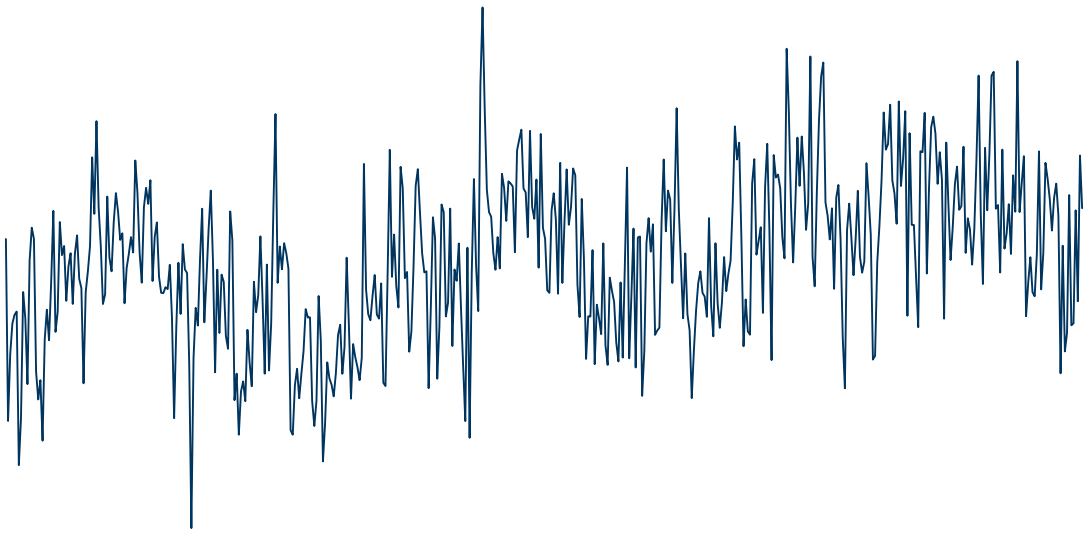
RUHR
UNIVERSITÄT
BOCHUM

RUB

Change-Point Tests For Long-Range Dependent Data

Strukturbruchtests für stark abhängige Daten

Aeneas Rooch



DISSERTATION

2012

Change-Point Tests For Long-Range Dependent Data

Strukturbruchtests für stark abhängige Daten

Aeneas Rooch

DISSERTATION

zur Erlangung des Doktorgrades der Naturwissenschaften
an der Fakultät für Mathematik der Ruhr-Universität Bochum

2012

Preface

This thesis is is entitled "Change-Point Tests For Long-Range Dependent Data" and deals with mathematical methods to analyse data.

An important issue in data analysis (such as the analysis of stock exchange prices or temperature measurements) is the question whether the observed process changes fundamentally at a certain time, for example whether after a certain point in time, the prices or the temperatures tend to be higher than before. But even if one observes an apparent change-point in the data – since all measurements are subject to random fluctuations, it is difficult to say if there really is a change in the structure or if it is just a random effect. *Change-Point tests* are mathematical methods which are designed to discriminate between structural changes and random effects and thus to detect change-points in random observations. In this work, I propose and analyse a class of such test methods for a specific and relevant case of data.

In many applications, observations cannot assumed to be independent (like the numbers when throwing a dice); for a realistic model one has to assume instead that each event influences the following ones (like the weather at one and at the following days). Although this dependence naturally declines by and by, in some fields (for example in econometrics, climate research, hydrology and information technology) processes occur where this decay is extremely slow: Even events from the distant past influence the present behaviour of the process. This is hard to imagine and to illustrate, but is has turned out that many processes like internet traffic and temperature measurements can be modelled by random time series with this *long-range dependence* or *long memory*.

On the titelpage, a sample of 500 long-range dependent data is shown (fractional Gaussian noise with Hurst parameter $H = 0.8$) which displays a typical characteristic of long-range dependent processes: with its periods of mostly large and its periods of mostly small observations, it appears to exhibit a periodic pattern, but there is none.

Unfortunately, usual methods for data analysis fail when the data is long-range dependent, and new techniques are needed. In this work, I propose and analyse some new mathematical methods to detect changes in observations with long memory.

I would like to express my gratitude to my supervisor Herold Dehling, who introduced me to these fascinating problems and whose expertise and understanding added considerably to my graduate experience.

Over the years, I have enjoyed the support of several collegues who discussed problems and shared insights with me, particularly I like to thank Matthias Kalus and Annett Püttmann (who provided useful tips and checked the results from my excursion through the jungle of complex analysis), Peter Otte (who made comments and suggestions on some challenging issues in functional analysis), Daniel Vogel (who helped me with his statistical experience to interpret some simulation results) and Martin Wendler (who had a fine idea for the proof of the 'sort and replace' method and for the direct approach to two-sample-U-statistics).

Certainly, a vast number of my extensive simulations would still be running without the kind support of Philipp Junker who easily and reliably provided computing capacity.

Finally, I owe deep gratitude to Jan Nagel for philosophical and entertaining debates and exchanges of knowledge, which helped enrich my PhD time, and after all for his careful proofreading of this thesis.

Contents

List of Symbols **ix**

1 What it is about **1**
1.1 Detecting change-points 4
1.2 Long-range dependence 5
1.3 Fractional Brownian motion and fractional Gaussian noise 9
1.4 Hermite polynomials 15
1.4.1 Definition and basic properties 15
1.4.2 Relation to LRD 16

2 The asymptotic behaviour of $\bar{X} - \bar{Y}$ **23**
2.1 Asymptotic equivalence and slowly varying functions 24
2.2 One divided sample 28
2.2.1 Asymptotic theory 28
2.2.2 Simulations 29
2.3 Two independent samples 30
2.3.1 Asymptotic theory 30
2.3.2 Simulations 32
2.4 Estimating the variance of $\bar{X}$ 33
2.4.1 Asymptotic behaviour of the variance-of-mean estimator 33
2.4.2 Simulations 36
2.5 Estimating the auto-covariance 39
2.5.1 Asymptotic behaviour of the auto-covariance estimator 41
2.5.2 Simulations 42
2.6 Estimating the variance of $\bar{X} - \bar{Y}$ 43
2.6.1 Asymptotic behaviour of the variance estimator for $\bar{X} - \bar{Y}$ . . . 43
2.6.2 Simulations 44

3 A "Wilcoxon-type" change-point test **49**
3.1 Setting the scene 50
3.2 Limit distribution under null hypothesis 51
3.3 Limit distribution in special situations 57

3.3.1 The Wilcoxon two-sample test . . . 57
3.3.2 Two independent samples . . . 57
3.4 Application . . . 59
3.4.1 "Wilcoxon-type" test . . . 59
3.4.2 "Difference-of-means" test . . . 61
3.5 "Difference-of-means" test under fGn . . . 63
3.6 Simulations . . . 66
3.6.1 Normally distributed data . . . 66
3.6.2 Symmetric, normal-tailed data . . . 71
3.6.3 Heavy-tailed data . . . 73

4 Power of some change-point tests 79
4.1 "Differences-of-means" test . . . 80
4.2 "Wilcoxon-type" test . . . 82
4.3 Asymptotic Relative Efficiency . . . 88
4.4 ARE for i.i.d. data . . . 92
4.5 ARE of Wilcoxon and Gauß test for the two-sample problem . . . 99
4.6 Simulations . . . 101
4.6.1 Power of change-point tests . . . 101
4.6.2 Power of two-sample tests, setting . . . 102
4.6.3 Power of two-sample tests, Gaussian observations . . . 105
4.6.4 Power of two-sample tests, Pareto(3,1) observations . . . 105

5 Change-point processes based on U-statistics 109
5.1 Special kernels . . . 110
5.2 General kernels . . . 115
5.3 Examples . . . 120
5.4 Simulations . . . 123

6 Change-point processes based on U-statistics, a direct approach 125
6.1 The limit distribution under the null hypothesis . . . 126
6.2 The limit distribution in special situations . . . 133
6.2.1 Hermite rank $m = 1$ . . . 133
6.2.2 Two independent samples . . . 134
6.3 Examples . . . 136
6.3.1 "Differences-of-means" test . . . 136
6.3.2 "Wilcoxon-type" test . . . 138
6.4 In search of handy criteria . . . 142

7 Solutions for estimation problems 157
7.1 The influence of an estimated Hurst parameter . . . 158
7.1.1 Methods of estimating in comparison . . . 158

7.1.2 Change-point tests with estimated Hurst parameter 160
7.1.3 Summary 162
7.2 Estimating the LRD parameter under a change in the mean 167
7.2.1 Adaption techniques 168
7.2.2 Simulations 171
7.2.3 Conclusion and outlook 177
7.3 Estimating the first Hermite coefficient 178
7.3.1 The sort and replace method 179
7.3.2 Simulations 182

A A short introduction into stochastic integration 185
A.1 Wiener integral 185
A.2 Itō integral 186
A.3 Itō process and Itō formula 190
A.4 Multiple Wiener-Itō integrals 192
A.4.1 Hermite polynomials and multiple Wiener-Itō integrals 195

B Additions 197
B.1 Proof of Theorem 2.3 197
B.2 An example illustrating Theorem 2.3 202
B.3 Bounded variation in higher dimensions 203
B.3.1 Definition, properties and examples 204
B.3.2 The case $h(x, y) = I_{\{x \leq y\}}$ 210

C Source code of the simulations 213
C.1 Estimating the variance of $\bar{X}$ (section 2.4) 213
C.2 Estimating the auto-covariance (section 2.5) 214
C.3 Estimating the variance of $\bar{X} - \bar{Y}$ (section 2.6) 214
C.4 "Differences-of-means" test (section 3.6) 215
C.5 $\bar{X} - \bar{Y}$ for one divided sample (section 2.2.2) 221
C.6 $\bar{X} - \bar{Y}$ for two independent samples (section 2.3.2) 221
C.7 Quantiles of $\sup |Z(\lambda) - \lambda Z(1)|$ (section 3.4) 223
C.8 "Wilcoxon-type" test (section 3.6) 224
C.9 Estimating the LRD parameter under a jump (section 7.2) 227
C.10 Estimating the first Hermite coefficient (section 7.3) 232

D Exact simulation results 235
D.1 $\bar{X} - \bar{Y}$ test (section 2.2 and 2.3) 235
D.1.1 One divided sample 235
D.1.2 Two independent samples 235
D.2 Estimation of the variance of $\bar{X}$ (section 2.4) 239
D.3 Change-point test comparison in (section 3.6) 242

D.3.1 Normally distributed data 242
D.3.2 Symmetric, normal-tailed data 242
D.3.3 Heavy-tailed data 244
D.4 Change-point tests with estimated Hurst parameter (section 7.1) 248
D.4.1 Normally distributed data 248
D.4.2 Heavy-tailed data 248
D.5 Estimating Hermite coefficients (section 7.3) 249
D.6 Wilcoxon's two-sample statistic (section 5.4) 251

List of Figures **253**

List of Tables **255**

Bibliography **257**

List of Symbols

α	level of a test (mostly, also: scalar, multi-index)
$\bar{X}$	mean of all observations $X_1, \ldots, X_n$
$\bar{X}_k^l$	mean of observations $X_k, \ldots, X_l$
$BB(\lambda)$	Brownian bridge
β	power of a test (mostly, also: scalar, multi-index)
$\delta_\tau(\lambda)$	function, characterises interplay between true change-point position τ and assumed position λ, page 80
$\Delta_R f$	d-increment of function f over rectangle R, page 205
η, η_i	Gaussian random variable
γ_k	auto-covariance, page 5
$\hat{h}$	Fourier transform of function h, also denoted as $\mathcal{F}(h)$
$\hat{f}(t_1, \ldots, t_d)$	symmetrization of f, page 194
$\mathcal{F}(h)$	Fourier transform of function h, also denoted as $\hat{h}$
$\mathcal{B}$	Borel σ-algebra
$\mathcal{F}$	sigma field
$\mathcal{F}^2(\mathbb{C}^d)$	Bargmann-Fock space, page 145
$\mathcal{F}_B$	sigma field, page 195
$\mathcal{G}^p(\mathbb{R}, \mathcal{N})$	subset of $L^p(\mathbb{R}, \mathcal{N})$, page 50
$\mathcal{G}^p(\mathbb{R}^2, \mathcal{N})$	subset of $L^p(\mathbb{R}^2, \mathcal{N})$, page 128
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
μ_i	mean of random variable X_i, page 50
$\otimes$	tensor product, page 195
$\psi(t)$	probability, needed when calculating ARE, page 91
$\psi^-(\beta)$	generalized inverse of ψ, page 91
$\sim$	asymptotically equivalent, page 24
$\sum'$	sum with special domain of summation, used in diagram formula, page 39
$\tilde{h}(x)$	auxiliary function, page 115
$\tilde{h}_k(x)$	k-th Hermite function, page 16
$\|f\|_{HK,[a,b]}$	Hardy-Krause variation of function f on (hyper-)rectangle $[a, b]$, page 208
$\|f\|_{V,[a,b]}$	total/Vitali variation of function f on (hyper-)rectangle $[a, b]$, page 206
$\widehat{\gamma}_h$	standard estimator for auto-covariances, page 41
$\widehat{\mathrm{Var}}\left[X^{(r)}\right]$	estimator for variance of $\bar{X}$, page 33

ξ, ξ_i	Gaussian random variable
A	alternative (in a change-point test, 'change-point'), page 50
$a(x), b(x)$	functions, components of special kernels $h(x, y)$, page 110
A_k	alternative (in a change-point test, 'change-point after k-th observation'), page 50
a_k	k-th Hermite coefficient (of a one-dimensional function), page 18
$A_{\tau,h_n}(n)$	local alternative (in a change-point test, 'change-point of height h_n after proportion of τ'), page 79
a_{kl}	(k, l)-th Hermite coefficient (of a two-dimensional function), page 127
$B_H(t)$	fractional Brownian motion, page 11
B_t	standard Brownian motion, page 11
$Bf(z)$	Bargman transform of function f, page 145
BV_{HK}	class of functions with bounded Hardy-Krause variation, page 208
BV_V	class of functions with bounded Vitali variation, page 208
c	constant
$C^k(S)$	space of k times countinously differentiable functions on space S
c_m	usual constant in limit theorems for LRD partial sums, page 19
d'_n	usual scaling for LRD partial sums, without constant c_m, page 128
D_n	"difference-of-means" test statistic, page 62
d_n	usual scaling for LRD partial sums, page 19
$D_{k,n}$	"difference-of-means" two-sample test statistic, page 62
$f(\lambda)$	spectral density, page 5
$F(x)$	c.d.f. of random variables X_i
$f(x)$	p.d.f. of random variables X_i
$F_{k+1,l}(x)$	e.d.f., based on observations $X_{k+1}, \ldots, X_l$, page 53
$F_k(x)$	e.d.f., based on observations $X_1, \ldots, X_k$, page 53
G	transformation, mostly a function in $\mathcal{G}^1$ or $\mathcal{G}^2$, page 50
H	null hypothesis (in a change-point test, 'no change-point'), page 50
h	kernel function
h, h_n	height of change-point, page 80
h_D, h_W	jump heights in the context of ARE, page 89
$h_k(x)$	k-th normalized Hermite function, page 16
$H_k^{(phy)}(x)$	k-th Hermite polynomial, "physicists' definition", page 16
$H_k(x)$	k-th Hermite polynomial, page 15
I_A	indicator function of set A
$J(x)$	short for $J_m(x)$
$J_q(x)$	q-th Hermite coefficient of class of functions $I_{\{G(\xi_i)\leq x\}} - F(x)$, page 52
L	slowly varying function, page 24
L^2_{ad}	class of functions, needed for Itō integrals, page 187
m	Hermite rank (of a two-dimensional function), page 127
m	Hermite rank, page 52

n_D, n_W	sample sizes in the context of ARE, page 89
O	Landau/big O notation
o	Landau/little-o notation
q_α	upper α-quantile (often: of $\sup_{0\le\lambda\le1} \lvert Z_1(\lambda) - \lambda\, Z_1(1)\rvert$), page 61
$U_{\mathrm{diff},\lambda,n}$	"difference-of-means" statistic, in the context of U-statistics, page 136
$U_{k,n}$	two-sample U-statistic, page 109
$U_{W,\lambda,n}$	Wilcoxon two-sample test statistic, in the context of U-statistics, page 138
$W_n(\lambda)$	"Wilcoxon-type" process, page 51
W'_n	"Wilcoxon-type" process for two independent LRD processes, page 57
$W_{k,n}$	Wilcoxon two-sample test statistic, page 51
$X_k^{(r)}$	mean of k-th block of length r, page 33
X_i	random variable, mostly $X_i = \mu_i + G(\xi_i)$, page 50
$Z(\lambda)$	short for $Z_m(\lambda)/m!$
$Z_m(t)$	m-th order Hermite process, page 19
ARE	asymptotic relative efficiency, page 89
c.d.f.	cumulative distribution function
D	long-range dependence parameter, page 5
e.d.f.	empirical distribution function
fBm	fractional Brownian motion, page 11
fGn	fractional Gaussian noise, page 12
H	Hurst exponent, page 5
LRD	long-range dependent, page 5
p.d.f.	probability density function
SRD	short-range/weak dependence, page 5

Chapter 1

What it is about

The goal in *inferential statistics* is to draw inferences about underlying regularities in an observed sample of measured data and to model the process which generated the data, accounting for randomness, in other words to filter out the main characteristics, the valuable information, in a vast amount of random observations – in order to characterize, to analyze or to forcast the process. An important issue of interest in all fields of inferential statistics is the detection of *change-points*, of unknown time instants at which the underlying regularities change, because any statistical inference naturally relies on large and homogenoeus samples of observations; a sudden change in the characteristics of the data may disturb and adulterate the inference and may lead to wrong conlusions. Moreover, often the change itself is of statistical interest, e.g. at monitoring patients in intensive care and production processes in industrial plants, in climate research and in gerenal in many kinds of data analysis, if in medicine, biology, geoscience, quality control, signal processing, financials or other fields.

The question of interest in change-point analysis is: *Is there a point at which the general behaviour of the observed data changes significantly?* Such changings can be modelled as a variation of certain parameters of the model which describes the process or as a gerenal change in the model.

If the process in which a change-point shall be detected is known and can be modelled, *parametric* change-point tests are applied which strongly rely on this knowledge of the framework. If little or none information about the measured data is available, *non-parametric* methods are used which on the one hand do not require that much a priori information about the data, but which may feature less accuracy on the other hand, because they give more scope to the data. There is one further big classification of change-point problems: Suppose we have observed some data and we suspect that there could have been a change in the location:

$$X_1, X_2, \ldots, X_k \sim F(\cdot)$$
$$X_{k+1}, X_{k+2}, \ldots, X_n \sim F(\cdot + \Delta)$$

Now we want to find out if there has really been a change or not, in other words we want to know if there is such a point k and $\Delta \neq 0$ or if we have $\Delta = 0$ throughout the whole sample. This is an *offline* problem: We have collected data and wish to test if there has been a change or not. On the contrary, an *online* problem is a situation in which the data are sequentially coming in, and we want to detect a change as soon as possible after it has occurred. If the point k is already known (for example by external reasons or indications), we only have to test if there has been a change at index k or not, and the problem reduces to a *two-sample problem.*

In this work I deal with non-parametric change-point problems for long-range dependent data, data which exhibit strong dependence even over long periods of time which causes unusual behaviour and makes statistical inferences difficult (Krämer, Sibbertsen and Kleiber, 2002, e.g.).

- In this chapter, I will introduce the main objects and problems that I treat in my work and qualify the research context. Mainly I present some **fundamental concepts and notions** from the field of long-range dependent statistics – what long-range depence actually is, where it occurs and what the essential limit theorems are.

- As a start and a careful approach to two-sample change-point tests, I consider in Chapter 2 the classical **two-sample Gauß test** under longe-range dependent data. By manual calculations, I derive its limit distribution

$$\sqrt{\frac{mn}{(m+n)^{2-D}L(m+n)}}\frac{\bar{X}_m - \bar{Y}_n}{\sigma_{\text{diff}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

  as $N \to \infty$ with $m = \lambda N$ and $n = (1-\lambda)N$ for a certain $\lambda \in [0,1]$. Moreover, I develop an **estimator for the variance of** $\bar{X}_m - \bar{Y}_n$. To this end, I analyse the classical estimator for the auto-covariance and an estimator for the variance of the sample mean, $\text{Var}[\bar{X}]$, which is based on aggregation over blocks. I assess the quality of the estimators in finite sample settings via a broad set of simulations.

- In Chapter 3, I develop a **"Wilcoxon-type" change-point test**, a non-parametric test which is based on the Wilcoxon two-sample test statistic. Using limit theorems of the empirical process, I derive its limit behaviour under the null hypothesis of no change, i.e. that

$$\frac{1}{n\,d_n}\sum_{i=1}^{[n\lambda]}\sum_{j=[n\lambda]+1}^{n}\left(I_{\{X_i \leq X_j\}} - \int_{\mathbb{R}} F(x)dF(x)\right), \quad 0 \leq \lambda \leq 1,$$

  converges in distribution towards the process

$$\frac{1}{m!}\left(Z_m(\lambda) - \lambda Z_m(1)\right)\int_{\mathbb{R}} J_m(x)dF(x), \quad 0 \leq \lambda \leq 1.$$

In an extensive simulation study, I compare the performance of the test with the classical "difference-of-means" change-point test which is based on the Gauß test (which is also known as CUSUM test). This part of my work is based on the article of Dehling, Rooch and Taqqu (2012).

- In Chapter 4, I go an important step further and determine the power of the change-point tests from Chapter 3 analytically: I derive the **limit behaviour of the "Wilcoxon-type" test and of the "difference-of-means" test under local alternatives**, i.e. under the sequence of alternatives

$$A_{\tau,h_n}(n):\ \mu_i = \begin{cases} \mu & \text{for } i = 1,\dots,[n\tau] \\ \mu + h_n & \text{for } i = [n\tau]+1,\dots,n, \end{cases}$$

  where $0 \le \tau \le 1$, in other words in a model where there is a jump of height h_n after a proportion of τ in the data, which decreases with increasing sample size n. Moreover, I compare both tests in this model by calculating their asymptotic relative efficiency and illustrate the findings by a further set of simulations. These results will be published in the article of Dehling, Rooch and Taqqu (2013).

- The methods used in Chapter 3 to handle the Wilcoxon statistic can be extended to treat the general process

$$U_{\lambda,n} = \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} \left(h(X_i, X_j) - \iint h(x_1, x_2)\, dF(x_1) dF(x_2) \right), \quad 0 \le \lambda \le 1,$$

  with general kernels $h(x, y)$ which satisfy some technical conditions (for the "Wilcoxon-type" change-point test statistic, choose $h(x, y) = I_{\{x \le y\}}$). So in Chapter 5, I present a **general approach to two-sample U-statistics of long-range dependent data**.

- In Chapter 6, I will follow a different approach to handle the above defined process $U_{\lambda,n}$ by a **direct Hermite expansion** of the kernel

$$h(x, y) = \sum_{k,l=0}^{\infty} \frac{a_{kl}}{k!\, l!} H_{kl}(x, y) = \sum_{k,l=0}^{\infty} \frac{a_{kl}}{k!\, l!} H_k(x) H_l(y).$$

  This is an alternative approach. Following this route, severe technical problems will arise. I will propose several **techniques to handle them**.

- In order to enhance the practicability of the two test produres of Chapter 3, I first analyse in Chapter 7 the **influence of an estimated long-range dependence parameter** on the test decisions in a simulation study, then I develop an **estimator for the first Hermite coefficient** $a_1 := E[\xi\, G(\xi)]$ in a broad

class of situations where only the $X_i := G(\xi_i)$ are observed, and not the ξ_i. I will demonstrate that

$$\widetilde{a_1} := \frac{1}{n}\sum_{i=1}^{n} \xi'_{(i)} X_{(i)} \xrightarrow{P} a_1,$$

where ξ'_i are i.i.d. standard normal random variables and check the suitability of this estimator via several simulations. Finally I deal with an inherent problem from statistical application: A change-point may easily lead to tampered long-range dependence estimation and spurious detection of long memory. I propose different methods to **estimate the long-range dependence parameter under a change in the mean** and analyse their quality by a large simulation study.

1.1 Detecting change-points

For a general survey about change-point analysis, see the books of Basseville and Nikiforov (1993), Brodsky and Darkhovsky (1993) and Csörgő and Horváth (1997). For the case of i.i.d. observations, Antoch et al. (2008) study for example rank tests in order to detect a change in the distribution of the data, which are a natural approach if the distribution of the data is unknown. There are also many results for weakly dependent observations (Ling, 2007; Aue et al., 2009; Wied, Krämer, Dehling, 2011), but for long-range dependent data, much less is known.

Giraitis, Leipus and Surgailis (1996) treat general change-point problems in which the marginal distribution of the observations changes after a certain time. They base their change-point tests on the difference of the empirical distribution function of the first and the remaining observations and estimate the location of the change-point considering the uniform distance and the L^2-distance between the two distribution functions. Horváth and Kokoszka (1997) analyse an estimator for the time of change in the mean of univariate Gaussian long-range dependent observations. Their estimator compares the mean of the observations up to a certain time with the overall mean of all observations; in the case of independent standard normal random variables, this estimator is just the MLE for the time of change. Wang (2008a) extends the results of Horváth and Kokoszka (1997) to linear models which are not necessarily Gaussian. Wang (2003) studies certain tests for a change in the mean of multivariate long-range dependent processes which have a representation as an instantaneous functional of a Gaussian long-range dependent process. He analyses the asymptotic properties of some tests which are based on the differences of means. Kokoszka and Leipus (1998) analyse CUSUM-type estimators for the time of change in the mean under weak assumptions on the dependence structure which also cover long-range dependence. Wang (2008b) studies Wilcoxon-type rank statistics for testing linear moving-average stationary sequences that exhibit long-range dependence and which have a common distribution

against the alternatives that there is a change in the distribution. There are also tests for a change in the model from short to long memory (Hassler and Scheithauer, 2011).

1.2 Long-range dependence

So called *long-range dependent (LRD)* processes are a special class of time series which exhibit an unusual behaviour: Although they look stationary overall, there often are periods with very large and periods with very small observations, but one cannot recognize a periodic pattern. In communication networks, one often observes bursty traffic which, astonishingly, cannot be smoothed by aggregation – this is an effect of LRD as well.

Moreover, the usual limit theorems do not hold: In the case of i.i.d. observations, the variance of the sample mean grows like the number of observations n, and this holds even for weakly correlated data, but for LRD processes the variance of the sample mean grows like n^D with a $D \in (0,1)$. And while sums of i.i.d. and weakly dependent observations, scaled with $n^{-1/2}$, are asymptotically normally distributed, sums of LRD observations may need a stronger scaling to converge, and the limit may be non-normal (see Theorem 1.1).

It has turned out that this strange behaviour can be explained by very slowly decaying correlations: If a time series has correlations that decay so slowly that they are not summable (and this is what we want to call LRD), then it shows these strange effects. More specifically, one often calls a process long-range dependent if its auto-correlation function obeys a power law, while a short-range dependent process possesses an auto-correlation function that decays exponentially fast. A rigorous definition is the following:

Definition 1.1 (Long-range dependence)**.** A second order stationary process $(X_i)_{i\geq 1}$ is called *long-range dependent* if its auto-covariances have the form

$$\boxed{\gamma_k = \mathrm{Cov}[X_i, X_{i+k}] \sim k^{-D} L(k),} \tag{1.1}$$

where $D \in (0,1)$, $a(x) \sim b(x)$ means $a(x)/b(x) \to 1$ as x tends to infinity and where L is some slowly varying function (i.e. a function that satisfies $\lim_{x\to\infty} L(ax)/L(x) = 1$ for all $a > 0$). The spectral density of such an LRD process is, under some technical conditions[1], given by

$$f(\lambda) \sim |\lambda|^{D-1} L(1/\lambda), \quad \text{as } \lambda \to 0^+.$$

The spectral density is unbounded at zero and obeys a power-law near the origin (while it is bounded in the case of short-range dependent processes). The exponent D is the LRD parameter. Equivalenty, the *Hurst exponent* $H = 1 - D/2$ is often used. As mentioned, we concentrate on the case $D \in (0,1)$, i.e. $H \in (1/2, 1)$; in this case, $(X_i)_{i\geq 1}$

[1] See for example the overview of Beran (2010, p. 26) and the references given there.

exhibits LRD. For $H = 1/2$, the variables are independent, and for $H \in (0, 1/2)$ they exhibit *short-range dependence* (SRD).

In LRD time series, correlations between two observations may be small, but even observations in the past affect present behaviour. This is why LRD is also called *long memory*.

We consider a classical historic example. In the 1950s, the British hydrologist Harold Hurst was interested in dam design and therefore in the long-term storage capacity of reservoirs. He studied the water flow in the Nile river (Hurst, 1951, 1955) and analysed a remarkable ancient data set, the annual minima of the water level in the Nile river at a gauge near Cairo between 622 and 1281, which is displayed in Figure 1.1. These data behave strange, and in fact, this can be explained by LRD: The auto-covariance function of the data, shown in Figure 1.2, decays at a power law rate. This long memory causes the wave-like shape of the time series: Extreme large observations entail other large observations, and extreme small observations entail other small ones. For more statistical evidence for this type of LRD in the Nile river data, see Beran (1994, p. 21).

As indicated by the phrase "this type of LRD", it is not mandatory to define LRD by the decay of correlations. It has the advantage to be a handy concept, but various other points of view are also reasonable when talking about LRD, since it is linked with non-stationary processes, ergodic theory, self-similar processes and fractionally differenced processes. Samorodnitsky (2007) discusses these concepts (and by the way notes that in literature there can be found eleven different definitions of what exactly LRD is).

Even though it causes strange and unusual behaviour, LRD has found a very large number of applications. Some, collected at random, are:

- Finance. Volatilities, roughly defined as the diffusion of price fluctuations, are LRD processes (Breidt, Crato and de Lima, 1998), and there is some evidence of a low LRD in stock market prices (Willinger, Taqqu and Teverovsky, 1999).

  There is little or no evidence for the presence of LRD in the big capital markets of the G-7 countries and in international stock market returns (Cheung and Lai, 1995), but long memory has been detected for example in the smaller Greek stock market (Barkoulas, Baum and Travlos, 1996).

  Baillie (1996) provides a survey of the major econometric work on LRD, fractional integration and their application in economics, including an extensive list of references. He finds substatial evidence that LRD processes describe well inflation rates.

  Cheung (1993) finds evidence for LRD in some exchange rates.

  In financial time series, observations are often uncorrelated, but the auto-correlation of their squares may be not summable; Beran (2010, p. 28) lists some references.

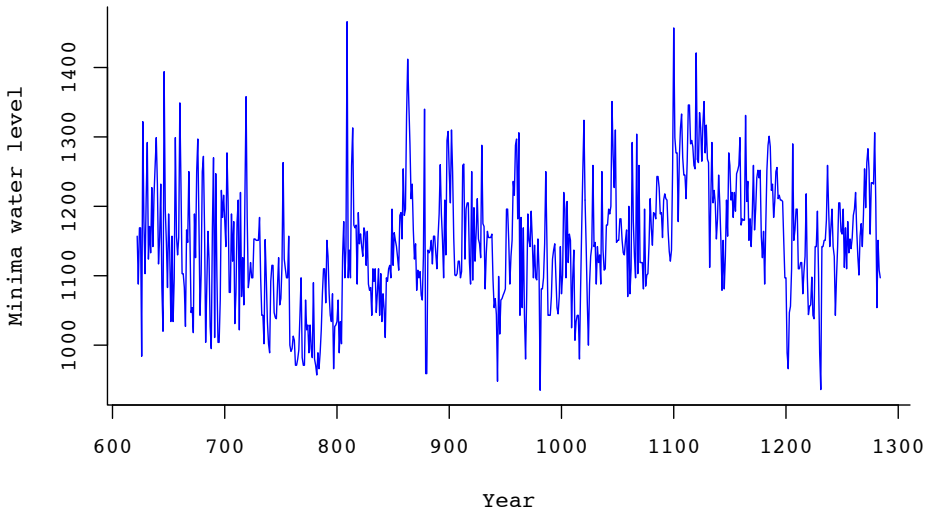


Figure 1.1: Annual minima of the water level in the Nile river near Cairo.

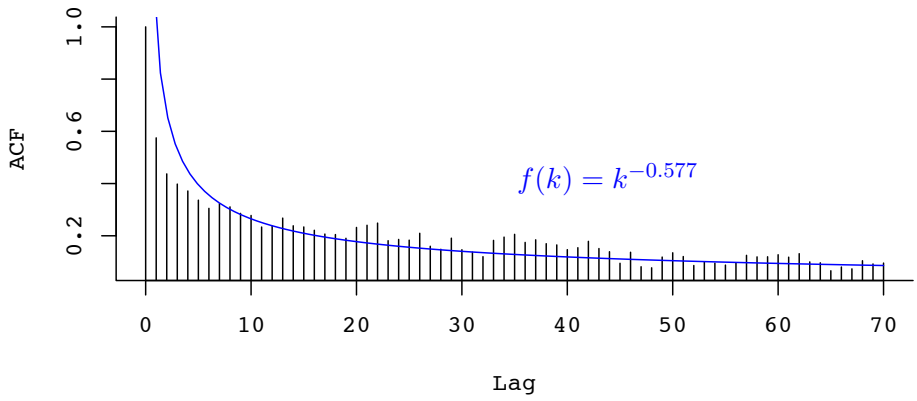


Figure 1.2: The empirical auto-correlation function of the Nile river data decays at a power law rate (and not, as usual for weakly dependent observations, exponentially fast).

When large orders are split up and executed incrementally and the size of such large orders follows a power law, then the signs of executed orders (to buy or to sell) have auto-correlations that exhibit a power-law decay (Lillo, Mike and Farmer, 2005).

- Network engineering. As the internet, an enormously complicated connection of networks, expands fastly and as the amount of memory-intensive content like videos increases, it is crucial to maintain a high networking performance. Here, LRD has a considerable impact on queueing performance and is a characteristic for some problems in data traffic engineering (Erramilli, Narayan and Willinger, 1996).

  Many papers focus on long memory in network data and on its impact on the networking performance, and indeed, LRD is an omnipresent property of data traffic both in local area networks and in wide area networks. Evidence for LRD can be found in many measurements of internet traffic like traffic load and packet arrival times; for examples and references see Li and Mills (1998) or Karagiannis, Faloutsos and Riedi (2002).

  LRD in network traffic can be explained by renewal processes that exhibit heavy-tailed interarrival distributions (Levy and Taqqu, 2000).

  A short overview about detection of LRD in internet traffic beside complex scaling and multifractal behaviour, periodicity, noise and trends is provided by Karagiannis, Molle and Faloutsos (2004), and Cappé et al. (2002) give a tutorial about statistical models for analyzing long-range dependence in network traffic data.

  Taqqu, Willinger and Sherman (1997) and Willinger et al. (1997) demonstrated that the superposition of many ON/OFF sources with strictly alternating ON- and OFF-periods and whose ON-periods or OFF-periods exhibit high variability or infinite variance can produce aggregate network traffic that exhibits self-similarity or LRD. This provides a physical explanation for the observed self-similar traffic patterns in high-speed network traffic.

- Physics. Particle diffusion in an electric current across two coupled superconductors shows LRD (Geisel, Nierwetberg and Zacherl, 1985), and the dynamics of aggregates of amphiphilic (both water-loving and fat-loving) molecules as well (Ott et al., 1990).

- Biology. Long-range power law correlation has been found in some DNA sequences and it is an issue in computational molecular biology (Peng et al., 1992, 1994; Buldyrev et al., 1995).

  LRD can be found in human coordination: If people are asked to sychronizing a movement (a fingertapping e.g.) to a periodic signal, the errors exhibit long memory, as Chen, Ding and Kelso (1997) have observed. They conjecture that this origins in random noise and sensory delay in the nervous system.

- Climate. LRD appears in surface air temperature: It can be detected in global data from the Intergovernmental Panel on Climate Change (Smith, 1993), and Caballero, Jewson and Brix (2002) show that the auto-covariance structure of observed temperature data can be reproduced by fractionally integrated time series; they explain the observed long memory by aggregation of several short-memory effects.

  Moreover, ground based observations and satellite measurements reveal that ozone and temperature fluctuations in short time intervals are correlated to those in large time intervals in a power law style (Varotsos and Kirk-Davidoff, 2006).

- And else. Many time series in political analysis (concerning variables like presidential approval or the monthly index of consumer sentiment) show LRD characteristics, as Lebo, Walker and Clarke (2000) write. They point out that many time series in political science are aggregated measures of single responses and that this aggregating of heterogenous individual-level information produces fractional dynamics.

  Long-range correlations appear also in human written language, beyond the short-range correlations which result from syntactic rules and apparently regardless of languages: They have been detected in Shakespeare's plays (Montemurro and Pury, 2002) and in novels in Korean langugage whose syntax differns strongly from the English one (Bhan et al., 2006).

A short overview about probabilistic foundations and statistical models for LRD data including extensive references is given by Beran (2010), while, even though older, Beran (1994) provides a more detailed survey. Taqqu, Teverovsky and Willinger (1995) and Taqqu and Teverovsky (1998) analyse a handful estimators that quantify the intensity of LRD in time series. Details on the models used for simulations in this work are presented in Section 1.3.

1.3 Two important LRD processes: fractional Brownian motion and fractional Gaussian noise

Now I will introduce two prominent and important examples of LRD time series, the *fractional Brownian motion (fBm)* and its incremental process, the *fractional Gaussian noise (fGn)*. To this end, we need some definitions. LRD processes are closely connected with self-similar processes, where a change of the time scale is equivalent to a change in the state space.

Definition 1.2 (Self-similar)**.** A continuous time process $(X_t)_{t\in\mathbb{R}}$ is called *self-similar*[2] *with index* H, if for all $a > 0$ and any integer $k \geq 1$

$$(X_{at_1}, X_{at_2}, \ldots, X_{at_k}) \stackrel{\mathcal{D}}{=} (a^H X_{t_1}, a^H X_{t_2}, \ldots, a^H X_{t_k}), \tag{1.2}$$

in other words if the finite-dimensional distributions of $(X_{at})_{t\in\mathbb{R}}$ are identical to the finite-dimensional distributions of $(a^H X_t)_{t\in\mathbb{R}}$.

As a consequence, typical sample paths of self-similar processes look qualitatively the same, irrespective of the time interval of observation, that is the picture stays structually the same, irrespective of if we look from distance or if we get closer.

The notion of self-similarity was introduced into statistics by Mandelbrot and van Ness (1968) and Mandelbrot and Wallis (1969a,b); Lamperti (1962) and Taqqu (1975) showed that self-similar processes occur naturally as limits of partial sums of stationary random variables, see also Pipiras and Taqqu (2011, Chap. 1.7). In nature, a fascinating richness of deterministic self-similarities can be observed, for example at leaves, montains and waves. In 1827, the scottish botanist Robert Brown (1773–1858) examined pollen particles suspended in water under a microscope and observed an erratic motion[3]. To his honour[4], the most simple and important self-similar process is called *Brownian motion*. Its definition is the following, see e.g. Beran (1994).

