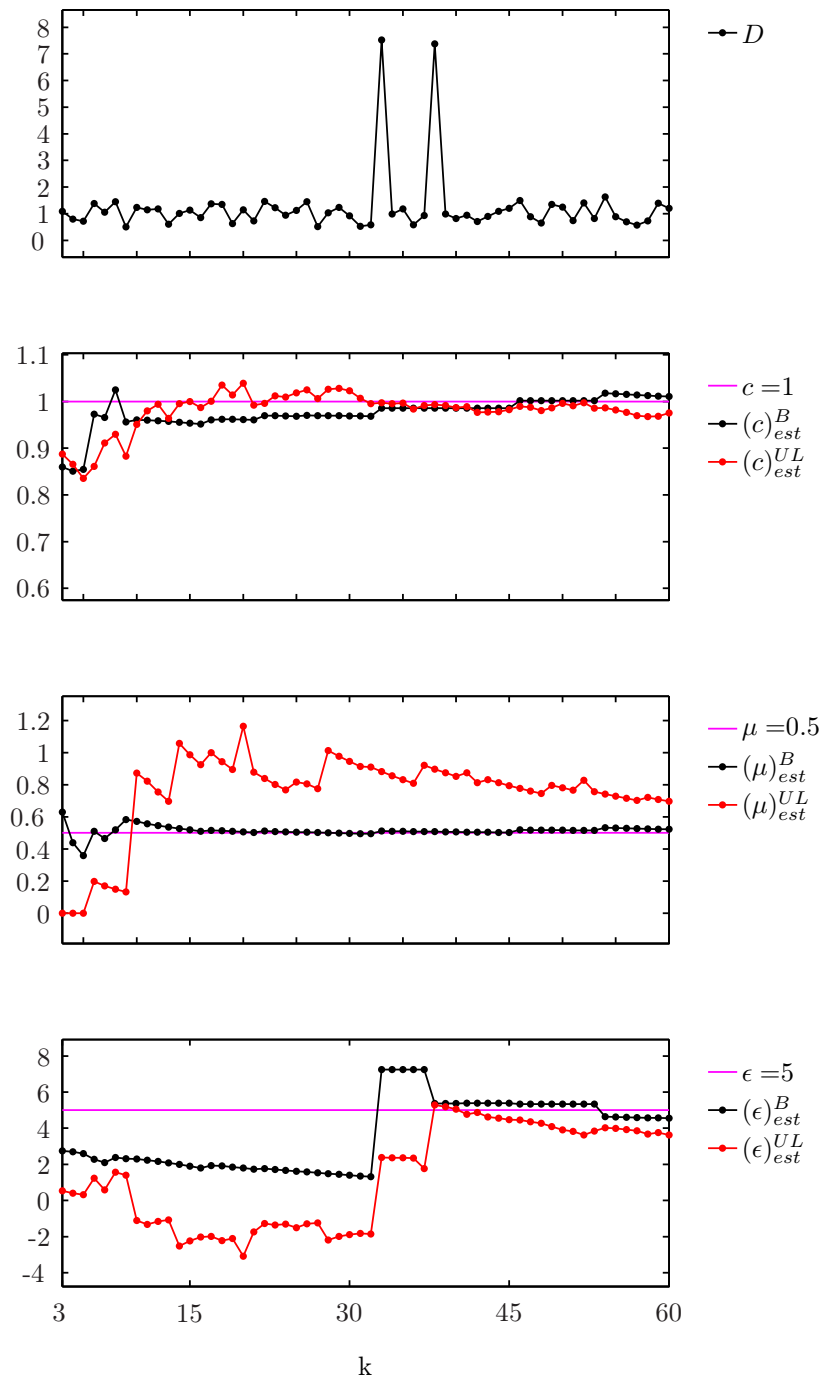
**Figure 6.1:** Comparison of end-values. A data sequence is generated according to the model of Chapter 2 section 2.4.2. The parameters are set at  $\alpha = 0.95$ ,  $c = 1$ ,  $\mu = 0.5$ ,  $\epsilon = 5$ . Red is for where the omit end-value rule was used, green for the zero end-value rule, blue for the replicate end-value rule and purple for the median end-value rule.

### 6.3 Results of Comparison

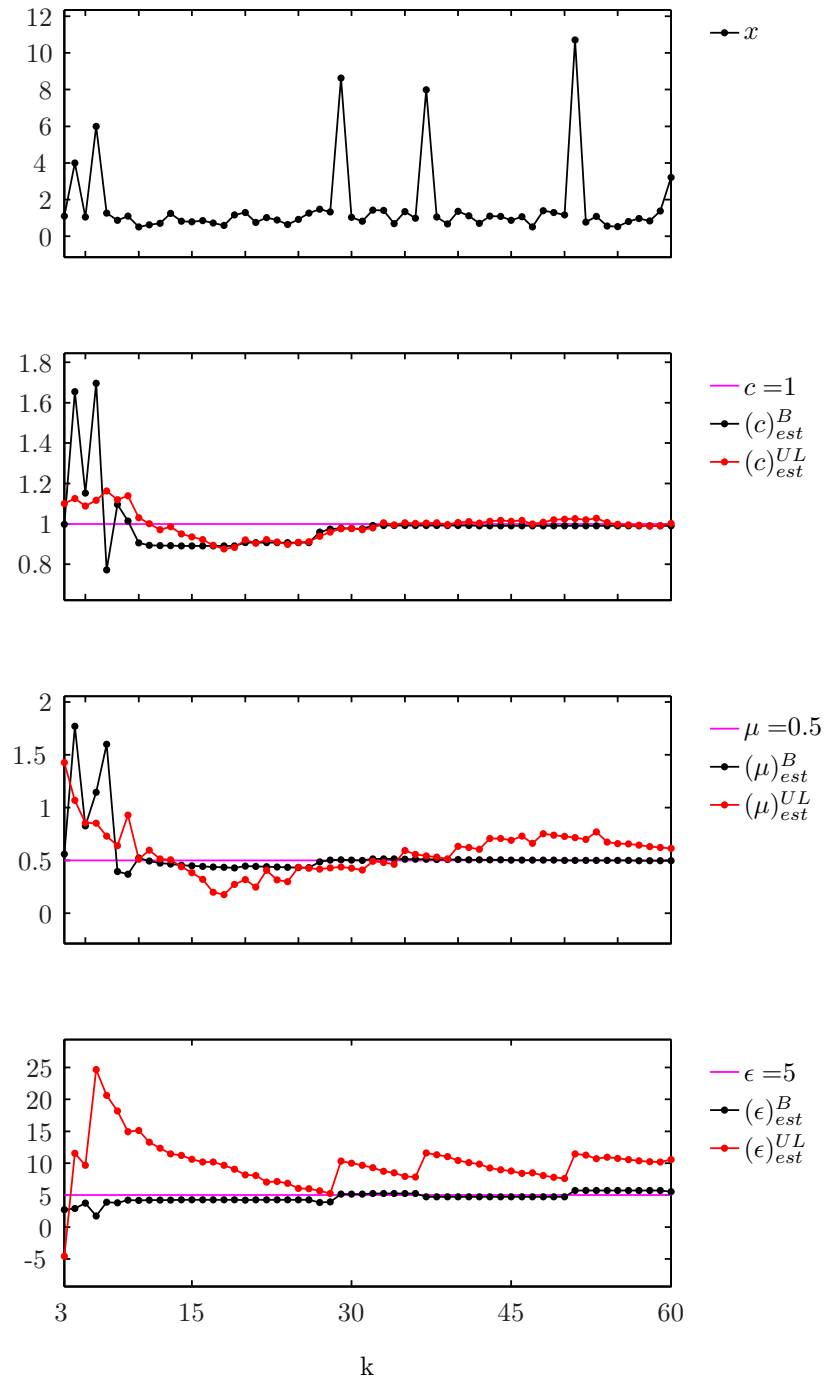
In Figs 6.2-6.4, two individual data sets and the resulting LULU and Bayesian estimators are shown. In Fig 6.5 and 6.6, averages of the MAEs and RMSEs of the estimators over many data simulations are shown.