Definition 1.3 (Brownian motion)**.** Let $B(t)$ be a stochastic process with continuous sample paths and such that

1. $B(t)$ is a Gaussian process,
2. $B(0) = 0$ almost surely,
3. $B(t)$ has independent increments, i.e. for all $t, s > 0$ and $0 \leq u \leq \min(t, s)$, $B(t) - B(s)$ is independent of $B(u)$,

[2]Because this definition refers to equality in distribution and the property can not be spotted at a single path of X_t, one should say "statistical self-similar", and to be entirely correct, one should say "statistical self-affine", because H does not need to be 1, so that the scaling in time and space to obtain equality in distribution may be different.

[3]Some scientists have doubted that Brown's microscopes were sufficient to observe these movements (D. H. Deutsch: Did Robert Brown Observe Brownian Motion: Probably Not, *Scientific American*, 1991, 265, p. 20), but already in the same year, a British microscopist has repeated Brown's experiment (B. J. Ford: Robert Brown, Brownian Movement, and Teethmarks on the Hatbrim, *The Microscope*, 1991, 39, p. 161–171) and finally a recent study which analysed Brown's original observations under historical, botanical, microscopical and physical aspects could resolve all doubt (P. Pearle, B. Collett, K. Bart, D. Bilderback, D. Newman, S. Samuels: What Brown saw and you can too. *Am. J. Phys.*, 2010, 78, p. 1278–1289).

[4]By the way, neither did Brown provide an explanation for the observed random motion, nor was he the first to discover it: The Dutch biologist and chemist Jan Ingenhousz (1730–1799) described in 1785 an irregular movement of coal dust on the surface of alcohol, thus he, not Brown, is the true discoverer of what came to be known as Brownian motion (P. W. van der Pas: The discovery of the Brownian motion, *Scientiarum Historia*, 1971, 13, p. 27–35).

4. $E[B(t) - B(s)] = 0$,
5. $\mathrm{Var}[B(t) - B(s)] = \sigma^2|t - s|$, for a certain $\sigma^2 \geq 0$.

Then $B(t)$ is called *Brownian motion*. If $\sigma^2 = 1$, it is *standard Brownian motion*.

Definition 1.4 (Stationary increments)**.** A process $(X_t)_{t\in\mathbb{R}}$ has *stationary increments* if the processes

$$(X_{t+c} - X_t)_{t\in\mathbb{R}}$$

have the same distribution, independent of the choice of $c \in \mathbb{R}$.

Now it is a beautiful result that there is only one unique Gaussian process which is self-similar and which has stationary increments: the so called *fractional Brownian motion (fBm)*, a generalisation of the above defined usual Brownian motion.

Definition 1.5 (Fractional Brownian motion)**.** Let $a > 0$ be a positive scaling constant and $B(t)$ a standard Brownian motion. Define a weight function w_H by

$$w_H(t,u) = \begin{cases} 0 & \text{for } t \leq u \\ (t-u)^{H-\frac{1}{2}} & \text{for } 0 \leq u < t \\ (t-u)^{H-\frac{1}{2}} - (-u)^{H-\frac{1}{2}} & \text{for } u < 0 \end{cases}.$$

For $0 < H < 1$ define the stochastic integral[5]

$$B_H(t) = a \int w_H(t,u)\, dB(u),$$

where the convergence of the integral is to be understood in the L^2 norm with respect to the Lebesgue measure on the real numbers. $B_H(t)$ is called *fractional Brownian motion (fBm) with self-similarity parameter H*.

For $H = \frac{1}{2}$ and $a = 1$, we obtain the regular Brownian motion. The main difference between fBm and regular Brownian motion is that, while the increments in Brownian motion are independent, they are dependent in fBm; this means for $H > \frac{1}{2}$, that if the previous steps have been increasing, it is likely that the next step will be increasing as well (for $H > \frac{1}{2}$, the increments of the process are positively correlated, while for $H < \frac{1}{2}$ they are negatively correlated and an increasing pattern in the previous steps will more likely cause a decreasing next step).

Standard fBm has the covariance function

$$\boxed{E[B_H(t)B_H(s)] = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |t-s|^{2H}).} \tag{1.3}$$

Since fBm has stationary increments, it gives rise to a new stationary process.

[5]Here, we encounter stochastic integrals, and in fact, the theory of LRD processes is strongly related to stochastic integration, as we will shortly see. I give a formal introduction into stochastic integration in Appendix A.

Definition 1.6 (Fractional Gaussian noise)**.** Let $B_H(t)$ be a fBm, as in Definition 1.5. Its incremental process

$$\boxed{\xi_k = B_H(k+1) - B_H(k), \quad k \in \mathbb{Z},}$$

is called *fractional Gaussian noise (fGn).*

Standard fGn has the auto-covariance function

$$\boxed{\gamma_k = \frac{1}{2}\left(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\right), \quad k \in \mathbb{Z},} \tag{1.4}$$

and for $H \neq \frac{1}{2}$ it holds

$$\boxed{\gamma_k \sim H(2H-1)k^{-(2-2H)} = \frac{(1-D)(2-D)}{2}k^{-D},} \tag{1.5}$$

see, e.g., Samorodnitsky and Taqqu (1994). (1.5) shows the above mentioned link between self-similarity and LRD: Fractional Gaussian noise, the incremet process of the only self-similar Gaussian process, is an LRD process, compare to (1.1). If $H = \frac{1}{2}$, fGn is regular white noise, and since it is Gaussian, it is i.i.d..

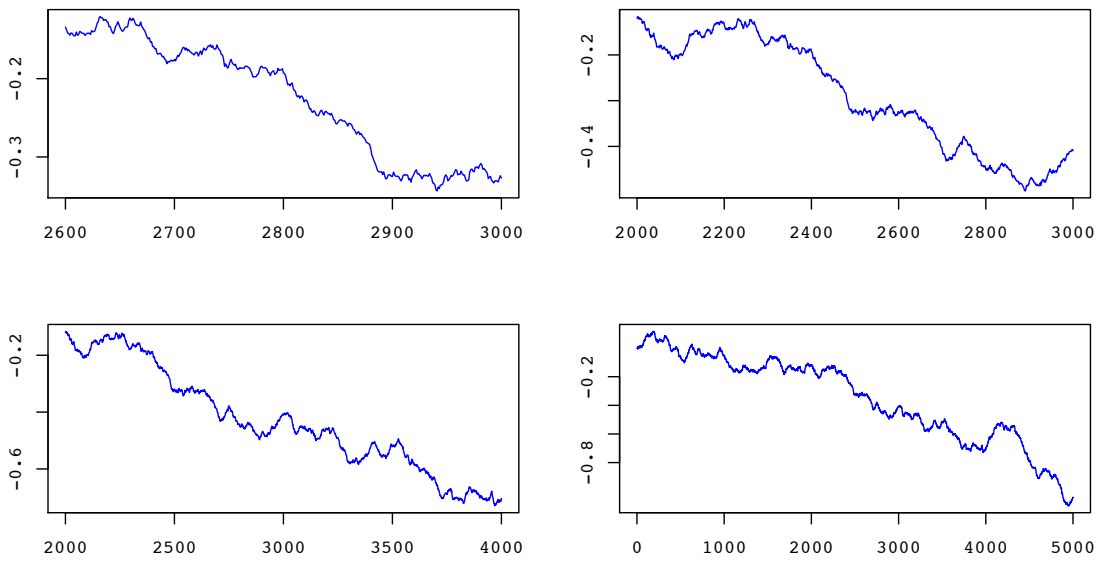


Figure 1.3: An example for self-similarity: Looks from different distances on one realization of fractional Brownian motion with Hurst parameter $H = 0.7$.

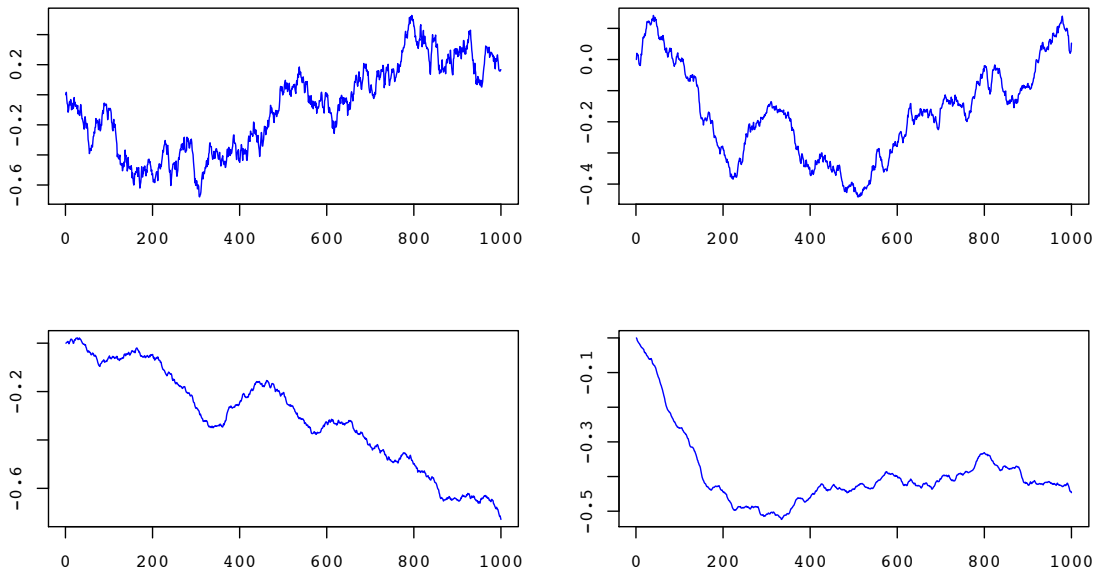


Figure 1.4: Fractional Brownian motion with different Hurst parameters, $H = 0.5, 0.7$ (top) and $H = 0.8, 0.9$ (bottom).

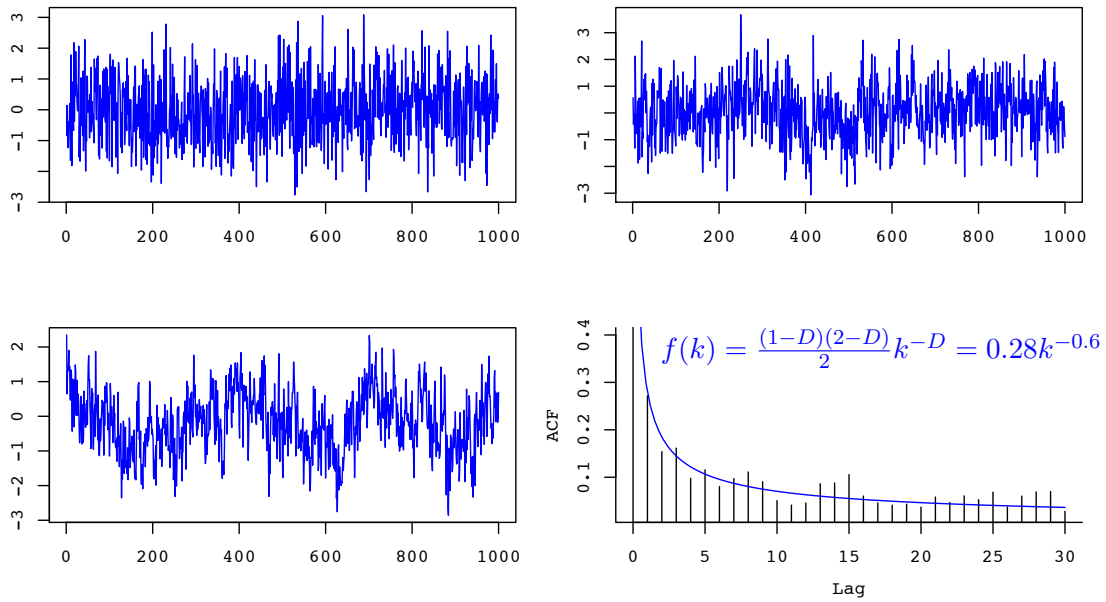


Figure 1.5: Fractional Gaussian noise with Hurst parameters $H = 0.5$ (top left, i.e. white noise), $H = 0.7$ (top right) and $H = 0.9$ (bottom left). At the bottom right, the empirical auto-correlation function of fGn with Hurst parameter $H = 0.7$ (i.e. $D = 0.6$) is shown.

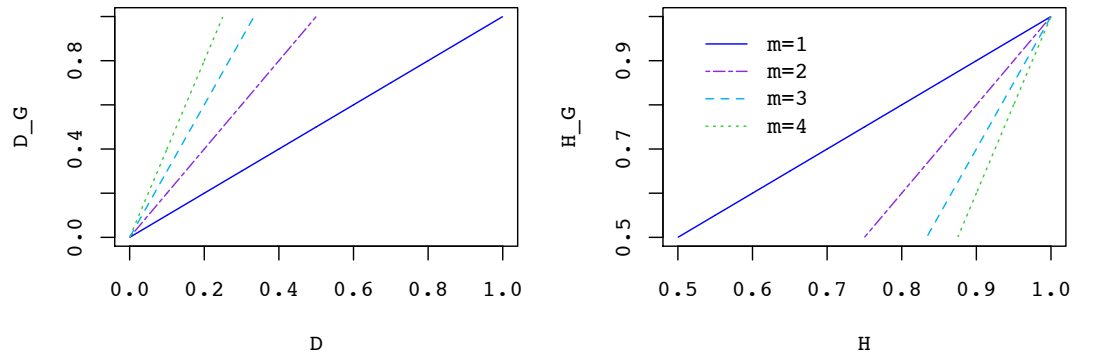


Figure 1.6: Relation between LRD behaviour of the ξ_i and $G(\xi_i)$: If the ξ_i are LRD with parameter $0 < D < 1$, the $G(\xi_i)$ are LRD with parameter $D_G = Dm$, if $0 < Dm < 1$ (left), respectively $1/2 < H < 1$, $H_G = (2 - 2H)m$ and $G(\xi_i)$ is LRD if $1 - 1/(2m) < H < 1$ (right).

1.4 Hermite polynomials

When tackling statistics of LRD observations, it has turned out to be useful to expand the statistic whose asymptotic behaviour one is interested in in so called Hermite polynomials. Since different definitions are used in different books and articles, it is not amiss to give a review here.

1.4.1 Definition and basic properties

Definition 1.7 (Hermite polynomials)**.** The functions

$$\boxed{H_k(x) = (-1)^k e^{x^2/2} \frac{d^k}{dx^k} e^{-x^2/2}, \qquad k = 0, 1, 2, \ldots}$$

are called *Hermite polynomials*.

The first few are given by

$$\begin{aligned} H_0(x) &= 1 \\ H_1(x) &= x \\ H_2(x) &= x^2 - 1 \\ H_3(x) &= x^3 - 3x \\ H_4(x) &= x^4 - 6x^2 + 3. \end{aligned}$$

Hermite polynomials have many interesting and important properties:

- $H_k(x)$ is a polynom (what else when it is called polynom?) of degree k with leading coefficient 1.
- $H_k(x), k = 0, 1, 2, \ldots$ form an orthogonal basis of the Hilbert space $L^2(\mathbb{R}, \mathcal{N})$, the space of real functions which are square-integrable with respect to the $\mathcal{N}(0, 1)$ density function:

  $$\langle H_i, H_j \rangle_{\mathcal{N}} = (2\pi)^{-1/2} \int_{-\infty}^{\infty} H_i(x) H_j(x) e^{-x^2/2}\, dx = \begin{cases} 0 & i \neq j \\ i! & i = j \end{cases}$$

  For a simple proof, see e.g. Pipiras and Taqqu (2011, Chap. 3.1).
- The Hermite polynomials in $L^2(\mathbb{R}^d, \mathcal{N})$, the Hilbert space of square-integrable functions on $\mathbb{R}^d$ with respect to the independent d-dimensional standard normal measure, can simply be defined as the product of d one-dimensional Hermite polynomials: $H_{k_1,\ldots,k_d}(x_1, \ldots, x_d) = H_{k_1}(x_1) \cdot \ldots \cdot H_{k_d}(x_d)$.
- $H_{k+1}(x) = xH_k(x) - kH_{k-1}(x)$
- $\frac{d}{dx} H_k(x) = kH_{k-1}(x)$

- One may admit an additional parameter ρ and define the Hermite polynomials by
$$H_k(x,\rho) = (-\rho)^k e^{x^2/2\rho} \frac{d^k}{dx^k} e^{-x^2/2\rho}, \qquad k = 0,1,2,\ldots,$$
the first few are
$$\begin{aligned} H_0(x,\rho) &= 1 \\ H_1(x,\rho) &= x \\ H_2(x,\rho) &= x^2 - \rho \\ H_3(x,\rho) &= x^3 - 3\rho x \\ H_4(x,\rho) &= x^4 - 6\rho x^2 + 3\rho^2. \end{aligned}$$

- There are many further identities involving Hermite polynomials; for an impressive overview and some references, see Weisstein (2010).

The definition above is sometimes called the "probabilists' definition" because of the normal weight. Widely spread is as well the "physicists' definition"
$$H_k^{(phy)}(x) = (-1)^k e^{x^2} \frac{d^k}{dx^k} e^{-x^2},$$
which defines an orthogonal basis of $L^2(\mathbb{R})$ with respect to the weight e^{-x^2}. Both definitions are related by $H_k^{(phy)}(x) = 2^{k/2} H_k(\sqrt{2}x)$.

Often Hermite functions
$$\tilde{h}_k(x) = H_k^{(phy)}(x) e^{-x^2/2}$$
or normalized Hermite functions
$$h_k(x) = \frac{1}{\sqrt{2^k k! \sqrt{\pi}}} H_k^{(phy)}(x) e^{-x^2/2}$$
are considered. The h_k, $k = 0,1,\ldots$ form an orthonormal basis for $L^2(\mathbb{R},\lambda)$.

1.4.2 Relation to LRD

We will now outline the important role of Hermite polynomials in the context of LRD. Let $\xi_1,\ldots,\xi_N$ be Gaussian random variables with mean 0, variance 1 and covariances (1.1); such variables exhibit long memory. Now consider the partial sum
$$S_N = \sum_{i=1}^{N} h(\xi_i),$$
where h is a centralized function in $L^2(\mathbb{R},\mathcal{N})$: h is measurable with $Eh(\xi_i) = 0$ and $Eh^2(\xi_i) < \infty$. We will see that the asymptotic behaviour of this sum is closely associated with the asymptotic behaviour of the Hermite polynomials $H_k(\xi_i)$.

For a start we assume that h itself is already a Hermite polynomial H_k. Now the growth of the partial sum is essentially governed by the fundamental property[6]

$$\boxed{\operatorname{Cov}\left[H_m(\xi_i), H_n(\xi_j)\right] = \begin{cases} m!\left(\operatorname{Cov}\left[\xi_i, \xi_j\right]\right)^m & m = n \\ 0 & \text{otherwise} \end{cases},} \tag{1.6}$$

and we have

$$\begin{aligned} \operatorname{Var}\left[\sum_{i=1}^{N} H_k(\xi_i)\right] &= \sum_{i=1}^{N} \operatorname{Var}[H_k(\xi_i)] + 2\sum_{j=1}^{N-1}(N-j)\operatorname{Cov}\left[H_k(\xi_1), H_k(\xi_{1+j})\right] \\ &= Nk! + 2k!\sum_{j=1}^{N-1}(N-j)\gamma_j^k \\ &= Nk! + 2k!\sum_{j=1}^{N-1}(N-j)j^{-Dk}L(j)^k. \end{aligned} \tag{1.7}$$

Now it is obvious that the limiting behaviour depends on the size of Dk. If $Dk > 1$, the sum is $O(N)$ and the variance of the partial sum grows asymptotically like N, just like in the weak dependent or in the independent case. But if $Dk < 1$, the situation is quite different: Like in (B.6) we find the asymptotic equivalence

$$\begin{aligned} \operatorname{Var}\left[\sum_{i=1}^{N} H_k(\xi_i)\right] &\sim Nk! + \frac{2k!}{(1-Dk)(2-Dk)}N^{2-Dk}L(N)^k \\ &= N^{2-Dk}L(N)^k k!\left(\frac{1}{N^{1-Dk}L(N)^k} + \frac{2}{(1-Dk)(2-Dk)}\right) \\ &\sim N^{2-Dk}L(N)^k k!\frac{2}{(1-Dk)(2-Dk)}. \end{aligned}$$

So we see the extraordinary property of long memory statistics: The size of the LRD parameter D and the degree of the Hermite polynomial k determine the growth of the variance of the partial sum – for $Dk > 1$ we observe usual SRD behaviour, while for $Dk < 1$ we observe a faster rate of growth (wich also can be taken as a definition of LRD[7]):

$$\boxed{\operatorname{Var}\left[\sum_{i=1}^{N} H_k(\xi_i)\right] \sim \begin{cases} Nk!C_{\mathrm{SRD}}, & \text{if } Dk > 1, \\ N^{2-Dk}L(N)^k\frac{2k!}{(1-Dk)(2-Dk)}, & \text{if } Dk < 1, \end{cases}} \tag{1.8}$$

[6] This follows from the diagram formula. With this formula one can compute expectations and cumulants of finite families of random variables, for example expectations of Hermite polynomials of Gaussian variables, but it is not a walk in the park. I will introduce a special version of the formula and give some examples in Section 2.5. Without going too much into theoretical details, one can see for example Beran (1994, eq. (3.18)) or Simon (1974, Th. 1.3), where the statement is proved for Wick powers, random variables with special properties; Hermite polynomials – in the probabilists' definition with respect to the normal density measure – are a special case of Wick powers. Results on Wick powers can also be found in Major (1981a, p. 9). A general introduction to the diagram formula is given by Surgailis (2003).

[7] This is the so called *LRD in the sense of Allen* (Pipiras and Taqqu, 2011, Chap. 1.1).

where C_{SRD} is a constant. (Moreover, Dk influences the limit distribution in a broader sense, as we will shortly see.) We will focus on the second case, because here the variance of the partial sum grows faster than with the usual rate N, and this is the case where also the limit distribution of the partial sum can be non-normal. The case $Dk = 1$ is special; here the usual central limit theorem holds, but the norming factor may be different than the usual $\sqrt{N}$, see e.g. the discussion of Theorem 8.3 in Major (1981a), the remark in Dobrushin and Major (1979, p. 30) or Breuer and Major (1983, p. 428).

Now we turn to a more generel kernel $h(x)$. We represent it by its Hermite expansion $h(x) = \sum_{k=1}^{\infty} a_k/k!\, H_k(x)$ (the equality is to be understood as convergence in L^2) with coefficients

$$a_k := \langle h, H_k \rangle_{\mathcal{N}} = (2\pi)^{-1/2} \int_{-\infty}^{\infty} h(x) H_k(x) e^{-x^2/2}\, dx.$$

Now the partial sum is

$$\begin{aligned}
\operatorname{Var}\left[\sum_{i=1}^{N} h(\xi_i)\right] &= \operatorname{Var}\left[\sum_{k=1}^{\infty} \frac{a_k}{k!} \sum_{i=1}^{N} H_k(\xi_i)\right] \\
&= \sum_{k=1}^{\infty} \frac{a_k^2}{k!^2} \operatorname{Var}\left[\sum_{i=1}^{N} H_k(\xi_i)\right] + \sum_{k\neq l}^{\infty} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{a_k a_l}{k!\; l!} \operatorname{Cov}\left[H_k(\xi_i), H_l(\xi_j)\right] \\
&= \sum_{k=1}^{\infty} \frac{a_k^2}{k!^2} \operatorname{Var}\left[\sum_{i=1}^{N} H_k(\xi_i)\right]
\end{aligned}$$

because of the orthogonality of the H_l, H_k. Now let m be the *Hermite rank* of $h(x)$, the smallest index among those $k \in \mathbb{N}$ with $a_k \neq 0$. The term with the belonging coefficient a_m dominates all others, because for an arbitrary $k > m$ we have by (1.8)

$$\frac{a_k^2 \operatorname{Var}\left[\sum_{j=1}^{N} H_k(\xi_j)\right]}{a_m^2 \operatorname{Var}\left[\sum_{j=1}^{N} H_m(\xi_j)\right]} \sim \frac{a_k^2 l!(1-Dm)(2-Dm)}{a_m^2 m!(1-Dl)(2-Dl)} N^{-D(k-m)} L(N)^{k-m} \to 0$$

as $N \to \infty$. Thus

$$\boxed{\operatorname{Var}\left[\sum_{j=1}^{N} h(\xi_j)\right] \sim \frac{a_m^2}{m!^2} \begin{cases} N m! C_{\mathrm{SRD}}, & \text{if } Dm > 1, \\ N^{2-Dm} L^m(N) c_m, & \text{if } Dm < 1 \end{cases}} \tag{1.9}$$

with $C_{\mathrm{SRD}}, c_m \in \mathbb{R}$ (to be more precise $c_m = \frac{2m!}{(1-Dm)(2-Dm)}$).

Finding the limit distribution of S_N is challenging. What makes functionals of LRD observations tricky is that they may have a non-normal limit. Although the most interesting cases, thankfully, follow a normal distribution, almost all functionals of long-range dependent observations have a limit which is neither normal nor easy (or possible) to write down in a closed form. Dobrushin and Major (1979) and, independently, Taqqu (1979) have derived a general representation for the limit: It can be expressed in terms of so called *multiple Wiener-Itō integrals*. In Appendix A, I explain the idea behind

these objects. We now quote a result from Dobrushin and Major (1979), which holds in greater generality, namely for processes, see also Major (1981b).

Theorem 1.1 (Non-central limit theorem for LRD processes)**.** *Let $Dm < 1$. Then as $N \to \infty$*

$$\boxed{\left\{ d_N^{-1} \sum_{i=1}^{\lfloor tN \rfloor} h(\xi_i) \right\}_{t\in[0,1]} \xrightarrow{\mathcal{D}} \left\{ \frac{a_m}{m!} Z_m(t) \right\}_{t\in[0,1]}} \tag{1.10}$$

with

$$Z_m(t) = K^{-m/2} c_m^{-1/2} \int_{\mathbb{R}^m}{}' \frac{e^{it\sum_{j=1}^m x_j} - 1}{i\sum_{j=1}^m x_j} \left(\prod_{j=1}^m |x_j|^{(D-1)/2} \right) dW(x_1) \dots dW(x_m) \tag{1.11}$$

$$d_N^2 = d_N^2(m) = c_m N^{2-Dm} L^m(N), \tag{1.12}$$

where i is the imaginary unit and

$$K = \int_{\mathbb{R}} e^{ix} |x|^{D-1} \, dx = 2\Gamma(D)\cos(D\pi/2),$$
$$c_m = \frac{2m!}{(1-Dm)(2-Dm)}.$$

Formula (1.11) denotes the multiple Wiener-Itō integral with respect to the random spectral measure W of the white-noise process, where $\int'$ means that the domain of integration excludes the hyperdiagonals $\{x_i = \pm x_j,\ i \neq j\}$, see also Dehling and Taqqu (1989, p. 1769). The constant of proportionality c_m ensures that $E[Z_m(1)]^2 = 1$. Taqqu (1979) or Pipiras and Taqqu (2011, Chap. 3.2) give another representation.

Technical remark. The limit is – as here – often denoted as $Z_m(t)$. However, one has to pay attention, because whether a special $Z_m(t)$ is normalized (so that $E[Z_m(1)]^2 = 1$) or not, differs from article to article (even when they are written by the same author).

For instance in the case of Hermite rank $m = 1$, the limit is

$$\begin{aligned} a_1 Z_1(t) &= a_1 \left(\frac{1}{2\Gamma(D)\cos(D\pi/2)} \right)^{1/2} \left(\frac{(1-D)(2-D)}{2} \right)^{1/2} \int \frac{e^{itx}-1}{ix} |x|^{(D-1)/2} \, dW(x) \\ &= a_1 \left(\frac{\Gamma(3-D)\sin(D\pi/2)}{2\pi} \right)^{1/2} \int \frac{e^{itx}-1}{ix} |x|^{(D-1)/2} \, dW(x) \end{aligned}$$

and with $D = 2 - 2H$

$$\begin{aligned} &= a_1 \left(\frac{H\Gamma(2H)\sin(\pi H)}{\pi} \right)^{1/2} \int \frac{e^{itx}-1}{ix} |x|^{-(H-1/2)} \, dW(x) \\ &= a_1 B_H(t) \end{aligned}$$

which is a fractional Brownian motion at time t. The first equality can be shown by Euler's reflection formula for the Gamma function, $\Gamma(z)\Gamma(1-z) = \pi/\sin(\pi z)$, and a trigonometric double-angle identitiy; the last equality is proved by Taqqu (2003, Prop. 9.2). So we receive

$$N^{-1+\frac{D}{2}} L(N)^{-1/2} \sum_{i=1}^{\lfloor tN \rfloor} h(\xi_i) \xrightarrow{\mathcal{D}} a_1 \left(\frac{2}{2H(2H-1)} \right)^{1/2} B_H(t) \tag{1.13}$$
$$= a_1 \left(\frac{2}{(1-D)(2-D)} \right)^{1/2} B_{1-D/2}(t),$$

which has been proven directly by Taqqu (1975, Cor. 5.1).

We have just seen how the Hermite rank of a function h influences the variance of the partial sum $S_N = \sum_{i=1}^{N} h(\xi_i)$. If the underlying Gaussian process $(\xi)_{i\geq 1}$ is LRD with parameter $D \in (0,1)$, the partial sum S_N inherits LRD-type behaviour if $Dm \in (0,1)$, where m is the Hermite rank of h. This rate of growth, N^{2-Dm} instead of N, is closely related with the definition of LRD as given in Definition 1.1, see the discussion by Pipiras and Taqqu (2011, Chap. 1.1), and in fact it can be taken as a definition for LRD. But here, we work with Definition 1.1, so to conclude this chapter, we will now show under which conditions a stationary Gaussian process $(\xi_i)_{i\geq 1}$ passes its LRD, in the sense of Definition 1.1, on to the transformed process $(X_i)_{i\geq 1}$, $X_i = G(\xi_i)$.

Let ξ, η be two standard normal random variables and $G_1, G_2 \in L^2(\mathbb{R}, \mathcal{N})$ two functions. We expand G_1 and G_2 in Hermite polynomials

$$G_1(x) = \sum_{k=0}^{\infty} \frac{a_{1,k}}{k!} H_k(x) \qquad\qquad G_2(x) = \sum_{l=0}^{\infty} \frac{a_{2,l}}{l!} H_l(x)$$

where $a_{i,k} = E[G_i(\xi) H_k(\xi)]$ is the associated k-th Hermite coefficient. With these expansions, (1.6) yields

$$\begin{aligned}
\operatorname{Cov}[G_1(\xi), G_2(\eta)] &= E[G_1(\xi) G_2(\eta)] - E[G_1(\xi)]\, E[G_2(\eta)] \\
&= \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \frac{a_{1,k}}{k!} \frac{a_{2,l}}{l!} E[H_k(\xi) H_l(\eta)] - E[G_1(\xi)]\, E[G_2(\eta)] \\
&= \sum_{k=0}^{\infty} \frac{a_{1,k}}{k!} \frac{a_{2,k}}{k!} k! \left(E[\xi\eta]\right)^k - E[G_1(\xi) H_0(\xi)]\, E[G_2(\eta) H_0(\eta)] \\
&= \sum_{k=0}^{\infty} \frac{a_{1,k}}{k!} \frac{a_{2,k}}{k!} k! \left(E[\xi\eta]\right)^k - a_{1,0} a_{2,0} \\
&= \sum_{k=1}^{\infty} \frac{a_{1,k}}{k!} \frac{a_{2,k}}{k!} k! \left(E[\xi\eta]\right)^k .
\end{aligned}$$

We have just proved the following

Proposition 1.2. *Consider a stationary Gaussian process $(\xi_i)_{i\geq 1}$ with mean 0 and variance 1 and auto-covariance as in* (1.1)*. Let $G \in L^2(\mathbb{R}, \mathcal{N})$ have Hermite rank m. Then the process $(X_i)_{i\geq 1} = (G(\xi_i))_{i\geq 1}$ has auto-covariances*

$$\begin{aligned}\gamma_G(k) &= \mathrm{Cov}[X_i, X_{i+k}] = \mathrm{Cov}[G(\xi_i), G(\xi_{i+k})] \\ &= \sum_{p=1}^{\infty} \frac{a_p^2}{p!} \left(E[\xi_i \xi_{i+k}]\right)^p ,\end{aligned}$$

and the first term in this expansion dominates the others, such that

$$\boxed{\gamma_G(k) \sim \frac{a_m^2}{m!} \left(L(k)k^{-D}\right)^m ,}$$

thus if $0 < Dm < 1$, $(G(\xi_i))_{i\geq 1}$ is LRD in the sense of Definition 1.1 with LRD parameter $D_G = Dm$ and slowly varying function $L_G(k) = a_m^2/m!L^m(k)$.

As a consequence, $(\xi_i)_{i\geq 1}$ may be LRD, but $(G(\xi_i))_{i\geq 1}$, e.g. $(\xi_i^2)_{i\geq 1}$ may be not. Figure 1.6 shows the relation between the LRD parameter D of the underlying ξ_i's and the LRD parameter D_G of the transformed $G(\xi_i)$'s.

Chapter 2

The asymptotic behaviour of $\bar{X} - \bar{Y}$

Before we investigate change-point tests, we consider the two-sample Gauß test which is traditionally used to detect a difference in the location of two samples of normal distributed observations. This test follows as a special case from the "difference-of-means" change-point test which we will discuss later (see section 3.4.2) – it is nothing else than the change-point test applied in a situation with a known change-point and with only Gaussian data –, but this example illustrates the impact of LRD on statistical procedures, the impact of persistent correlations, and as a basic and direct approach, it may lead to an intuitive understanding of the subject.

So suppose we have observed some data and we suspect that after the m-th observation there has been a change in the mean:

$$\begin{aligned} X_1, X_2, \ldots, X_m &\sim \mathcal{N}(\mu_1, \sigma^2) \\ X_{m+1}, X_{m+2}, \ldots, X_{m+n} &\sim \mathcal{N}(\mu_2, \sigma^2) \end{aligned} \tag{2.1}$$

We want to find out if there really has been a change or not: We wish to test the nullhypothesis $H : \mu_1 = \mu_2$ against the alternative $A : \mu_1 \neq \mu_2$. A natural idea to do this is to compare the means of both samples. To be in line with standard notation, we call the second sample the Y-sample ($Y_k := X_{m+k}$) and consider the data

$$\begin{aligned} X_1, X_2, \ldots, X_m &\sim \mathcal{N}(\mu_1, \sigma^2) \\ Y_1, Y_2, \ldots, Y_n &\sim \mathcal{N}(\mu_2, \sigma^2). \end{aligned}$$

In the i.i.d. case and with unknown variance σ^2, a common test for the problem (H, A) is the t-test which is based on the difference of the means

$$T = \sqrt{\frac{mn}{m+n}} \frac{\bar{X} - \bar{Y}}{s_{X,Y}}$$

with the weighted variance

$$s_{X,Y}^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2},$$

where s_x^2 and s_y^2 are the empirical variances of the X- and the Y-sample. From introductory mathematical statistics we know that under H, $T \sim t_{m+n-2}$, so we reject H on a significance level of α, if $|T| > t_{\alpha/2,\, n+m-2}$, where $t_{\alpha,q}$ is the upper α-quantile of Student's t-distribution with q degrees of freedom.

We will examine the LRD case with observations from a stationary Gaussian process $(X_i)_{i\geq 1}$ with mean 0, variance 1 and auto-covariance (1.1). One should expect that $\bar{X} - \bar{Y}$ shows a different behaviour in this case, since long memory requires a stronger normalization factor and often leads to non-normal limit distributions. We shall see that this intuition is right. But first, we need some preliminaries to handle the covariance structure for it is given as an asymptotic equivalence and it contains a slowly varying function.

Finally, we will estimate the variance of $\bar{X} - \bar{Y}$ in this chapter. To this end, we will propose an estimator for the variance of the mean $\bar{X_N} = \sum_{i=1}^N X_i$ and an estimator for the auto-covariances $\gamma_k = \mathrm{Cov}[X_i, X_{i+k}]$ in a sample of N observations $X_1, \ldots, X_N$, as N tends to ∞. We base our estimator for the variance of $\bar{X} - \bar{Y}$ on these two methods and demonstrate that all these estimators are asymptotically unbiased.

2.1 Asymptotic equivalence and slowly varying functions

Definition 2.1 (Asymptotic equivalence)**.** Two real functions f and g are called *asymptotically equivalent* as $x \to \infty$, written as $f \sim g$, if

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1.$$

Asymptotically equivalent functions have the same limit, if it exists, and the same growth behaviour as x increases.

Definition 2.2 (Slowly varying function)**.** A real function $L : (0,\infty) \to (0,\infty)$ is called *slowly varying* (at infinity) if for all $a > 0$

$$\boxed{\lim_{x\to\infty} \frac{L(ax)}{L(x)} = 1.}$$

Any function with $\lim_{x\to\infty} L(x) = b \in (0,\infty)$ and any power of the logarithm $L(x) = \log^c x$, $c \in \mathbb{R}$ is slowly varying. Usual powers $f(x) = x^c$ are not slowly varying.

Technical remark. In this work, we always consider Gaussian processes $(X_i)_{i\geq 1}$ with mean 0, variance 1 and covariances as in (1.1), i.e. $\gamma_k = \mathrm{Cov}[X_i, X_{i+k}] \sim k^{-D}L(k)$, where $D \in (0,1)$ is the parameter for the long memory and L is a slowly varying function. It makes no difference if we write $\gamma_k = k^{-D}L(k)$ or $\gamma_k \sim k^{-D}L(k)$, because L is not exactly specified. The asymptotic equivalence notation makes clearer that we admit covariances whose behaviour we know only asymptotically.

L does not only model disturbance or uncertaincy, it is rather there to ensure that the covariance matrix of the process is positive-semidefinite – what a covariance matrix has to be. $\gamma_k = k^{-D}$ alone does not provide a covariance matrix because it may lead to a non-positive-semidefinite matrix, as the following example shows:

For three observations X_1, X_2, X_3, $\gamma_k = k^{-D}$ and $D = 0.6$ we obtain as 'covariance matrix'

$$\Sigma = \left(\gamma_{|k-l|}\right)_{k,l=1}^{3} = \begin{pmatrix} 1 & 2^{-3/5} \\ 1 & 1 \\ 2^{-3/5} & 1 \end{pmatrix}$$

which has the (rounded) eigenvalues 2.782, 0.340, −0.122. But for $\gamma_k = \log k \cdot k^{-D}$ we obtain a true covariance matrix

$$\Sigma = \left(\gamma_{|k-l|}\right)_{k,l=1}^{3} = \begin{pmatrix} 1 & 0 & \frac{\log 2}{2^{3/5}} \\ 0 & 0 & \\ \frac{\log 2}{2^{3/5}} & 0 & 1 \end{pmatrix}$$

which now has the nice (rounded) eigenvalues 1.457, 1, 0.543.

Lemma 2.1 (Asymptotic equivalence in sums)**.** *Let $(a_n), (b_n), (\alpha_n), (\beta_n)$ be sequences of real numbers, $g \neq 1$ a constant and $\alpha_n/\beta_n > 0$ for large n.*
(i) If $\alpha_n/\beta_n \to g$ for $n \to \infty$, then

$$\boxed{a_n \sim \alpha_n, b_n \sim \beta_n \quad \Rightarrow \quad (a_n \pm b_n) \sim (\alpha_n \pm \beta_n), \qquad \text{as } n \to \infty.} \tag{2.2}$$

(ii) If $\sum_{k=1}^{n} \alpha_k$ is unbounded and strictly monotonic increasing from a certain index n_0, then

$$\boxed{a_n \sim \alpha_n \quad \Rightarrow \quad \sum_{k=1}^{n} a_k \sim \sum_{k=1}^{n} \alpha_k, \qquad \text{as } n \to \infty.} \tag{2.3}$$

(iii) If $(a_{k,n}), (\alpha_{k,n})$ are two sequences which depend on the indices n and $k \leq n$ and satisfy

$$\sum_{k=1}^{n-1} (a_{k,n} - a_{k,n-1}) + a_{n,n} \sim \sum_{k=1}^{n-1} (\alpha_{k,n} - \alpha_{k,n-1}) + \alpha_{n,n}, \qquad \text{as } n \to \infty,$$

and if $\sum_{k=1}^{n} \alpha_{k,n}$ is unbounded and strictly monotonic increasing from a certain index n_0, then

$$\boxed{\sum_{k=1}^{n} a_{k,n} \sim \sum_{k=1}^{n} \alpha_{k,n}, \qquad \text{as } n \to \infty.} \tag{2.4}$$

Proof. (i) Let $\varepsilon > 0$ be given, but small enough that $|g - 1| > 2\varepsilon$. Choose $n_0 \in \mathbb{N}$ so big that $|(a_n - \alpha_n)/\alpha_n|$ and $|(b_n - \beta_n)/\beta_n|$ are both smaller than $\varepsilon^2/(2+g)$, that $|\alpha_n/\beta_n - 1| > \varepsilon$ and $|\alpha_n/\beta_n - g| < 1$ and that α_n/β_n is positive for all $n \geq n_0$ (we

can find such an index n_0 because from a certain point forward, a_n/α_n and b_n/β_n stay close to 1 and α_n/β_n stays close to $g \neq 1$). Then we have for all $n \geq n_0$

$$\begin{aligned}\left|\frac{(a_n - b_n) - (\alpha_n - \beta_n)}{\alpha_n - \beta_n}\right| &\leq \frac{||a_n - \alpha_n| + |b_n - \beta_n||}{|\alpha_n - \beta_n|} \\ &\leq \frac{\varepsilon^2}{2+g}\frac{||\alpha_n| + |\beta_n||}{|\alpha_n - \beta_n|} \\ &= \frac{\varepsilon^2}{2+g}\frac{\left|\frac{\alpha_n}{\beta_n} + 1 - g + g\right|}{\left|\frac{\alpha_n}{\beta_n} - 1\right|} \\ &\leq \frac{\varepsilon^2}{2+g}\frac{|1+g|+1}{\varepsilon} = \varepsilon,\end{aligned}$$

and this means that $\frac{a_n - b_n}{\alpha_n - \beta_n} \to 1$, in other words $(a_n - b_n) \sim (\alpha_n - \beta_n)$. In the exact same manner we can verify $(a_n + b_n) \sim (\alpha_n + \beta_n)$.

(ii) This is a special case of (iii), when $(a_{k,n}), (\alpha_{k,n})$ depend only on k.

(iii) When we have two sequences of real numbers (r_n) and (s_n), (s_n) is strictly increasing and unbounded, and we know the convergence of the difference quotient $(r_n - r_{n-1})/(s_n - s_{n-1}) \to c$, then the Stolz-Cesàro theorem (Heuser, 2003, Th. 27.3) ensures that $r_n/s_n \to c$ as well. Take $r_n = \sum_{k=1}^{n} a_{k,n}$ and $s_n = \sum_{k=1}^{n} \alpha_{k,n}$, then

$$\lim_{n\to\infty} \frac{r_n - r_{n-1}}{s_n - s_{n-1}} = \lim_{n\to\infty} \frac{\sum_{k=1}^{n-1}(a_{k,n} - a_{k,n-1}) + a_{n,n}}{\sum_{k=1}^{n-1}(\alpha_{k,n} - \alpha_{k,n-1}) + \alpha_{n,n}} = 1,$$

and we receive the desired result with $c = 1$. A review of the proof reveals that the Stolz-Cesàro theorem even holds if s_n is not strictly monotone right from the start, but only from a certain point on. □

Example. As a counterexample where the conditions of Lemma 2.1 are not fulfilled, consider for (i)

$$\begin{aligned} a_n &= n+1 & \alpha_n &= n \\ b_n &= n & \beta_n &= n + \frac{1}{n}; \end{aligned}$$

clearly $a_n \sim \alpha_n$ and $b_n \sim \beta_n$, but $a_n - b_n = 1$ is not equivalent to $\alpha_n - \beta_n = -\frac{1}{n}$. For (ii), consider any sequence a_n with $\sum_{k=1}^{\infty} a_k < \infty$ and choose a sequence $a'_n < a_n$ with $a'_n = o(a_n)$. Define $\alpha_n := a_n + a'_n$.

Lemma 2.2 (Some properties of slowly varying functions)**.** *Consider a slowly varying function $L(x)$.*
(i)

$$\boxed{L(x+a) \sim L(x) \qquad \text{as } x \to \infty, \text{ for all fixed } a \in \mathbb{R}^+} \tag{2.5}$$

(ii) If L is locally bounded (that means bounded on every compact set) and we have $\gamma_k \sim \frac{c}{\Gamma(1-D)} k^{-D} L(k) =: g_k$ with $D \in (0,1)$ and a constant c, then

$$\boxed{\sum_{k=1}^{n} \gamma_k \sim \frac{c}{\Gamma(2-D)} n^{1-D} L(n).} \tag{2.6}$$

Proof. (i) By Karamata's Representation Theorem (Bingham et al., 1989, Th. 1.3.1), we can write $L(x) = c(x) \exp\left\{\int_d^x \frac{\varepsilon(u)}{u}\, du\right\}$ with an arbitrary $d > 0$ (for instance $d = 1$), $c(\cdot)$ mesurable and $c(x) \to c \in (0, \infty), \varepsilon(x) \to 0$ as $x \to \infty$. We therefore have for large n and an arbitrarily small $\varepsilon > 0$

$$\begin{aligned}\frac{L(x+a)}{L(x)} &= \frac{c(x+a)}{c(x)} \exp\left\{\int_x^{x+a} \frac{\varepsilon(u)}{u}\, du\right\} \\ &\leq \frac{c(x+a)}{c(x)} \exp\left\{\varepsilon \int_x^{x+a} \frac{1}{u}\, du\right\} = \frac{c(x+a)}{c(x)} \left(\frac{x+a}{x}\right)^{\varepsilon} \to 1\end{aligned}$$

and

$$\frac{L(x+a)}{L(x)} \geq \frac{c(x+a)}{c(x)} \exp\left\{-\varepsilon \int_x^{x+a} \frac{1}{u}\, du\right\} = \frac{c(x+a)}{c(x)} \left(\frac{x+a}{x}\right)^{-\varepsilon} \to 1,$$

so all in all $L(x+a)/L(x) \to 1$.

(ii) Define the step function $u(x) := \gamma_n$ for $n \leq x < n+1$. We obtain from the Uniform Convergence Theorem (Bingham et al., 1989, Th. 1.2.1), which provides that $L(\lambda x)/L(x) \to 1$ (as $x \to \infty$) uniformly on each compact set of λ in $(0, \infty)$, that $L(x) \sim L(n)$ for $n \leq x < n+1$ and thus

$$u(x) \sim \frac{c x^{-D}}{\Gamma(1-D)} L(x).$$

Now we know by Karamata's Theorem that asymptotic relations are integrable (Bingham et al., 1989, Prop. 1.5.8):

$$\int_a^x t^{-\beta} L(t)\, dt \sim L(x) \int_a^x t^{-\beta}\, dt \sim \frac{x^{1-\beta}}{1-\beta} L(x) \qquad \text{as } x \to \infty$$

if $L(x)$ is locally bounded in $[a, \infty)$ and $\beta < 1$. So we can figure out the asymptotic behaviour of $\sum_{k=1}^{n} \gamma_k$ as follows:

$$\sum_{k=1}^{n} \gamma_k = \int_1^n u(t)\, dt \sim \frac{c n^{1-D}}{(1-D)\Gamma(1-D)} L(n)$$

□

Technical remarks. (a) As L is defined on the positive real line $(0, \infty)$, $L(x-a)$ may formally not be defined, but we are interested in asymptotic behaviour for $x \to \infty$, so we can admit (fixed) negative a in the lemma as well. If this is desired, change the variables $y = x - a$, and the proof covers $L(x-a) \sim L(x)$ as well.

(b) As long as it is a fixed point, it is irrelevant for the asymptotic behaviour where the sum $\sum_{k=1}^{n} \gamma_k$ starts. Technically we must take care that L is locally bounded on the whole domain of summation. To ensure this, we want L to be locally bounded everywhere. For practical use this is not a serious restriction.

2.2 One divided sample

At first we consider the asymptotic behaviour of the Gauß test for two samples like in (2.1), roughly speaking one LRD time series that is cut into two samples. (Subsequent, we will consider two independent samples which both exhibit the same LRD.)

2.2.1 Asymptotic theory

Theorem 2.3. *Let $(X_i)_{i\geq 1}$ be a stationary Gaussian process with mean 0*[1]*, variance 1 and covariances*

$$\gamma_k = \mathrm{Cov}[X_i, X_{i+k}] \sim \frac{c}{\Gamma(1-D)} k^{-D} L(k) =: g_k$$

with c a constant, $D \in (0,1)$ and L a slowly varying function. Assume that we have a series of N observations which is cut into two pieces:

$$X_1, X_2, \ldots, X_m \quad \textit{and} \quad X_{m+1}, X_{m+2}, \ldots, X_{m+n}$$

with $N = m + n$. It is $m = [\lambda N]$ and $n = [(1-\lambda)N]$ for a $\lambda \in (0,1)$. We call the second sample the Y-sample ($Y_k := X_{m+k}$). Then for the difference of both sample means holds

$$\boxed{\sqrt{\frac{mn}{(m+n)^{2-D} L(m+n)}} \frac{\bar{X} - \bar{Y}}{\sigma_{diff}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)} \tag{2.7}$$

with

$$\begin{aligned}
\sigma^2_{diff} &= \frac{2c}{\Gamma(3-D)} \left(\lambda^{1-D}(1-\lambda) + \lambda(1-\lambda)^{1-D} - 1 + \lambda^{2-D} + (1-\lambda)^{2-D} \right) \\
&= \frac{2c}{\Gamma(3-D)} \left(\lambda^{1-D} + (1-\lambda)^{1-D} - 1 \right).
\end{aligned}$$

Technical remark. a) It is natural to write the limit theorem in terms of the single sample sizes m and n, but this can be misleading, because m and n change with λ, while the above expression erroneously suggests that only the variance σ^2_{diff} is a function of λ. Theorem 2.3 can also be written as:

$$N^{D/2} L(N)^{-1/2} \frac{\bar{X} - \bar{Y}}{\sigma'_{\mathrm{diff}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

with

$$\begin{aligned}
\sigma'^2_{\mathrm{diff}} &= \frac{\sigma^2_{\mathrm{diff}}}{\lambda(1-\lambda)} \\
&= \frac{2c}{\Gamma(3-D)} \frac{\lambda^{1-D}(1-\lambda) + \lambda(1-\lambda)^{1-D} - 1 + \lambda^{2-D} + (1-\lambda)^{2-D}}{\lambda(1-\lambda)} \\
&= \frac{2c}{\Gamma(3-D)} \frac{\left(\lambda^{1-D} + (1-\lambda)^{1-D} - 1\right)}{\lambda(1-\lambda)}.
\end{aligned}$$

[1] In a test problem for a change-point, this corresponds to the null hypothesis that there is no change in the mean of the data: All observations have the same mean, and we can assume that this common mean is 0. Here, we derive the (non-degenerate) asymptotic distribution of the test statistic under this null hypothesis. In contrast, if there is a change in the mean, i.e. if $E[X_i]$ is the same for all i, $E[Y_j]$ is the same for all j, but $E[X_i] \neq E[Y_j]$ for all i, j, then the variance blows up the mean of the statistic as well, so the expression blows up to infinity – which gives notice of the change-point.

This variance σ'^2_{diff} has, in contrast to σ^2_{diff}, the expected property that it has a minimum in $\lambda = 0.5$.

b) It may be advantageous to write σ^2_{diff} and σ'^2_{diff} in the not summarized, longer form: The difference between the case of two independent samples and the case of one divided sample can easily be seen. In the case of two independent samples, three summands of the longer representation of σ^2_{diff} and σ'^2_{diff} are lacking.

The limit behaviour of the Gauß test statistic is a special case of the "difference-of-means" change-point test which we will treat in section 3.4.2 and which has been analyzed by Csörgő and Horváth (1997, Chap. 4.3), so we source the direct proof out to Appendix B (the proof is substantially only calculating the variance of the test statistic, but one has to take into account asymptotic equivalences and slowly varing functions; even though this is laborious, it is down-to-earth analysis, so it gives a feel for LRD).

2.2.2 Simulations

We will now investigate the finite sample performance of the statistic from Theorem 2.3 in a simulation study. To this end, I have simulated $10,000$ time series of fGn[2] with Hurst parameter H (respectively $D = 2-2H$) and length $N = 10, 50, 100, 500, 1000, 2000$. In this model, the auto-covariances are

$$\begin{aligned}\gamma_k &\sim \left(1 - \frac{D}{2}\right)(1-D)k^{-D} \\ &= \frac{c}{\Gamma(1-D)}k^{-D}L(k) \qquad \text{with } L(k) \equiv 1 \text{ and } c = \frac{1}{2}\Gamma(3-D),\end{aligned}$$

so Theorem 2.3 states in this situation that

$$T := \sqrt{\frac{mn}{(m+n)^{2-D}}}\frac{\bar{X} - \bar{Y}}{\sigma_{\text{diff}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

with

$$\sigma^2_{\text{diff}} = \left(\lambda^{1-D}(1-\lambda) + \lambda(1-\lambda)^{1-D} - 1 + \lambda^{2-D} + (1-\lambda)^{2-D}\right).$$

Since T is a linear statistic of X and Y, the convergence is more than a general convergence in distribution: In some sense, only the variance needs to converge to 1, the distribution is always normal. For each of the $10,000$ time series, I have calculated T, and based on these values, I have computed the sample variance of T. All sample variances are pretty close to 1, no matter for which split-up or for which length of the series. The exact results are given in Appendix D.1; the `R`-source code is given in Appendix C.5. In Figure 2.1, the estimated density of T is shown, compared to the standard normal density.

To get a feel for how the variance of $\bar{X} - \bar{Y}$ behaves for different choices of λ and H, we write

$$T := \sqrt{\frac{mn}{(m+n)^{2-D}}}\frac{\bar{X} - \bar{Y}}{\sigma_{\text{diff}}} = \sqrt{\frac{\lambda(1-\lambda)}{N^{-D}}}\frac{\bar{X} - \bar{Y}}{\sigma_{\text{diff}}}$$

[2] See Section 1.3. I have simulated fGn via a routine in the `fArma` package in `R` which uses a fast Fourier Transform, based on the `SPLUS` code presented by Beran (1994).

and then look at the sample variance of the unnormalized statistic $N^{D/2}(\bar{X} - \bar{Y})$ versus its limit $\sigma'_{\text{diff}} = \sigma_{\text{diff}}/\sqrt{\lambda(1-\lambda)}$. This is shown in Figure 2.2. As one expects, the variance is minimal for equal sample sizes ($\lambda = 0.5$) and it is smaller the stronger the dependencies are. This can be explained intuitively: In a not so strongly dependent series, the Y-observations have more freedom to behave different from the X's, but when they are very strong dependend, both samples rather do the same, so the variance is small. The same holds for the sample sizes: If one sample is greater, the smaller one (which is positively correlated) has less strength to countervail the fluctuations of the first one.

2.3 Two independent samples

Now instead of one sample of observations which is cut into two pieces, we consider two single stationary Gaussian processes $(X_i)_{i\geq 1}$ and $(Y_j)_{j\geq 1}$ which are independent of each other (each X_i is independent of any selection of the Y_j's), but which have the same long memory, and we study the performance of the difference of the means in this situation.

2.3.1 Asymptotic theory

Theorem 2.4. *Let $(X_i)_{i\geq 1}$, $(Y_j)_{j\geq 1}$ be two stationary Gaussian processes which are independent of each other, both with mean 0, variance 1 and covariances*

$$\operatorname{Cov}[X_i, X_{i+k}] = \operatorname{Cov}[Y_j, Y_{j+k}] = \gamma_k \sim \frac{c}{\Gamma(1-D)} k^{-D} L(k),$$

where c is a constant, $D \in (0,1)$ and L is a slowly varying function. Assume that we observe $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ and set $m + n = N$ and $\lambda = m/N$, $1 - \lambda = n/N$. In this situation, for the difference of both sample means holds

$$\boxed{\sqrt{\frac{mn}{(m+n)^{2-D} L(m+n)}} \frac{\bar{X} - \bar{Y}}{\sigma_{diff,2}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)} \tag{2.8}$$

with

$$\sigma^2_{diff,2} := \frac{2c}{\Gamma(3-D)} \left(\lambda^{1-D}(1-\lambda) + \lambda(1-\lambda)^{1-D} \right).$$

Proof. The X_i and Y_j are commonly Gaussian with mean 0, the numerator $\bar{X} - \bar{Y}$ is an affine linear transformation and therefore Gaussian with mean 0 as well. We analyse its asymptotic variance

$$\begin{aligned} \operatorname{Var}[\bar{X} - \bar{Y}] &= \frac{1}{m^2} \operatorname{Var}\left[\sum_{i=1}^{m} X_i\right] + \frac{1}{n^2} \operatorname{Var}\left[\sum_{j=1}^{n} Y_j\right] - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \operatorname{Cov}[X_i, Y_j] \\ &= \frac{1}{m^2} \operatorname{Var}\left[\sum_{i=1}^{m} X_i\right] + \frac{1}{n^2} \operatorname{Var}\left[\sum_{j=1}^{n} Y_j\right], \end{aligned}$$

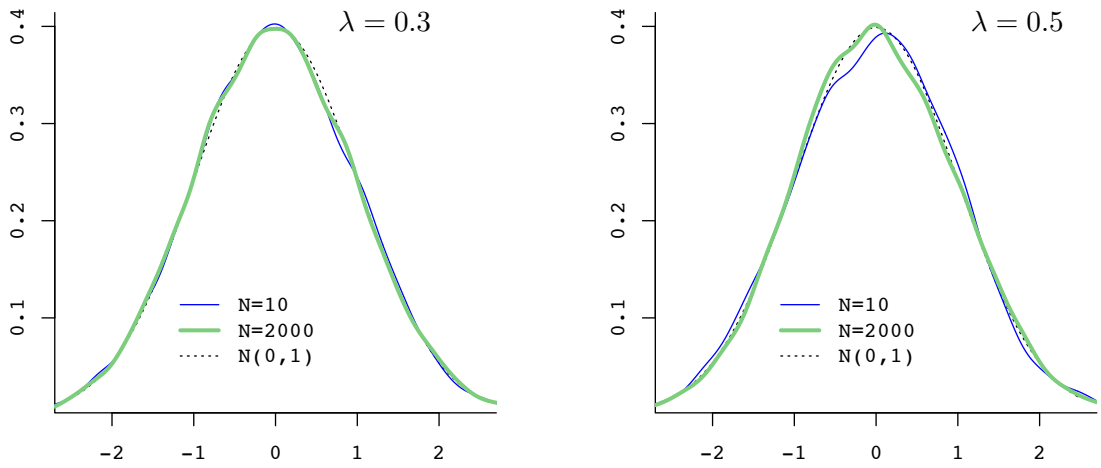


Figure 2.1: Density of $\bar{X} - \bar{Y}$, scaled and normalized. fGn with $k = 10,000$ and $H = 0.7$ ($D = 0.6$).

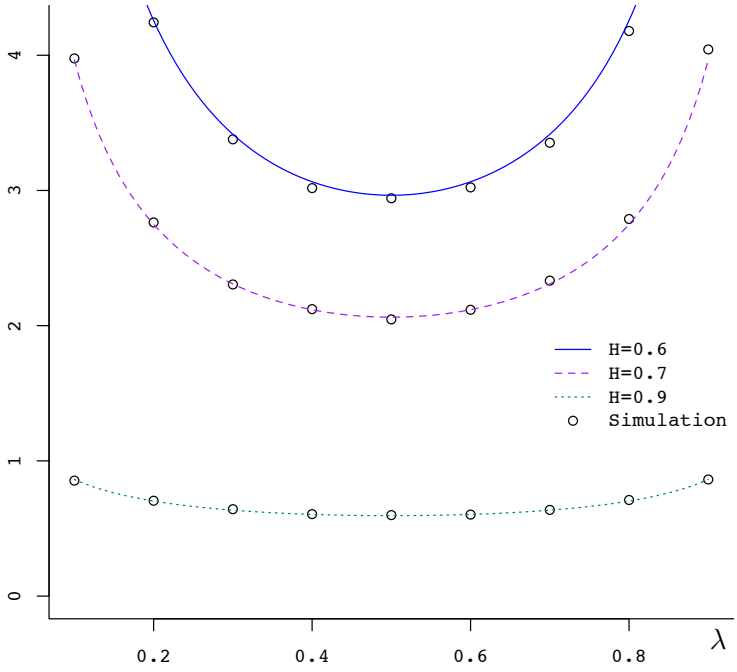


Figure 2.2: Unnormalized, but scaled variance of $\bar{X} - \bar{Y}$. fGn with $k = 10,000$ and $N = 2000$ (lines: asymptotic variance, points: simulation results).

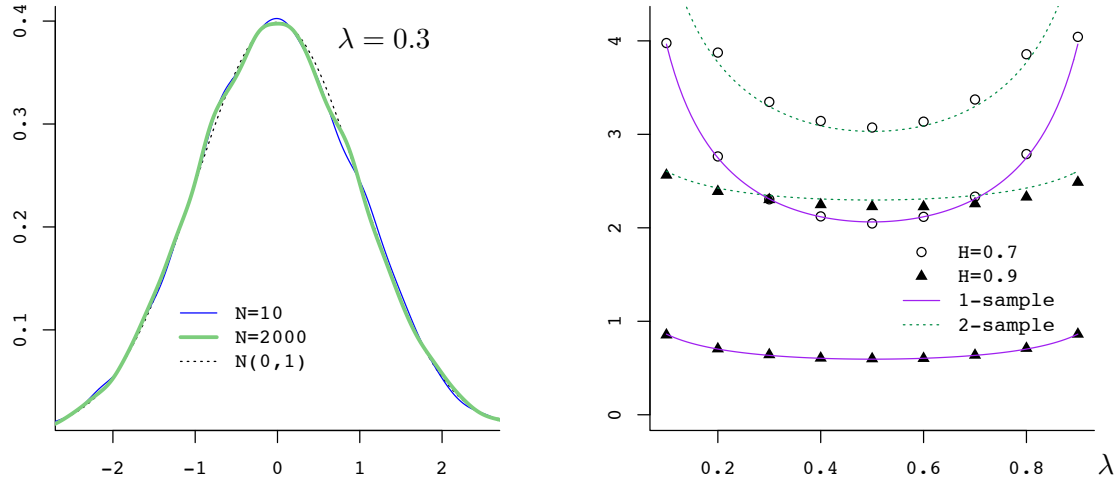


Figure 2.3: Estimated density of the $\bar{X} - \bar{Y}$-statistic in two-sample case (left), unnormalized variance of $\bar{X} - \bar{Y}$ for the one-sample and the two-sample case (right; lines: asymptotic variance, points: simulation results). fGn with $k = 10,000$, $H = 0.7$ (left), $N = 2000$ (right).

and so we obtain, picking up results from the proof before:

$$\begin{aligned}
&\lim_{N\to\infty} \frac{mn}{(m+n)^{2-D}} \frac{1}{L(m+n)} \operatorname{Var}\left[\bar{X} - \bar{Y}\right] \\
&\quad = \lim_{N\to\infty} \frac{mn}{(m+n)^{2-D}} \frac{1}{L(m+n)} \left[\frac{1}{m^2} m\gamma_0 + \frac{1}{m^2} 2cm^{2-D} \frac{L(m)}{\Gamma(3-D)} + \frac{1}{n^2} n\gamma_0 \right. \\
&\qquad \left. + \frac{1}{n^2} 2cn^{2-D} \frac{L(n)}{\Gamma(3-D)} \right] \\
&\quad = \lim_{N\to\infty} \frac{\lambda(1-\lambda)}{N^{-D} L(N)} \left[\frac{\gamma_0}{\lambda N} + 2c\lambda^{-D} N^{-D} \frac{L(\lambda N)}{\Gamma(3-D)} + \frac{\gamma_0}{(1-\lambda)N} \right. \\
&\qquad \left. + 2c(1-\lambda)^{-D} N^{-D} \frac{L((1-\lambda)N)}{\Gamma(3-D)} \right] \\
&\quad = \frac{2c}{\Gamma(3-D)} \lambda(1-\lambda) \left(\lambda^{-D} + (1-\lambda)^{-D} \right)
\end{aligned}$$

□

2.3.2 Simulations

I have simulated 10,000 pairs of fGn time series. Theorem 2.4 states

$$T := \sqrt{\frac{mn}{(m+n)^{2-D}}} \frac{\bar{X} - \bar{Y}}{\sigma_{\text{diff},2}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

with

$$\sigma^2_{\text{diff},2} := \frac{2c}{\Gamma(3-D)} \left(\lambda^{1-D}(1-\lambda) + \lambda(1-\lambda)^{1-D} \right).$$

As before, all sample variances are pretty close to their asymptotic value of 1, even for small und unequal sample sizes. The exact results are given in Appendix D.1, while Figure 2.3 provides an overview. The R-source code of the simulation is given in Appendix C.6.

When we plot the unnormalized variance of $\bar{X} - \bar{Y}$ as in the simulations before, we see that the variance is smaller in the one-sample case. Intuitively, this is not a surprise: Both samples are mutually positively correlated, so they behave similar and their difference does not fluctuate that much as in the case of two independent fGn samples.

2.4 Estimating the variance of $\bar{X}$

In a next step we want to estimate $\mathrm{Var}[\bar{X} - \bar{Y}]$ in Theorem 2.3. For a start we concentrate on the case when N observations $X_1, \ldots, X_N$ are given and we want to estimate $\mathrm{Var}[\bar{X}]$. (We still we consider a stationary Gaussian process $(X_i)_{i\geq 1}$ with $E[X_i] = 0$, $E[X_i^2] = 1$ and auto-covariance function (1.1).) The problem here is that we want to estimate the variance of a variable that we observe only one single time. So we enforce several observations by an artifical segmenting of the observed data: We divide the original time series $X_1, \ldots, X_N$ into N/r blocks of length r (for simplicity of notation we assume that N/r is an integer) and take the sample mean over each block, i.e.

$$\boxed{X_k^{(r)} := \frac{1}{r} \sum_{i=(k-1)r+1}^{kr} X_i, \qquad 1 \leq k \leq N/r.}$$

These sample means $X_k^{(r)}$ define new variables, and we can estimate their variance by the standard sample variance

$$\boxed{\widehat{\mathrm{Var}}\left[X^{(r)}\right] := \widehat{\mathrm{Var}}\left[X_k^{(r)}\right] := \frac{1}{N/r} \sum_{k=1}^{N/r} \left(X_k^{(r)}\right)^2 - \left(\frac{1}{N/r} \sum_{k=1}^{N/r} X_k^{(r)}\right)^2.} \tag{2.9}$$

This sample variance is an estimator of $\mathrm{Var}[X_k^{(r)}]$ and it is widely used as an estimate of the Hurst exponent $H = 1 - D/2$, respectively as an estimate to detect long range dependence in a time series, known as the *aggregated variance method*, see for example Teverovsky and Taqqu (1997).

2.4.1 Asymptotic behaviour of the variance-of-mean estimator

The covariances $\gamma_{i,j} = E[X_i X_j]$ depend only on the lag $|i - j|$, thus

$$\sum_{\substack{i,j=(k-1)r+1 \\ i\neq j}}^{kr} \gamma_{i,j} = 2 \sum_{\substack{i,j=(k-1)r+1 \\ i<j}}^{kr} \gamma_{i,j} = 2 \sum_{i=1}^{r} (r-i)\gamma_i,$$

and so

$$E\left[\left(X_k^{(r)}\right)^2\right] = E\left[\frac{1}{r^2}\sum_{i=(k-1)r+1}^{kr} X_i^2 + \frac{1}{r^2}\sum_{\substack{i,j=(k-1)r+1 \\ i\neq j}}^{kr} X_i X_j\right] = \frac{1}{r}E\left[X_1^2\right] + \frac{2}{r^2}\sum_{i=1}^{r}(r-i)\gamma_i. \tag{2.10}$$

Furthermore for $l > k$

$$\begin{aligned}
E\left[X_k^{(r)} X_l^{(r)}\right] &= E\left[\frac{1}{r^2}\sum_{i=(k-1)r+1}^{kr} X_i \cdot \sum_{j=(l-1)r+1}^{lr} X_j\right] \\
&= \frac{1}{r^2}\sum_{i=(k-1)r+1}^{kr}\sum_{j=(l-1)r+1}^{lr}\gamma_{i,j} = \frac{1}{r^2}\sum_{i=1}^{r}\sum_{j=((l-1)-(k-1))r+1}^{(l-k+1)r}\gamma_{i,j} \\
&= \frac{1}{r^2}\sum_{i=1}^{r}\sum_{j=(l-k)r+1-i}^{(l-k+1)r-i}\gamma_j \\
&= \frac{1}{r^2}\left(\sum_{j=0}^{r-1}\gamma_{j+(l-k)r} + \sum_{j=-1}^{r-2}\gamma_{j+(l-k)r} + \ldots + \sum_{j=1-r}^{0}\gamma_{j+(l-k)r}\right) \\
&= \frac{1}{r^2}\left(\sum_{i=0}^{r-1}(r-i)\gamma_{i+(l-k)r} + \sum_{i=1}^{r-1}(r-i)\gamma_{-i+(l-k)r}\right),
\end{aligned}$$

and so

$$\begin{aligned}
E\left[\frac{1}{N/r}\sum_{k=1}^{N/r} X_k^{(r)}\right]^2 &= \frac{1}{(N/r)^2}\left\{\sum_{k=1}^{N/r} E\left[X_k^{(r)}\right]^2 + 2\sum_{k<l} E\left[X_k^{(r)} X_l^{(r)}\right]\right\} \\
&= \frac{1}{(N/r)^2}\left\{\frac{N}{r}\left(\frac{1}{r}E\left[X_1^2\right] + \frac{2}{r^2}\sum_{i=1}^{r}(r-i)\gamma_i\right)\right. \\
&\quad \left. + \frac{2}{r^2}\sum_{k<l}\left(\sum_{i=0}^{r-1}(r-i)\gamma_{i+(l-k)r} + \sum_{i=1}^{r-1}(r-i)\gamma_{-i+(l-k)r}\right)\right\} \\
&= \frac{1}{(N/r)^2}\left\{\frac{N}{r}\left(\frac{1}{r}E\left[X_1^2\right] + \frac{2}{r^2}\sum_{i=1}^{r}(r-i)\gamma_i\right)\right. \\
&\quad \left. + \frac{2}{r^2}\sum_{p=1}^{N/r-1}\left(\frac{N}{r} - p\right)\left(\sum_{i=0}^{r-1}(r-i)\gamma_{i+pr} + \sum_{i=1}^{r-1}(r-i)\gamma_{-i+pr}\right)\right\},
\end{aligned} \tag{2.11}$$

where we have used for the last transformation that all summands in the sum over $k < l$ only depend on the difference $p := l - k$. All in all, from (2.10) and (2.11) we obtain

$$\begin{aligned} E\left[\widehat{\mathrm{Var}}\left[X^{(r)}\right]\right] &= E\left[\frac{1}{N/r}\sum_{k=1}^{N/r}\left(X_k^{(r)}\right)^2 - \left(\frac{1}{N/r}\sum_{k=1}^{N/r}X_k^{(r)}\right)^2\right] \\ &= \frac{1}{r}\sigma^2 + \frac{2}{r^2}\sum_{i=1}^{r}(r-i)\gamma_i - \frac{1}{(N/r)^2}\left\{\frac{N}{r}\left(\frac{1}{r}\sigma^2 + \frac{2}{r^2}\sum_{i=1}^{r}(r-i)\gamma_i\right)\right. \\ &\quad \left. + \frac{2}{r^2}\sum_{p=1}^{N/r-1}\left(\frac{N}{r}-p\right)\left(\sum_{i=0}^{r-1}(r-i)\gamma_{i+pr} + \sum_{i=1}^{r-1}(r-i)\gamma_{-i+pr}\right)\right\}, \end{aligned} \tag{2.12}$$

whose limiting behaviour we will now investigate.

Lemma 2.5. *With an appropriate scaling, $\widehat{\mathrm{Var}}[X^{(r)}]$ is an asymptotically unbiased estimator for $\mathrm{Var}\left[\bar{X}_N\right]$: When $r, N \to \infty$ such that $r = o(N)$, then*

$$\boxed{E\left[\left(\frac{r}{N}\right)^D \frac{L(N)}{L(r)}\widehat{\mathrm{Var}}\left[X^{(r)}\right]\right] \sim \mathrm{Var}\left[\bar{X}_N\right].}$$

Proof. By a short standard transformation we see

$$\mathrm{Var}\left[\bar{X}_r\right] = \frac{1}{r}\sigma^2 + \frac{2}{r^2}\sum_{i=1}^{r}(r-i)\gamma_i,$$

and by (1.8)

$$\mathrm{Var}\left[\bar{X}_r\right] \sim \frac{2}{(1-D)(2-D)}r^{-D}L(r)$$

(keep in mind that all limiting processes in this proof are meant to happen as $r, N \to \infty$ with $r = o(N)$); c_γ is a constant factor. So the first two summands in (2.12), scaled with $(r/N)^D$, behave like this:

$$\left(\frac{r}{N}\right)^D \frac{L(N)}{L(r)}\mathrm{Var}\left[\bar{X}_r\right] \sim \left(\frac{r}{N}\right)^D \frac{L(N)}{L(r)}\frac{2}{(1-D)(2-D)}r^{-D}L(r) \sim \mathrm{Var}\left[\bar{X}_N\right]. \tag{2.13}$$

Similary, we have for the third summand

$$\begin{aligned} \left(\frac{r}{N}\right)^D \frac{L(N)}{L(r)}\frac{1}{(N/r)^2}\frac{N}{r}\left(\frac{1}{r}\sigma^2 + \frac{2}{r^2}\sum_{i=1}^{r}(r-i)\gamma_i\right) &= \frac{r^{1+D}}{N^{1+D}}\frac{L(N)}{L(r)}\mathrm{Var}\left[\bar{X}_r\right] \\ &= O\left(\frac{rL(N)}{N^{1+D}}\right) \to 0. \end{aligned} \tag{2.14}$$

The last summand in (2.12) is bounded by a term that vanishes asymptotically, and so it must vanish, too (we are summing up positive covariances with positive weights, so the natural lower bound is 0).

$$\begin{aligned}
&\frac{2}{r^2}\sum_{p=1}^{N/r-1}\left(\frac{N}{r}-p\right)\left(\sum_{i=0}^{r-1}(r-i)\gamma_{i+pr}+\sum_{i=1}^{r-1}(r-i)\gamma_{-i+pr}\right)\\
&\quad=\frac{2}{r^2}\sum_{p=1}^{N/r}\left(\frac{N}{r}-p\right)\left(\sum_{i=-(r-1)}^{r-1}(r-|i|)\gamma_{i+pr}\right)\\
&\quad\le\frac{2}{r^2}\sum_{p=1}^{N/r}\frac{N}{r}\left(\sum_{i=-(r-1)}^{r-1}r\gamma_{i+pr}\right)\le\frac{2N}{r^2}\sum_{p=1}^{N/r}\left(\sum_{i=r(p-1)+1}^{r(p+1)}\gamma_i\right)=\frac{N^2}{r^3}o(r),
\end{aligned}$$

because $\gamma_r \to 0$ as $r \to \infty$, and so by the Stolz-Cesàro theorem $\frac{1}{r}\sum_{i=r(p-1)+1}^{r(p+1)}\gamma_i \to 0$ uniformly in p, respectively $\sum_{i=r(p-1)+1}^{r(p+1)}\gamma_i = o(r)$ independent of the value of p, which we will now demonstrate. By Lemma 2.1

$$\begin{aligned}
\sum_{i=r(p-1)+1}^{r(p+1)}\gamma_i &= \sum_{i=1}^{r(p+1)}\gamma_i-\sum_{i=1}^{r(p-1)}\gamma_i\\
&\sim c\left((r(p+1))^{1-D}L(r(p+1))-(r(p-1))^{1-D}L(r(p-1))\right)\\
&\le c' r^{1-D}L(r)
\end{aligned}$$

since $((p-1)/(p+1))^{1-D} \neq 1$ and $\left((p+1)^{1-D}-(p-1)^{1-D}\right) \le 2^{1-D}$ for all $p \ge 1$. Finally, note that with prefactor $(N/r)^{-2}$ and scaling $(r/N)^D L(N)/L(r)$, the last summand in (2.12) converges to 0, and the lemma is proved. □

2.4.2 Simulations

To see how well $\widehat{\mathrm{Var}}[X^{(r)}]$, as defined in (2.9) and appropriately scaled, estimates $\mathrm{Var}[\bar{X}_N]$ for finite sample sizes N, we take a look at some simulations. I have simulated a time series $X_1, \ldots, X_N$ of fGn with Hurst parameter $H = 1 - D/2$. To this time series I have applied the estimator (2.9) for different block sizes[3] r: Relative block sizes of $1/50 \cdot N$, $1/10 \cdot N$ and $1/5 \cdot N$, fixed block sizes of 10, 50 and 100 and block sizes $r = N^\beta$ for $\beta = 0.1$, 0.3, 0.5, 0.7 and 0.9. Each of these simulation was repeated 10,000 times and the results were averaged.