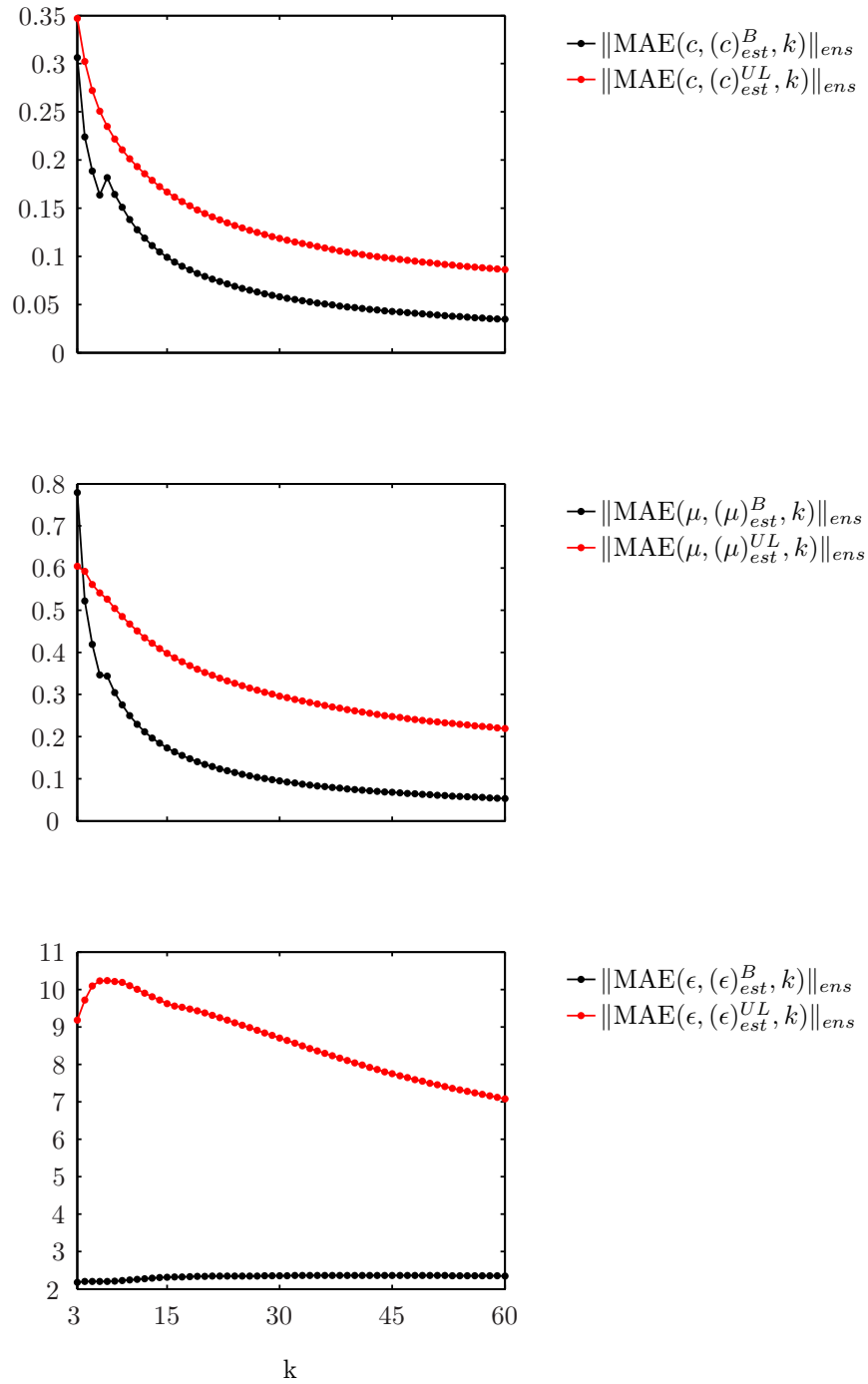
**Figure 6.2:** Direct comparison of the LULU and Bayesian estimators. The figures share the same horizontal axis. The true values of the parameters are represented by straight lines. Parameters were set at  $\alpha = 0.975$ ,  $M_1 = Q_1 = 0.01$  and  $M_2 = Q_2 = 20$



**Figure 6.3:** Direct comparison of the LULU and Bayesian estimators. The figures share the same horizontal axis. The true values of the parameters are represented by straight lines. Parameters were set at  $\alpha = 0.95$ ,  $M_1 = Q_1 = 0.01$  and  $M_2 = Q_2 = 50$