[3]The relative and the fixed block sizes do not fulfil the requirements of the Lemma in the previous section: $r, N \to \infty$ with $r = o(N)$. However, in practice we will not have infinitely many data, but one single fixed sample size N, and then of course one chooses one single block length r – by guess, by taking a fraction or root or whatever –, so it is definitely interesting to see how the estimator performs with these kinds of block sizes.

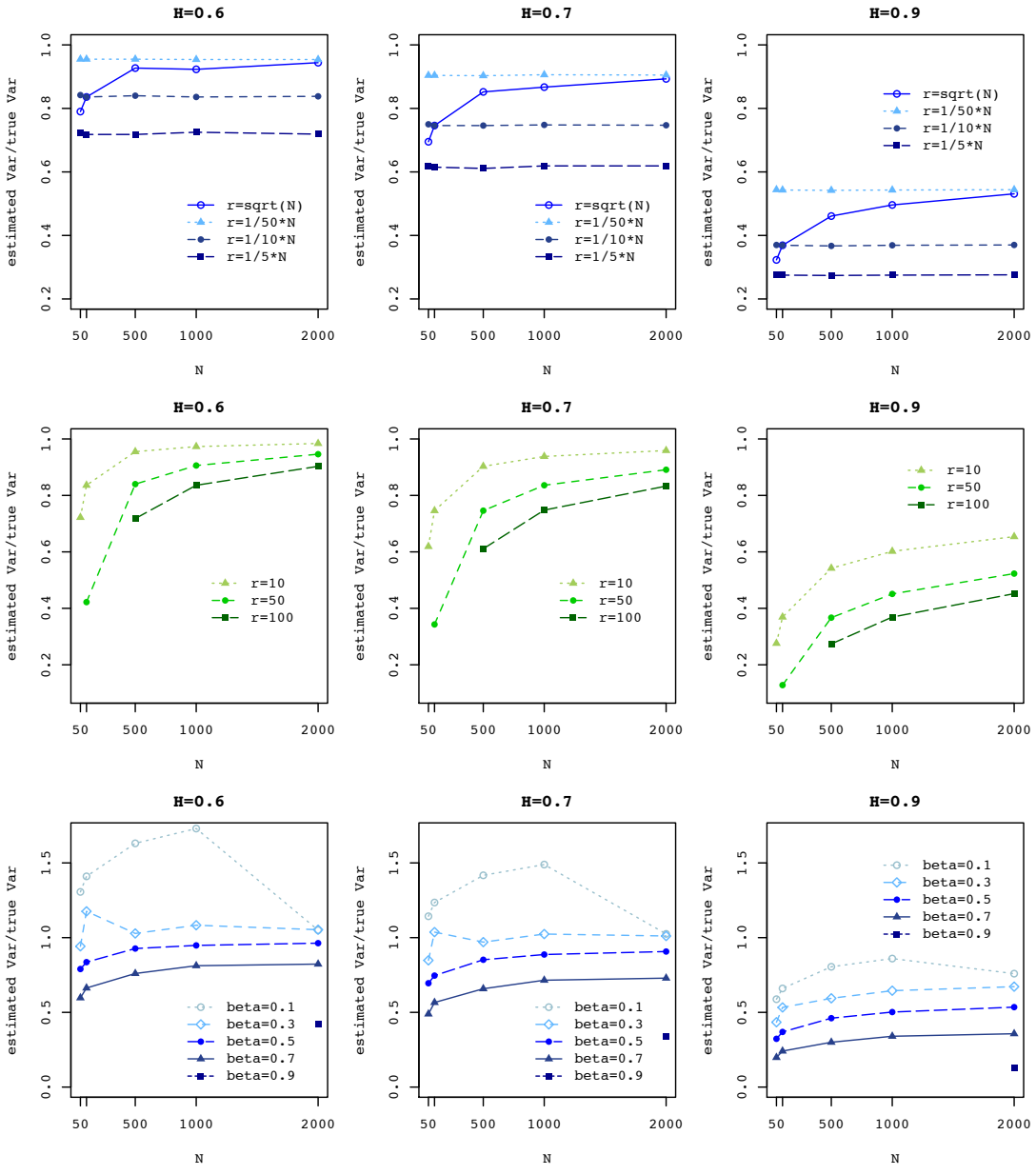


Figure 2.4: Relative simulation results $\widehat{\mathrm{Var}}[X^{(r)}]/\mathrm{Var}[\bar{X}_N]$ of variance estimation of $\bar{X}$, each value averaged over 10,000 simulations, for different Hurst parameters H, sample sizes N and block size r (top: relative, middle: fixed, bottom: $r = N^\beta$).

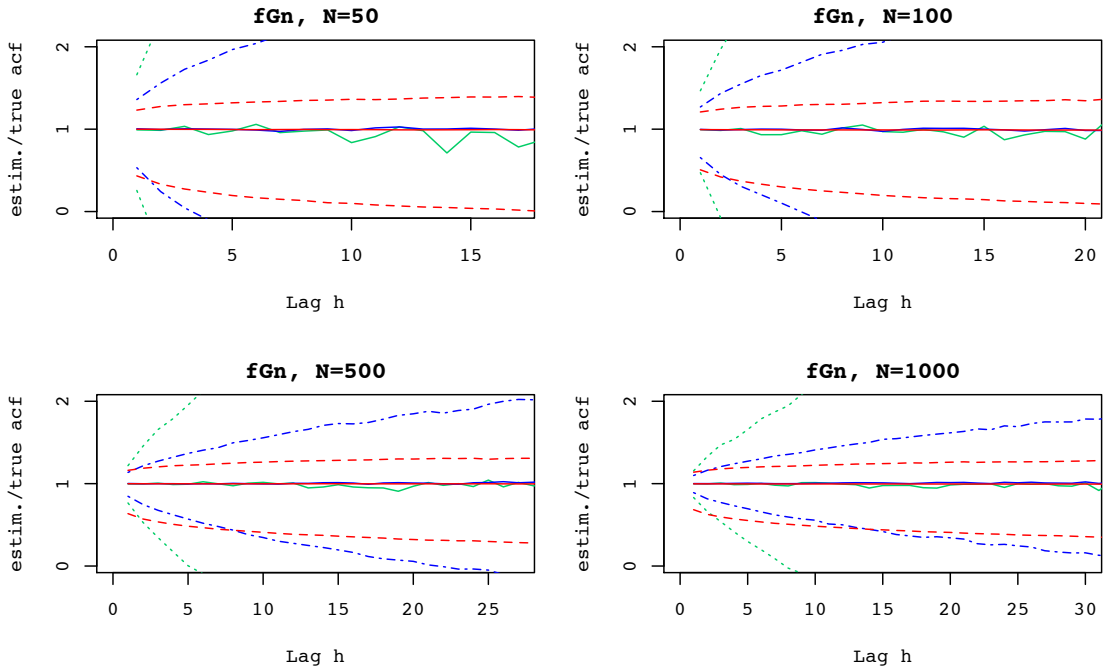


Figure 2.5: Auto-covariance estimator $\widehat{\gamma}_h$, average and confidence belts (lower and upper quartile) based on 10,000 repetitions, for different sample sizes and different Hurst parameters $H = 0.6$ (· · ·), $H = 0.7$ (– · –), $H = 0.9$ (– – –).

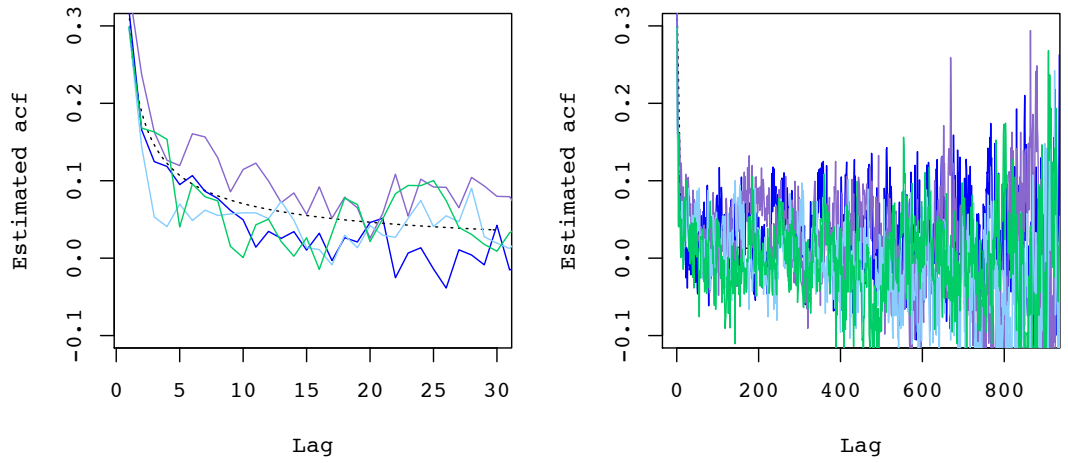


Figure 2.6: Auto-covariance estimator $\widehat{\gamma}_h$, based on four different realizations of fGn ($N = 1000$, $H = 0.7$), compared to the true auto-covariance (dotted). Clearly, the estimator gets worse for larger lags.

The simulation results, i.e. the estimated variance of $\bar{X}$ relative to the true variance[4], $\widehat{\mathrm{Var}}[X^{(r)}]/\mathrm{Var}[\bar{X}_N]$, are shown in Figure 2.4. The exact simulation results are given in Appendix D.2.

Some values have not been computed because the number of blocks was smaller than 2 or equal or greater than N, which does not make sense. Apparently, relative block sizes do not perform better with increasing sample size N. As one expects, dividing the sample in only 5 blocks yields a very poor result. It gets better when the number of blocks is 50.

If the blocks are of size $\sqrt{N}$, the estimator becomes better with increasing sample sizes (from $N = 50$ to $N = 2000$ it gains about 20 percentage points).

Fixed block sizes (the second table of each trio) get better with increasing sample size N as well. In ordinary dimension (a few hundred observations), a block size of $r = 10$ is quite respectable.

Polynomial growing block sizes also gain accuracy when N grows. While block sizes $r = N^{0.1}$ are unpractical (they tend to overestimate the true variances, in parts with some ten percent), $r = N^{0.3}$ seems to be reliable: For all sample sizes it gives a good approximation to the true variance.

Let us finally look how the estimators handle different levels of long memory. All types of block sizes yield poorer results when H increases. For $H = 0.9$, the estimated variance is alarmingly often less than 50% of the true variance. So estimating the variance of the mean by block techniques is obviously the more difficult, the stronger the dependencies are.

2.5 Estimating the auto-covariance

In the following, we will need some higher and mixed moments. They can be calculated via the diagram formula, see e.g. Major (1981a, Cor. 5.5) or Surgailis (2003).

Lemma 2.6 (Diagram formula)**.** *For zero mean Gaussian random variables $X_1, \ldots, X_p$, $p \geq 2$, with $E[X_j^2] = 1$ and $\gamma_{i,j} := E[X_i X_j]$, $i, j = 1, \ldots, p$ it holds*

$$E\left[H_{k_1}(X_1) \cdots H_{k_p}(X_p)\right] = \begin{cases} \frac{k_1! \cdots k_p!}{2^q q!} {\sum}' \gamma_{i_1,j_1} \cdots \gamma_{i_q,j_q} & k_1 + \ldots + k_p = 2q \text{ and} \\ & 0 \leq k_1, \ldots, k_q \leq q \\ 0 & \text{otherwise} \end{cases},$$

where ${\sum}'$ denotes a sum over all indices $i_1, j_1, \cdots i_q, j_q$ with

- $i_1, j_1, \ldots, i_q, j_q \in \{1, 2, \ldots, p\}$
- $i_1 \neq j_1, \ldots, i_q \neq j_q$

[4]Since the auto-covariances γ_k of fGn are known, see (1.4), $\mathrm{Var}[\bar{X}_N] = n\gamma_0 + 2\sum_{k=1}^{N-1}(N-k)\gamma_k$ can be calculated.

- *k_1 indices are 1, k_2 indices are 2, ..., k_p indices are p.*

Example. We will consider some examples in order to illustrate the use of the diagram formula and to prepare the upcoming calculations.

1. Let i, j, k, l be pairwise different indices.

$$E[X_i X_j X_k X_l] = \frac{1}{8}\left(8\gamma_{i,j}\gamma_{k,l} + 8\gamma_{i,k}\gamma_{j,l} + 8\gamma_{i,l}\gamma_{j,k}\right), \tag{2.15}$$

respectively in the special case when $i \neq j$, $k = i + h \neq j$ and $l = j + h \neq i$ for $h > 0$

$$E[X_i X_j X_{i+h} X_{j+h}] = \frac{1}{2}\left(\gamma_{i,j}^2 + \gamma_h^2 + \gamma_{i,j+h}\gamma_{j,i+h}\right).$$

$$\begin{aligned} E[X_i^2 X_j^2] &= E[(X_i^2 - 1)(X_j^2 - 1)] + E\left[X_i^2\right] + E\left[X_j^2\right] - 1 \\ &= E\left[H_2(X_i)H_2(X_j)\right] + 1 = \frac{1}{2}4\gamma_{i,j}^2 + 1 \end{aligned} \tag{2.16}$$

$$\begin{aligned} E[X_i X_j X_k^2] &= E\left[H_1(X_i)H_1(X_j)H_2(X_k)\right] + E\left[X_i X_j\right] \\ &= \frac{1}{2^2}8\gamma_{i,k}\gamma_{j,k} + \gamma_{i,j} = 2\gamma_{i,k}\gamma_{j,k} + \gamma_{i,j} \end{aligned} \tag{2.17}$$

2. Consider Hermite polynomials H_k, $k \in \mathbb{N}$.

$$E\left[H_{k_i}(X_i)H_{k_j}(X_j)\right] = \begin{cases} \frac{k_i!k_j!}{2^q q!}\sum{}' \gamma_{i_1,j_1}\cdots\gamma_{i_q,j_q} & k_i + k_j = 2q \text{ and} \\ & 0 \leq k_i, k_j \leq q \\ 0 & \text{otherwise.} \end{cases}$$

If $k_i \neq k_j$, $E[H_{k_i}(X_i)H_{k_j}(X_j)] = 0$, since if $k_i + k_j$ are even, i.e. $k_i + k_j = 2q$ for a number $q \in \mathbb{N}$, then either k_i, k_j are both even or both are odd. Let w.l.o.g. $k_i > k_j$, then $k_i = k_j + 2k$, $k \in \mathbb{N}$. It follows $k_i + k_j = 2k_j + 2k = 2q$, thus $k_j + k = q$, and that means $k_i = k_j + 2k = q + k > q$.

If $k_i = k_j =: k$, we have $q = k$ and

$$E\left[H_{k_i}(X_i)H_{k_j}(X_j)\right] = E\left[H_k(X_i)H_k(X_j)\right] = \frac{k!}{2^k}\sum{}' \gamma_{i_1,j_1}\cdots\gamma_{i_q,j_q},$$

where $\sum'$ denotes a sum over all indices $i_1, j_1, \cdots i_q, j_q$ such that k_1 indices are 1, k_2 indices are 2, ..., k_p indices are p, in other words k indices are i and k indices are j. Since the indices must be pairwise different, $i_1 \neq j_1, \ldots, i_q \neq j_q$, the sum is $\sum'\gamma_{i,j}\cdots\gamma_{i,j}$ and each summand consists of k factors. How many such summands are possible? Each factor is $\gamma_{i,j}$ or $\gamma_{j,i}$, so each factor has two possibilities, thus there are 2^k possible summands in $\sum'$. So we obtain

$$E\left[H_{k_i}(X_i)H_{k_j}(X_j)\right] = k!\gamma_{i,j}^k.$$

We have just established (1.6).

3. We will now verify (1.7) as well, applying the just proved formula (1.6):

$$E\left[\sum_{i=1}^{N} H_k(X_i)\right]^2 = E\left[\sum_{i=1}^{N} H_k(X_i)H_k(X_i)\right] + E\left[\sum_{i\neq j}^{N} H_k(X_i)H_k(X_j)\right]$$
$$= Nk! + \sum_{i\neq j}^{N} k!\gamma_{i,j}^k,$$

which yields (1.7) with $\gamma_k = \gamma_{i,i+k} = \mathrm{Cov}[X_i, X_{i+k}] = k^{-D}L(k)$.

2.5.1 Asymptotic behaviour of the auto-covariance estimator

We will use the diagram formula to prove

Theorem 2.7. *The standard estimator for the auto-covariances* $\gamma_h = \mathrm{Cov}\,[X_i, X_{i+h}]$

$$\boxed{\widehat{\gamma}_h = \frac{1}{N-h}\sum_{i=1}^{N-h} X_i X_{i+h}} \tag{2.18}$$

is asymptotically consistent for all fixed $h \in \mathbb{N}$ *as* $N \to \infty$*:*

$$\boxed{E\,[\widehat{\gamma}_h] = \gamma_h, \quad \mathrm{Var}\,[\widehat{\gamma}_h] \to 0}$$

This is comforting to know.

Proof. Unbiasedness is immediate. The asymptotic zero variance requires some more effort. First consider the case $h = 0$. Then

$$\mathrm{Var}[\widehat{\gamma_0}] = \frac{1}{N^2}\,\mathrm{Var}\left[\sum_{i=1}^{N} X_i^2\right] = \frac{1}{N^2}\,\mathrm{Var}\left[\sum_{i=1}^{N} H_2(X_i)\right]$$

and this converges to 0 due to (1.8). If $h > 0$, we set $M := N - h$. We will show that

$$\mathrm{Var}\,[\widehat{\gamma}_h] \leq O(M^{-1}) + O(M^{-D}L(M)), \tag{2.19}$$

so that $\mathrm{Var}\,[\widehat{\gamma}_h] \to 0$ not uniformly in h as $N \to \infty$, but in some sense uniformly over certain sets of values for N, h. We will make use of this in a later proof. Now write

$$\mathrm{Var}\,[\widehat{\gamma}_h] = \frac{1}{M^2}\left(E\left[\sum_{i=1}^{M} X_i X_{i+h}\right]^2 - (M\gamma_h)^2\right)$$
$$= \frac{1}{M^2}\left(\sum_{i=1}^{M} E\left[X_i^2 X_{i+h}^2\right] + 2\sum_{i<j}^{M} E\left[X_i X_{i+h} X_j X_{j+h}\right]\right) - \gamma_h^2. \tag{2.20}$$

The first sum is bounded by $O(M)$. We will now show that the second sum, scaled with M^{-2}, converges to γ_h^2. In order to apply the diagram formula, we have to distinguish two cases: $i + h = j$ and $i + h \neq j$.

$$2\sum_{i<j}^{M} E\left[X_i X_{i+h} X_j X_{j+h}\right] = 2\sum_{i=1}^{M-h} E\left[X_i X_j^2 X_{j+h}\right] + 2\sum_{i<j, j\neq i+h}^{M} E\left[X_i X_{i+h} X_j X_{j+h}\right]$$
$$= O(M) + 2\sum_{k=1,k\neq h}^{M-1} (M-k)(\gamma_k^2 + \gamma_h^2 + \gamma_{k+h}\gamma_{|k-h|}),$$

where in the last step we have used that the covariances in the sum over $i < j$ depend only on the lag $k = j - i$, and below the diagonal $\{i = j\}$ there are $(M-k)$ indices (i, j) with lag k. Now $\gamma_k = E[X_i X_{i+k}] \leq (E[X_i^2]E[X_{i+k}^2])^{1/2} = 1$, and so we can roughly estimate

$$\sum_{k=1,k\neq h}^{M-1} (M-k)\gamma_k^2 \leq \sum_{k=1}^{M} (M-k)\gamma_k \sim c \cdot M^{2-D} L(M)$$

where c is a constant, due to (B.6) which follows from Karamata's Theorem, similar to (2.6). The same upper bound holds for $\sum_{k=1,k\neq h}^{M-1}(M-k)\gamma_{k+h}\gamma_{|k-h|}$, and so both sums are $o(M^2)$. Finally note that

$$2\sum_{k=1,k\neq h}^{M-1} (M-k) \leq 2\sum_{k=1}^{M} (M-k) = 2\left(M^2 - \frac{M(M+1)}{2}\right),$$

thus

$$\mathrm{Var}\left[\widehat{\gamma}_h\right] \leq O(M^{-1}) + O(M^{-D}L(M)) + \left|2\left(1 - \frac{M(M+1)}{2M^2}\right) - 1\right|,$$

which proves (2.19). □

2.5.2 Simulations

By Theorem 2.7, the standard auto-covariance estimator

$$\widehat{\gamma}_h = \frac{1}{N-h}\sum_{i=1}^{N-h} X_i X_{i+h},$$

as defined in (2.18), is asymptotically unbiased. We will now investigate its performance in finite sample situations by a simulation study. To this end, we have simulated samples of fGn of length N (with $N = 50, 100, 500, 1000$). fGn has the exact auto-covariances (1.4). For each sample we have calculated the estimated auto-covariances $\widehat{\gamma}_h$, $h = 1, \ldots, N-1$. Each simulation was repeated 10,000 times; so based on these 10,000 repetitions, for each lag h, we have calculated an average value of $\widehat{\gamma}_h$ and the lower and upper quartile. The results are presented in Figure 2.5, where the simulation results are shown, relative to the true auto-covariance (1.4).

The larger the sample size N is, the less fluctuates the average and the narrower is the confidence belt. The estimation is better for larger LRD parameters H, but it tends to underestimate the true value. The maximum lag at which to calculate the auto-covariance was restricted to

$$h_{\max} := \min(10 \log_{10}(N), N - 1), \tag{2.21}$$

where N is the number of observations (this is the default setting for estimating auto-covariances in `R`), since the estimator gets worse for large lags h, because the larger the lag, the less data is available on which the estimation is based. This is illustrated by Figure 2.6 where $\widehat{\gamma}_h$ is shown, based on different realizations of fGn.

2.6 Estimating the variance of $\bar{X} - \bar{Y}$

Now we are prepared to estimate the variance of $\bar{X}_m - \bar{Y}_n$ in the situation of Theorem 2.3. We will estimate

$$\operatorname{Var}\left[\bar{X}_m - \bar{Y}_n\right] = \operatorname{Var}\left[\bar{X}_m\right] + \operatorname{Var}\left[\bar{Y}_n\right] - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=m+1-i}^{m+n-i} \gamma_j,$$

by

$$\boxed{\begin{aligned}\widehat{\operatorname{Var}}\left[\bar{X}_m - \bar{Y}_n\right] := &\left(\frac{r}{m}\right)^D \frac{L(m)}{L(r)} \widehat{\operatorname{Var}}\left[X^{(r)}\right] + \left(\frac{r}{n}\right)^D \frac{L(n)}{L(r)} \widehat{\operatorname{Var}}\left[Y^{(r)}\right] \\ &- \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=m+1-i}^{m+n-i} \widehat{\gamma}_j,\end{aligned}} \tag{2.22}$$

where $\widehat{\operatorname{Var}}\left[X^{(r)}\right]$ and $\widehat{\operatorname{Var}}\left[Y^{(r)}\right]$ are defined in (2.9) and $\widehat{\gamma}_j$ is defined in (2.18).

2.6.1 Asymptotic behaviour of the variance estimator for $\bar{X} - \bar{Y}$

Theorem 2.8. *The estimator* (2.22) *is asymptotically unbiased:*

$$\boxed{E\left[\widehat{\operatorname{Var}}\left[\bar{X}_m - \bar{Y}_n\right]\right] \sim \operatorname{Var}\left[\bar{X}_m - \bar{Y}_n\right]}$$

with $m = [\lambda N]$, $n = [(1 - \lambda)N]$, $\lambda \in (0, 1)$, *and* $r = o(N)$ *as* $r, N \to \infty$.

Proof. We already know from the preceding theorems that

$$\begin{aligned} E\left[\left(\frac{r}{m}\right)^D \frac{L(m)}{L(r)} \widehat{\operatorname{Var}}\left[X^{(r)}\right]\right] &\sim \operatorname{Var}\left[\bar{X}_m\right] \\ E\left[\left(\frac{r}{n}\right)^D \frac{L(n)}{L(r)} \widehat{\operatorname{Var}}\left[Y^{(r)}\right]\right] &\sim \operatorname{Var}\left[\bar{Y}_n\right] \\ E\left[\frac{2}{mn} \sum_{i=1}^{m} \sum_{j=m+1-i}^{m+n-i} \widehat{\gamma}_j\right] &= \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=m+1-i}^{m+n-i} \gamma_j, \end{aligned}$$

so in order to establish the first statement of the theorem, we only have to show the asymptotic equivalence of the sum of the expressions on the left hand side and of the sum of the expressions on the right-hand side. We employ Lemma 2.1 twice. At first, if $\lambda \neq 1/2$

$$\frac{\operatorname{Var}\left[\bar{X}_m\right]}{\operatorname{Var}\left[\bar{Y}_n\right]} \sim \frac{m^{-D}}{n^{-D}} = \left(\frac{1-\lambda}{\lambda}\right)^D = \left(\frac{1}{\lambda} - 1\right)^D \neq 1$$

by (1.8), like on page 35, so we obtain

$$E\left[\left(\frac{r}{m}\right)^D \frac{L(m)}{L(r)} \widehat{\operatorname{Var}}\left[X^{(r)}\right] + \left(\frac{r}{n}\right)^D \frac{L(n)}{L(r)} \widehat{\operatorname{Var}}\left[Y^{(r)}\right]\right] \sim \operatorname{Var}\left[\bar{X}_m\right] + \operatorname{Var}\left[\bar{Y}_n\right]. \quad (2.23)$$

If $\lambda = 1/2$, the sample of observations is cut in half, and because of the stationarity of the process $(X_i)_{i\geq 1}$, both halfs $(X_1, \ldots, X_{[N/2]})$ and $(X_{[N/2]+1}, \ldots, X_N)$ have the same asymptotic probabilistic properties. So

$$E\left[\left(\frac{r}{n}\right)^D \frac{L(n)}{L(r)} \widehat{\operatorname{Var}}\left[Y^{(r)}\right]\right] \sim E\left[\left(\frac{r}{m}\right)^D \frac{L(m)}{L(r)} \widehat{\operatorname{Var}}\left[X^{(r)}\right]\right]$$
$$\operatorname{Var}\left[\bar{X}_m\right] \sim \operatorname{Var}\left[\bar{Y}_n\right],$$

and (2.23) holds as well.

Second, we obtain the equivalence of all three summands (recall (B.5) on page 198); we have

$$\begin{aligned}
&\lim_{N\to\infty} \frac{\operatorname{Var}\left[\bar{X}_m\right] + \operatorname{Var}\left[\bar{Y}_n\right]}{\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=m+1-i}^{m+n-i}\gamma_j} \\
&\quad = \lim_{N\to\infty} \frac{2C(m^{-D}L(m) + n^{-D}L(n))}{\frac{2}{nm}N^{2-D}CL(N)\cdot(1-\lambda^{2-D}-(1-\lambda)^{2-D})} \\
&\quad = \frac{\lambda^{-D} + (1-\lambda)^{-D}}{\frac{1}{\lambda(1-\lambda)}(1-\lambda^{2-D}-(1-\lambda)^{2-D})},
\end{aligned}$$

because $L(\lambda N)/L(N) \to 1$ for any $\lambda \in (0,1)$. We will now show that this always exceeds 1, such that the condition of Lemma 2.1 is fulfilled. $\lambda^{2-D} > \lambda^2$ and $\lambda^{-D} > 1$ for all $\lambda \in (0,1)$, thus

$$\lim_{N\to\infty} \frac{\operatorname{Var}\left[\bar{X}_m\right] + \operatorname{Var}\left[\bar{Y}_n\right]}{\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=m+1-i}^{m+n-i}\gamma_j} > \frac{2\lambda(1-\lambda)}{1-\lambda^2-(1-\lambda)^2} = 1.$$

□

2.6.2 Simulations

We have just shown that the estimator $\widehat{\operatorname{Var}}\left[\bar{X}_m - \bar{Y}_n\right]$ for $\operatorname{Var}\left[\bar{X}_m - \bar{Y}_n\right]$, as defined in (2.22), is asymptotically unbiased. We will now take a look at its finite sample behaviour. For this purpose, we have simulated N observations of standard fGn with Hurst parameter H (for $N = 500, 1000$ and $H = 0.7, 0.9$) and divided the sample

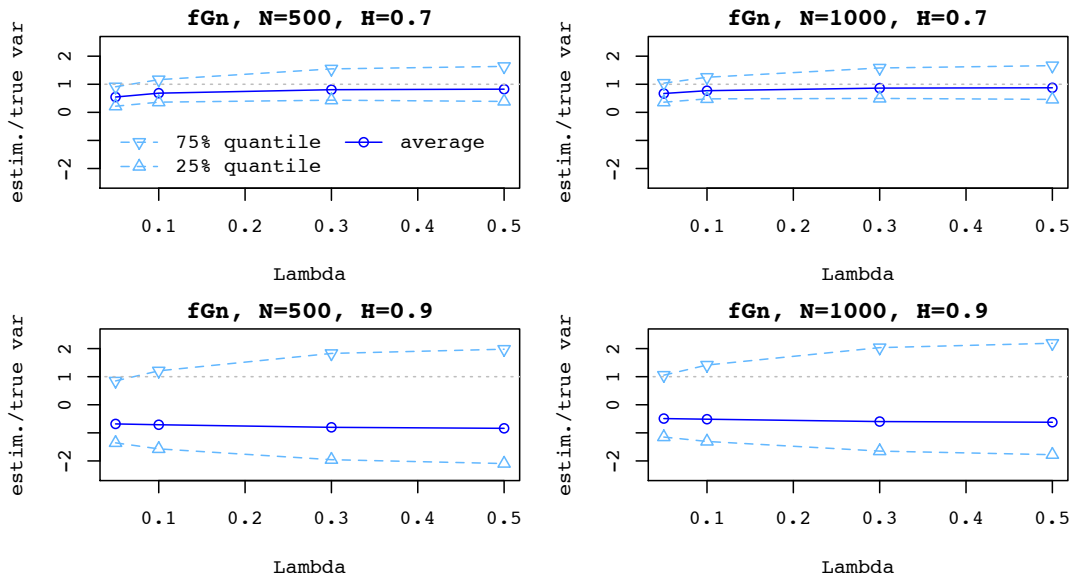


Figure 2.7: $\widehat{\mathrm{Var}}\left[\bar{X}_m - \bar{Y}_n\right] / \mathrm{Var}[\bar{X}_m - \bar{Y}_n]$ for different sample lengths N and LRD parameters N, based on the usual auto-covariance estimator $\widehat{\gamma}_h$.

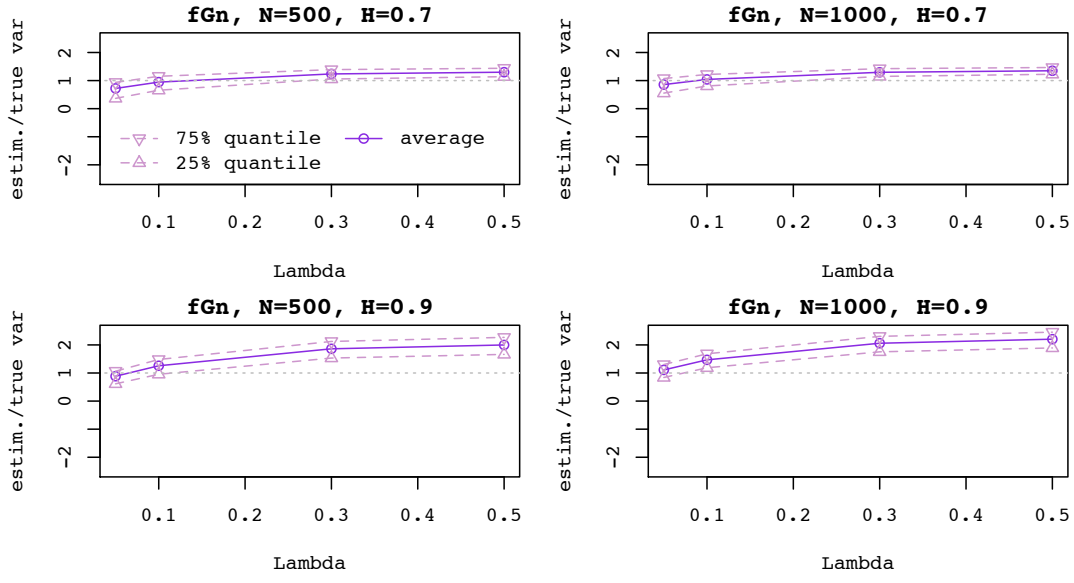


Figure 2.8: $\widehat{\mathrm{Var}}\left[\bar{X}_m - \bar{Y}_n\right] / \mathrm{Var}[\bar{X}_m - \bar{Y}_n]$ for different sample lengths N and LRD parameters N, based on a trimmed auto-covariance estimator $\widehat{\gamma}_{h,\mathrm{trim}}$ (the usual auto-covariance estimator $\widehat{\gamma}_h$, but set to 0 for large lags).

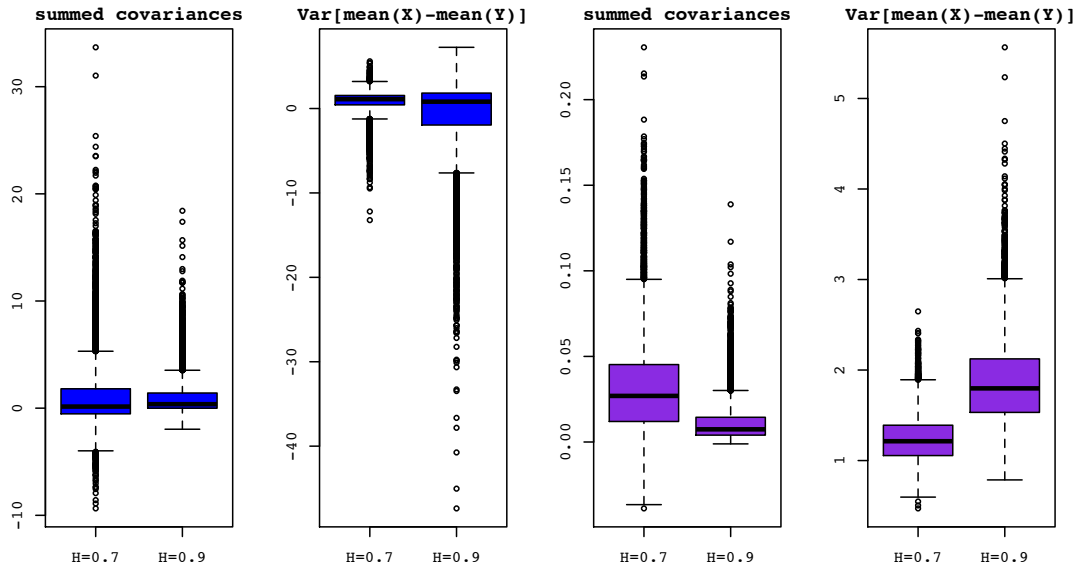


Figure 2.9: Relative estimation $\sum \widehat{\gamma}_h / \sum \gamma_h$ (1) and the resulting estimation $\widehat{\mathrm{Var}}\left[\bar{X}_m - \bar{Y}_n\right] / \mathrm{Var}[\bar{X}_m - \bar{Y}_n]$ (2) for fGn with split point $\lambda = 0.3$. $\mathrm{Var}[\bar{X}_m]$ and $\mathrm{Var}[\bar{Y}_n]$ have been estimated with $\widehat{\mathrm{Var}}[X^{(r)}]$, using polynomial block size with $\beta = 0.3$. (3) and (4) show the same, but $\widehat{\gamma}_{h,\mathrm{trim}}$ was used instead of $\widehat{\gamma}_h$.

after the $[\lambda N]$-th observation (for $\lambda = 0.05, 0.1, 0.3, 0.5$, i.e. for sample size $N = 500$ after observation $m = 25, 50, 150, 250$ and for sample size $N = 1000$ after observation $m = 50, 100, 300, 500$). In our simulation, we have compared these simulations of $\widehat{\mathrm{Var}}\left[\bar{X}_m - \bar{Y}_n\right]$ with the respective true variance

$$\begin{aligned}
\mathrm{Var}\left[\bar{X}_m - \bar{Y}_n\right] &= \frac{1}{m^2}\,\mathrm{Var}\left[\sum_{i=1}^{m} X_i\right] + \frac{1}{n^2}\,\mathrm{Var}\left[\sum_{j=1}^{n} Y_j\right] - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=m+1-i}^{N-i}\gamma_j \\
&= \frac{1}{m^2}\left(m + 2\sum_{k=1}^{m-1}(m-k)\gamma_k\right) + \frac{1}{n^2}\left(n + 2\sum_{k=1}^{n-1}(n-k)\gamma_k\right) \\
&\quad - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=m+1-i}^{N-i}\gamma_j,
\end{aligned}$$

where the γ_k, the exact auto-covariances of fGn, are known and thus can be calculated, see (1.4). The result of this comparison is shown in Figure 2.7.

$\widehat{\mathrm{Var}}\left[\bar{X}_m - \bar{Y}_n\right]$ tends to underestimate the true variance $\mathrm{Var}[\bar{X}_m - \bar{Y}_n]$, for $H = 0.9$ much more than for $H = 0.7$. This can be explained as follows: For $H = 0.7$, the estimation of $\mathrm{Var}[\bar{X}_m]$ and $\mathrm{Var}[\bar{Y}_n]$ is reliable (remember Figure 2.4), so this is due to an overestimation of $\sum \gamma_h$ by $\sum \widehat{\gamma}_h$. The left half of Figure 2.9 shows the effect of summing up the single estimations $\widehat{\gamma}_h$: For a certain set of parameters, it displays the boxplot of $\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=m+1-i}^{m+n-i}\widehat{\gamma}_j$ relative to the true value $\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=m+1-i}^{m+n-i}\gamma_j$ (first plot, each for $H = 0.7, 0.9$) and the resulting estimate $\widehat{\mathrm{Var}}\left[\bar{X}_m - \bar{Y}_n\right]$ relative to

the true value $\mathrm{Var}\left[\bar{X}_m - \bar{Y}_n\right]$ (second plot, each for $H = 0.7, 0.9$). Indeed, we observe a huge variance for $\sum \widehat{\gamma}_h$ with more upper outliers than lower ones. When $H = 0.9$, then $\mathrm{Var}[\bar{X}_m]$ and $\mathrm{Var}[\bar{Y}_n]$ are underestimated (as shown in Figure 2.4), and moreover, as Figure 2.9 demonstrates, $\sum \widehat{\gamma}_h$ heavily overestimates $\sum \gamma_h$, and we have only upper outliers. In consequence, recall the definition (2.22), $\widehat{\mathrm{Var}}\left[\bar{X}_m - \bar{Y}_n\right]$ gets much too small.

An improvement can simply be obtained by adjusting the auto-covariance estimator $\widehat{\gamma}_h$ which is involved in $\widehat{\mathrm{Var}}\left[\bar{X}_m - \bar{Y}_n\right]$ and which is bad for large lags, as we have just illustrated in Figure 2.6; a natural idea is to trimm it. If we estimate the auto-covariances γ_h by

$$\widehat{\gamma}_{h,\mathrm{trim}} = \begin{cases} \widehat{\gamma}_h & \text{if } h \leq h_{\max} \\ 0 & \text{else} \end{cases},$$

with $h_{\max}$ as in (2.21), and use this as auto-covariance estimator in $\widehat{\mathrm{Var}}\left[\bar{X}_m - \bar{Y}_n\right]$, the estimation gets better: As shown in Figure 2.8, the confidence belts become narrower, and especially for $H = 0.9$, the average gets closer to the true variance (even though it now exceeds it). The right half of Figure 2.9 confirms that trimming improves the estimation of $\sum \gamma_h$ by $\sum \widehat{\gamma}_h$: Now, we do not obtain this broad range of outliers, and as a consequence $\widehat{\mathrm{Var}}\left[\bar{X}_m - \bar{Y}_n\right]$ yields better results.

Chapter 3

A "Wilcoxon-type" change-point test

In this chapter we develop a non-parametric change-point test for changes in the mean of LRD processes that have a representation as an instantaneous functional of a stationary Gaussian process. We aim at a test which is based on the Wilcoxon two-sample test statistic, roughly speaking a "Wilcoxon-type" test.

Wilcoxon's rank statistic can be represented as a U-statistic, so a natural approach is to rely on the work of Dehling and Taqqu (1989) who derived a limit theorem for one-sample U-statistics of LRD data. Unfortunately, the technical requirements forbid to extend the method as a whole to the Wilcoxon two-sample test statistic; the crucial point is that the kernel of the U-statistic has to have bounded variation, but $h(x, y) = I_{\{x \leq y\}}$, which is the kernel that leads to Wilcoxon's two-sample test statistic, does not have this property.

In this chapter, we will thus make a different approach: We represent the Wilcoxon two-sample test statistic as a functional of the one-dimensional empirical process and apply to this the limit theorem of Dehling and Taqqu (1989). This part of this work is based on the article of Dehling, Rooch and Taqqu (2012).

In what follows, we define our test and derive its asymptotic distribution under the null hypothesis that no change occured. In a subsequent simulation study, we also compare its power with the power of a test which is based on differences of means in a simulation study. We will see that for Gaussian data, the non-parametric change-point test has only a slightly smaller power, while for heavy-tailed data it outperforms the "difference-of-means" test. Motivated by these observations, we will compare the power of both tests analytically; this is carried out in the next chapter.

3.1 Setting the scene

We consider observations $(X_i)_{i\geq 1}$ of the type

$$\boxed{X_i = \mu_i + \epsilon_i,}$$

where $(\mu_i)_{i\geq 1}$ are unknown means and $(\epsilon_i)_{i\geq 1}$ is an instantaneous functional of a stationary Gaussian process with non-summable covariances, i.e.

$$\boxed{\epsilon_i = G(\xi_i), \quad i \geq 1,}$$

where $(\xi_i)_{i\geq 1}$ is a mean-zero Gaussian process with $E(\xi_i^2) = 1$ and long-range dependence, that is, with auto-covariance function (1.1). G is a measurable transformation which fulfills some technical requirements.

Definition 3.1 (Centralized/normalized $L^p(\mathbb{R})$-functions)**.** Let $\xi \sim \mathcal{N}(0,1)$ be a standard normal random variable. We define

$$\mathcal{G}^1 = \mathcal{G}^1(\mathbb{R},\mathcal{N}) := \{G : \mathbb{R} \to \mathbb{R} \text{ measurable} \mid E[G(\xi)] = 0\} \subset L^1(\mathbb{R},\mathcal{N}),$$

the class of (with respect to the standard normal measure) centralized and integrable functions, and

$$\mathcal{G}^2 = \mathcal{G}^2(\mathbb{R},\mathcal{N}) := \{G : \mathbb{R} \to \mathbb{R} \text{ measurable} \mid E[G(\xi)] = 0, E[G^2(\xi)] = 1\} \subset L^2(\mathbb{R},\mathcal{N}),$$

the class of (with respect to the standard normal measure) normalized and square-integrable functions.

Any transformation $G : \mathbb{R} \to \mathbb{R}$ which is measurable with mean zero and finite variance under standard normal measure can be normalized by dividing the standard deviation, so it can be considered as a function in $\mathcal{G}^2$. In the following, G is either in $\mathcal{G}^1$ or in $\mathcal{G}^2$. So after all, the observations $(X_i)_{i\geq 1}$ are always assumed to be an LRD stationary process with mean zero.

Based on the observations $X_1, \ldots, X_n$, we wish to test the hypothesis

$$\boxed{H : \mu_1 = \ldots = \mu_n} \tag{3.1}$$

that there is no change in the means of the data against the alternative

$$\boxed{A : \mu_1 = \ldots = \mu_k \neq \mu_{k+1} = \ldots = \mu_n \text{ for some } k \in \{1, \ldots, n-1\}} \tag{3.2}$$

that there is an index after which the level shifts. We shall refer to this test problem as (H, A).

A usual approach to change-point tests is to start with a two-sample problem where the change-point is known; in this case, for a given $k \in \{1, \ldots, n-1\}$, the alternative is

$$\boxed{A_k : \mu_1 = \ldots = \mu_k \neq \mu_{k+1} = \ldots = \mu_n.}$$

For such a test problem (H, A_k), a commonly used non-parametric test is the Wilcoxon two-sample rank test. It rejects for large and small values of the test statistic

$$W_{k,n} = \sum_{i=1}^{k} \sum_{j=k+1}^{n} I_{\{X_i \leq X_j\}}$$

which counts the number of times the second part of the sample exceeds the first part of the sample. This is the Mann-Whitney representation of Wilcoxon's rank test statistic as a U-statistic (Lehmann, 1975). From this two-sample problem (H, A_k) with a known change-point, one passes over to the change-point problem (H, A), where one does not know k, by considering a functional of the vector of test statistics $W_{1,n}, \ldots, W_{n-1,n}$. A common procedure is to reject the null hypothesis for large values of $W_n = \max_{k=1,\ldots,n-1} |W_{k,n}|$. This is what we will do (but of course, other functions of $W_{1,n}, \ldots, W_{n-1,n}$ are also possible).

In order to set the critical values of the test, we need to know the asymptotic distribution of W_n under the null hypothesis of no change (the exact distribution is hard to obtain, so we have to settle for an asymptotic test), so we want to obtain the asymptotic distribution for a large sample size and consider the process

$$W_{[n\lambda],n} = \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} I_{\{X_i \leq X_j\}}, \quad 0 \leq \lambda \leq 1,$$

parametrized by λ. After centering and scaling, where F denotes the c.d.f. of the $X_i = G(\xi_i)$ and d_n is defined in (1.12), we obtain the process

$$\boxed{W_n(\lambda) = \frac{1}{n\, d_n} \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} \left(I_{\{X_i \leq X_j\}} - \int_{\mathbb{R}} F(x) dF(x) \right), \quad 0 \leq \lambda \leq 1.} \tag{3.3}$$

Technical remark. The centering constant $\int_{\mathbb{R}} F(x)dF(x)$ equals $E[I_{\{X_1 \leq X_1'\}}]$, where X_1' is an independent copy of X_1. It is the proper normalization as the dependence between X_i and X_j vanishes asymptotically when $|j - i| \to \infty$. Of course, this is only an heuristic argument, but in the course of the proof we will see that it is just the right centering constant. If the distribution function F is continuous, $\int_{\mathbb{R}} F(x)dF(x) = \frac{1}{2}$.

Under the null hypothesis of no change in the mean, the distribution of $W_n(\lambda)$ does not depend on the common mean $\mu := \mu_1 = \ldots = \mu_n$, so we may assume without loss of generality that $\mu = 0$ and hence that $X_i = G(\xi_i)$.

3.2 The limit distribution under the null hypothesis

Our aim is to analyse the asymptotic distribution of the process $(W_n(\lambda))_{0 \leq \lambda \leq 1}$. Since it has the representation of a U-statistic, a natural idea to handle it is to extend the

method of Dehling and Taqqu (1989) who treat one-sample U-statistics of LRD data, but as mentioned in the introduction of the chapter, this approach fails due to the technical requirements: The kernel $h(x,y) = I_{\{x<y\}}$ does not have bounded variation (we prove this in few words on page 210; an overview about the concept of bounded variation in higher dimensions, which may be needed for this, is given in Appendix B.3). Nevertheless, to analyse the asymptotic distribution of the process $(W_n(\lambda))_{0\le\lambda\le1}$, we can apply the empirical process invariance principle of Dehling and Taqqu (1989) to the sequence $(G(\xi_i))_{i\ge1}$, and this approach succeeds.

We consider the Hermite expansion

$$I_{\{G(\xi_i)\le x\}} - F(x) = \sum_{q=1}^{\infty} \frac{J_q(x)}{q!} H_q(\xi_i),$$

where H_q is the q-th order Hermite polynomial and where $J_q(x) = E[H_q(\xi_i) I_{\{G(\xi_i)\le x\}}]$ is the belonging Hermite coefficient (depending on x).

Definition 3.2 (Hermite rank of class of functions)**.** We define the *Hermite rank of the class of functions* $\{I_{\{G(\xi_i)\le x\}} - F(x),\, x\in\mathbb{R}\}$ by

$$m := \min\{q \ge 1 : J_q(x) \ne 0 \text{ for some } x \in \mathbb{R}\}. \tag{3.4}$$

Theorem 3.1. *Suppose that* $(\xi_i)_{i\ge1}$ *is a stationary Gaussian process with mean zero, variance* 1 *and auto-covariance function* (1.1) *with* $0 < D < \frac{1}{m}$*. For* $G \in \mathcal{G}^1$*, we define*

$$X_k = G(\xi_k).$$

Assume that X_k *has a continuous distribution function* F*. Let* m *denote the Hermite rank of the class of functions* $I_{\{G(\xi_i)\le x\}} - F(x)$*,* $x\in\mathbb{R}$*, as defined in* (3.4)*, and let* $d_n > 0$ *satisfy* (1.12)*. Then*

$$\boxed{\frac{1}{n\,d_n} \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} \left(I_{\{X_i\le X_j\}} - \int_{\mathbb{R}} F(x)dF(x) \right), \quad 0\le\lambda\le1,} \tag{3.5}$$

converges in distribution towards the process

$$\boxed{\frac{1}{m!}\left(Z_m(\lambda) - \lambda Z_m(1)\right) \int_{\mathbb{R}} J_m(x)dF(x), \quad 0\le\lambda\le1.}$$

Details to the process $(Z_m(\lambda))_{\lambda\ge0}$ are given in Section 1.4.2 and in the paper of Dehling and Taqqu (1989). The process is a so called *Hermite process*, and it is self-similar with parameter[1]

$$H = 1 - \frac{mD}{2} \in \left(\frac{1}{2}, 1\right).$$

$Z_m(\lambda)$ is not Gaussian when $m \ge 2$. When $m = 1$, $Z_1(\lambda)$ is the standard Gaussian fBm $B_H(\lambda)$.

[1] Do not confuse the index H, which is called *Hurst parameter*, with the other H's used in this work to denote for example *hypothesis* and *Hermite polynomial*.

Proof. We introduce the empirical distribution functions of the first k and the last $(n-k)$ observations, respectively:

$$F_k(x) = \frac{1}{k}\sum_{i=1}^{k} I_{\{X_i \le x\}}$$
$$F_{k+1,n}(x) = \frac{1}{n-k}\sum_{i=k+1}^{n} I_{\{X_i \le x\}}$$

Thus we have $[n\lambda]F_{[n\lambda]}(x) = \sum_{i=1}^{[n\lambda]} I_{\{X_i \le x\}}$ and $(n-[n\lambda])F_{[n\lambda]+1,n}(x) = \sum_{i=[n\lambda]+1}^{n} I_{\{X_i \le x\}}$, and we obtain the following representation:

$$\begin{aligned}
&\sum_{i=1}^{[n\lambda]}\sum_{j=[n\lambda]+1}^{n}\left(I_{\{X_i \le X_j\}} - \int_{\mathbb{R}} F(x)dF(x)\right) \qquad (3.6)\\
&\quad = [n\lambda]\sum_{j=[n\lambda]+1}^{n}\left(F_{[n\lambda]}(X_j) - \int_{\mathbb{R}} F(x)dF(x)\right)\\
&\quad = [n\lambda](n-[n\lambda])\left(\int_{\mathbb{R}} F_{[n\lambda]}(x)dF_{[n\lambda]+1,n}(x) - \int_{\mathbb{R}} F(x)dF(x)\right)\\
&\quad = [n\lambda](n-[n\lambda])\int_{\mathbb{R}}(F_{[n\lambda]}(x) - F(x))dF_{[n\lambda]+1,n}(x)\\
&\qquad + [n\lambda](n-[n\lambda])\int_{\mathbb{R}} F(x)d(F_{[n\lambda]+1,n} - F)(x)\\
&\quad = [n\lambda](n-[n\lambda])\int_{\mathbb{R}}(F_{[n\lambda]}(x) - F(x))dF_{[n\lambda]+1,n}(x)\\
&\qquad - [n\lambda](n-[n\lambda])\int_{\mathbb{R}}(F_{[n\lambda]+1,n}(x) - F(x))dF(x).
\end{aligned}$$

In the final step, we have used integration by parts, namely $\int_{\mathbb{R}} G\,dF = 1 - \int_{\mathbb{R}} F\,dG$, if F and G are two distribution functions. We now apply the empirical process non-central limit theorem of Dehling and Taqqu (1989) which states that

$$\left(d_n^{-1}[n\lambda](F_{[n\lambda]}(x) - F(x))\right)_{x\in[-\infty,\infty],\,\lambda\in[0,1]} \xrightarrow{\mathcal{D}} (J(x)Z(\lambda))_{x\in[-\infty,\infty],\,\lambda\in[0,1]}, \qquad (3.7)$$

where

$$J(x) = J_m(x) \quad \text{and} \quad Z(\lambda) = Z_m(\lambda)/m!.$$

By the Dudley-Wichura version of Skorohod's representation theorem (Shorack and Wellner, 1986, Th. 2.3.4) we may assume without loss of generality that convergence holds almost surely with respect to the supremum norm on the function space $D([0,1]\times[-\infty,\infty])$, i.e.

$$\sup_{\lambda,x}\left|d_n^{-1}[n\lambda](F_{n\lambda}(x) - F(x)) - J(x)Z(\lambda)\right| \longrightarrow 0 \quad a.s.. \qquad (3.8)$$

Technical remark. The theorem states: If we have weak convergence, we can find an equivalent process which converges almost surely. In detail, let (M,d) denote a metric space and let $\mathcal{M}_d^B$ be the σ-field generated by the collection of all open balls $B = \{y \mid d(x,y) < r\} \subset M$ for some $x \in M$ and some $r > 0$. Now if we have weak convergence of a random process $X_n \to_w X_0$ on $(M, \mathcal{M}_d^B, d)$, then we can find equivalent variables $X_n =_{\mathcal{D}} X_n'$ (which may live on another probability space) which converge almost surely, as long as the limit variable is concentrated on a separable subset of M: $d(X_n', X_0') \to_{\text{a.s.}} 0$ as $n \to \infty$, as long as $P_0(M_s) = 1$ for a $\mathcal{M}_d^B$-measurable subset $M_s \subset M$ that is d-seperable, i.e. that has a countable subset d-dense in M.

First, notice that the equivalent process may have another common distribution. But we are only interested in weak convergence of the process, so we do not need to care about it.

Second, the convergence in (3.7) happens in $D((-\infty,\infty)\times[0,1])$ which is not separable, but the limit $J(x)Z(\lambda)$ can be regarded as an element of $C((-\infty,\infty)\times[0,1])$ which is. To see this, note that $Z(\lambda)$ can be assumed to be a version with continuous sample paths, and that $J(x)$ does not have jumps as long as X_k has a continuous c.d.f. $F(t) = P(G(Y) \le t)$:

$$J(t) = \int I_{\{G(y)\le t\}} H_q(y)\, d\Phi(y) = \int_{\{y|G(y)\le t\}} H_q(y)\, dP_Y(y) = \int_{\{\omega|G(Y(\omega))\le t\}} H_q(Y(\omega))\, dP(\omega),$$

so the domain of integration (more precisely: its measure under P) changes continuously with t, and H_q is smooth.

(3.8) is a result on the e.d.f. of the first $[\lambda n]$ observations. We can deduce a similar result on the e.d.f. of the remaining $n - [\lambda n]$ observations $X_{[n\lambda]+1}, \ldots, X_n$:

$$\sup_{\lambda,x} \left| d_n^{-1}(n-[n\lambda])(F_{[n\lambda]+1,n}(x) - F(x)) - J(x)(Z(1) - Z(\lambda)) \right| \longrightarrow 0 \quad a.s.. \tag{3.9}$$

To see this, consider the empirical processes of all n and of the initial $[\lambda n]$ observations,

$$n\left(F_n(x) - F(x)\right) = \sum_{i=1}^{n} \left(I_{\{X_i \le x\}} - F(x)\right)$$
$$[n\lambda]\left(F_{[n\lambda]}(x) - F(x)\right) = \sum_{i=1}^{[n\lambda]} \left(I_{\{X_i \le x\}} - F(x)\right).$$

By definition, we have

$$(n-[n\lambda])\left(F_{[n\lambda]+1,n}(x) - F(x)\right) = \sum_{i=[n\lambda]+1}^{n} \left(I_{\{X_i \le x\}} - F(x)\right),$$

and this equals (take the difference on both sides between the two lines above)

$$n\left(F_n(x) - F(x)\right) - [n\lambda]\left(F_{[n\lambda]}(x) - F(x)\right),$$

so we can deduce (3.9) in this way:

$$\begin{aligned}
&\sup_{\lambda,x} d_n^{-1}(n-[\lambda n]) \left| \left(F_{[n\lambda+1,n]}(x) - F(x)\right) - J(x)\left(Z(1) - Z(\lambda)\right) \right| \\
&\quad \le \sup_{\lambda,x} d_n^{-1} \left(n\left|(F_n(x) - F(x)) - J(x)Z(1)\right| + [n\lambda]\left|\left(F_{[n\lambda]}(x) - F(x)\right) - J(x)Z(\lambda)\right| \right),
\end{aligned}$$

and both terms vanish by (3.8).

Thus we get, for the first term in the right-hand side of (3.6),

$$\frac{1}{n\,d_n}[n\lambda](n-[n\lambda])\int_{\mathbb{R}}(F_{[n\lambda]}(x)-F(x))dF_{[n\lambda]+1,n}(x)-(1-\lambda)\int_{\mathbb{R}}J(x)Z(\lambda)dF(x) \tag{3.10}$$

$$\begin{aligned}
&= \frac{n-[n\lambda]}{n}\int_{\mathbb{R}} d_n^{-1}[n\lambda](F_{[n\lambda]}(x)-F(x))dF_{[n\lambda]+1,n}(x) - \frac{n-[n\lambda]}{n}\int_{\mathbb{R}}J(x)Z(\lambda)dF(x)\\
&\quad + \left\{\frac{n-[n\lambda]}{n}-(1-\lambda)\right\}\int_{\mathbb{R}}J(x)Z(\lambda)dF(x)\\
&= \frac{n-[n\lambda]}{n}\int_{\mathbb{R}}\left\{d_n^{-1}[n\lambda](F_{[n\lambda]}(x)-F(x))-J(x)Z(\lambda)\right\}dF_{[n\lambda]+1,n}(x)\\
&\quad + \frac{n-[n\lambda]}{n}\int_{\mathbb{R}}J(x)Z(\lambda)d(F_{[n\lambda]+1,n}-F)(x)\\
&\quad + \left\{\frac{n-[n\lambda]}{n}-(1-\lambda)\right\}\int_{\mathbb{R}}J(x)Z(\lambda)dF(x).
\end{aligned}$$

The first term on the right-hand side converges to 0 almost surely by (3.8). The third term converges to zero as

$$\sup_{0\le\lambda\le1}\left|\frac{n-[n\lambda]}{n}-(1-\lambda)\right|\to 0,$$

because with $n\lambda-1\le[n\lambda]\le n\lambda+1$, we obtain

$$\left|\frac{n-[n\lambda]}{n}-(1-\lambda)\right| = \left|\lambda-\frac{[n\lambda]}{n}\right| \le \left|\lambda-\frac{n\lambda+1}{n}\right| = \frac{1}{n}.$$

Regarding the second term, note that $\int_{\mathbb{R}}J(x)dF(x)=E(J(X_i))$ and hence

$$\begin{aligned}
&\frac{n-[n\lambda]}{n}\int_{\mathbb{R}}J(x)Z(\lambda)d(F_{[n\lambda]+1,n}-F)(x)\\
&= Z(\lambda)\frac{1}{n}\sum_{i=[n\lambda]+1}^{n}(J(X_i)-EJ(X_i))\\
&= Z(\lambda)\left\{\frac{1}{n}\sum_{i=1}^{n}(J(X_i)-E(J(X_i)))-\frac{1}{n}\sum_{i=1}^{[n\lambda]}(J(X_i)-E(J(X_i)))\right\}.
\end{aligned}$$

By the ergodic theorem[2] , $\frac{1}{n}\sum_{i=1}^{n}(J(X_i) - E(J(X_i))) \to 0$, almost surely. Hence $\sum_{i=1}^{n}(J(X_i) - E(J(X_i))) = o(n)$ and thus

$$\max_{1\le k\le n}\left|\sum_{i=1}^{k}(J(X_i) - E(J(X_i)))\right| = o(n), \text{ as } n\to\infty, \quad a.s.,$$

otherwise $\frac{1}{n}\sum_{i=1}^{n}(J(X_i) - E(J(X_i)))$ would possess a subsequence which does not convergece to 0. Hence (3.10) converges to zero almost surely, uniformly in $\lambda \in [0,1]$.

Regarding the second term on the right-hand side of (3.6) we obtain

$$\begin{aligned}&\frac{1}{n\,d_n}[n\lambda](n-[n\lambda])\int_{\mathbb{R}}(F_{[n\lambda]+1,n} - F(x))dF(x) - \lambda\int_{\mathbb{R}}J(x)(Z(1)-Z(\lambda))dF(x) \quad (3.11)\\ &= \frac{[n\lambda]}{n}\int_{\mathbb{R}}\left\{d_n^{-1}(n-[n\lambda])(F_{[\lambda n]+1,n}(x) - F(x)) - J(x)(Z(1)-Z(\lambda))\right\}dF(x)\\ &\quad - \left(\lambda - \frac{[n\lambda]}{n}\right)\int_{\mathbb{R}}J(x)(Z(1)-Z(\lambda))dF(x).\end{aligned}$$

Both terms on the right-hand side converge to zero a.s., uniformly in $\lambda \in [0,1]$. For the first term, this follows from (3.9). For the second term, this holds since $\sup_{0\le\lambda\le1}\left|\frac{[n\lambda]}{n} - \lambda\right| \to 0$, as $n\to\infty$, as shown above. Using (3.6) and the fact that the right-hand sides of (3.10) and (3.11) converge to zero uniformly in $0 \le \lambda \le 1$, we have proved that the normalized Wilcoxon two-sample test statistic (3.5) converges in distribution towards

$$\int_{\mathbb{R}}(1-\lambda)Z(\lambda)J(x)dF(x) - \int_{\mathbb{R}}\lambda(Z(1)-Z(\lambda))J(x)dF(x), \quad 0\le\lambda\le1,$$

which equals

$$(Z(\lambda) - \lambda Z(1))\int_{\mathbb{R}}J(x)dF(x),$$

and thus we have established Theorem 3.1. □

In this proof we have used an integration by parts in order to express our test statistic as a functional of the empirical process. Recall that a similar integration by parts technique was used by Dehling and Taqqu (1989, 1991); but here, we use a one-dimensional integration by parts formula, whereas Dehling and Taqqu use a two-dimensional integration by parts, because the latter would not work here, since the kernel $I_{\{x\le y\}}$ does not have locally bounded variation, see section B.3.2, page 210.

[2] A stationary, mean zero Gaussian process $(\xi_i)_{i\ge1}$ is ergodic if and only if its spectral measure $F(d\lambda)$ is continous, i.e. that the spectral measure of each point λ is zero (Rozanov, 1967, Ex. 6.2). Moreover, if $(\xi_i)_{i\ge1}$ is a stationary, ergodic process and $f:\mathbb{R}^{\mathbb{N}}\to\mathbb{R}$ a measurable function, the process $(X_i)_{i\ge1}$ defined as $X_i = f(\xi_i,\xi_{i+1},\xi_{i+2},\ldots)$ is ergodic, too, especially if $f(t_1,t_2,\ldots) = f(t_1)$ is only a function of the first variable. To see this, let A be shift-invariant with respect to the measure P_X. Then $P_X(A) = P_\xi(\{(\xi_i)\,|\,(X_i) = (f(\xi_i)) \in A\})$ and $\{(\xi_i)\,|\,(X_i) = (f(\xi_i)) \in A\}$ is shift-invariant:

$$\{(\xi_i)\,|\,(X_i) = (f(\xi_i)) \in A\} = \{(\xi_i)\,|\,(f(\xi_{i-1})) \in A\} = \{(\xi_{i+1})\,|\,(f(\xi_i)) \in A\},$$

Thus $P_X(A)$ is 0 or 1, and thus, $(X_i)_{i\ge1}$ is ergodic.

3.3 The limit distribution in special situations

3.3.1 The Wilcoxon two-sample test

As a corollary to Theorem 3.1, we obtain for *fixed* $\lambda \in [0,1]$ the asymptotic distribution of the Wilcoxon two-sample test, where the two samples are

$$X_1, \ldots, X_{[n\lambda]}$$
$$X_{[n\lambda]+1}, \ldots, X_n.$$

After centering and scaling, the corresponding test statistic is

$$U_n = \frac{1}{n\, d_n} \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} \left(I_{\{X_i \le X_j\}} - \int_{\mathbb{R}} F(x) dF(x) \right).$$

Corollary (to Theorem 3.1)**.** *U_n converges in distribution to*

$$\boxed{\frac{1}{m!}\big(Z_m(\lambda) - \lambda\, Z_m(1)\big) \int_{\mathbb{R}} J_m(x) dF(x)}$$

which equals

$$\frac{1}{m!}\big((1-\lambda) Z_m(\lambda) - \lambda (Z_m(1) - Z_m(\lambda))\big) \int_{\mathbb{R}} J_m(x) dF(x).$$

3.3.2 Two independent samples

Now we assume that we observe samples from two independent LRD processes $(X_i)_{i\ge 1}$ and $(X'_i)_{i\ge 1}$ with identical joint distributions. In this case, the samples are

$$X_1, \ldots, X_{[n\lambda]}$$
$$X'_1, \ldots, X'_{n-[n\lambda]}.$$

The two-sample Wilcoxon test statistic for the problem (H, A), see (3.1) and (3.2) and compare to (3.3), for this case is

$$W'_n = \frac{1}{n\, d_n} \sum_{i=1}^{[n\lambda]} \sum_{j=1}^{n-[n\lambda]} \left(I_{\{X_i \le X'_j\}} - \int_{\mathbb{R}} F(x) dF(x) \right).$$

By going through the proof above and making appropriate changes where needed, we can derive a result on the normalized two-sample Wilcoxon test statistic for independent LRD samples:

Theorem 3.2. *W'_n converges in distribution towards the process*

$$\boxed{\frac{1}{m!}\left((1-\lambda) Z_m(\lambda) - \lambda Z'_m(1-\lambda)\right) \int_{\mathbb{R}} J_m(x) dF(x),}$$

where $(Z'_m(\lambda))_{0\le\lambda\le 1}$ is an independent copy of $(Z_m(\lambda))_{0\le\lambda\le 1}$.

Note that the limit distributions in the two models (Theorem 3.1 and Theorem 3.2) are different, as the joint distribution of $(Z_m(\lambda), Z_m(1) - Z_m(\lambda))$ is different from that of $(Z_m(\lambda), Z'_m(1-\lambda))$. This is a result of the fact that the Hermite process does not have independent increments. This is in contrast to the short-range dependent case, where the Wilcoxon two-sample test statistic has the same distribution in both models (Dehling and Fried, 2010). Roughly speaking, the dependence washes away in the limit for short-range dependence, but not for long-range dependence.

Proof. Instead of the e.d.f. $F_{[\lambda n]+1,n}$ of the last $n - [\lambda n]$ observations we now have to consider the e.d.f. of the $n - [\lambda n]$ observations of the second sample,

$$F_{n-[\lambda n]}(x) = \frac{1}{n - [\lambda n]} \sum_{i=1}^{n-[\lambda n]} I_{\{X'_i \le x\}},$$

which has the same probabilistic properties as the e.d.f. of the first sample because of the identical joint distribution (and thus, we can denote it by the same symbol, if we bear in mind, that it does not rely on the same random variables X_k, but on copies of them). We obtain

$$\begin{aligned}
&\frac{1}{n\, d_n} \sum_{i=1}^{[\lambda n]} \sum_{j=1}^{n-[\lambda n]} \left(I_{\{X_i \le X'_j\}} - \int F(x)\, dF(x) \right) \\
&\quad = \frac{[\lambda n](n - [\lambda n])}{n\, d_n} \int \left(F_{[\lambda n]} - F \right)(x)\, dF_{n-[\lambda n]}(x) \\
&\qquad - \frac{[\lambda n](n - [\lambda n])}{n\, d_n} \int \left(F_{n-[\lambda n]} - F \right)(x)\, dF(x).
\end{aligned}$$

Exactly as in the proof for one divided sample, one can show that the first term on the right-hand side converges to $(1-\lambda)Z(\lambda) \int J(x)\, dF(x)$. For the second term, note that

$$\begin{aligned}
&d_n^{-1}(n - [\lambda n]) \left(F_{n-[\lambda n]} - F \right)(x) \\
&\quad = d_n^{-1}[(1-\lambda)n] \left(F_{[(1-\lambda)n]}(x) - F(x) \right) + d_n^{-1} \left([(1-\lambda)n] - (n - [\lambda n]) \right) F(x) \\
&\qquad - d_n^{-1} \left([(1-\lambda)n] F_{[(1-\lambda)n]}(x) - (n - [\lambda n]) F_{n-[\lambda n]}(x) \right).
\end{aligned}$$

The first term converges uniformly in x and λ to $J(x)Z'(1-\lambda)$, due to (3.8), where Z' is an independent copy of $(Z(\lambda))_{0 \le \lambda \le 1}$. The second term vanishes uniformly, since $\left([(1-\lambda)n] - (n - [\lambda n]) \right) F(x)$ is elementarily bounded between -2 and 2. And in the third term we have by definition

$$[(1-\lambda)n] F_{[(1-\lambda)n]}(x) - (n - [\lambda n]) F_{n-[\lambda n]}(x) = \sum_{i=n-[\lambda n]+1}^{[(1-\lambda)n]} I_{\{X'_i \le x\}},$$

a sum that has not more than two summands (in fact, it is exactly one for all n and λ). □

3.4 Application

We still consider the model $X_i = \mu_i + G(\xi_i)$, $i = 1, \ldots, n$, where $(\xi_i)_{i\geq 1}$ is a mean-zero Gaussian process with $\mathrm{Var}[\xi_i] = 1$ and auto-covariance function (1.1) and a transformation $G : \mathbb{R} \to \mathbb{R}$, $G \in \mathcal{G}^1$ or $G \in \mathcal{G}^2$. We assume that the X_i have a continuous c.d.f. F. We wish to test the hypothesis

$$H : \mu_1 = \ldots = \mu_n$$

against the alternative

$$A : \mu_1 = \ldots = \mu_k \neq \mu_{k+1} = \ldots = \mu_n \text{ for some } k \in \{1, \ldots, n-1\}.$$

In what follows, we will develop two tests for the test problem (H, A): a "Wilcoxon-type" test, based on the well-known Wilcoxon's rank test and Theorem 3.1, and a "difference-of-means" test.

3.4.1 "Wilcoxon-type" test

The change-point test based on Wilcoxon's rank test will reject the null hypothesis for large values of

$$\boxed{W_n = \frac{1}{n\, d_n} \max_{1\leq k\leq n-1} \left| \sum_{i=1}^{k} \sum_{j=k+1}^{n} \left(I_{\{X_i \leq X_j\}} - \frac{1}{2} \right) \right|.} \tag{3.12}$$

It is intuitively clear that a strictly monotonely increasing transformation G does not disturb the order of the underlying data $(\xi_i)_{i\geq 1}$. As a consequence, W_n stays the same – no matter if we apply it to the X_i's or the original fGn ξ_i's. This stays true for strictly monotonely decreasing transformations which just invert the order of the underlying data $(\xi_i)_{i\geq 1}$. We state this fact as

Lemma 3.3. *The test statistic W_n is invariant under strictly monotone transformations of the data, i.e.*

$$\boxed{\max_{1\leq k\leq n-1} \left| \sum_{i=1}^{k} \sum_{j=k+1}^{n} \left(I_{\{G(X_i)\leq G(X_j)\}} - \frac{1}{2} \right) \right| = \max_{1\leq k\leq n-1} \left| \sum_{i=1}^{k} \sum_{j=k+1}^{n} \left(I_{\{X_i \leq X_j\}} - \frac{1}{2} \right) \right|}$$

for all strictly monotone functions $G : \mathbb{R} \to \mathbb{R}$.

Proof. If G is strictly increasing, this is obvious, as $G(X_i) \leq G(X_j)$ if and only if $X_i \leq X_j$. If G is strictly decreasing, $G(X_i) \leq G(X_j)$ if and only if $X_j \leq X_i$, and thus

$$I_{\{G(X_i)\leq G(X_j)\}} = 1 - I_{\{X_i \leq X_j\}}.$$

Hence we get

$$\sum_{i=1}^{k} \sum_{j=k+1}^{n} \left(I_{\{G(X_i)\leq G(X_j)\}} - \frac{1}{2} \right) = -\sum_{i=1}^{k} \sum_{j=k+1}^{n} \left(I_{\{X_i \leq X_j\}} - \frac{1}{2} \right),$$

and the lemma is proved. □

Since $X_i = \mu_i + G(\xi_i)$, under the null hypothesis that all μ_i are the same, the test statistic is

$$W_n = \frac{1}{n\, d_n} \max_{1\le k\le n-1} \left| \sum_{i=1}^{k} \sum_{j=k+1}^{n} \left(I_{\{G(\xi_i)\le G(\xi_j)\}} - \frac{1}{2} \right) \right|. \tag{3.13}$$

Theorem 3.1 and the continuous mapping theorem yield that, under the null hypothesis, W_n converges in distribution, as $n \to \infty$, to

$$\sup_{0\le\lambda\le 1} \left| \frac{Z_m(\lambda)}{m!} - \lambda \frac{Z_m(1)}{m!} \right| \left| \int_{\mathbb{R}} J_m(x) dF(x) \right|.$$

In order to set critical values for the asymptotic test based on W_n, we need to calculate the distribution of this expression.

In what follows, we will assume that G is a strictly monotone function. In this case, combining (3.13) and Lemma 3.3 we get that

$$W_n = \frac{1}{n\, d_n} \max_{1\le k\le n-1} \left| \sum_{i=1}^{k} \sum_{j=k+1}^{n} \left(I_{\{\xi_i\le\xi_j\}} - \frac{1}{2} \right) \right|.$$

Note that in this case,

$$J_1(x) = E(\xi I_{\{\xi\le x\}}) = \int_{-\infty}^{x} t\varphi(t)dt = -\varphi(x),$$

where $\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ denotes the standard normal density function. In the last step, we have used the fact that $\varphi'(t) = -t\,\varphi(t)$. Thus $J_1(x) \neq 0$ for all x and hence the class of functions $\{I_{\{\xi\le x\}},\, x \in \mathbb{R}\}$ has Hermite rank $m = 1$. Moreover, as F is the normal distribution function we obtain

$$\int_{\mathbb{R}} J_1(x) dF(x) = -\int_R \varphi(x)\varphi(x)dx = \int_{\mathbb{R}} -(\varphi(s))^2 ds = -\frac{1}{2\sqrt{\pi}}. \tag{3.14}$$

We have thus proved the following theorem.

Theorem 3.4. *Let $(\xi_i)_{i\ge1}$ be a stationary Gaussian process with mean zero, variance* 1 *and auto-covariance function* (1.1) *with $0 < D < 1$. Moreover, let $G \in \mathcal{G}^1$ be a strictly monotone function and define $X_i = \mu_i + G(\xi_i)$. Then, under the null hypothesis H: $\mu_1 = \ldots = \mu_n$, the test statistic W_n, as defined in* (3.12)*, converges in distribution towards*

$$\boxed{\frac{1}{2\sqrt{\pi}} \sup_{0\le\lambda\le1} |Z_1(\lambda) - \lambda Z_1(1)|,}$$

where $(Z_1(\lambda))_{\lambda\ge0}$ denotes the standard fBm process with Hurst parameter $H = 1 - D/2 \in (1/2, 1)$.

We have evaluated the distribution of $\sup_{0\le\lambda\le1} |Z(\lambda) - \lambda Z(1)|$ in the following way:

1. Create one vector of a fBm $Z(t)$ at times $t = 0, 0.001, 0.002, \ldots, 0.999, 1$. (We have done this using a routine in the `fArma` package in `R` which simulates fBm from a numerical approximation of the stochastic integral.)

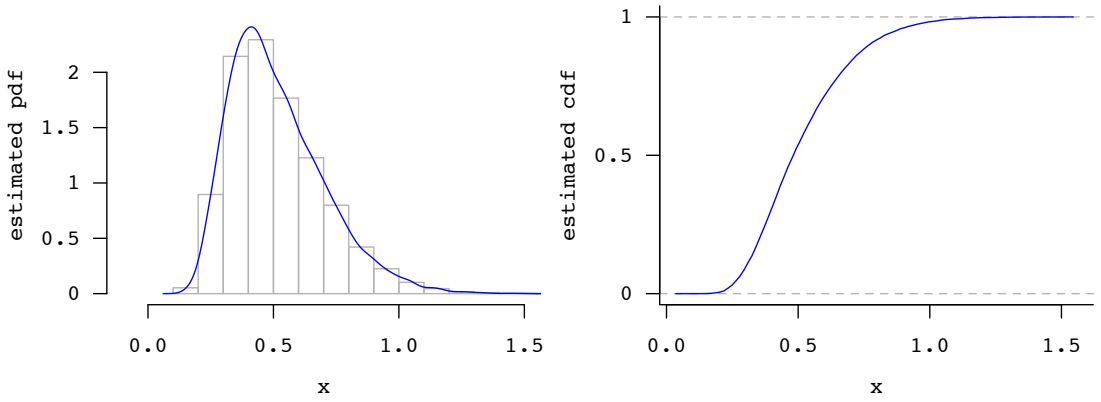


Figure 3.1: Probability density (left) and distribution function (right) of $\sup_{0\le\lambda\le1}|Z(\lambda)-\lambda Z(1)|$; calculations based on 10,000 simulation runs.