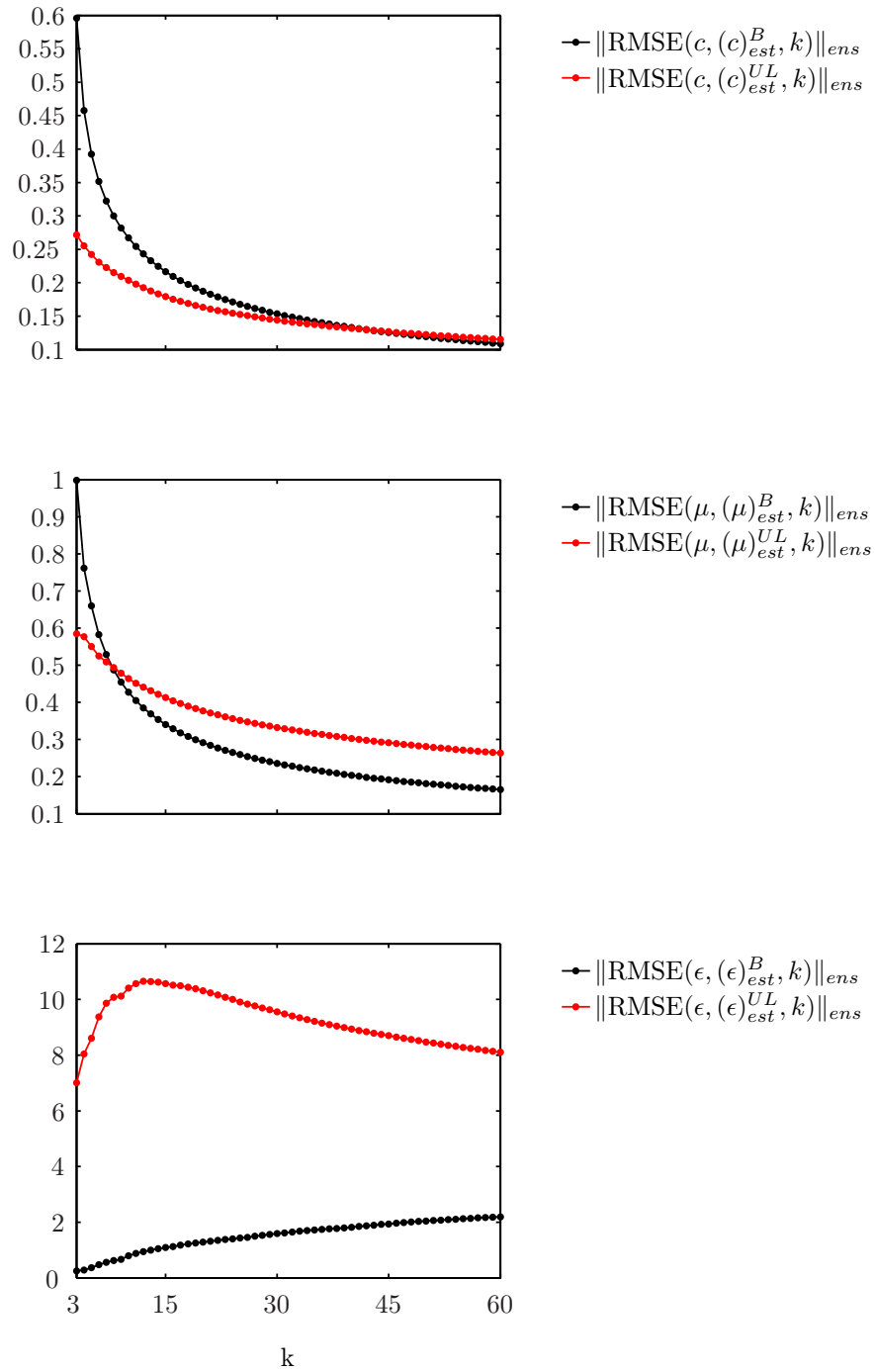


**Figure 6.4:** Direct comparison of the LULU and Bayesian estimators. The figures share the same horizontal axis. The true values of the parameters are represented by straight lines. Parameters were set at  $\alpha = 0.95$ ,  $M_1 = Q_1 = 0.01$  and  $M_2 = Q_2 = 50$



**Figure 6.5:** Ensemble averages of the MAEs. The figures share the same horizontal axis. An ensemble average was calculated using 200 sets of data. Parameters were set at  $\alpha = 0.98$ ,  $c = 1.5$ ,  $\mu = 0.5$ ,  $\epsilon = 5$ ,  $M_1 = Q_1 = 0.01$ ,  $M_2 = Q_2 = 50$ .





**Figure 6.6:** Ensemble averages of the RMSEs. The figures share the same horizontal axis. An ensemble average was calculated using 200 sets of data. Parameters were set at  $\alpha = 0.98$ ,  $c = 1.5$ ,  $\mu = 0.5$ ,  $\epsilon = 5$ ,  $M_1 = Q_1 = 0.01$ ,  $M_2 = Q_2 = 50$ .

The following qualitative description of the comparison is given. With  $\alpha$  known, the Bayesian estimators fair better than when compared with the results of the previous chapter where  $\alpha$  was assumed unknown. Even without the occurrence of an outlier datum they have the parameters  $c$  and  $\mu$  more or less fixed after only a few data points have been gathered (see Fig. 6.2 and 6.3). When an outlier datum does eventually appear, only the estimate for  $\epsilon$  (the scale parameter of the outlier distribution  $h$ ) is updated. The LULU-estimators also fair well and seem to need only a few more data points to get within a reasonably ‘good’ estimate. Notice that both  $(\epsilon)_{est}^{UL}$  and  $(\epsilon)_{est}^B$  are updated accordingly when an outlier point appears in data set. Also notice that both  $(c)_{est}^{UL}$  and  $(c)_{est}^B$  are robust to the outlier events (see Fig 6.2 and 6.3). Studying many such data sets it appears as though the Bayesian estimators are more accurate with fewer data points available, although at times, depending on exactly which data set was drawn, the LULU-estimates for  $c$  and  $\mu$  do much better. This occurs when we have outliers appearing in the first few measurements as can be seen in Fig 6.4.

As was expected, the ensemble averages of the MAE (see Fig 6.4) shows that the Bayesian estimates do better on average than the LULU estimates. The ensemble averages of the RMSE for the estimates of  $c$  and  $\mu$  show that the Bayesian estimators are outperformed by the LULU estimates for a smaller data set (see Fig. 6.6). This can be understood as follows. The RMSE takes large errors more seriously (an error twice as big is taken four times as serious) than it does small errors. In the few cases in the ensemble (realizations of the data) where outliers occur when the data set is still small (see for example Fig 6.4 for one such realization) the Bayesian estimators are ‘thrown off’ (the Bayesian machinery believes the outliers have come from  $g$ ) while the LULU estimators are not. It is these RMSEs that enter the ensemble average which eventually show the LULU-estimators as being more accurate (on average) for smaller data sets.

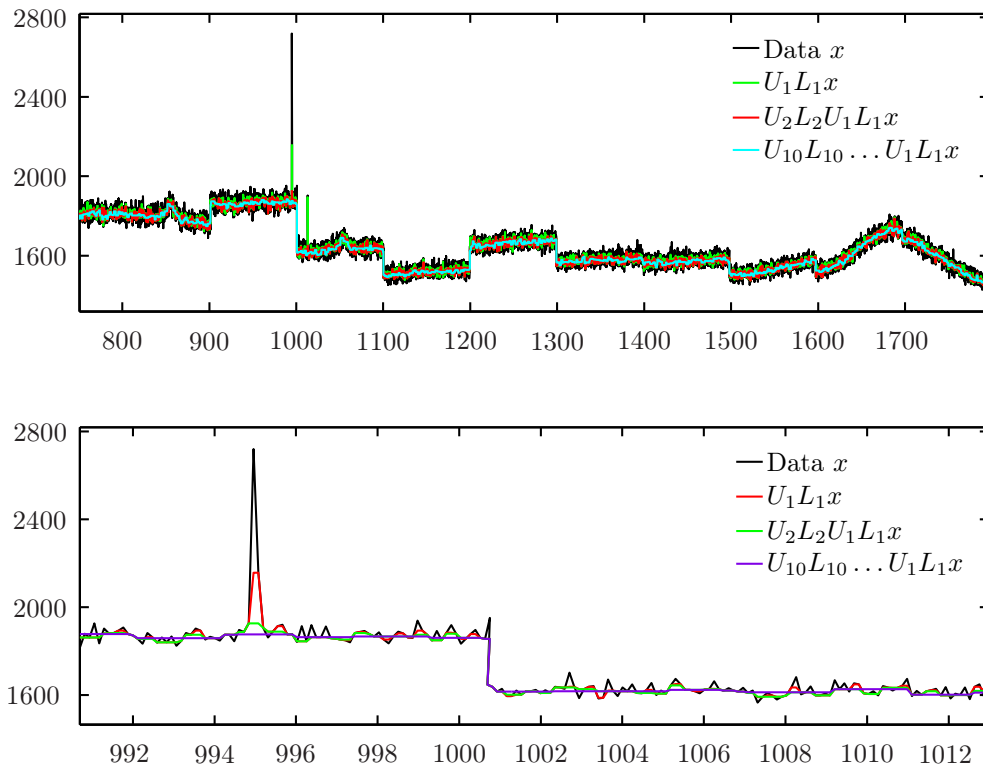
Overall, LULU’s estimates of the parameters of distribution  $g$  are good and only lagging slightly behind those achieved by the Bayesian estimates. They are however far simpler to calculate: One only needs the assumption that  $g$  is symmetric about its location parameter in order to use  $\overline{U_1 L_1 x}$  as the estimate in which case it has been shown to be robust when the outlier events are rare ( $\beta \approx 0$ ). Further, the scale parameter may be estimated using averages of pulses attained after smoothing. The only calculation needed to get these scale parameter estimates involves integrals over the cdf of the model distribution. As previous examples show, these are very easy to calculate. On the other hand the Bayesian solution is very difficult to attain analytically and slow to compute numerically if one compares with the much simpler LULU procedure.

The LULU-estimators developed in this thesis are shown to be a practical compromise to the specific problem at hand.