2. For each of these times $t = 0, 0.001, 0.002, \ldots, 0.999, 1$, calculate $|Z(t) - tZ(1)|$.

3. Store the maximum of these 1001 values; interpret this as one observation of $\sup_{0\le\lambda\le1}|Z(\lambda)-\lambda Z(1)|$.

By repeating these steps 10.000 times, one obtains a sample of observations. Their empirical upper α-quantile estimates the α-quantile of $\sup_{0\le\lambda\le1}|Z(\lambda)-\lambda Z(1)|$. Formally, we have simulated 10,000 realizations of a standard fBm $(Z^{(j)}(t))_{0\le t\le1}$, $1 \le j \le 10{,}000$, $t = \frac{i}{1000}$, $0 \le i \le 1000$, and for each realization, we have calculated $M_j := \max_{1\le i\le1000}|Z^{(j)}(\frac{i}{1000}) - \frac{i}{1000}Z^{(j)}(1)|$ as a numerical approximation to $\sup_{0\le\lambda\le1}|Z^{(j)}(\lambda) - \lambda\, Z^{(j)}(1)|$. The empirical distribution

$$F_M(x) := \frac{1}{10{,}000}\#\{1 \le j \le 10{,}000 \,:\, M_j \le x\}$$

of these 10,000 maxima was used as approximation to the distribution of $\sup_{0\le\lambda\le1}|Z(\lambda)-\lambda Z(1)|$; see Figure 3.1 for the estimated probability density and the empirical distribution function, for the Hurst parameter $H = 0.7$, and see Appendix C.7 for the source code. We have calculated the corresponding upper α-quantiles

$$\boxed{q_\alpha := \inf\{x \,:\, F_M(x) \ge 1-\alpha\}} \tag{3.15}$$

for $H = 0.6, 0.7, 0.9$ (that is, $D = 2-2H = 0.8, 0.6, 0.2$); see Table 3.1.

3.4.2 "Difference-of-means" test

As an alternative, we also consider a test based on differences of means of the observations. We consider the test statistic

$$\boxed{D_n := \max_{1\le k\le n-1}|D_{k,n}|\,,} \tag{3.16}$$

H / α	0.10	0.05	0.01
0.6	0.98	1.10	1.34
0.7	0.77	0.87	1.06
0.9	0.38	0.44	0.54

Table 3.1: Upper α-quantiles of $\sup_{0\leq\lambda\leq 1}|Z(\lambda)-\lambda Z(1)|$, where Z is a standard fBm, for different LRD parameters H, based on $10,000$ repetitions.

where

$$D_{k,n} := \frac{1}{n\,d_n}\sum_{i=1}^{k}\sum_{j=k+1}^{n}(X_i - X_j)\,.$$

One would reject the null hypothesis (3.1) if D_n is large. To obtain the asymptotic distribution of the test statistic, we apply the functional non-central limit theorem of Taqqu (1979) and obtain that

$$\begin{aligned}\frac{1}{n\,d_n}\sum_{i=1}^{[n\lambda]}\sum_{j=[n\lambda]+1}^{n}(X_i - X_j) &= \frac{1}{n\,d_n}\left((n-[\lambda n])\sum_{i=1}^{[n\lambda]}G(\xi_i) - [\lambda n]\sum_{j=[n\lambda]+1}^{n}G(\xi_j)\right)\\ &\xrightarrow{w} a_m\left((1-\lambda)\frac{Z_m(\lambda)}{m!} - \lambda\left(\frac{Z_m(1)}{m!} - \frac{Z_m(\lambda)}{m!}\right)\right)\\ &= a_m\frac{1}{m!}(Z_m(\lambda) - \lambda\, Z_m(1)),\end{aligned}$$

where m denotes the Hermite rank of $G(\xi)$ and where $a_m = E[G(\xi)H_m(\xi)]$ is the Hermite coefficient. Applying the continuous mapping theorem, we obtain the following theorem concerning the asymptotic distribution of the D_n.

Theorem 3.5. *Let $(\xi_i)_{i\geq 1}$ be a stationary Gaussian process with mean zero, variance 1 and auto-covariance function (1.1) with $0 < D < 1/m$. Moreover, let $G \in \mathcal{G}^2$ and define $X_i = \mu_i + G(\xi_i)$. Then, under the null hypothesis $H:\ \mu_1 = \ldots = \mu_n$, the test statistic D_n, as defined in (3.16), converges in distribution towards*

$$\boxed{\frac{|a_m|}{m!}\sup_{0\leq\lambda\leq 1}|Z_m(\lambda) - \lambda Z_m(1)|,}$$

where $(Z_m(\lambda))$ denotes the m-th order Hermite process with Hurst parameter $H = 1 - Dm/2 \in (1/2, 1)$.

Note that Horváth and Kokoszka (1997) analyse an estimator for the time of change in the mean of Gaussian LRD observations. Their estimator compares the mean of the observations up to a certain time with the overall mean of all observations which in the case of independent standard normal random variables is just the MLE for the time of

change. The test of Horváth and Kokoszka (1997) is the same as our "difference-of-means" test because

$$\sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} (X_i - X_j) = n \sum_{i=1}^{[\lambda n]} (X_i - \bar{X}_n).$$

Note that they admit a further scaling parameter, but they only deal with Gaussian observations.

Corollary. *For a strictly monotone function G, the Hermite rank is 1 and the test statistic D_n converges under the null hypothesis towards*

$$\boxed{|a_1| \sup_{0 \le \lambda \le 1} |Z_1(\lambda) - \lambda Z_1(1)|,}$$

where Z_1 is the standard fBm $B_H(\lambda)$ with index $H = 1 - D/2$.

Proof. If G is strictly increasing, we obtain

$$E[G(\xi)H_1(\xi)] = \int_{\mathbb{R}} G(s)H_1(s)\varphi(s)ds = \int_0^\infty s\,\varphi(s)(G(s) - G(-s))ds > 0.$$

Similarly, for a stricly decreasing function G we obtain $E[G(\xi)H_1(\xi)] < 0$. So in both cases the Hermite rank is 1. □

Note that in this case of a strictly monotone transformation G, up to a norming constant, the limit distribution of the "difference-of-means" test is the same as for the test based on Wilxocon's rank statistic.

3.5 "Difference-of-means" test under fGn

In the next section, we will analyse the level and the power of both tests, the "Wilcoxon-type" test and the "difference-of-means" test, in a finite sample situation by a simulation study. But before, we will concentrate on a special case in which we can analytically calculate a lower bound for the power of the "difference-of-means" test: We assume that we observe fGn, i.e. we consider the model

$$X_i = \mu_i + \xi_i, \quad i = 1, \ldots, n.$$

In this situation, the distribution of the "difference-of-means" test statistic D_n, as defined in (3.16), is not explicitly known, but one can calculate the exact distribution of

$$D_{k,n} = \frac{1}{n\,d_n} \sum_{i=1}^{k} \sum_{j=k+1}^{n} (X_i - X_j).$$

To this end, recall that fGn can be obtained by differencing fBm, that is $\xi_k = B_H(k) - B_H(k-1)$, where $(B_H(t))_{t\geq 0}$ is the standard fBm with Hurst parameter H, which we denoted Z_1. Its covariance, see (1.3), is given by

$$E\left[Z_1(\lambda_1)\, Z_1(\lambda_2)\right] = \frac{1}{2}\left(\lambda_1^{2H} + \lambda_2^{2H} - |\lambda_1 - \lambda_2|^{2H}\right).$$

Consider the alternative that there is a jump of height h after the $[\lambda n]$-th observation:

$$H_{\lambda,h}:\ E[X_i] = 0 \text{ for } i = 1, \ldots, [n\lambda] \text{ and } E[X_i] = h \text{ for } i = [n\lambda]+1, \ldots, n \qquad (3.17)$$

We shall compute the exact distribution of $D_{k,n}$ under $H_{\lambda,h}$ and thus obtain a lower bound for the power of D_n, since

$$P(D_n \geq q_\alpha) \geq P(D_{[n\lambda],n} \geq q_\alpha),$$

where $\{D_n \geq q_\alpha\}$ is the rejection region and q_α is given in (3.15).

First note that $d_n = n^H$, because of (1.5) and (1.12), and thus $nd_n = n^{1+H}$, where H is again the Hurst coefficient. $D_{k,n}$ has a normal distribution with mean

$$E[D_{k,n}] = \frac{1}{n^{1+H}} \sum_{i=1}^{k} \sum_{j=k+1}^{n} (E[X_i] - E[X_j]).$$

Thus a small calculation shows that

$$E[D_{k,n}] = \begin{cases} -\frac{1}{n^{1+H}} k\,(n - [n\lambda])\, h & \text{if } \ k \leq [n\lambda] \\ -\frac{1}{n^{1+H}} (n-k)\,[n\lambda]\, h & \text{if } \ k \geq [n\lambda]. \end{cases}$$

Note that $\max_{1\leq k\leq n-1} |E[D_{k,n}]| = |E[D_{[n\lambda],n}]| = \frac{1}{n^{1+H}}(n - [n\lambda])\,[n\lambda]\, h \sim n^{1-H}\lambda\,(1-\lambda)\, h$. Since the variance of $D_{k,n}$ is not changed by the level shift, we get

$$\begin{aligned} \mathrm{Var}[D_{k,n}] &= \mathrm{Var}\left[\frac{1}{n^{1+H}} \sum_{i=1}^{k} \sum_{j=k+1}^{n} (\xi_i - \xi_j)\right] \\ &= \mathrm{Var}\left[\frac{1}{n^{1+H}} \left((n-k)\sum_{i=1}^{k} \xi_i - k \sum_{j=k+1}^{n} \xi_j\right)\right] \\ &= \mathrm{Var}\left[\frac{1}{n^{1+H}} \left((n-k)B_H(k) - k(B_H(n) - B_H(k))\right)\right] \\ &= \mathrm{Var}\left[\frac{1}{n^{1+H}} \left(n\, B_H(k) - kB_H(n)\right)\right]. \end{aligned}$$

By the self-similarity of fractional Brownian motion, we finally get

$$\begin{aligned}
\mathrm{Var}[D_{k,n}] &= \mathrm{Var}\left[B_H\left(\frac{k}{n}\right) - \frac{k}{n}B_H(1)\right] \\
&= \mathrm{Var}\left[B_H\left(\frac{k}{n}\right)\right] + \frac{k^2}{n^2}\,\mathrm{Var}[B_H(1)] - 2\frac{k}{n}\,\mathrm{Cov}\left[B_H\left(\frac{k}{n}\right), B_H(1)\right] \\
&= \left(\frac{k}{n}\right)^{2H} + \frac{k^2}{n^2} - \frac{k}{n}\left(\left(\frac{k}{n}\right)^{2H} + 1 - \left(1-\frac{k}{n}\right)^{2H}\right) \\
&= \left(\frac{k}{n}\right)^{2H} + \left(\frac{k}{n}\right)^{2} - \left(\frac{k}{n}\right)^{2H+1} - \frac{k}{n} + \frac{k}{n}\left(1-\frac{k}{n}\right)^{2H} \\
&= \left(\frac{k}{n}\right)^{2H}\left(1-\frac{k}{n}\right) - \frac{k}{n}\left(1-\frac{k}{n}\right) + \frac{k}{n}\left(1-\frac{k}{n}\right)^{2H}.
\end{aligned}$$

Defining

$$\sigma^2(\lambda) = \lambda^{2H}(1-\lambda) - \lambda(1-\lambda) + \lambda(1-\lambda)^{2H},$$

we thus obtain

$$\mathrm{Var}[D_{k,n}] = \sigma^2(k/n).$$

The variance is maximal for $k = n/2$, in which case we obtain for $H = 0.7$ e.g.

$$\mathrm{Var}[D_{n/2,n}] = \frac{1}{2^{2H}} - \frac{1}{4} \approx 0.13.$$

The distribution of $D_{k,n}$ gives a lower bound for the power of the "difference-of-means" test at the alternative $H_{\lambda,h}$ considered above. We have

$$\begin{aligned}
P(D_n \geq q_\alpha) &\geq P(|D_{[n\lambda],n}| \geq q_\alpha) \\
&= P(D_{[n\lambda],n} \leq -q_\alpha) + P(D_{[n\lambda],n} \geq q_\alpha) \\
&= P\left(\frac{D_{[n\lambda],n} + n^{1-H}\lambda(1-\lambda)h}{\sqrt{\sigma^2(\lambda)}} \leq \frac{-q_\alpha + n^{1-H}\lambda(1-\lambda)h}{\sqrt{\sigma^2(\lambda)}}\right) \\
&+ P\left(\frac{D_{[n\lambda],n} + n^{1-H}\lambda(1-\lambda)h}{\sqrt{\sigma^2(\lambda)}} \geq \frac{q_\alpha + n^{1-H}\lambda(1-\lambda)h}{\sqrt{\sigma^2(\lambda)}}\right) \\
&\approx \Phi\left(\frac{-q_\alpha + n^{1-H}\lambda(1-\lambda)h}{\sqrt{\sigma^2(\lambda)}}\right) + 1 - \Phi\left(\frac{q_\alpha + n^{1-H}\lambda(1-\lambda)h}{\sqrt{\sigma^2(\lambda)}}\right),
\end{aligned}$$

where Φ is the c.d.f. of a standard normal random variable. E.g., for $H = 0.7$, $\lambda = \frac{1}{2}$, we get $q_{0.05} = 0.87$ using Table 3.1 and thus

$$P(D_n \geq q_{0.05}) \geq \Phi\left(\frac{-0.87 + n^{0.3}h/4}{\sqrt{\sigma^2(1/2)}}\right) \approx \Phi(-2.42 + 0.70\,h\,n^{0.3}).$$

In this way, for sample size $n = 500$ and level shift $h = 1$ we get $\Phi(2.07) \approx 0.98$ as lower bound on the power of the "difference-of-means" test. For the same sample size, but $h = 0.5$, we get the lower bound $\Phi(-0.18) \approx 0.43$. Compare this to the simulation results in Table D.11.

3.6 Simulations

In this section, we will present the results of a simulation study which compares the "Wilcoxon-type" rank test (3.12) with the "difference-of-means" test (3.16). We first analyse whether the tests reach their asymptotic level when applied in a finite sample setting, for sample sizes ranging from $n = 10$ to $n = 1,000$. Secondly, we compare the power of the two tests for sample size $n = 500$ at various different alternatives

$$A_k : \quad \mu_1 = \ldots = \mu_k \neq \mu_{k+1} = \ldots = \mu_n.$$

We let both the break point k and the level shift $h := \mu_{k+1} - \mu_k$ vary. Specifically, we choose $k = 25, 50, 150, 250$ and $h = 0.5, 1, 2$.

As data, we have taken simulated realizations $\xi_1, \ldots, \xi_n$ of a fGn process with Hurst parameter H, respectively $D = 2 - 2H$, see (1.5) and (1.4); to these fGn data, we have applied several transformations G, ranging from mild ones which produce symmetric data to boisterous ones which generate heavy tails. We have repeated each simulation $10,000$ times.

3.6.1 Normally distributed data

In our first simulations, we took

$$G(t) = t,$$

so that $(X_i)_{i\geq 1}$ is fGn. F is then the c.d.f. Φ of a standard normal random variable. Since G is strictly increasing, Theorem 3.4 yields that, under the null hypothesis, W_n has approximately the same distribution as

$$\frac{1}{2\sqrt{\pi}} \sup_{0\leq\lambda\leq 1} |Z_1(\lambda) - \lambda Z_1(1)|.$$

Since G is the first Hermite polynomial, its Hermite rank is $m = 1$ and the associated Hermite coefficient is $a_1 = 1$. Hence, Theorem 3.5 yields that, under the null hypothesis, the test statistic D_n has approximately the same distribution as

$$\sup_{\lambda\in[0,1]} |Z_1(\lambda) - \lambda Z_1(1)| .$$

We have calculated asymptotic critical values for both tests by using the upper 5%-quantiles of $\sup_{\lambda\in[0,1]} |Z_1(\lambda) - \lambda Z_1(1)|$, as given in Table 3.1. Thus the "Wilcoxon-type" test rejected the null hypothesis when $W_n \geq \frac{1}{2\sqrt{\pi}} q_\alpha$, while the "difference-of-means" test rejected when $D_n \geq q_\alpha$, where q_α is given in (3.15).

We have checked whether the tests reach their asymptotic level of 5% and counted the number of (false) rejections of the null hypothesis in $10,000$ simulations, where the null hypothesis was true. We see in Figure 3.3 that both tests perform well already for moderate sample sizes of $n = 50$, with the notable exception of the "Wilcoxon-type" test when $H = 0.9$, i.e. when we have very strong dependence. In that case, the

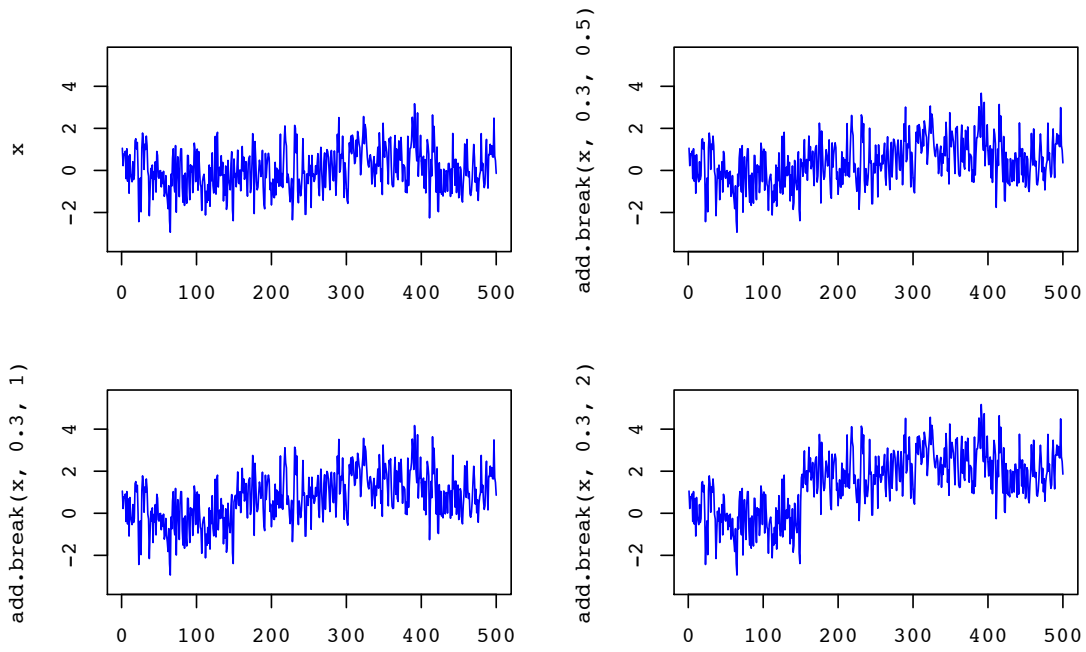


Figure 3.2: fGn without breaks (top left) and with a jump after observation 150 (this is $[\lambda n]$ with $\lambda = 0.3$) of height $h = 0.5$ (top right), $h = 1$ (bottom left) and $h = 2$ (bottom right).

convergence in Theorem 3.4 appears to be very slow so that the asymptotic critical values are misleading when applied in a finite sample setting. The exact simulation results are given in Table D.10 in Appendix D.

In order to analyze how well the tests detect break points, we have introduced a level shift h at time $[n\lambda]$, i.e. we consider the time series

$$X_i = \begin{cases} \xi_i & \text{for } i = 1, \dots, [n\lambda] \\ \xi_i + h & \text{for } i = [n\lambda] + 1, \dots, n \end{cases},$$

i.e. the alternative $H_{\lambda,h}$ as in (3.17). We have done this for several choices of λ and h, for sample size $n = 500$. As Figure 3.4 shows, both tests detect breaks very well – and the better, the larger the level shift is and the more in the middle the shift takes place. When the break occurs in the middle, both tests perform equally well. Breaks at the beginning are better detected by the "difference-of-means" test. Exact values are given in Table D.11 in Appendix D.

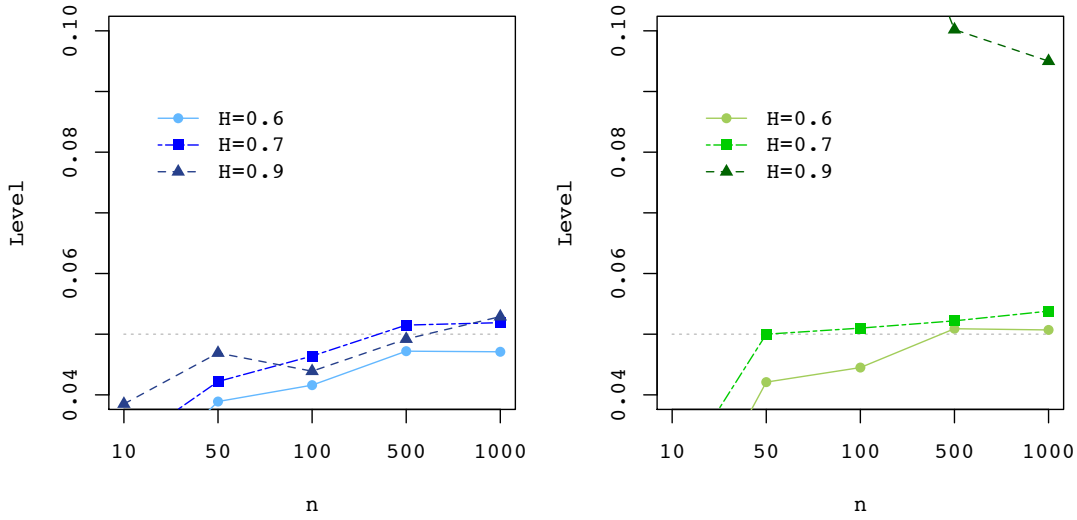


Figure 3.3: Level of "difference-of-means" test (left) and level of "Wilcoxon-type" test (right) for fGn time series with LRD parameter H; 10,000 simulation runs. Both tests have asymptotically level 5%.

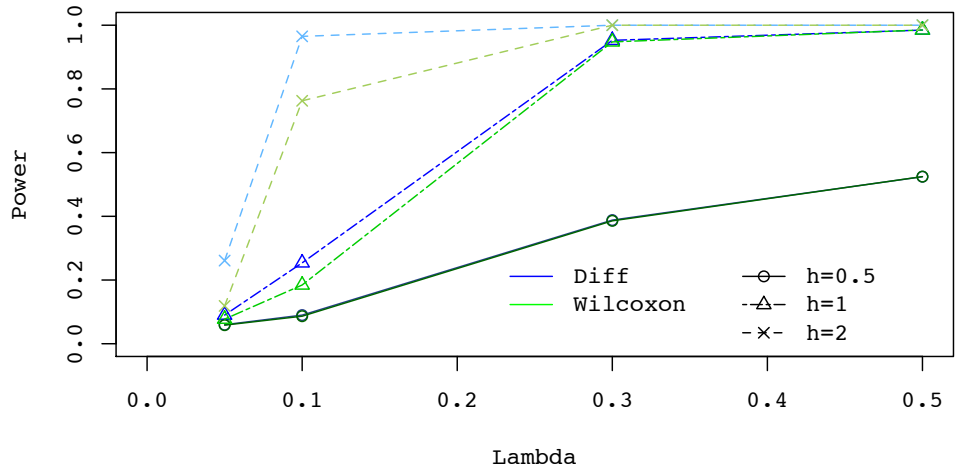


Figure 3.4: Power of "difference-of-means" test (blue) and power of "Wilcoxon-type" test (green) for $n = 500$ observations of fGn with LRD parameter $H = 0.7$, different break points $[\lambda n]$ and different level shifts h. Both tests have asymptotically level 5%. The calculations are based on 10,000 simulation runs.

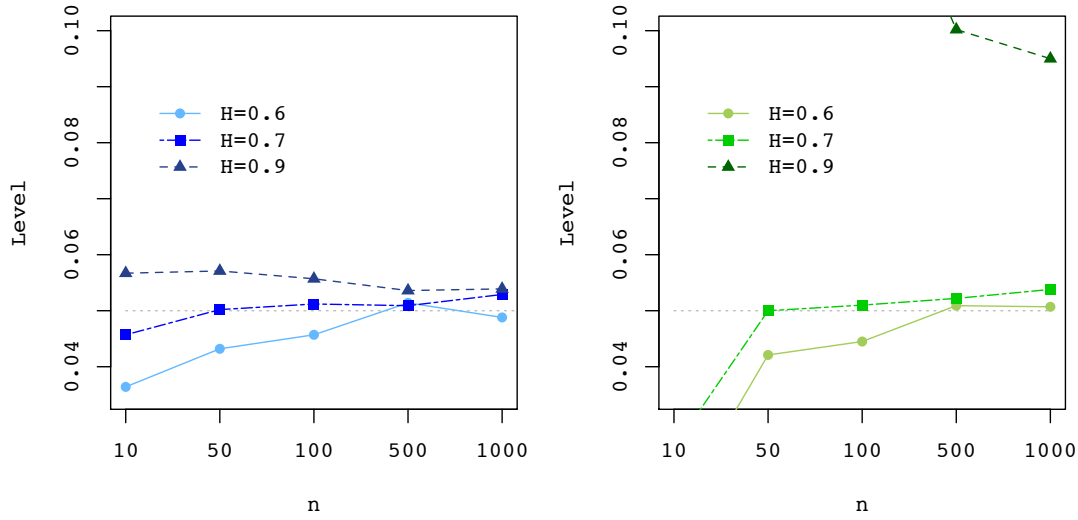


Figure 3.5: Level of "difference-of-means" test (left) and level of "Wilcoxon-type" test (right) for standardised Laplace(0,4)-transformed fGn with LRD parameter H; $10,000$ simulation runs. Both tests have asymptotically level 5%.

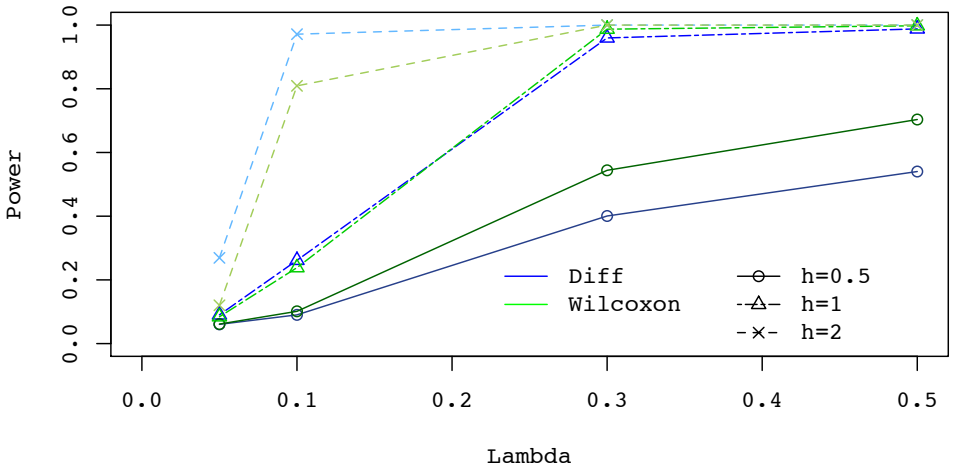


Figure 3.6: Power of "difference-of-means" test (blue) and power of "Wilcoxon-type" test (green) for $n = 500$ observations of standardised Laplace(0,4)-transformed fGn with LRD parameter $H = 0.7$, different break points $[\lambda n]$ and different level shifts h. Both tests have asymptotically level 5%. The calculations are based on 10,000 simulation runs.

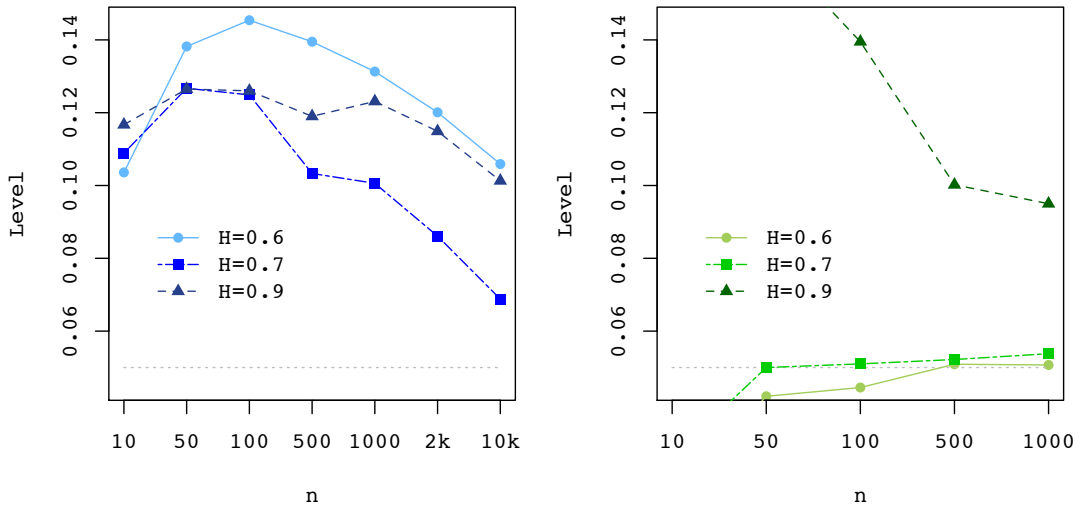


Figure 3.7: Level of "difference-of-means" test (left) and level of "Wilcoxon-type" test (right) for standardised Pareto(3,1)-transformed fGn with LRD parameter H; 10,000 simulation runs. Both tests have asymptotically level 5%.

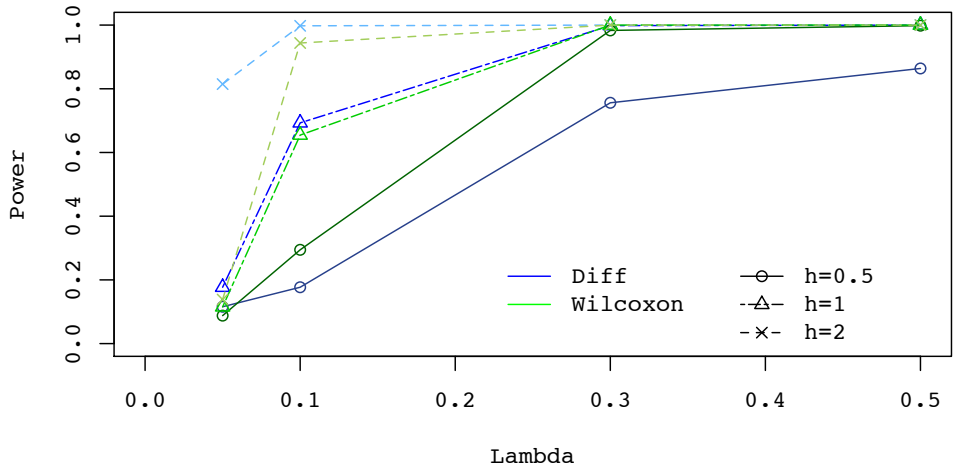


Figure 3.8: Power of "difference-of-means" test (blue) and power of "Wilcoxon-type" test (green) for $n = 500$ observations of standardised Pareto(3,1)-transformed fGn with LRD parameter $H = 0.7$, different break points $[\lambda n]$ and different level shifts h. Both tests have asymptotically level 5%. The calculations are based on 10,000 simulation runs.

3.6.2 Symmetric, normal-tailed data

In the second set of simulations, we generated Laplace (or double exponential) distributed data. The Laplace(μ, b) distribution has c.d.f. and p.d.f.

$$\begin{aligned} F_{\mu,b}(x) &= \begin{cases} \frac{1}{2}\exp\left(\frac{x-\mu}{b}\right) & \text{if } x < \mu \\ 1 - \frac{1}{2}\exp\left(-\frac{x-\mu}{b}\right) & \text{if } x \geq \mu \end{cases} \\ &= \frac{1}{2}\left[1 + \operatorname{sgn}(x-\mu)\left(1 - \exp\left(\frac{-|x-\mu|}{b}\right)\right)\right] \\ f_{\mu,b}(x) &= \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right) \end{aligned}$$

where μ is the location (mean, median and mode) and b is a shape parameter influencing the variance:

$$E[X] = \mu \qquad \operatorname{Var}[X] = 2b^2$$

The Laplace distribution is symmetric and normal-tailed. In order to obtain Laplace(μ, b)-distributed $X = G(\xi)$, we take G to be the quantile transform

$$G(t) = \mu - b\operatorname{sgn}\left(\Phi(t) - \frac{1}{2}\right)\log\left(1 - 2\left|\Phi(t) - \frac{1}{2}\right|\right).$$

Note that $\Phi(\xi_i)$ is $U[0,1]$ distributed, that $\operatorname{sgn}\left(\Phi(t) - \frac{1}{2}\right) = \operatorname{sgn}(t)$ and that

$$\begin{aligned} \{G(s) \leq x\} &= \left\{\mu - b\operatorname{sgn}(s)\log\left(1 - 2\left|\Phi(s) - \frac{1}{2}\right|\right) \leq x\right\} \\ &= \left(\left\{\log\left(1 - 2\left(\Phi(s) - \frac{1}{2}\right)\right) \geq -\frac{x-\mu}{b}\right\} \cap \{s \geq 0\}\right) \\ &\quad \cup \left(\left\{1 + 2\left(\Phi(s) - \frac{1}{2}\right) \leq \exp\left(\frac{x-\mu}{b}\right)\right\} \cap \{s < 0\}\right) \\ &= \left(\left\{1 - \frac{1}{2}\exp\left(-\frac{x-\mu}{b}\right) \geq \Phi(s)\right\} \cap \{s \geq 0\}\right) \\ &\quad \cup \left(\left\{\Phi(s) \leq \frac{1}{2}\exp\left(\frac{x-\mu}{b}\right)\right\} \cap \{s < 0\}\right). \end{aligned}$$

Now observe that on $\{s \geq 0\}$ it holds $\Phi(s) \geq 1/2$ and thus $\left\{1 - \frac{1}{2}\exp\left(-\frac{x-\mu}{b}\right) \geq \Phi(s)\right\}$ is non-empty for $x \geq \mu$. Analogously, on $\{s < 0\}$ it holds $\Phi(s) < 1/2$ and so the set $\left\{\Phi(s) \leq \frac{1}{2}\exp\left(\frac{x-\mu}{b}\right)\right\}$ is non-empty for $x < \mu$. From this, we obtain

$$\begin{aligned} P(G(\xi_i) \leq x) &= \int_{\left\{1 - \frac{1}{2}\exp\left(-\frac{x-\mu}{b}\right) \geq \Phi(s)\right\} \cap \{x \geq \mu\}} d\Phi(s) + \int_{\left\{\Phi(s) \leq \frac{1}{2}\exp\left(\frac{x-\mu}{b}\right)\right\} \cap \{x < \mu\}} d\Phi(s) \\ &= \begin{cases} \frac{1}{2}\exp\left(\frac{x-\mu}{b}\right) & \text{if } x < \mu \\ 1 - \frac{1}{2}\exp\left(-\frac{x-\mu}{b}\right) & \text{if } x \geq \mu \end{cases} \\ &= F_{\mu,b}(x). \end{aligned}$$

We choose $\mu = 0$ and $b = 4$ (but since we are interested in standardised data, the choice of b is irrelevant). We standardise and consider the transformation

$$G_{\mathrm{st}}(t) = \frac{-b \operatorname{sgn}\left(\Phi(t) - \frac{1}{2}\right) \log\left(1 - 2\left|\Phi(t) - \frac{1}{2}\right|\right)}{\sqrt{2b^2}} = \begin{cases} \frac{1}{\sqrt{2}} \log(2\Phi(t)) & \text{if } t \le 0 \\ -\frac{1}{\sqrt{2}} \log\left(2(1 - \Phi(t))\right) & \text{else} \end{cases}.$$

Note that in this case, the p.d.f. of the data is

$$f_{\mathrm{st}}(x) = f_{0,2^{-1/2}}(x) = \frac{1}{\sqrt{2}} \exp\left(-|\sqrt{2}x|\right).$$

We will now compare the "Wilcoxon-type" test and the "difference-of-means" test. Since G_{st} is strictly monotone, the Hermite rank of the class of functions $I_{\{G_{\mathrm{st}}(\xi_i) \le x\}} - F(x)$ and the Hermit rank of G_{st} are both 1, by the arguments following Lemma 3.3 and Theorem 3.5. Theorem 3.4 yields that, under the null hypothesis, W_n has approximately the same distribution as

$$\frac{1}{2\sqrt{\pi}} \sup_{0 \le \lambda \le 1} |Z_1(\lambda) - \lambda Z_1(1)|.$$

For the "difference-of-means" test we need the first Hermite coefficient of G_{st}. With numerical integration and using that $tG_{\mathrm{st}}(t)$ and $\phi(t)$ are axially symmetric we obtain:

$$a_1 = E[\xi G_{\mathrm{st}}(\xi)] = 2 \cdot \frac{1}{\sqrt{2}} \int_{-\infty}^{0} t \log(2\Phi(t)) \phi(t)\, dt \approx 0.981344$$

Hence, Theorem 3.5 yields that, under the null hypothesis, D_n has approximately the same distribution as

$$0.981 \sup_{\lambda \in [0,1]} |Z_1(\lambda) - \lambda Z_1(1)|.$$

As before with the upper 5%-quantiles of $\sup_{\lambda \in [0,1]} |Z_1(\lambda) - \lambda Z_1(1)|$ from Table 3.1, the "Wilcoxon-type" test rejected the null hypothesis when $W_n \ge \frac{1}{2\sqrt{\pi}} q_\alpha$, while the "difference-of-means" test rejected when $D_n \ge 0.981 q_\alpha$, where q_α is given in (3.15).

We see in Figure 3.5 (and in Table D.12) that for $H = 0.9$, that is for very strongly dependent data, the "Wilcoxon-type" test converges slowly while the level of the "difference-of-means" is rather close to its asymptotic limit of 5%[3]. As Fig-

[3]Note that the level of the test under Laplace distributed data (and as well under the Pareto distributed data which we will consider in a short while) is exactly the same as under fGn, as a consequence of the invariance of the "Wilcoxon-type" test, see Lemma 3.3. Of course, the power of the "Wilcoxon-type" test here is not the same as for untransformed fGn. This is caused by the interaction of the transform G and the shift: G does not disturb the order of the observations, but it alters the amount of the increments; when we now shift a part of the time series, the order of the resulting data may be different from the order of fGn after the same shift. As illustration, consider for simplicity a situation when four fGn observations are ascending: $\xi_1 < \xi_2 < \xi_3 < \xi_4$. Then any positive shift will not change the value of the test-statistic W_n since it does not change the order of the data ($W_n = 2/(n\, d_n)$). But if we apply a strictly decreasing transformation G so that $G(\xi_1) > G(\xi_2) > G(\xi_3) > G(\xi_4)$, and if the shift afterwards lifts the last two observations up, but only so far that $G(\xi_3) + h > G(\xi_2) > G(\xi_4) + h$, then the order of the data is mixed up which results in a different W_n ($W_n = 1.5/(n\, d_n)$).

ure 3.6 shows (exact values are given in Table D.13), the "Wilcoxon-type" test is better in detecting small jumps, both tests perform similar when the jump is of the same order of magnitude as the variance of the data, and the "difference-of-means" test is better when the jump is big.