## Chapter 7

# Application to Laser Spectroscopy Data

The RAMAN spectroscopy data analysed here was kindly provided by the Laser Physics group at Stellenbosch University. A plot of the original data appears in the introduction and is repeated in Fig. 7.1 below. The full data set is shown in the upper panel. Upon initial inspection, the measurement appears to be of a signal that remains constant (although with obvious trend at times) for certain periods along the horizontal axis before a step is made to a different level.



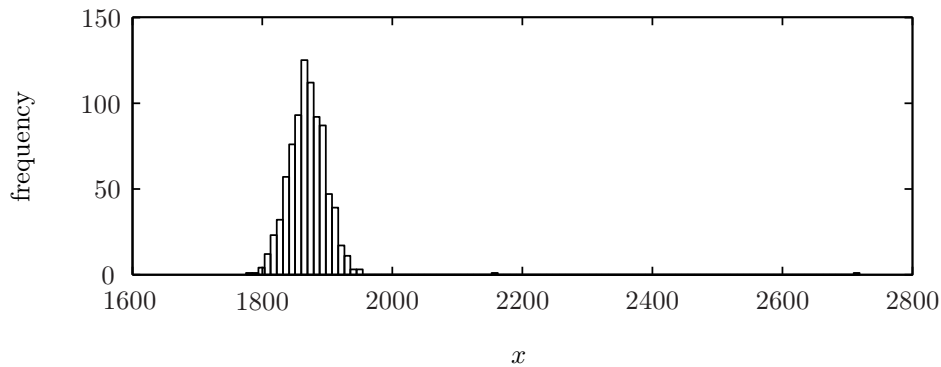
**Figure 7.1:** Raman Spectroscopy data with smoothed sequences  $U_1L_1x$ ,  $U_2L_2U_1L_1x$  and  $U_{10}L_{10} \dots U_1L_1x$ .

The lower panel displays a smaller subset of the data which contains outliers as well as a step. As can be seen, the advantages of smoothing with  $U_n L_n$  is that the outliers are removed while the steps and trend in the data are preserved.

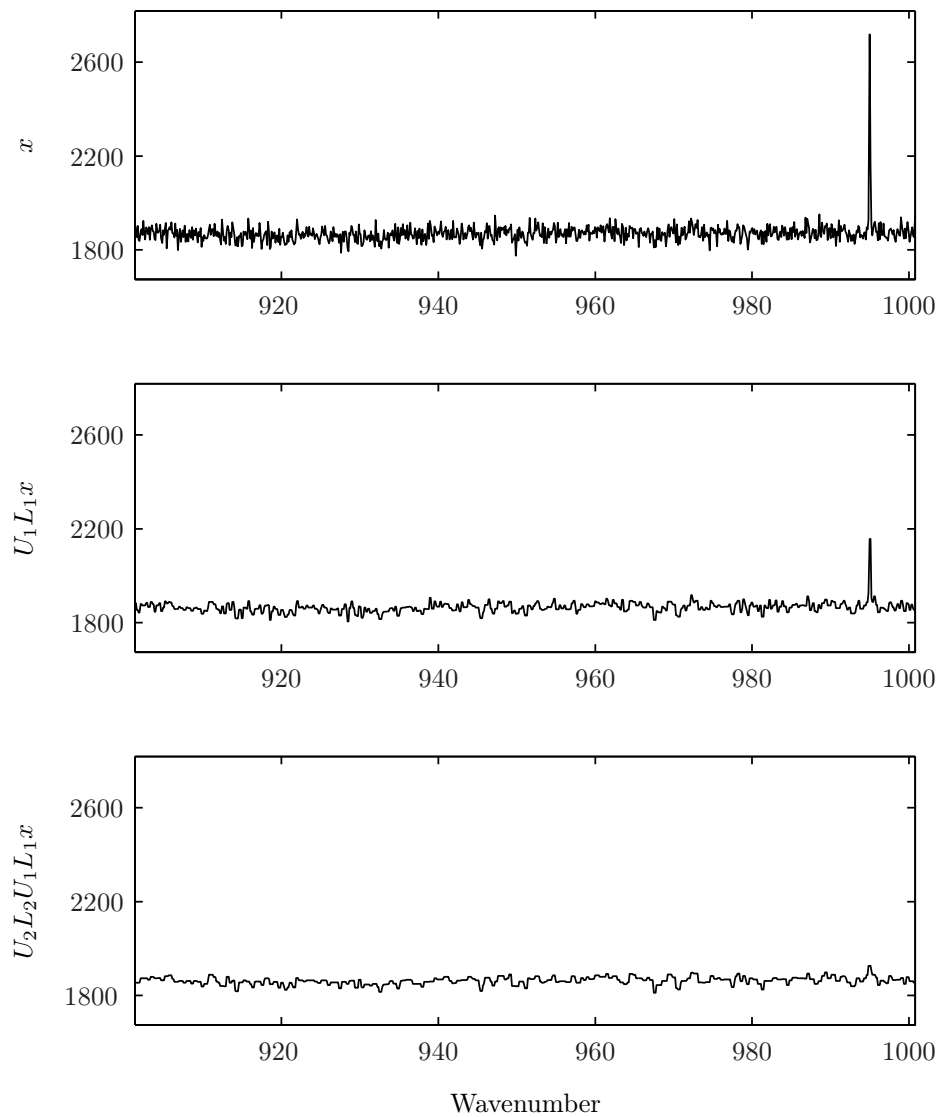
We opt to concentrate on the more or less constant section of the data which displays the obvious outliers in the measurements roughly between entries 900 and 1000 (see upper panel of Fig. 7.1 and Fig. 7.3). As can be seen after successive smoothing with  $U_n L_n$  up to level  $n = 2$  (see Fig 7.3), the outlier effect caused by the cosmic ray that strikes the detector results in two successive outlying measurements being made. **This implies that the data is no longer i.i.d.**, and although it can always be smoothed to any level desired, it is not to be analysed by the current theory due to previous assumptions. In order to proceed, we propose to *thin* the data by taking every second data point to create two sequences  $y$  and  $z$ , each containing one outlier (see Fig 7.4 where  $y$  and  $z$  as well as  $U_1 L_1 y$  and  $U_1 L_1 z$  are plotted). The original sequence  $x = x_{i=1}^n$  can thus be constructed as follows

$$x(i) = \begin{cases} y(i) & , \text{ if } i \text{ is odd} \\ z(i) & , \text{ if } i \text{ is even} . \end{cases} \quad (7.1)$$

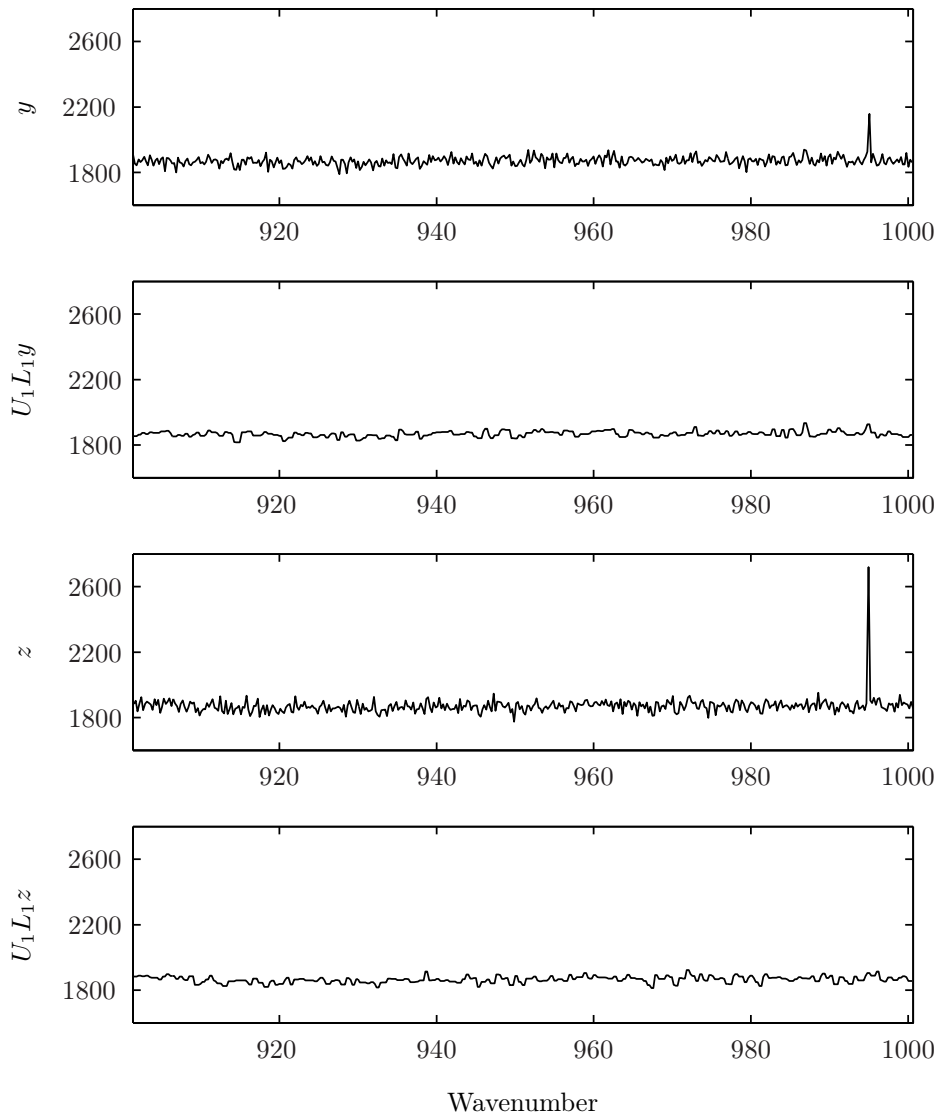
An estimate of a certain parameter can then be attained from an average of those estimates attained from each data set which is analysed separately. The thinned data is plotted below (see Fig 7.4). Considering the histogram of the data (Fig 7.2), the regularly occurring noise looks Gaussian. We would like to use a distribution that has finite bounds, and it could probably be well approximated by a 2nd order 4-spline. For simplicity, let us however choose a triangular distribution together with a uniform distribution (for the impulsive noise).



**Figure 7.2:** Histogram of one section of the RAMAN data. Notice the single entries near 2160 and 2710.



**Figure 7.3:** Section of Raman Spectroscopy data containing outliers. Original data represented by  $x$  with accompanying smoothed sequences  $U_1 L_1 x$  and  $U_2 L_2 U_1 L_1 x$ .



**Figure 7.4:** Thinned data with smoothed sequences  $U_1L_1y$  and  $U_1L_1z$ .

Choose the model distribution  $f = \alpha g + \beta h$  with  $\beta = 1 - \alpha$  and the triangular distribution

$$g(x) = \begin{cases} \frac{\mu+x}{\mu^2} & x \in [-\mu, 0] \\ \frac{\mu-x}{\mu^2} & x \in (0, \mu] \end{cases}, \quad h(x) = \begin{cases} \frac{1}{2\epsilon} & x \in [\mu, \mu + 2\epsilon] \\ 0 & \text{elsewhere} \end{cases}. \quad (7.2)$$

Their respective cdf's  $G(x) = \int_{-\infty}^x dt g(t)$  and  $H(x) = \int_{-\infty}^x dt h(t)$  are

$$G(x) = \begin{cases} 0 & x < -\mu \\ \frac{x^2}{2\mu^2} + \frac{x}{\mu} + \frac{1}{2} & x \in [-\mu, 0] \\ -\frac{x^2}{2\mu^2} + \frac{x}{\mu} + \frac{1}{2} & x \in (0, \mu] \\ 1 & x > \mu \end{cases}, \quad H(x) = \begin{cases} 0 & x < \mu \\ \frac{x-\mu}{2\epsilon} & x \in [\mu, \mu + 2\epsilon] \\ 1 & x > \mu + 2\epsilon \end{cases}. \quad (7.3)$$

Using the results of Chapter 3 where  $\langle x - L_1x \rangle$  and  $\langle L_1x - U_1L_1x \rangle$  were calculated for the general mixture model with non-overlapping support, we only have to set  $k = \mu$  and  $l = \mu + 2\epsilon$  and perform the integrals to find that

$$\langle x - L_1x \rangle = \underbrace{\left(-\frac{13}{20}\alpha^3 + \frac{23}{30}\alpha^2\right)}_{=\rho_{11}}\mu + \underbrace{\left(-\frac{1}{2}\alpha^3 + \frac{1}{6}\alpha^2 + \frac{1}{6}\alpha + \frac{1}{6}\right)}_{=\rho_{12}}\epsilon, \quad (7.4)$$

and

$$\begin{aligned} &\langle L_1x - U_1L_1x \rangle \\ &= \underbrace{\left(-\frac{527}{1008}\alpha^5 + \frac{1451}{630}\alpha^4 - \frac{39}{10}\alpha^3 + \frac{46}{15}\alpha^2 - \alpha\right)}_{=\eta_{22}}\mu + \underbrace{\left(-\frac{1}{3}\alpha^5 + \frac{19}{15}\alpha^4 - \frac{26}{15}\alpha^3 + \frac{14}{15}\alpha^2 - \frac{1}{15}\alpha - \frac{1}{15}\right)}_{=\eta_{21}}\epsilon. \end{aligned} \quad (7.5)$$

Assuming  $\alpha$  is known, point estimators for the scale parameters  $\mu$  and  $\epsilon$  can be constructed as was shown in Chapter 2. They are given by

$$\begin{bmatrix} (\mu)_{est} \\ (\epsilon)_{est} \end{bmatrix} = \frac{1}{\rho_{11}\eta_{22} - \rho_{12}\eta_{21}} \begin{bmatrix} \eta_{22} & -\rho_{12} \\ -\eta_{21} & \rho_{11} \end{bmatrix} \begin{bmatrix} \overline{x - L_1x} \\ \overline{L_1x - U_1L_1x} \end{bmatrix}. \quad (7.6)$$

Let the estimators of the parameter  $\mu$  and  $\epsilon$  for the sequences  $y$  and  $z$  be indicated by  $(\mu)_{est}^y$  and  $(\mu)_{est}^z$ , and  $(\epsilon)_{est}^y$  and  $(\epsilon)_{est}^z$  respectively. Then our estimates for  $\mu$  and  $\epsilon$  are

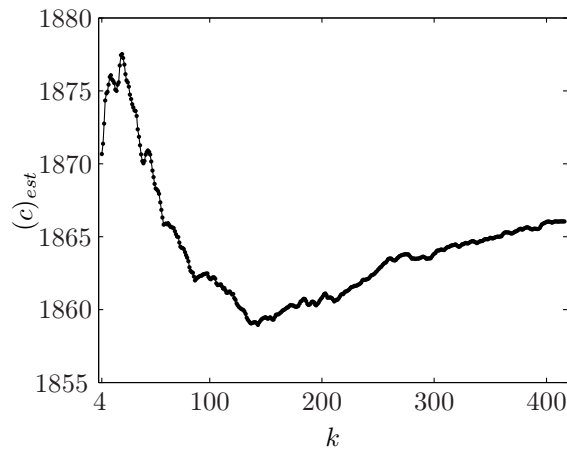
$$(\mu)_{est} = \frac{(\mu)_{est}^y + (\mu)_{est}^z}{2} \quad (7.7)$$

$$(\epsilon)_{est} = \frac{(\epsilon)_{est}^y + (\epsilon)_{est}^z}{2}. \quad (7.8)$$