3.6.3 Heavy-tailed data

In the third set of simulations we took Pareto distributed data. The Pareto(β, k) distribution has distribution function

$$F_{\beta,k}(x) = \begin{cases} 1 - \left(\frac{k}{x}\right)^{\beta} & \text{if } x \geq k \\ 0 & \text{else,} \end{cases}$$

where the scale parameter k is the smallest possible value for x and where β is a shape parameter. It has a finite expected value when $\beta > 1$ and finite variance when $\beta > 2$. The expected value and the variance are given by

$$\mu = E[X] = \frac{\beta k}{\beta - 1}, \quad \beta > 1$$
$$\sigma^2 = \mathrm{Var}[X] = \frac{\beta k^2}{(\beta-1)^2(\beta-2)}, \quad \beta > 2.$$

In order to obtain Pareto(β, k)-distributed $X = G(\xi)$, we take G to be the quantile transform, i.e. $G(t) = k\left(\Phi(t)\right)^{-1/\beta}$ where Φ denotes the standard normal c.d.f., so that for $x \geq k$

$$P(X_i \geq x) = P(G(\xi_i) \geq x) = P\left((\Phi(\xi_i))^{-1/\beta} \geq \frac{x}{k}\right) = P\left(\xi_i \leq \Phi^{-1}\left(\left(\frac{k}{x}\right)^{\beta}\right)\right) = \left(\frac{k}{x}\right)^{\beta}.$$

Since we want the X_i to be standardized to have mean 0 and variance 1, we will in fact take

$$G(t) = \left(\frac{\beta k^2}{(\beta-1)^2(\beta-2)}\right)^{-1/2} \left(k(\Phi(t))^{-1/\beta} - \frac{\beta k}{\beta - 1}\right) \tag{3.18}$$

The corresponding distribution function is then

$$F_{\beta,k,\mathrm{st}}(z) = \begin{cases} 1 - \left(\frac{k}{\sigma z + \mu}\right)^{\beta} & \text{if } z \geq \frac{k-\mu}{\sigma} \\ 0 & \text{else} \end{cases} \tag{3.19}$$

and its density function is

$$f_{\beta,k,\mathrm{st}}(z) = \begin{cases} k^{\beta}\beta\sigma(\sigma z + \mu)^{-\beta-1} & \text{if } z \geq \frac{k-\mu}{\sigma} \\ 0 & \text{else} \end{cases}. \tag{3.20}$$

Pareto$(3,1)$ **Data:** We first performed simulations with Pareto$(3,1)$ data, i.e. heavy-tailed data with finite variance. In this case, $\beta = 3$, $k = 1$ and we have $E[X] = \frac{3}{2}$ and $\mathrm{Var}[X] = \frac{3}{4}$. For a better comparison with the simulations involving fGn, we also standardize the data, i.e. we consider, see (3.18),

$$G(t) = \frac{1}{\sqrt{3/4}} \left((\Phi(t))^{-1/3} - \frac{3}{2} \right).$$

The probability density function of the standardized X is given by (see (3.20)),

$$f_{3,1,\mathrm{st}}(x) = \begin{cases} 3\sqrt{\frac{3}{4}} \left(\sqrt{\frac{3}{4}}\, x + \frac{3}{2} \right)^{-4} & \text{if } x \geq -\sqrt{\frac{1}{3}} \\ 0 & \text{else.} \end{cases}$$

G is strictly decreasing, and by the above results following Lemma 3.3 and Theorem 3.5, the Hermite rank of the class of functions $\{I_{\{G(\xi_i) \leq x\}} - F(x),\ x \in \mathbb{R}\}$ is $m = 1$, the Hermite rank of G itself is $m = 1$ and $|\int_{\mathbb{R}} J_1(x)\, dF(x)| = (2\sqrt{\pi})^{-1}$, see (3.14). Numerical integration yields

$$a_1 = E[\xi G(\xi)] = \sqrt{\frac{4}{3}} \int_{-\infty}^{\infty} s\, \Phi(s)^{-1/3} \varphi(s)\, ds \approx -0.6784.$$

Figure 3.7 shows the observed level of the tests, for various sample sizes and various Hurst parameters (for exact simulation results, see Table D.14). For sample sizes up to $n = 1,000$, the "difference-of-means" test has a level larger than 10%. We conjecture that this is due to the slow convergence in Theorem 3.5, which is supported by the simulation results with sample sizes $n = 2,000$ and $n = 10,000$; see Table D.14.

Figure 3.8 gives the observed power of the "difference-of-means" test and of the "Wilcoxon-type" test, for sample size $n = 500$ and various values of the break points and height of level shift. The results show that the "Wilcoxon-type" test has larger power than the "difference-of-means" test for a small level shift h, but that the "difference-of-means" test outperforms the "Wilcoxon-type" test for larger level shifts. For the exact simulation results, see Table D.15.

However, the above comparison is not meaningful, since the "difference-of-means" test has a realized level of approximately 10% while the "Wilcoxon-type" test has level close to 5%, see Figure 3.7, respectively Table D.14. Thus we have calculated the finite sample 5%-quantiles of the distribution of the "difference-of-means" test, using a Monte-Carlo simulation, see Table 3.2. For example, for $n = 500$ and $H = 0.7$, the corresponding critical value is 0.70. Thus we reject the null hypothesis of no break point if the "difference-of-means" test statistic is greater than 0.70.

The value of 0.70 should be contrasted to the asymptotic ($n = \infty$) value of 0.59. (This asymptotic value was obtained as follows: According to Theorem 3.5, D_n is under the null hypothesis asymptotically distributed like $|a_1| \sup_{0 \leq \lambda \leq 1} |Z(\lambda) - \lambda Z(1)|$, so the asymptotic upper α-quantiles of D_n can be calculated as $|a_1| q_\alpha$, where q_α is the upper α-quantile of the distribution of $\sup_{0 \leq \lambda \leq 1} |Z(\lambda) - \lambda Z(1)|$, as tabulated in Table 3.1.)

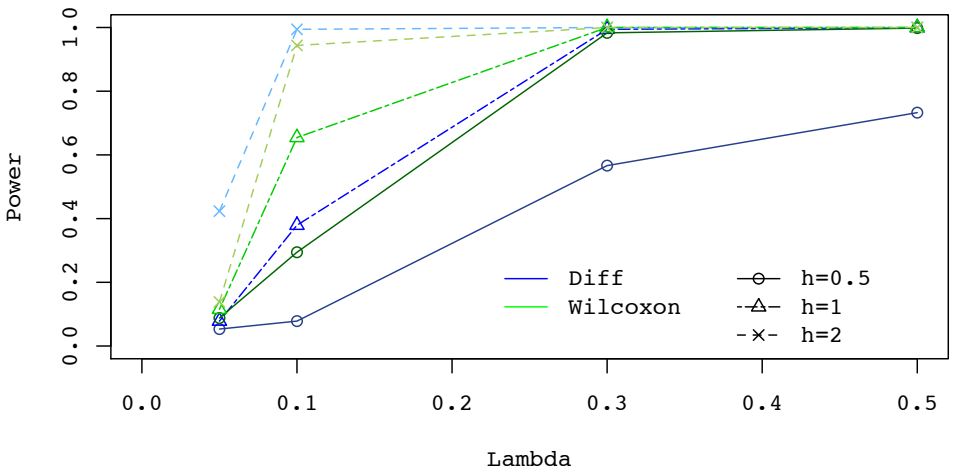


Figure 3.9: Power of "difference-of-means" test (blue) and power of "Wilcoxon-type" test (green), based on the finite sample quantiles, for $n = 500$ observations of standardised Pareto(3,1)-transformed fGn with LRD parameter $H = 0.7$, different break points $[\lambda n]$ and different level shifts h. Both tests have asymptotically level 5%. The calculations are based on 10,000 simulation runs.

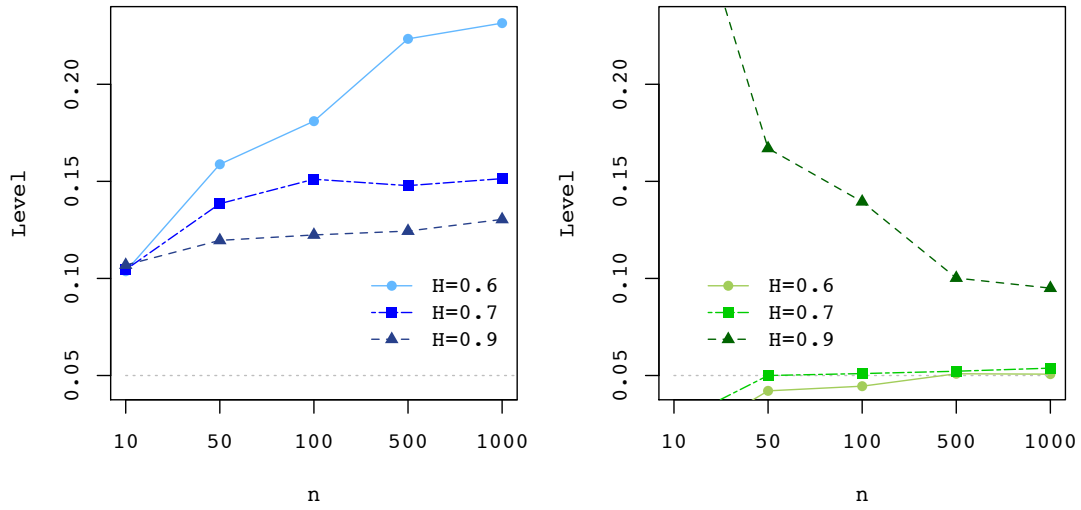


Figure 3.10: Level of "difference-of-means" test (left) and level of "Wilcoxon-type" test (right) for standardised Pareto(2,1)-transformed fGn with LRD parameter H; $10,000$ simulation runs. Both tests have asymptotically level 5%.

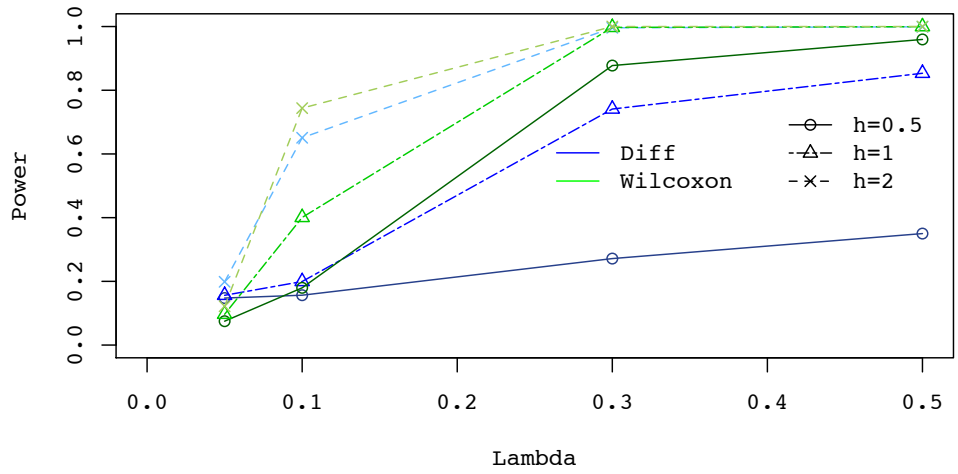


Figure 3.11: Power of "difference-of-means" test (blue) and power of "Wilcoxon-type" test (green) for $n = 500$ observations of standardised Pareto(2,1)-transformed fGn with LRD parameter $H = 0.7$, different break points $[\lambda n]$ and different level shifts h. Both tests have asymptotically level 5%. The calculations are based on 10,000 simulation runs.

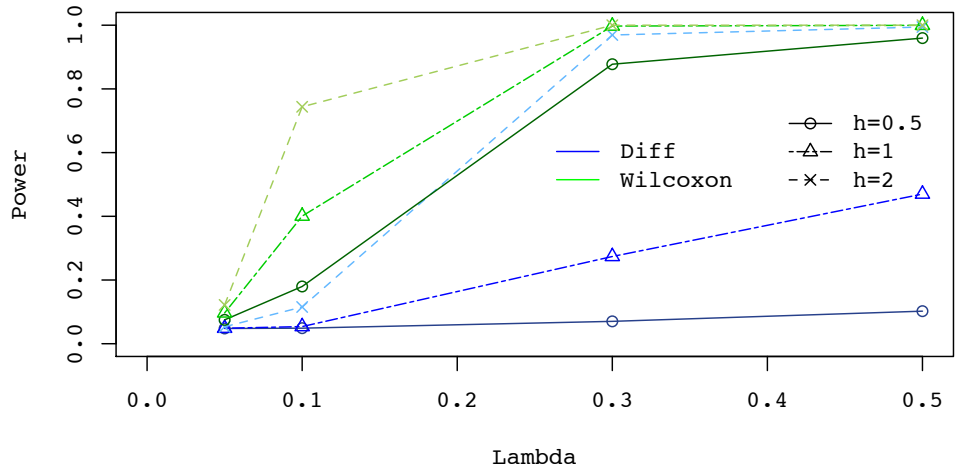


Figure 3.12: Power of "difference-of-means" test (blue) and power of "Wilcoxon-type" test (green), based on the finite sample quantiles, for $n = 500$ observations of standardised Pareto(2,1)-transformed fGn with LRD parameter $H = 0.7$, different break points $[\lambda n]$ and different level shifts h. Both tests have asymptotically level 5%. The calculations are based on 10,000 simulation runs.

H / n	10	50	100	500	1,000	2,000	10,000	∞
0.6	1.02	1.04	1.02	0.93	0.90	0.88	0.85	0.75
0.7	0.84	0.82	0.79	0.70	0.68	0.65	0.62	0.59
0.9	0.47	0.47	0.44	0.43	0.42	0.41	0.38	0.30

Table 3.2: 5%-quantiles of the finite sample distribution of the "difference-of-means" test under the null hypothesis for Pareto(3,1)-transformed fGn with different LRD parameter H and different sample sizes n. The calculations are based on 10,000 simulation runs.

We have then calculated the power of the "difference-of-means" test in a further simulation, with $n = 500$, $H = 0.7$ and the finite sample quantile critical value of 0.70 instead of the asymptotic value of 0.59 (see Table 3.2). Table D.16 in Appendix D shows the power of the test. We can now compare the results of the "Wilcoxon-type" test with the finite sample "difference-of-means" test results; this is shown in Figure 3.9. We see that the "Wilcoxon-type" test has better power than the "difference-of-means" test, except for large level shifts at an early time. Such changes are detected more often by the "difference-of-means" test. For exact values, compare the right-hand side of Table D.15 with Table D.16.

Pareto$(2, 1)$ **Data:** We now choose $k = 1$ and $\beta = 2$, so that the X have finite expectation, but infinite variance. In order to have centered data, we take

$$G(t) = \frac{1}{\sqrt{\Phi(t)}} - 2.$$

We will now compare both tests, i.e. the "Wilcoxon-type" test and the "difference-of-means" test, although the latter can strictly speaking not be applied because it requires data with finite variance, respectively $G \in \mathcal{G}^2 \subset L^2(\mathbb{R}, \mathcal{N})$. By Theorem 3.4, under the null hypothesis of no change, the "Wilcoxon-type" test statistic W_n converges in distribution towards

$$\frac{1}{2\sqrt{\pi}} \sup_{0 \le \lambda \le 1} |Z_1(\lambda) - \lambda Z_1(1)|.$$

As a consequence of Lemma 3.3, even the finite sample distribution of W_n is the same as for normally distributed data. Figure 3.10 gives the measured level of the "Wilcoxon-type" test (the asymptotic level is 5%) and Figure 3.11 suggests it has good power, especially for small shifts in the middle of the observations.

Let us now consider the "difference-of-means" test. Again note that, strictly speaking, Theorem 3.5 cannot be applied in the case of the Pareto data with $\beta = 2$ because it requires the variance of the data to be finite. It is interesting nevertheless to use the asymptotic test suggested by Theorem 3.5. Since G is strictly monotone, the Hermite

rank of G is $m = 1$ as well, by the Corollary following Theorem 3.5. Using numerical integration, we have calculated

$$a_1 = E\left[\xi G(\xi)\right] = \int_{-\infty}^{\infty} s(\Phi(s)^{-1/2} - 2)\varphi(s)\, ds = \int_{-\infty}^{\infty} s\,\Phi(s)^{-1/2}\varphi(s)\, ds \approx -1.40861.$$

We clearly see in Figure 3.10 that the "difference-of-means" test very often falsely rejects the null hypothesis, that is detects breaks where there are none, while the "Wilcoxon-type" test is robust. Figure 3.11 shows that both tests have good power, but again, the "Wilcoxon-type" test is clearly better, especially for small shifts in the middle of the observations. For an exact comparison, see Table D.17 and Table D.15 in Appendix D.

As in the situation with Pareto(3,1) distributed data, the "difference-of-means" test rejects the null hypothesis much too often (see Figure 3.10, it does not reach its asymptotic level of 5%, it does not even seem to converge), so the power comparison is again not meaningful.

So we have calculated the finite sample 5%-quantiles of the distribution of the "difference-of-means" test here as well; the results are shown in Table D.16. Based on these critical values, the test can be compared to the "Wilcoxon-type" test. In this comparison, the "difference-of-means" test shows an even poorer performance, see Figure 3.12.

Chapter 4

Power of some change-point tests

In Chapter 3, we have studied two tests for the change-point problem (H, A) as defined in (3.1) and (3.2): The first test, the "Wilcoxon-type" test, rejects the null hypothesis H for large values of the statistic

$$W_n = \frac{1}{n\, d_n} \max_{1 \le k \le n-1} \left| \sum_{i=1}^{k} \sum_{j=k+1}^{n} \left(I_{\{X_i \le X_j\}} - \frac{1}{2} \right) \right|, \tag{4.1}$$

the second test, the "difference-of-means" test rejects H for large values of

$$D_n := \frac{1}{n\, d_n} \max_{1 \le k \le n-1} \left| \sum_{i=1}^{k} \sum_{j=k+1}^{n} (X_j - X_i) \right|. \tag{4.2}$$

Note that for reasons of comparability, we like to define the kernel of D_n with a negative sign here, in contrast to definition (3.16). In Chapter 3, we have derived the limit distribution of both tests under the null hypothesis that no change occured and we have simulated the behaviour of both tests under several alternatives (for data with a jump in the mean of different heights and at various positions). Now we compute the power of the above tests analytically under a local alternative. This chapter is based on the article of Dehling, Rooch and Taqqu (2013).

We consider the following sequence of alternatives

$$\boxed{A_{\tau,h_n}(n):\ \mu_i = \begin{cases} \mu & \text{for } i = 1, \dots, [n\tau] \\ \mu + h_n & \text{for } i = [n\tau]+1, \dots, n, \end{cases}} \tag{4.3}$$

where $0 \le \tau \le 1$, in other words we consider a model where there is a jump of height h_n after a proportion of τ in the data. Note that the height h_n of the level shift depends on the sample size n; we will have to choose h_n in a certain way to obtain a non-degenerate limit distribution.

4.1 Power of the "difference-of-means" test under local alternatives

We first investigate the asymptotic distribution of the process

$$D_n(\lambda) := \frac{1}{n\,d_n} \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} (X_j - X_i), \quad 0 \le \lambda \le 1, \tag{4.4}$$

under $A_{\tau,h_n}(n)$, as defined in (4.3). Since the statistic $D_n(\lambda)$ splits up the data into two blocks at time $[n\lambda]+1$, while the local alternative $A_{\tau,h_n}(n)$ involves a jump at $[n\tau]+1$, we expect to obtain an interplay between λ and τ in the limit.

Under the general alternative A, as defined in (3.2), we obtain

$$D_n(\lambda) = \frac{1}{n\,d_n} \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} (G(\xi_j) - G(\xi_i)) + \frac{1}{n\,d_n} \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} (\mu_j - \mu_i), \tag{4.5}$$

and under the local alternative $A_{\tau,h_n}(n)$ (which is included in A), the second term is

$$\frac{1}{n\,d_n} \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} (\mu_j - \mu_i) = \begin{cases} \frac{h_n}{n\,d_n}[\lambda n](n - [\tau n]) & \text{for } \lambda \le \tau \\ \frac{h_n}{n\,d_n}(n - [\lambda n])[\tau n] & \text{for } \lambda \ge \tau. \end{cases} \tag{4.6}$$

We can write this in a shorter form, using the function $\delta_\tau : [0,1] \to \mathbb{R}$ which we define as

$$\boxed{\delta_\tau(\lambda) = \begin{cases} \lambda(1-\tau) & \text{for } \lambda \le \tau \\ (1-\lambda)\tau & \text{for } \lambda \ge \tau \end{cases};} \tag{4.7}$$

note that $\delta_\tau(\lambda)$ takes its maximum value $\tau(1-\tau)$ at $\lambda = \tau$. Now we obtain for large n

$$\frac{1}{n\,d_n} \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} (\mu_j - \mu_i) \sim \frac{n\,h_n}{d_n}\,\delta_\tau(\lambda).$$

Thus, in order for the second term in (4.5) to converge as $n \to \infty$, we have to choose the level shift $h_n \sim c\,d_n/n$, where c is a constant. When n is large, this is exactly the order of the level shift that can be detected with a non-trivial power (i.e. a power which is neither 0 nor 1).

Theorem 4.1. *Let $G \in \mathcal{G}^2$ be a transformation with Hermite rank m, as defined in Section 1.4.2, and let $(\xi_i)_{i\ge1}$ be a stationary Gaussian process with mean zero, variance 1 and auto-covariance function as in (1.1) with $0 < D < \frac{1}{m}$. For observations $X_i = \mu_i + G(\xi_i)$, under the local alternative $A_{\tau,h_n}(n)$ with*

$$h_n = \frac{d_n}{n}\,c \tag{4.8}$$

for an arbitrary constant c, the process $(D_n(\lambda))_{0\le\lambda\le 1}$ converges in distribution to the process

$$\boxed{\frac{a_m}{m!}\left(\lambda Z_m(1) - Z_m(\lambda)\right) + c\,\delta_\tau(\lambda),} \tag{4.9}$$

where $(Z_m(\lambda))_{\lambda\ge 0}$ denotes the m-th order Hermite process with Hurst parameter $H = 1 - \frac{Dm}{2} \in (\frac{1}{2}, 1)$, and where a_m is the m-th Hermite coefficient of G, as defined in Section 1.4.2.

Proof. We use decomposition (4.5). The first term on the right-hand side has the same distribution as $D_n(\lambda)$ under the null hypothesis, and thus it converges in distribution to $\frac{a_m}{m!}(\lambda Z_m(1) - Z_m(\lambda))$, see Theorem 3.5. Regarding the second term, we obtain by (4.8) and (4.6)

$$\begin{aligned}\frac{1}{n\,d_n}\sum_{i=1}^{[n\lambda]}\sum_{j=[n\lambda]+1}^{n}(\mu_j - \mu_i) &= \begin{cases} \frac{c}{n^2}[\lambda n](n - [\tau n]) & \text{for } \lambda \le \tau \\ \frac{c}{n^2}(n - [\lambda n])[\tau n] & \text{for } \lambda \ge \tau. \end{cases} \\ &\to c\,\delta_\tau(\lambda),\end{aligned}$$

uniformly in $\lambda \in [0,1]$, as $n \to \infty$. □

Remark. (i) For $c = 0$, which means $h_n = 0$, the local alternative $A_{\tau,h_n}(n)$ is identical to the null hypothesis H, and in fact, Theorem 4.1 reproduces the limit distribution of D_n under the null hypothesis. Thus, Theorem 4.1 is a generalization of Theorem 3.5.
(ii) Under the local alternative, i.e. when $c \neq 0$, the limit process is the sum of a fractional bridge process (like in Theorem 3.1 and Theorem 3.5) and the deterministic function $c\,\delta_\tau$.
(iii) As a corollary to Theorem 4.1, we can determine the asymptotic distribution of test statistics based on the process $(D_n(\lambda))_{0\le\lambda\le 1}$ by applying the continuous mapping theorem. For example, we get that D_n, as defined in (4.2) converges in distribution to

$$\sup_{0\le\lambda\le 1}|\lambda Z_m(1) - Z_m(\lambda) + c\,\delta_\tau(\lambda)|.$$

This limit distribution depends on the constant c. For $c = 0$, we obtain the limit distribution under the null hypthesis, see Theorem 3.5. Increasing the value of $|c|$ leads to a shift of the distribution to the right.
(iv) For a given break position $\tau \in [0,1]$, the function $\delta_\tau(\lambda)$ takes its maximum value in $\lambda = \tau$, and this maximum value equals $\tau(1-\tau)$. Thus, for values of τ close to 0 and close to 1, $\tau(1-\tau)$ is close to 0, and thus the effect of adding the term $c\delta_\tau(\lambda)$ is rather small. As a result, the power of the test is small at level shifts that occur very early or very late in the process.
(v) The higher the level shift and the closer λ is to τ, the easier it is to detect the level shift.

Theorem 4.1 can be applied in order to make power calculations for the change-point test that rejects for large values of D_n. If we denote by q_α the upper α-quantile

of the distribution of $\sup_{0\le\lambda\le 1}|\lambda Z(1)-Z(\lambda)|$, where $Z(\lambda)=Z_m(\lambda)/m!$, our test will reject the null hypothesis H when $D_n \ge a_m q_\alpha$, where a_m is the m-th Hermite coefficient of G, see Theorem 3.5. By construction, this test has asymptotic level α, i.e. the test rejects H with probability α if the null hypothesis H holds. If n is large and

$$h = h_n = \frac{d_n}{n}c,$$

the power of the test at the alternative $A_{\tau,h}(n)$ is by (4.9) approximately given by

$$P\left(\sup_{0\le\lambda\le 1}|\lambda Z_m(1)-Z_m(\lambda)+c\,\delta_\tau(\lambda)|\ge q_\alpha\right).$$

We may apply Theorem 4.1 as well in order to determine the size of a level shift at time $[\tau n]$ that can be detected with a given probability β. To this end, we calculate c such that

$$P\left(\sup_{0\le\lambda\le 1}|\lambda Z_m(1)-Z_m(\lambda)+c\,\delta_\tau(\lambda)|\ge q_\alpha\right)=\beta.$$

Choosing then $h=\frac{d_n}{n}c$, we get that the asymptotic power of the test at the alternative $A_{\tau,h}(n)$ is equal to β. Thus, a level shift of this size can be detected with probability β with a test that has level α.

4.2 Power of the "Wilcoxon-type" test under local alternatives

Theorem 4.2. *Suppose that $(\xi_i)_{i\ge1}$ is a stationary Gaussian process with mean zero, variance 1 and auto-covariance function (1.1) with $0\le D<\frac{1}{m}$. Moreover, let $G\in\mathcal{G}^1$, and assume that $G(\xi_i)$ has continuous distribution function $F(x)$. Let m denote the Hermite rank of the class of functions $I_{\{G(\xi_i)\le x\}}-F(x)$, $x\in\mathbb{R}$. Then, under A_{τ,h_n}, as defined in (4.3), if $h_n\to 0$ as $n\to\infty$*

$$\boxed{\frac{1}{n\,d_n}\sum_{i=1}^{[n\lambda]}\sum_{j=[n\lambda]+1}^{n}\left(I_{\{X_i\le X_j\}}-\frac{1}{2}\right)-\frac{n}{d_n}\delta_\tau(\lambda)\int_{\mathbb{R}}(F(x+h_n)-F(x))\,dF(x),} \tag{4.10}$$

indexed by $0\le\lambda\le 1$, converges in distribution towards the process

$$\boxed{\frac{1}{m!}(Z_m(\lambda)-\lambda Z_m(1))\int_{\mathbb{R}}J_m(x)\,dF(x),\quad 0\le\lambda\le 1,}$$

where $(Z_m(\lambda))_{\lambda\ge0}$ denotes the m-th order Hermite process with Hurst parameter $H=1-\frac{Dm}{2}\in(\frac12,1)$ and where $J_m(x)=E\left[H_m(\xi_i)I_{\{G(\xi_i)\le x\}}\right]$.

Remark. (i) Note that we make no assumption about the exact order of the sequence $(h_n)_{n\ge1}$. Theorem 4.2 holds under the very general assumption that $h_n\to0$ as $n\to\infty$.

(ii) If we choose $(h_n)_{n\geq 1}$ as in (4.8), the centering constants in (4.10) converge, provided some technical assumptions are satisfied. To see this, observe that

$$\begin{aligned}\frac{n}{d_n}\delta_\tau(\lambda)\int_{\mathbb{R}}(F(x+h_n)-F(x))\,dF(x) &= \frac{n\,h_n}{d_n}\delta_\tau(\lambda)\int_{\mathbb{R}}\frac{F(x+h_n)-F(x)}{h_n}\,dF(x)\\ &\to c\,\delta_\tau(\lambda)\int_{\mathbb{R}} f(x)\,dF(x)\\ &= c\,\delta_\tau(\lambda)\int_{\mathbb{R}} f^2(x)\,dx.\end{aligned}$$

The convergence in the next to last step requires some justification; it holds for example if F is differentiable with bounded derivative $f(x)$.

Corollary. *Suppose that $(\xi_i)_{i\geq 1}$ is a stationary Gaussian process with mean zero, variance 1 and auto-covariance function as in (1.1) with $0 \leq D < \frac{1}{m}$. Moreover, let $G \in \mathcal{G}^1$, and assume that $G(\xi_k)$ has a distribution function $F(x)$ with bounded density $f(x)$. Let m denote the Hermite rank of the class of functions $I_{\{G(\xi_i)\leq x\}} - F(x)$, $x \in \mathbb{R}$. Then, under the sequence of alternatives A_{τ,h_n}, as defined in (4.3), with $h_n = c\frac{d_n}{n}$ we obtain that*

$$\boxed{\frac{1}{n\,d_n}\sum_{i=1}^{[n\lambda]}\sum_{j=[n\lambda]+1}^{n}\left(I_{\{X_i\leq X_j\}}-\frac{1}{2}\right)} \tag{4.11}$$

converges in distribution to the process

$$\boxed{\frac{1}{m!}(Z_m(\lambda)-\lambda Z_m(1))\int_{\mathbb{R}} J_m(x)\,dF(x) + c\delta_\tau(\lambda)\int_{\mathbb{R}} f^2(x)\,dx, \quad 0\leq\lambda\leq 1.}$$

Proof of Theorem 4.2. We use the same techniques as in the proof of Theorem 3.1, where we derived the limit distribution of the "Wilcoxon-type" test statistic under the null hypothesis; more precisely we will decompose the test statistic (4.10) into a term whose distribution is the same both under the null hypothesis as well as under the alternative, and a second term which, after proper centering, converges to zero.

We express our test statistic as a functional of the e.d.f. $F_k(x)$ of the first k observations $G(\xi_1),\ldots,G(\xi_k)$ and of the e.d.f. $F_{k,l}(x)$ which is based on the observations $G(\xi_k),\ldots,G(\xi_l)$. Recall that under the local alternative, we have

$$X_i = \begin{cases} G(\xi_i)+\mu & \text{for } i = 1,\ldots,[n\tau]\\ G(\xi_i)+\mu+h_n & \text{for } i = [n\tau]+1,\ldots,n.\end{cases}$$

Thus we obtain for $\lambda \leq \tau$

$$\begin{aligned}
&\sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} I_{\{X_i \leq X_j\}} \\
&= \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{[n\tau]} I_{\{G(\xi_i)+\mu \leq G(\xi_j)+\mu\}} + \sum_{i=1}^{[n\lambda]} \sum_{j=[n\tau]+1}^{n} I_{\{G(\xi_i)+\mu \leq G(\xi_j)+\mu+h_n\}} \\
&= \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{[n\tau]} I_{\{G(\xi_i) \leq G(\xi_j)\}} + \sum_{i=1}^{[n\lambda]} \sum_{j=[n\tau]+1}^{n} I_{\{G(\xi_i) \leq G(\xi_j)+h_n\}} \\
&= \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} I_{\{G(\xi_i) \leq G(\xi_j)\}} + \sum_{i=1}^{[n\lambda]} \sum_{j=[n\tau]+1}^{n} \left(I_{\{G(\xi_i) \leq G(\xi_j)+h_n\}} - I_{\{G(\xi_i) \leq G(\xi_j)\}} \right) \\
&= \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} I_{\{G(\xi_i) \leq G(\xi_j)\}} + \sum_{i=1}^{[n\lambda]} \sum_{j=[n\tau]+1}^{n} I_{\{G(\xi_j) < G(\xi_i) \leq G(\xi_j)+h_n\}}. \qquad (4.12)
\end{aligned}$$

In the same way, we obtain for $\lambda \geq \tau$

$$\begin{aligned}
&\sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} I_{\{X_i \leq X_j\}} \\
&= \sum_{i=1}^{[n\tau]} \sum_{j=[n\lambda]+1}^{n} I_{\{G(\xi_i)+\mu \leq G(\xi_j)+\mu+h_n\}} + \sum_{i=[n\tau]+1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} I_{\{G(\xi_i)+\mu+h_n \leq G(\xi_j)+\mu+h_n\}} \\
&= \sum_{i=1}^{[n\tau]} \sum_{j=[n\lambda]+1}^{n} I_{\{G(\xi_i) \leq G(\xi_j)+h_n\}} + \sum_{i=[n\tau]+1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} I_{\{G(\xi_i) \leq G(\xi_j)\}} \\
&= \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} I_{\{G(\xi_i) \leq G(\xi_j)\}} + \sum_{i=1}^{[n\tau]} \sum_{j=[n\lambda]+1}^{n} \left(I_{\{G(\xi_i) \leq G(\xi_j)+h_n\}} - I_{\{G(\xi_i) \leq G(\xi_j)\}} \right) \\
&= \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} I_{\{G(\xi_i) \leq G(\xi_j)\}} + \sum_{i=1}^{[n\tau]} \sum_{j=[n\lambda]+1}^{n} I_{\{G(\xi_j) < G(\xi_i) \leq G(\xi_j)+h_n\}}. \qquad (4.13)
\end{aligned}$$

We recognize that for the first term on the right-hand side of (4.12) and (4.13), we each get the asymptotic distribution via Theorem 3.1. Thus, in order to prove Theorem 4.2, it suffices to show that the following two terms,

$$\frac{1}{n\, d_n} \sup_{0 \leq \lambda \leq \tau} \left| \sum_{i=1}^{[n\lambda]} \sum_{j=[n\tau]+1}^{n} I_{\{G(\xi_j) < G(\xi_i) \leq G(\xi_j)+h_n\}} - n^2 \lambda(1-\tau) \int_{\mathbb{R}} (F(x+h_n) - F(x))\, dF(x) \right| \qquad (4.14)$$

and

$$\frac{1}{n\, d_n} \sup_{\tau \leq \lambda \leq 1} \left| \sum_{i=1}^{[n\tau]} \sum_{j=[n\lambda]+1}^{n} I_{\{G(\xi_j) < G(\xi_i) \leq G(\xi_j)+h_n\}} - n^2 \tau(1-\lambda) \int_{\mathbb{R}} (F(x+h_n) - F(x))\, dF(x) \right| \qquad (4.15)$$

both converge to zero in probability. We first show this for (4.14). Observe that

$$\begin{aligned}
&\sum_{i=1}^{[n\lambda]} \sum_{j=[n\tau]+1}^{n} I_{\{G(\xi_j)<G(\xi_i)\le G(\xi_j)+h_n\}} - n^2\lambda(1-\tau)\int_{\mathbb{R}} (F(x+h_n)-F(x))\,dF(x)\\
&= [n\lambda] \sum_{j=[n\tau]+1}^{n} \big(F_{[n\lambda]}(G(\xi_j)+h_n) - F_{[n\lambda]}(G(\xi_j))\big)\\
&\qquad - n^2\lambda(1-\tau)\int_{\mathbb{R}} (F(x+h_n)-F(x))\,dF(x)\\
&= [n\lambda](n-[n\tau])\int_{\mathbb{R}} \big(F_{[n\lambda]}(x+h_n) - F_{[n\lambda]}(x)\big)\,dF_{[n\tau]+1,n}(x)\\
&\qquad - n^2\lambda(1-\tau)\int_{\mathbb{R}} (F(x+h_n)-F(x))\,dF(x)\\
&= [n\lambda](n-[n\tau])\Big(\int_{\mathbb{R}} (F_{[n\lambda]}(x+h_n) - F_{[n\lambda]}(x))\,dF_{[n\tau]+1,n}(x)\\
&\qquad - \int_{\mathbb{R}} (F(x+h_n)-F(x))\,dF(x)\Big)\\
&\qquad + \big([n\lambda](n-[n\tau]) - n^2\lambda(1-\tau)\big)\int_{\mathbb{R}} (F(x+h_n)-F(x))\,dF(x).
\end{aligned}$$

Note that by the basic estimate $n\lambda - 1 \le [n\lambda] \le n\lambda + 1$ we obtain

$$\begin{aligned}
&[n\lambda](n-[n\tau]) - n^2\lambda(1-\tau)\\
&= [n\lambda]n(1-\tau) + [n\lambda](n-[n\tau]-n(1-\tau)) - n^2\lambda(1-\tau)\\
&= n\lambda n(1-\tau) + n(1-\tau)\left([n\lambda]-n\lambda\right) + [n\lambda]\left(n-[n\tau]-n(1-\tau)\right) - n^2\lambda(1-\tau)\\
&= [n\lambda]\left(n\tau-[n\tau]\right) + (1-\tau)n\left([n\lambda]-n\lambda\right)\\
&\le [n\lambda] + (1-\tau)n = O(n).
\end{aligned}$$