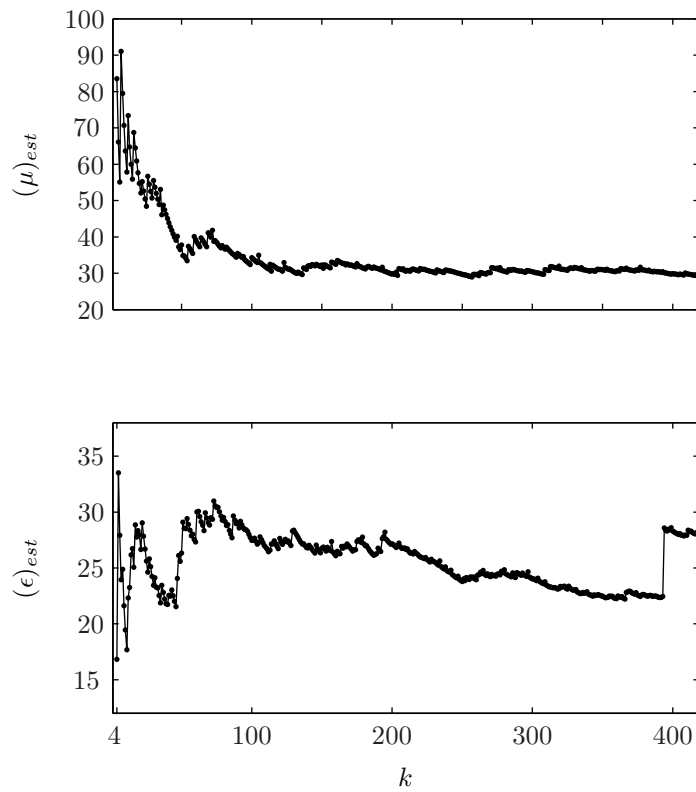
Our estimate for the location parameter is taken as

$$(c)_{est} = \frac{\overline{U_1L_1y} + \overline{U_1L_1z}}{2}. \quad (7.9)$$

The results for  $\alpha = 1/400$  are shown in Figs. 7.5 and 7.6 below.



**Figure 7.5:** Estimation of location parameter  $c$ .



**Figure 7.6:** Estimation of scale parameters  $\mu$  and  $\epsilon$ .

The final value of the average  $\overline{U_1 L_1 x}$  is found to be 1864.6 where the average of the data with the outliers forcefully removed is calculated at 1866. The closeness of these two estimates gives us confidence that our method for estimating the scale parameter works well. The estimator for  $\mu$  looks as though it gravitates toward a confident estimate and we would not expect it to change much if even more data was available. The estimator for  $\epsilon$  receives a ‘jolt’ as it benefits from the occurrence of the two outliers.



## Chapter 8

# Summary and Conclusions

The purpose of this thesis is to investigate the problem of outlier noise and in particular strongly asymmetric cases of such noise. In pursuit of a simple model which contains ‘regular’ noise distributed according to  $g(x)$  plus the occasional outlier distributed according to  $h(x)$ , we have constructed a combined model noise distribution  $f(x) = \alpha g(x) + \beta h(x)$ , with  $0 < \alpha = 1 - \beta < 1$ .

In Chapter 2, a theorem is presented that establishes a simple way of attaining expectation values for the positive and negative pulse heights of the first resolution level of a DPT from integrals involving the cumulative distribution function of  $f$ . This has opened up an avenue for analysing the expectation values of smoothed sequences, in particular  $L_1U_1$  and  $U_1L_1$ . We have shown that under certain conditions

$$\langle U_1L_1x \rangle = (\text{terms involving } g) + O(\beta^2).$$

The same holds for  $\langle L_1U_1x \rangle$ . When  $\beta \approx 0$ , the average  $\overline{U_1L_1x}$  is shown to be a robust estimator for the location parameter of  $g$  since any contribution from  $h$  is strongly suppressed. Furthermore, if  $h$  is generally not well-known due to the assumed lack of data, simulations have shown that there is a weak dependence on the actual shape of  $h$ . This has been reserved for thorough investigation in future research.

The theorem also provides a procedure for estimating the scale parameters of  $f$ . These results are confirmed by simulation and the LULU-estimators are shown to perform their task well.

Making use of the fact that  $L_1x \leq U_1L_1x \leq M_1x \leq L_1U_1x \leq U_1x$ , the accompanying theory of the LULU approach to the problem of nonlinear smoothing as applied to the problem of outliers provides a deeper understanding why the popular median smoother  $M_1$  is so successful, while a full theory of this success is still lacking.

In the second part of this thesis, a Bayesian perspective is adopted and is shown to provide insight into the logic of the problem. Certain practical case studies are examined where assumptions on the outcome spaces of  $g$  and  $h$  and the prior bounds on the parameters lead to a truncation of the  $2^n$  number of terms in the binomial expansion of the likelihood function. A way of organising the outcome spaces of the parameters by using window functions proved practical and enables as far as possible exact analytical expressions for the moments and posterior distributions of the parameters. A particular example is worked out in detail.

While the information provided by an analytical Bayesian solution proves to be exhaustive, the Bayesian estimators are difficult to obtain analytically and slow to compute by Monte Carlo in-

tegration.

A qualitative comparison shows that the LULU-estimators are far easier to calculate and perform admirably when compared to Bayesian estimators. They are shown to be reasonable and practical compromises in practical problems.

# Appendices

## Appendix A

### Window functions

In Chapters 4 and 5, various step functions are used frequently. They help simplify expressions, and keep track of the correct boundaries of integration of all the various parameters involved. Introduced here are various window functions which help with economy of writing, and for quick interpretation of results, as combinations of step functions can become long and cumbersome to read.

The familiar Heaviside step function, or theta function, is defined as

$$\theta(a - z) = \begin{cases} 1 & z \leq a \\ 0 & \text{otherwise.} \end{cases} \quad (\text{A.1})$$

Now define the window function:

$$U[z | a, b] = \theta(z - a)\theta(b - z) = \begin{cases} 1 & z \in [a, b] \\ 0 & \text{otherwise,} \end{cases} \quad (\text{A.2})$$

$$U(z | a, b) = \begin{cases} 1 & z \in (a, b) \\ 0 & \text{otherwise,} \end{cases} \quad (\text{A.3})$$

$$U[z | a, b) = \begin{cases} 1 & z \in [a, b) \\ 0 & \text{otherwise.} \end{cases} \quad (\text{A.4})$$

Notice that  $U[z | a, b]$  implies  $z \in [a, b]$ , and thus

$$\int dz U[z | a, b] = \int_a^b dz. \quad (\text{A.5})$$

From the definition of the theta function we have that:

$$\theta(a_1 - z)\theta(a_2 - z) \dots \theta(a_n - z) = \theta(\min(a_1, a_2, \dots, a_n) - z), \quad (\text{A.6})$$

$$\theta(z - a_1)\theta(z - a_2) \dots \theta(z - a_n) = \theta(z - \max(a_1, a_2, \dots, a_n)). \quad (\text{A.7})$$

The following result will also be useful:

$$\prod_{i=1}^n U[z_i | a, b] = \theta(\min(z_1, z_2, \dots, z_n) - a)\theta(b - \max(z_1, z_2, \dots, z_n)). \quad (\text{A.8})$$

To prove this write out the left hand side using definition (A.2), rearrange appropriately to apply results (A.6) and (A.7).

## Appendix B

# Preparing the window functions for integration

Similarly to what was done for the uniform distribution,  $U_1$ ,  $U_2$  and  $U_3$  (from Eq. (5.70) in section 5.3.1) have to be prepared for integrations over  $X$ ,  $L_1$  and  $L_2$ . In doing so we are choosing the correct boundaries over which to integrate.

Using result (A.8),  $U_1$ ,  $U_2$  and  $U_3$  become

$$U_1 = \theta(x_1 - X + L_1)\theta(X - x_n)U[X, L_1, L_2], \quad (\text{B.1})$$

$$U_2 = \theta(x_1 - X + L_1)\theta(X - x_{n-k})\theta(x_{n-k+1} - X)\theta(X + L_2 - x_n)U[X, L_1, L_2], \quad (\text{B.2})$$

$$U_3 = \theta(x_1 - X)\theta(X + L_2 - x_n)U[X, L_1, L_2], \quad (\text{B.3})$$

where from (5.67)

$$U[X, L_1, L_2] = U[X | K_1, K_2]U[L_1 | M_1, M_2]U[L_2 | Q_1, Q_2].$$

Concentrating on  $U_1$  for the moment, use (7.60) and the definition of the window function (A.2) to write  $U_1$  in terms of theta functions:

$$U_1 = \theta(x_1 - X + L_1)\theta(X - x_n)\theta(X - K_1)\theta(K_2 - X)\theta(L_1 - M_1)\theta(M_2 - L_1)\theta(L_2 - Q_1)\theta(Q_2 - L_2). \quad (\text{B.4})$$

In order to calculate  $B$ ,  $C$  and  $D$ , we want to prepare this term for an integration over  $X$ , but from the experience gained on the work done for the uniform distribution, we know we have to be careful with how we restrict  $L_1$ . From  $\theta(x_1 - X + L_1)$  we have that  $L_1 \geq X - x_1$ . But since from  $\theta(X - K_1)$  and  $\theta(X - x_n)$  we have the inequality  $X \geq \max(K_1, x_n)$ , we have to insist that  $L_1 \geq \max(K_1, x_n) - x_1$  while integration over  $X$ . Inserting  $\theta(L_1 - (\max(K_1, x_n) - x_1))$  we get

$$U_1 = \theta(x_1 - X + L_1)\theta(X - x_n)\theta(X - K_1)\theta(K_2 - X)\theta(L_1 - M_1)\theta(L_1 - (\max(K_1, x_n) - x_1)) \\ \times \theta(M_2 - L_1)\theta(L_2 - Q_1)\theta(Q_2 - L_2), \quad (\text{B.5})$$

and after rearranging in order to gather  $X$  as the subject of a window function the above becomes

$$U_1 = U[X | \max(K_1, x_n), \min(K_2, x_1 + L_1)]U[L_1 | \max(M_1, \max(K_1, x_n) - x_1), M_2]U[L_2 | Q_1, Q_2]. \quad (\text{B.6})$$

If  $A$  is to be calculated, prepare (7.66) for an integration over  $L_1$ . We instead reason as follows. From  $\theta(x_1 - X + L_1)$  we have that  $X \leq x_1 + L_1$ , but since  $M_2 \geq L_1$  we must insist that  $X \leq x_1 + M_2$

whilst integrating over  $L_1$ . Thus, after inserting  $\theta(M_2 + x_1 - X)$  and gathering  $L_1$  as the subject of a window function,  $U_1$  becomes

$$U_1 = U[X \mid \max(K_1, x_n), \min(K_2, x_1 + M_2)]U[L_1 \mid \max(M_1, X - x_1), M_2]U[L_2 \mid Q_1, Q_2]. \quad (\text{B.7})$$

Writing  $U_2$  with integration over  $X$  in mind is slightly trickier. Write  $U_2$  in terms of  $\theta$  functions:

$$U_2 = \theta(x_1 - X + L_1)\theta(X - x_{n-k})\theta(x_{n-k+1} - X)\theta(X + L_2 - x_n)\theta(X - K_1)\theta(K_2 - X) \\ \times \theta(L_1 - M_1)\theta(M_2 - L_1)\theta(Q_2 - L_2)\theta(L_2 - Q_1). \quad (\text{B.8})$$

From  $\theta(x_1 - X + L_1)$  we have the inequality  $L_1 \geq X - x_1$ , but from  $\theta(X - x_{n-k})$  and  $\theta(X - K_1)$  we have that  $X \geq \max(x_{n-k}, K_1)$ . Thus while integrating out  $X$  we must insist that  $L_1 \geq \max(x_{n-k}, K_1) - x_1$  and have to insert  $\theta(L_1 - (\max(x_{n-k}, K_1) - x_1))$  into (7.70) above. From  $\theta(X + L_2 - x_n)$  we have the inequality  $L_2 \geq x_n - X$ . From  $\theta(K_2 - X)$  and  $\theta(x_{n-k+1} - X)$  we have the inequality  $L_2 \geq x_n - K_2$  and  $L_2 > x_n - x_{n-k+1}$ . We again absorb this restriction into a single inequality  $L_2 \geq x_n - \min(K_2, x_{n-k+1})$ . Inserting  $\theta(L_1 - (\max(x_{n-k}, K_1) - x_1))$  and  $\theta(L_2 - (x_n - \min(K_2, x_{n-k+1})))$  into  $U_2$ , and rearranging to gather  $X$  as the subject of a window function we get:

$$U_2 = U[X \mid \max(K_1, x_{n-k}, x_n - L_2), \min(K_2, x_{n-k+1}, x_1 + L_1)] \\ \times U[L_1 \mid \max(M_1, \max(K_1, x_{n-k}) - x_1), M_2]U[L_2 \mid \max(Q_1, x_n - \min(K_2, x_{n-k+1})), Q_2]. \quad (\text{B.9})$$

For integration over  $L_1$  or  $L_2$ , from  $\theta(x_1 + L_1 - X)$  we have  $x_1 + L_1 \geq X$ , but since  $M_2 \geq L_1$  we have that  $x_1 + M_2 \geq X$  and insert  $\theta(x_1 + M_2 - X)$ . From  $\theta(X + L_2 - x_n)$  we have  $X \geq x_n - L_2$ , but since  $Q_2 \geq L_2$  we have that  $X \geq x_n - Q_2$  and insert  $\theta(X - (x_n - Q_2))$ . The result is

$$U_2 = U[X \mid \max(K_1, x_{n-k}, x_n - Q_2), \min(K_2, x_{n-k+1}, x_1 + M_2)] \\ \times U[L_1 \mid \max(M_1, X - x_1), M_2]U[L_2 \mid \max(Q_1, x_n - X), Q_2]. \quad (\text{B.10})$$

In terms of  $\theta$  functions,  $U_3$  reads

$$U_3 = \theta(X + L_2 - x_n)\theta(X - K_1)\theta(K_2 - X)\theta(x_1 - X)\theta(L_2 - Q_1)\theta(Q_2 - L_2)U[L_1 \mid M_1, M_2]. \quad (\text{B.11})$$

From  $\theta(X + L_2 - x_n)$  we have that  $L_2 \geq x_n - X$ , but since  $\min(x_1, K_2) \geq X$ , we have that  $L_2 \geq x_n - \min(x_1, K_2)$ . Inserting  $\theta(L_2 - (x_n - \min(x_1, K_2)))$  takes care of this restriction:

$$U_3 = U[X \mid \max(K_1, x_n - L_2), \min(K_2, x_1)]U[L_1 \mid M_1, M_2]U[L_2 \mid \max(Q_1, x_n - \min(x_1, K_2)), Q_2]. \quad (\text{B.12})$$

For integration over  $L_1$  or  $L_2$ , from  $\theta(X + L_2 - x_n)$  we have that  $X \geq x_n - L_2$ , but since  $X \geq Q_2$ , we have that  $X \geq x_n - Q_2$ . Inserting  $\theta(X - (x_n - Q_2))$  takes care of this restriction:

$$U_3 = U[X \mid \max(K_1, x_n - Q_2), \min(K_2, x_1)]U[L_1 \mid M_1, M_2]U[L_2 \mid \max(Q_1, x_n - X), Q_2]. \quad (\text{B.13})$$

To summarize, if we want to calculate  $B$ ,  $C$  or  $D$ , the terms  $U_1$ ,  $U_2$  and  $U_3$  are chosen as shown in (B.6), (B.9) and (B.12) respectively:

$$U_1 = U[X \mid \max(K_1, x_n), \min(K_2, x_1 + L_1)]U[L_1 \mid \max(M_1, \max(K_1, x_n) - x_1), M_2]U[L_2 \mid Q_1, Q_2] \\ U_2 = U[X \mid \max(K_1, x_{n-k}, x_n - L_2), \min(K_2, x_{n-k+1}, x_1 + L_1)] \\ \times U[L_1 \mid \max(M_1, \max(K_1, x_{n-k}) - x_1), M_2]U[L_2 \mid \max(Q_1, x_n - \min(K_2, x_{n-k+1})), Q_2] \\ U_3 = U[X \mid \max(K_1, x_n - L_2), \min(K_2, x_1)]U[L_1 \mid M_1, M_2]U[L_2 \mid \max(Q_1, x_n - \min(x_1, K_2)), Q_2]. \quad (\text{B.14})$$

If we want to calculate  $A$ , the terms  $U_1$ ,  $U_2$  and  $U_3$  are chosen as shown in (B.7), (B.10) and (B.13) respectively:

$$\begin{aligned}
 U_1 &= U[X \mid \max(K_1, x_n), \min(K_2, x_1 + M_2)]U[L_1 \mid \max(M_1, X - x_1), M_2]U[L_2 \mid Q_1, Q_2] \\
 U_2 &= U[X \mid \max(K_1, x_{n-k}, x_n - Q_2), \min(K_2, x_{n-k+1}, x_1 + M_2)] \\
 &\quad \times U[L_1 \mid \max(M_1, X - x_1), M_2]U[L_2 \mid \max(Q_1, x_n - X), Q_2]. \\
 U_3 &= U[X \mid \max(K_1, x_n - Q_2), \min(K_2, x_1)]U[L_1 \mid M_1, M_2]U[L_2 \mid \max(Q_1, x_n - X), Q_2].
 \end{aligned}
 \tag{B.15}$$

## List of References

- [1] Anguelov, R. and Fabris-Rotelli, I. (2010). LULU operators and discrete pulse transform for multi-dimensional arrays. *IEEE Transactions on Image Processing*, vol. 19:11, pp. 3012–3023.
- [2] Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, vol. 53, pp. 370–418.
- [3] Bernoulli, J. (1713). *Ars Conjectandi*. Thurnisiorum, Basel.
- [4] Conradie, W., de Wet, T. and Jankowitz, M. (2006). Exact and asymptotic distributions of LULU smoothers. *Journal of Computational and Applied Mathematics*, vol. 186, pp. 253–267.
- [5] Cox, R. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, vol. 14:1.
- [6] de Finetti, B. (1937). La prevision: ses lois logiques, ses sources subjectives. *Ann. Inst. H.*, vol. 7, pp. 1–68.
- [7] Fabris-Rotelli, I. (2012). *Applications of the Discrete Pulse Transform in Image Analysis*. Ph.D. thesis, Dept. Mathematics and Applied Mathematics, University of Pretoria, Pretoria, South Africa.
- [8] Fabris-Rotelli, I. and van der Walt, S. (2009). The discrete pulse transform in two dimensions. In: Nicolls, F. (ed.), *Proceedings of the Twentieth Annual Symposium of the Pattern Recognition Association of South Africa*. Stellenbosch. ISBN 978-0-7992-2356-9.
- [9] Feller, W. (1950). *An Introduction to Probability theory and its Applications*. Wiley and Sons, New York.
- [10] Feller, W. (1966). *An Introduction to Probability theory and its Applications, Volume 2*. Wiley and Sons, New York.
- [11] Fienberg, S. (1989). Idempotent one-sided approximation of median smoothers. *Journal of Approximation Theory*, vol. 58, pp. 151–163.
- [12] Jankowitz, M. (2007). *Some Statistical Aspects of LULU smoothers*. Ph.D. thesis, Dept. Statistics and Actuarial Science, University of Stellenbosch, Stellenbosch, South Africa.
- [13] Jaynes, E. (2003). *Probability Theory, The Logic of Science*. Cambridge University Press.
- [14] Jeffreys, H. (1939). *Theory of Probability*. Clarendon Press, Oxford.
- [15] Kendall, M. and Stewart, A. (1977). *The Advanced Theory of Statistics, Vol 1*. Hafner Publishing Co., New York.
- [16] Laplace, P. (1812). *Theorie analytique des probabilites*. Ve. Courcier, Paris.
- [17] Laurie, D. (2011). The roadmaker’s algorithm for the discrete pulse transform. *IEEE Transactions on Image Processing*, vol. 20:2, pp. 361–371.



- [18] Mallows, C. (1980). Some theory of nonlinear smoothers. *The Annals of Statistics*, vol. 8:4, pp. 695–715.
- [19] Rohwer, C. (1989). Idempotent one-sided approximation of median smoothers. *Journal of Approximation Theory*, vol. 58, pp. 151–163.
- [20] Rohwer, C. (1999). Projections and separators. *Quaestiones Mathematicae*, vol. 22, pp. 219–230.
- [21] Rohwer, C. (2002). Multiresolution analysis with pulses. In: M.D., B. and D.A, M. (eds.), *Advanced Problems in Constructive Approximation*, International Series of Numerical Mathematics, 142nd edn, pp. 165–186. Birkhäuser-Verlag, Basel.
- [22] Rohwer, C. (2002). Variation reduction and LULU-smoothing. *Quaestiones Mathematicae*, vol. 25:2, pp. 163–176.
- [23] Rohwer, C. (2004). Fully trend preserving operators. *Quaestiones Mathematicae*, vol. 27:3, pp. 217–229.
- [24] Rohwer, C. (2005). *Nonlinear Smoothing and Multiresolutional Analysis*. Birkhäuser.
- [25] Rohwer, C. (2007). The estimation of moments of an unknown error distribution in the discrete pulse transform. *Numerical Algorithms*, vol. 45, pp. 239–251.
- [26] Rohwer, C. and Laurie, D. (2006). The discrete pulse transform. *SIAM Journal on Mathematical Analysis*, vol. 38:3, pp. 1012–1034.
- [27] Rohwer, C. and Toerien, L. (1991). Locally monotone robust approximation of sequences. *Journal of Computational and Applied Mathematics*, vol. 36, pp. 399–408.
- [28] Rohwer, C. and Wild, M. (2002). Natural alternatives for one dimensional median filtering. *Quaestiones Mathematicae*, vol. 25:2, pp. 135–162.
- [29] Rossman, A., Short, T. and Parks, M. (1998). Bayes estimators for the continuous uniform distribution. *Journal of Statistics Education*, vol. 6:3.
- [30] Sivia, D. (2006). *Data Analysis, A Bayesian Tutorial*. Oxford University Press.
- [31] Velleman, P. (1977). Robust nonlinear data smoothers: Definitions and recommendations. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74:2, pp. 434–436.
- [32] Velleman, P. (1980). Definition and comparison of robust nonlinear data smoothing algorithms. *Journal of the American Statistical Association*, vol. 75:371, pp. 609–615.