Together with $|\int_{\mathbb{R}}(F(x+h_n)-F(x))dF(x)| \le 1$, this yields

$$\frac{1}{n\,d_n}\big([n\lambda](n-[n\tau]) - n^2\lambda(1-\tau)\big)\int_{\mathbb{R}} (F(x+h_n)-F(x))\,dF(x) = O\left(\frac{1}{d_n}\right) \to 0,$$

as $n \to \infty$. Hence, in order to show that (4.14) converges to zero in probability, it suffices to show that

$$\frac{[n\lambda](n-[n\tau])}{n\,d_n}\left(\int_{\mathbb{R}} (F_{[n\lambda]}(x+h_n) - F_{[n\lambda]}(x))dF_{[n\tau]+1,n}(x) - \int_{\mathbb{R}} (F(x+h_n)-F(x))dF(x)\right) \tag{4.16}$$

converges to zero in probability. In order to prove this, we rewrite the difference of the integrals in (4.16) as

$$\begin{aligned}
&\int_{\mathbb{R}} \big(F_{[n\lambda]}(x+h_n) - F_{[n\lambda]}(x)\big)\, dF_{[n\tau]+1,n}(x) - \int_{\mathbb{R}} (F(x+h_n) - F(x))\, dF(x) \qquad (4.17)\\
&= \int_{\mathbb{R}} \big(F_{[n\lambda]}(x+h_n) - F_{[n\lambda]}(x)\big) - (F(x+h_n) - F(x))\, dF_{[n\tau]+1,n}(x)\\
&\quad + \int_{\mathbb{R}} (F(x+h_n) - F(x))\, d(F_{[n\tau]+1,n} - F)(x)\\
&= \int_{\mathbb{R}} \big(F_{[n\lambda]}(x+h_n) - F_{[n\lambda]}(x)\big) - (F(x+h_n) - F(x))\, dF_{[n\tau]+1,n}(x)\\
&\quad - \int_{\mathbb{R}} \big(F_{[n\tau]+1,n}(x) - F(x)\big)\, d(F(x+h_n) - F(x)),
\end{aligned}$$

where we have used integration by parts in the final step. Thus, in order to prove that (4.16) converges to zero, it suffices to show that the following two terms converge in probability, as $n \to 0$,

$$\frac{1}{d_n}[n\lambda] \int_{\mathbb{R}} \big((F_{[n\lambda]}(x+h_n) - F_{[n\lambda]}(x)) - (F(x+h_n) - F(x))\big)\, dF_{[n\tau]+1,n}(x) \to 0 \qquad (4.18)$$

$$\frac{1}{d_n}(n - [n\tau]) \int_{\mathbb{R}} \big(F_{[n\tau]+1,n}(x) - F(x)\big)\, d(F(x+h_n) - F(x)) \to 0. \qquad (4.19)$$

In order to prove (4.18) and (4.19), we now employ the empirical process non-central limit theorem (3.8), just like in the proof for the limit under the null hypothesis. Note that by definition of the e.d.f., for any $\lambda \le \tau$

$$([n\tau] - [n\lambda])(F_{[n\lambda]+1,[n\tau]}(x) - F(x)) = [n\tau](F_{[n\tau]}(x) - F(x)) - [n\lambda](F_{[n\lambda]}(x) - F(x)).$$

Hence, we may deduce from (3.8) the following limit theorem for the e.d.f. of the observations $X_{[n\lambda]+1}, \ldots, X_{[n\tau]}$,

$$\sup_{0\le\lambda\le\tau,\, x\in\mathbb{R}} \big| d_n^{-1}([n\tau] - [n\lambda])(F_{[n\lambda]+1,[n\tau]}(x) - F(x)) - J(x)(Z(\tau) - Z(\lambda))\big| \to 0, \qquad (4.20)$$

almost surely. For $\tau = 1$, this is

$$\sup_{0\le\lambda\le 1,\, x\in\mathbb{R}} \big| d_n^{-1}(n - [n\lambda])(F_{[n\lambda]+1,n} - F(x)) - J(x)(Z(1) - Z(\lambda))\big| \to 0, \qquad (4.21)$$

almost surely. Now we return to (4.18), which is to show, and write

$$\begin{aligned}
&\left| \int_{\mathbb{R}} \frac{1}{d_n} [n\lambda] \left((F_{[n\lambda]}(x+h_n) - F_{[n\lambda]}(x)) - (F(x+h_n) - F(x)) \right) dF_{[n\tau]+1,n}(x) \right| \qquad (4.22)\\
&\leq \left| \int_{\mathbb{R}} (J(x+h_n) - J(x)) Z(\lambda)\, dF_{[n\tau]+1,n}(x) \right| \\
&\quad + \sup_{x\in\mathbb{R}, 0\leq\lambda\leq 1} \left| \frac{1}{d_n} [n\lambda] \left((F_{[n\lambda]}(x+h_n) - F_{[n\lambda]}(x)) - (F(x+h_n) - F(x)) \right) \right. \\
&\quad \left. - (J(x+h_n) - J(x)) Z(\lambda) \right| \\
&\leq \left| \int_{\mathbb{R}} (J(x+h_n) - J(x))\, dF_{[n\tau]+1,n}(x) \right| \sup_{0\leq\lambda\leq 1} |Z(\lambda)| \\
&\quad + \sup_{x\in\mathbb{R}, 0\leq\lambda\leq 1} \left| \frac{1}{d_n} [n\lambda] \left((F_{[n\lambda]}(x+h_n) - F_{[n\lambda]}(x)) - (F(x+h_n) - F(x)) \right) \right. \\
&\quad \left. - (J(x+h_n) - J(x)) Z(\lambda) \right|.
\end{aligned}$$

The second term on the right-hand side converges to zero by (3.8). Concerning the first term, note that

$$J(x) = \int_{\mathbb{R}} I_{\{G(y)\leq x\}} H_m(y)\varphi(y)\, dy = -\int_{\mathbb{R}} I_{\{x\leq G(y)\}} H_m(y)\varphi(y)\, dy, \qquad (4.23)$$

where $\varphi(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$ denotes the standard normal density function. For the second identity, we have used the fact that $G(\xi)$, by assumption, has a continuous distribution and that $\int_{\mathbb{R}} H_m(y)\varphi(y)\, dy = 0$ for $m \geq 1$. Using (4.23), we thus obtain

$$\begin{aligned}
\int_{\mathbb{R}} J(x)\, dF_{[n\tau]+1,n}(x) &= -\int_{\mathbb{R}} \int_{\mathbb{R}} I_{\{x\leq G(y)\}} H_m(y)\varphi(y)\, dy\, dF_{[n\tau]+1,n}(x) \qquad (4.24)\\
&= -\int_{\mathbb{R}} \left(\int_{\mathbb{R}} I_{\{x\leq G(y)\}}\, dF_{[n\tau]+1,n}(x) \right) H_m(y)\varphi(y)\, dy \\
&= -\int_{\mathbb{R}} F_{[n\tau]+1,n}(G(y)) H_m(y)\varphi(y)\, dy,
\end{aligned}$$

and, using analogous arguments,

$$\int_{\mathbb{R}} J(x+h_n)\, dF_{[n\tau]+1,n}(x) = -\int_{\mathbb{R}} F_{[n\tau]+1,n}(G(y) - h_n)\, H_m(y)\varphi(y)\, dy. \qquad (4.25)$$

By the Glivenko-Cantelli theorem, applied to the stationary, ergodic process $(G(\xi_i))_{i\geq 1}$, we get $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0$, almost surely. Since

$$F_{[n\tau]+1,n}(x) = \frac{n}{n - [n\tau]} F_n(x) - \frac{[n\tau]}{n - [n\tau]} F_{[n\tau]}(x),$$

and thus

$$\begin{aligned}
\left| F_{[n\tau]+1,n}(x) - F(x) \right| \leq{}& \left| \frac{n}{n - [n\tau]} (F_n(x) - F(x)) \right| + \left| \frac{[n\tau]}{n - [n\tau]} (F_{[n\tau]}(x) - F(x)) \right| \\
&+ \left| \left(\frac{[n\tau]}{n - [n\tau]} - \frac{n}{n - [n\tau]} \right) F(x) + F(x) \right|,
\end{aligned}$$

we get that, almost surely,

$$\sup_{x\in\mathbb{R}} \left|F_{[n\tau]+1,n}(x) - F(x)\right| \to 0. \tag{4.26}$$

Returning to the first term on the right-hand side of (4.22), we obtain, using (4.24) and (4.25),

$$\begin{aligned}
&\left|\int_{\mathbb{R}} (J(x+h_n) - J(x))\ dF_{[n\tau]+1,n}(x)\right| \\
&= \left|\int_{\mathbb{R}} \left(F_{[n\tau]+1,n}(G(y)-h_n) - F_{[n\tau]+1,n}(G(y))\right) H_m(y)\varphi(y)\ dy\right| \\
&\le \int_{\mathbb{R}} |F(G(y)-h_n) - F(G(y))|\, |H_m(y)|\, \varphi(y)\ dy \\
&\qquad + 2\sup_x \left|F_{[n\tau]+1,n}(x) - F(x)\right| \int_{\mathbb{R}} |H_m(y)|\, \varphi(y)\ dy.
\end{aligned}$$

Both terms on the right-hand side converge to zero; the second one by (4.26), the first one by continuity of F and Lebesgue's dominated convergence theorem, since F is bounded by 1. In both cases, we have made use of the fact that $\int |H_m(y)|\varphi(y)\, dy < \infty$. Thus we have finally established (4.18). In order to prove (4.19), we observe that

$$\begin{aligned}
&\frac{1}{d_n}(n-[n\tau]) \int_{\mathbb{R}} \left(F_{[n\tau]+1,n}(x) - F(x)\right)\ d(F(x+h_n) - F(x)) \\
&\le \left|\int_{\mathbb{R}} J(x)(Z(1) - Z(\tau)) d(F(x+h_n) - F(x))\right| \\
&\qquad + \sup_{x\in\mathbb{R}} \left|\frac{1}{d_n}(n-[n\tau])(F_{[n\tau]+1,n}(x) - F(x)) - J(x)(Z(1)-Z(\tau))\right| \\
&\le \left|\int_{\mathbb{R}} J(x)\ d(F(x+h_n) - F(x))\right|\ |Z(1) - Z(\tau)| \\
&\qquad + \sup_{x\in\mathbb{R}} \left|\frac{1}{d_n}(n-[n\tau])(F_{[n\tau]+1,n}(x) - F(x)) - J(x)(Z(1)-Z(\tau))\right|.
\end{aligned}$$

The second term on the right-hand side converges to zero by (4.21). Concerning the first term, note that

$$\int_{\mathbb{R}} J(x)\ d(F(x+h_n) - F(x)) = E\left[J(G(\xi_i) - h_n) - J(G(\xi_i))\right].$$

Applying Lebesgue's dominated convergence theorem and making use of the fact that, by assumption, J is continuous, we obtain that $\int_{\mathbb{R}} J(x)\, d(F(x+h_n) - F(x)) \to 0$. In this way, we have finally proved that (4.14) converges to zero, in probability. By similar arguments, we can prove this for (4.15), which finally ends the proof of Theorem 4.2. □

4.3 Asymptotic Relative Efficiency

In this section, we calculate the *asymptotic relative efficiency* (ARE) of the "Wilcoxon-type" test, based on (4.1), over the "difference-of-means" test, based on (4.2). To do

so, we calculate the number of observations needed to detect a small level shift h at time $[\tau n]$ with a test of given (asymptotic) level α and given (asymptotic) power β, both for the "Wilcoxon-type" test and the "difference-of-means" test, and denote these numbers by n_W and n_D, respectively. We then define the asymptotic relative efficiency of the "Wilcoxon-type" test over the "difference-of-means" test by

$$\boxed{ARE(W,D) = \lim_{h\to 0} \frac{n_D}{n_W}.} \tag{4.27}$$

It will turn out that this limit exists and that the ARE does not depend on the choice of τ, α, β. An ARE less than 1 means that the "Wilcoxon-type" test needs on large scale more observations than the "difference-of-means" test in order to detect a given jump on the same level with the same power; this is what we call less efficient.

As preparation, we first calculate a quantity that is related to the ARE, namely the ratio of the sizes of level shifts that can be detected by the two tests, based on the same number of observations n, again for given values of τ, α, β. We denote the corresponding level shifts by $h_W(n)$ and $h_D(n)$, respectively, assuming that these numbers depend on n in the following way

$$h_W(n) = c_W \frac{d_n}{n}$$
$$h_D(n) = c_D \frac{d_n}{n},$$

as specified in Theorem 4.1 and Theorem 4.2. In order to simplify the following considerations, we take a one-sided change-point test, thus rejecting the hypothesis of no change-point for large values of

$$\max_{1\le k\le n-1} \sum_{i=1}^{k} \sum_{j=k+1}^{n} (X_j - X_i)$$

and

$$\max_{1\le k\le n-1} \sum_{i=1}^{k} \sum_{j=k+1}^{n} I_{\{X_i \le X_j\}},$$

respectively. These are the appropriate tests when testing against the alternative of a non-negative level shift. In order to obtain tests that have asymptotically level α, the "difference-of-means" test and the "Wilcoxon-type" test reject the null hypothesis when the test statistics

$$\frac{m!}{n\, d_n\, a_m} \max_{1\le k\le n-1} \sum_{i=1}^{k} \sum_{j=k+1}^{n} (X_j - X_i) \tag{4.28}$$

and

$$\frac{m!}{n\, d_n \int_{\mathbb{R}} J_m(x) dF(x)} \max_{1\le k\le n-1} \sum_{i=1}^{k} \sum_{j=k+1}^{n} \left(I_{\{X_i \le X_j\}} - \frac{1}{2} \right) \tag{4.29}$$

exceed the upper α quantile q_α of the distribution of $\sup_{0\le\lambda\le1}(Z_m(\lambda)-\lambda Z_m(1))$, see Theorem 3.5 and Theorem 3.1 and keep in mind that we changed the sign of the kernel of the "difference-of-means" test statistic here. Under the sequence of alternatives $A_{\tau,h_D(n)}$ and $A_{\tau,h_W(n)}$, respectively, the asymptotic distribution of the above test statistics is given by

$$\sup_{0\le\lambda\le1}\left(Z_m(\lambda)-\lambda Z_m(1)+\frac{c_D\,m!}{a_m}\delta_\tau(\lambda)\right)$$

and

$$\sup_{0\le\lambda\le1}\left(Z_m(\lambda)-\lambda Z_m(1)+\frac{c_W\,m!\int f^2(x)dx}{\int_{\mathbb{R}}J_m(x)dF(x)}\delta_\tau(\lambda)\right),$$

respectively, see Theorem 4.1 and Theorem 4.2 (or Corollary 4.2). Thus, the asymptotic power of these tests, based on (4.28) and (4.29), is given by the following formulae,

$$P\left(\sup_{0\le\lambda\le1}\left(Z_m(\lambda)-\lambda Z_m(1)+\frac{c_D\,m!}{a_m}\delta_\tau(\lambda)\right)\ge q_\alpha\right) \tag{4.30}$$

and

$$P\left(\sup_{0\le\lambda\le1}\left(Z_m(\lambda)-\lambda Z_m(1)+\frac{c_W\,m!\int f^2(x)dx}{\int_{\mathbb{R}}J_m(x)dF(x)}\delta_\tau(\lambda)\right)\ge q_\alpha\right). \tag{4.31}$$

Thus, if we want the two tests to have identical power, we have to choose c_D and c_W in such a way that

$$\frac{c_D\,m!}{a_m}\delta_\tau(\lambda)=\frac{c_W\,m!\int f^2(x)dx}{\int_{\mathbb{R}}J_m(x)dF(x)}\delta_\tau(\lambda),$$

which yields

$$\boxed{\frac{h_D(n)}{h_W(n)}=\frac{c_D}{c_W}=\frac{a_m\int_{\mathbb{R}}f^2(x)dx}{\int_{\mathbb{R}}J_m(x)dF(x)}.}$$

This quantity gives the ratio of the height of a level shift that can be detected by a "difference-of-means" test over the height that can be detected by a "Wilcoxon-type" test, when both tests are assumed to have the same level α, the same power β and the shifts are taking place at the same time $[n\tau]$. In addition, we assume that the tests are based on the same number of observations n, which is supposed to be large.

Example. For the case of Gaussian data, i.e. when $G(t)=t$, we have $m=1$, $a_1=-1$ (the minus sign arises because we consider for the "difference-of-means" test statistic the kernel $y-x$, instead of $x-y$ like in Section 3.4.2) and $\int_{\mathbb{R}}J_1(x)dF(x)=-\frac{1}{2\sqrt{\pi}}$, see (3.14). Thus we obtain

$$\frac{h_D(n)}{h_W(n)}=\frac{c_D}{c_W}=2\sqrt{\pi}\int_{\mathbb{R}}\frac{1}{2\pi}e^{-x^2}dx=\int_{\mathbb{R}}\frac{1}{\sqrt{\pi}}e^{-x^2}dx=1. \tag{4.32}$$

Hence, both tests can asymptotically, as $n\to\infty$, detect level shifts of the same height. This is surprising, because according to conventional statistical wisdom, a Gauß-type test such as the "difference-of-means" test should outperform a rank test when the underlying data have a normal distribution; this is at least the case for independent observations.

In order to calculate the ARE now, we need to consider the ratio of the sample sizes n_W and n_D corresponding to a given level shift h_n. We will thus study the probability

$$\psi(t) := P\left(\sup_{0\le\lambda\le 1}\left(Z_m(\lambda) - \lambda Z_m(1) + t\,\delta_\tau(\lambda)\right) \ge q_\alpha\right)$$

as a function of t, for fixed values of τ and α. The function ψ is monotonically increasing. We define the generalized inverse

$$\psi^-(\beta) := \inf\{t \ge 0 : \psi(t) \ge \beta\}.$$

Thus, we get

$$P\left(\sup_{0\le\lambda\le 1}\left(Z_m(\lambda) - \lambda Z_m(1) + \psi^-(\beta)\,\delta_\tau(\lambda)\right) \ge q_\alpha\right) \ge \beta, \tag{4.33}$$

and, in fact, for given τ, α and β, $\psi^-(\beta)$ is the smallest number having this property.

With the help of the function $\psi^-(\beta)$, we can apply Theorem 4.1 and Theorem 4.2 in order to calculate the number of observations needed to detect a level shift of a given height. By comparing (4.30) and (4.33), we can detect a level shift of size h at time $[n\tau]$ with the "difference-of-means" test of (asymptotic) level α and power β based on n observations, if $h_D(n) = \frac{d_n}{n}\,c_D$, where c_D satisfies $\frac{c_D\,m!}{a_m} = \psi^-(\beta)$. Hence we obtain that $h_D(n)$ has to satisfy

$$h_D(n) = \frac{d_n}{n}\frac{a_m}{m!}\psi^-(\beta).$$

Similarly, by comparing (4.31) and (4.33), we get for the "Wilcoxon-type" test that n has to satisfy

$$h_W(n) = \frac{d_n}{n}\frac{\int_{\mathbb{R}} J_m(x)\,dF(x)}{m!\int_{\mathbb{R}} f^2(x)\,dx}\psi^-(\beta).$$

Solving these two equations for n and denoting the resulting numbers of observations by n_D and n_W, respectively, we obtain

$$n_D = \left(\frac{h_D\,m!}{\psi^-(\beta)L^{m/2}(n)a_m}\right)^{-2/Dm}$$
$$n_W = \left(\frac{h_W\,m!\int_{\mathbb{R}} f^2(x)\,dx}{\psi^-(\beta)L^{m/2}(n)\int_{\mathbb{R}} J_m(x)\,dF(x)}\right)^{-2/Dm},$$

and thus we have established the following theorem.

Theorem 4.3. *Suppose that $(\xi_i)_{i\ge1}$ is a stationary Gaussian process with mean zero, variance 1 and auto-covariance function (1.1) with $0 \le D < \frac{1}{m}$. Moreover, let $G \in \mathcal{G}^2$, and assume that $G(\xi_i)$ has continuous distribution function $F(x)$. Let m denote the Hermite rank of the class of functions $I_{\{G(\xi_i)\le x\}} - F(x)$, $x \in \mathbb{R}$. Then, the ARE of the "Wilcoxon-type" test over the "difference-of-means" test, as defined in (4.27), is given by*

$$\boxed{ARE(W,D) = \left(\frac{a_m\int_{\mathbb{R}} f^2(x)dx}{\int_{\mathbb{R}} J_m(x)dF(x)}\right)^{2/Dm}}. \tag{4.34}$$

Here, a_m is the m-th Hermite coefficient of $G(\cdot)$, as defined in Section 1.4.2, and $J_m(x)$ is the m-th Hermite coefficient of the class $I_{\{G(\cdot)\le x\}} - F(x)$, $x \in \mathbb{R}$. The "Wilcoxon-type" test and the "difference-of-means" test are based on (4.28) *and* (4.29)*, respectively.*

Example. (i) In the case of Gaussian observations, i.e. $G(t) = t$, we have $m = 1$, $a_1 = -1$, $f(x) = \varphi(x) = (2\pi)^{-1/2} e^{-x^2/2}$ and $\int_{\mathbb{R}} J(x)\, dF(x) = -(2\sqrt{\pi})^{-1}$, like in (3.14), so after all we obtain

$$ARE(W, D) = 1.$$

(ii) When we consider the transformation

$$G(t) = \frac{1}{\sqrt{3/4}} \left((\Phi(t))^{-1/3} - \frac{3}{2} \right),$$

we obtain, according to Section 3.6.3, standardized Pareto(3,1) data with p.d.f. and c.d.f.

$$f_{3,1,\mathrm{st}}(x) = \begin{cases} 3\sqrt{\frac{3}{4}} \left(\sqrt{\frac{3}{4}}\, x + \frac{3}{2} \right)^{-4} & \text{if } x \ge -\sqrt{\frac{1}{3}} \\ 0 & \text{else} \end{cases}$$

and

$$F_{3,1,\mathrm{st}}(x) = \begin{cases} 1 - \left(\frac{1}{\sqrt{3/4 t + 3/2}} \right)^3 & t \ge \frac{1-3/2}{\sqrt{3/4}} \\ 0 & \text{else} \end{cases},$$

and the Hermite rank m equals 1. With numerical integration we obtain

$$ARE(W, D) \approx (-2.678)^{2/D}.$$

We will illustrate these findings by a set of computer simulations in Section 4.6.

4.4 ARE for i.i.d. data

We have shown that in the case of LRD data, the ARE of the "Wilcoxon-type" test and the "difference-of-means" test is 1 for Gaussian data. In this section, we will compare this surprising result with the case of i.i.d. Gaussian data. We will find that in this case, the ARE is $\frac{3}{\pi}$, i.e. the "Wilcoxon-type" test is less efficient than the "difference-of-means" test.

We consider i.i.d. observations $X_1, \ldots, X_n$ with $X_i \sim \mathcal{N}(0, 1)$ and the U-statistic

$$U_k = \sum_{i=1}^{k} \sum_{j=k+1}^{n} h(X_i, X_j).$$

As kernel we will choose $h_D(x,y) = y - x$ and $h_W(x,y) = I_{\{x \le y\}} - \frac{1}{2}$, in other words we consider

$$U_k^{(D)} = \sum_{i=1}^{k} \sum_{j=k+1}^{n} h_D(X_i, X_j) = \sum_{i=1}^{k} \sum_{j=k+1}^{n} (X_j - X_i),$$
$$U_k^{(W)} = \sum_{i=1}^{k} \sum_{j=k+1}^{n} h_W(X_i, X_j) = \sum_{i=1}^{k} \sum_{j=k+1}^{n} \left(I_{\{X_i \le X_j\}} - \frac{1}{2} \right).$$

Both kernels h_D, h_W are antisymmetric, i.e. they satisfy $h(x,y) = -h(y,x)$, so in order to determine the limit behaviour of $U_k^{(D)}$ and $U_k^{(W)}$, we can apply the limit theorems of Csörgő and Horváth (1988).

Theorem 4.4. *Let $X_1, \ldots, X_n$ be i.i.d. random variables with $X_i \sim \mathcal{N}(0,1)$. Under the null hypothesis of no change in the mean, it holds*

$$\boxed{\sup_{0 \le \lambda \le 1} \left| \frac{1}{n^{3/2}} U_{[\lambda n]}^{(D)} - BB_{1,n}(\lambda) \right| = o_P(1)} \tag{4.35}$$

and

$$\boxed{\sup_{0 \le \lambda \le 1} \left| \frac{1}{n^{3/2} \sqrt{\frac{1}{12}}} U_{[\lambda n]}^{(W)} - BB_{2,n}(\lambda) \right| = o_P(1),} \tag{4.36}$$

where $(BB_{i,n}(\lambda))_{0 \le \lambda \le 1}$, $i = 1, 2$, is a sequence of Brownian bridges with mean $E[BB_{i,n}(\lambda)] = 0$ and auto-covariance $E[BB_{i,n}(s)\, BB_{i,n}(t)] = \min(s,t) - st$.

Proof. By Theorem 4.1 of Csörgő and Horváth (1988), it holds under the null hypothesis H that

$$\sup_{0 \le \lambda \le 1} \left| \frac{1}{n^{3/2} \sigma} U_{[\lambda n]} - BB_n(\lambda) \right| = o_P(1)$$

where $(BB_n(\lambda))_{0 \le \lambda \le 1}$ is a sequence of Brownian bridges like $BB_{1,n}$ and $BB_{2,n}$ above and where $\sigma^2 = E[\tilde{h}^2(X_1)]$ with $\tilde{h}(t) = E[h(t, X_1)]$. The kernel h has to fulfill $E[h^2(X_1, X_2)] < \infty$ which is the case for $h_D(x,y) = y - x$ and $h_W(x,y) = I_{\{x \le y\}} - \frac{1}{2}$ and Gaussian X_i. □

Theorem 4.5. *Let $X_1, \ldots, X_n$ be i.i.d. random variables with $X_i \sim \mathcal{N}(0,1)$. Under the sequence of alternatives $A_{\tau, h_n}(n)$, as defined in (4.3), and with $h_n = \frac{1}{\sqrt{n}} c$, where c is a constant, it holds*

$$\boxed{\left(\frac{1}{n^{3/2}} U_{[\lambda n]}^{(D)} \right)_{0 \le \lambda \le 1} \to (BB_1(\lambda) + c\delta_\tau(\lambda))_{0 \le \lambda \le 1}} \tag{4.37}$$

and

$$\boxed{\left(\frac{1}{n^{3/2}\sqrt{\frac{1}{12}}}U^{(W)}_{[\lambda n]}\right)_{0\le\lambda\le 1} \to \left(BB_2(\lambda)+\frac{c}{2\sqrt{\pi}\cdot\sqrt{\frac{1}{12}}}\delta_\tau(\lambda)\right)_{0\le\lambda\le 1}} \tag{4.38}$$

in distribution, where $(BB_i(\lambda))_{0\le\lambda\le 1}$ *is a Brownian bridge,* $i=1,2$*, and* $\delta_\tau(\lambda)$ *is defined in* (4.7)*.*

Proof. First, we prove (4.37). Like for the case of LRD observations, we decompose the statistic, so that we obtain under the sequence of alternatives $A_{\tau,h_n}(n)$

$$\frac{1}{n^{3/2}}U^{(D)}_{[\lambda n]} = \frac{1}{n^{3/2}}\sum_{i=1}^{[\lambda n]}\sum_{j=[\lambda n]+1}^{n}(\epsilon_j-\epsilon_i)+\frac{1}{n^{3/2}}\sum_{i=1}^{[\lambda n]}\sum_{j=[\lambda n]+1}^{n}(\mu_j-\mu_i).$$

By Theorem 4.4, the first term on the right-hand side converges to a Brownian bridge $BB(\lambda)$. For the second term we have like in the proof for LRD observations

$$\frac{1}{n^{3/2}}\sum_{i=1}^{[\lambda n]}\sum_{j=[\lambda n]+1}^{n}(\mu_j-\mu_i)\sim\sqrt{n}h_n\delta_\tau(\lambda),$$

and in order for the right-hand side to converge, we have to choose $h_n=\frac{1}{\sqrt{n}}c$.

Now prove (4.38). Again like for LRD observations, we decompose the statistic into one term that converges like under the null hypothesis and one term which converges to a constant. Under the sequence of alternatives $A_{\tau,h_n}(n)$ and for the case $\lambda\le\tau$, this decomposition is

$$\frac{1}{n^{3/2}}U^{(W)}_{[\lambda n]} = \frac{1}{n^{3/2}}\sum_{i=1}^{[\lambda n]}\sum_{j=[\lambda n]+1}^{n}\left(I_{\{\epsilon_i\le\epsilon_j\}}-\frac{1}{2}\right)+\frac{1}{n^{3/2}}\sum_{i=1}^{[\lambda n]}\sum_{j=[\tau n]+1}^{n}I_{\{\epsilon_j<\epsilon_i\le\epsilon_j+h_n\}}. \tag{4.39}$$

The first term converges uniformly to a Brownian Bridge, like under the null hypothesis. We will show that, if the observations $\epsilon_i=G(\xi_i)$ are i.i.d. with c.d.f. F which has two bounded derivatives $F'=f$ and F'', the second term converges uniformly to $c\lambda(1-\tau)\int_{\mathbb{R}}f^2(x)\,dx$, which is $c\delta_\tau(\lambda)\int_{\mathbb{R}}f^2(x)\,dx$ for the case $\lambda\le\tau$. In the case of standard normally distributed $G(\xi_i)$, i.e. for $F=\Phi$ and $f=\varphi$, this function is $c(2\sqrt{\pi})^{-1}\delta_\tau(\lambda)$. To this end, we consider the following sequence of Hoeffding decompositions for the sequence of kernels $h_n(x,y)=I_{\{y<x\le y+h_n\}}$:

$$h_n(x,y)=\vartheta_n+h_{1,n}(x)+h_{2,n}(y)+h_{3,n}(x,y) \tag{4.40}$$

Let $X, Y \sim F$ be i.i.d. random variables. Then we define

$$\begin{aligned}\vartheta_n &:= E[h_n(X,Y)] \\ &= P(Y \le X \le Y + h_n) \\ &= \int_{\mathbb{R}} \left(\int_y^{y+h_n} f(x)\, dx \right) f(y)\, dy \\ &= \int_{\mathbb{R}} (F(y+h_n) - F(y))\, f(y)\, dy \\ &= h_n \int_{\mathbb{R}} \frac{F(y+h_n) - F(y)}{h_n} f(y)\, dy \\ &\sim h_n \int_{\mathbb{R}} f^2(y)\, dy,\end{aligned}$$

where in the last step we have used that $F(y+h_n) - F(y)/h_n \to f(y)$ and the dominated convergence theorem. Moreover, we set

$$\begin{aligned}h_{1,n}(x) &:= E[h_n(x,Y)] - \vartheta_n \\ &= E[I_{\{Y<x\le Y+h_n\}}] - \vartheta_n \\ &= F(x) - F(x-h_n) - \vartheta_n\end{aligned}$$

and

$$\begin{aligned}h_{2,n}(y) &:= E[h_n(X,y)] - \vartheta_n \\ &= E[I_{\{y<X\le y+h_n\}}] - \vartheta_n \\ &= F(y+h_n) - F(y) - \vartheta_n\end{aligned}$$

and

$$\begin{aligned}h_{3,n}(x,y) &:= h_n(x,y) - h_{1,n}(x) - h_{2,n}(y) - \vartheta_n \\ &= I_{\{y<x\le y+h_n\}} - F(x) + F(x-h_n) + \vartheta_n - F(y+h_n) + F(y).\end{aligned}$$

We will now show that

$$\sup_{0\le\lambda\le\tau} \frac{1}{n^{3/2}} \left| \sum_{i=1}^{[\lambda n]} \sum_{j=[\tau n]+1}^{n} (h_{1,n}(\epsilon_i) + h_{2,n}(\epsilon_j) + h_{3,n}(\epsilon_i, \epsilon_j)) \right| \to 0 \tag{4.41}$$

in probability, and from this it follows by the sequence of Hoeffding decompositions (4.40) that

$$\sup_{0\le\lambda\le\tau} \frac{1}{n^{3/2}} \left| \sum_{i=1}^{[\lambda n]} \sum_{j=[\tau n]+1}^{n} (h_n(\epsilon_i, \epsilon_j) - \vartheta_n) \right| \to 0$$

i.e. that the second term in (4.39) converges uniformly to

$$\lim_{n\to\infty} \frac{1}{n^{3/2}} [\lambda n](n - [\tau n])\vartheta_n = \lambda(1-\tau)c \int_{\mathbb{R}} f^2(x)\, dx.$$

We use the triangle inequality and show the uniform convergence to 0 for each of the three terms in (4.41) seperately. Since the parameter λ occurs only in the floor function value $[\lambda n]$, the supremum is in fact a maximum, and the $h_{1,n}(\epsilon_i)$ are i.i.d. random variables, so we can use Kolmogorov's inequality. We obtain for the first term in (4.41)

$$P\left(\sup_{0\le\lambda\le\tau}\frac{n-[\tau n]}{n^{3/2}}\left|\sum_{i=1}^{[\lambda n]}h_{1,n}(\epsilon_i)\right|>s\right)\le\frac{1}{s^2}\frac{n^2(1-\tau)^2}{n^3}\sum_{i=1}^{[\tau n]}\operatorname{Var}[h_{1,n}(\epsilon_i)]. \tag{4.42}$$

Now consider an independent copy ϵ of the ϵ_i and the Taylor expansion of F around the value of ϵ,

$$F(t)=F(\epsilon)+F'(\epsilon)(t-\epsilon)+\frac{F''(\epsilon+\delta)}{2}(t-\epsilon)^2,$$

where the last term is the Lagrange remainder and thus $\epsilon+\delta$ is between ϵ and t. Then we obtain

$$\begin{aligned}\frac{1}{h_n^2}\operatorname{Var}[h_{1,n}(\epsilon)]&=\operatorname{Var}\left[\frac{F(\epsilon)-F(\epsilon-h_n)}{h_n}\right]\\&=\operatorname{Var}\left[f(\epsilon)+F''(\epsilon+\delta)\frac{h_n}{2}\right]\\&=\operatorname{Var}\left[f(\epsilon)\right]+\operatorname{Var}\left[F''(\epsilon+\delta)\frac{h_n}{2}\right]\\&\quad+2\left(E[f(\epsilon)F''(\epsilon+\delta)h_n]-E[f(\epsilon)]\,E[F''(\epsilon+\delta)h_n]\right),\end{aligned}$$

and since $f=F'$ and F'' are bounded by assumption, we receive $\operatorname{Var}[h_{1,n}(\epsilon)]=O(h_n^2)$ and the right-hand side of (4.42) converges to 0 as n increases.

In the same manner, we obtain

$$P\left(\sup_{0\le\lambda\le\tau}\frac{[\lambda n]}{n^{3/2}}\left|\sum_{j=[\tau n]+1}^{n}h_{2,n}(\epsilon_j)\right|>s\right)\le\frac{1}{s^2}\frac{n^2\lambda^2}{n^3}\sum_{j=1}^{n}\operatorname{Var}[h_{2,n}(\epsilon_j)]\to 0. \tag{4.43}$$

Finally, we have to show that

$$\sup_{0\le\lambda\le\tau}\frac{1}{n^{3/2}}\left|\sum_{i=1}^{[\lambda n]}\sum_{j=[\tau n]+1}^{n}h_{3,n}(\epsilon_i,\epsilon_j)\right|\to 0 \tag{4.44}$$

in probability. We set temporarily $l:=[\lambda n]$ and $T:=[\tau n]$ and obtain from Chebyshev's inequality

$$P\left(\max_{0\le l\le T}\frac{1}{n^{3/2}}\left|\sum_{i=1}^{l}\sum_{j=T+1}^{n}h_{3,n}(\epsilon_i,\epsilon_j)\right|>s\right)\le\frac{1}{s^2}\operatorname{Var}\left[\max_{0\le l\le T}\frac{1}{n^{3/2}}\sum_{i=1}^{l}\sum_{j=T+1}^{n}h_{3,n}(\epsilon_i,\epsilon_j)\right].$$

Now for any collection of random variables $Y_1, \ldots, Y_k$, it holds $E[\max\{Y_1^2, \ldots Y_k^2\}] \leq \sum_{i=1}^k EY_i^2$, such that

$$\frac{1}{s^2}\operatorname{Var}\left[\max_{0\leq l\leq T}\frac{1}{n^{3/2}}\sum_{i=1}^{l}\sum_{j=T+1}^{n}h_{3,n}(\epsilon_i,\epsilon_j)\right] \leq \frac{1}{s^2}\frac{1}{n^3}\sum_{l=1}^{T}\operatorname{Var}\left[\sum_{i=1}^{l}\sum_{j=T+1}^{n}h_{3,n}(\epsilon_i,\epsilon_j)\right]$$
$$= \frac{1}{s^2}\frac{1}{n^3}\sum_{l=1}^{T}\sum_{i=1}^{l}\sum_{j=T+1}^{n}\operatorname{Var}\left[h_{3,n}(\epsilon_i,\epsilon_j)\right],$$

where in the last step we have used that $h_{3,n}(\epsilon_i, \epsilon_j)$ are uncorrelated by Hoeffding's decomposition. Now for two i.i.d. random variables ϵ, η, we have, like above with the Taylor expansion of F:

$$\begin{aligned}
\operatorname{Var}\left[h_{3,n}(\epsilon,\eta)\right] &= \operatorname{Var}\left[I_{\{\eta<\epsilon\leq\eta+h_n\}} - F(\epsilon) + F(\epsilon-h_n) + \vartheta_n - F(\eta+h_n) + F(\eta)\right] \\
&= \operatorname{Var}\left[I_{\{\eta<\epsilon\leq\eta+h_n\}} - h_n\left(f(\epsilon)+O_P(h_n)\right) + h_n\left(f(\eta)+O_P(h_n)\right)\right] \\
&= \operatorname{Var}\left[I_{\{\eta<\epsilon\leq\eta+h_n\}}\right] + \operatorname{Var}\left[h_n\left(f(\epsilon)+f(\eta)+O_P(h_n)\right)\right] \\
&\quad + 2\operatorname{Cov}\left[I_{\{\eta<\epsilon\leq\eta+h_n\}}, h_n\left(f(\epsilon)+f(\eta)+O_P(h_n)\right)\right] \\
&\leq (\vartheta_n - \vartheta_n^2) + h_n^2 O(1) + 2\sqrt{(\vartheta_n-\vartheta_n^2)\cdot h_n^2 O(1)} \\
&= O(h_n).
\end{aligned}$$

We have just shown that

$$P\left(\max_{0\leq l\leq T}\frac{1}{n^{3/2}}\left|\sum_{i=1}^{l}\sum_{j=T+1}^{n}h_{3,n}(\epsilon_i,\epsilon_j)\right| > s\right) \leq \frac{1}{s^2}O(h_n),$$

which proves (4.44). So we have proven (4.39) for the case $\lambda \leq \tau$. The proof for $\lambda > \tau$ is similar. □

Now the stage is set to calculate the ARE of the "Wilcoxon-type" test based on $U_{[\lambda n]}^{(W)}$ and the "difference-of-means" test based on $U_{[\lambda n]}^{(D)}$, as defined in the section about the ARE in the LRD case. Let q_α denote the upper α-quantile of the distribution of $\sup_{0\leq\lambda\leq 1} BB(\lambda)$. By Theorem 4.5, the power of both tests is given by

$$P\left(\sup_{0\leq\lambda\leq 1}\left(BB(\lambda) + c_D\delta_\tau(\lambda)\right) \geq q_\alpha\right) \tag{4.45}$$

and

$$P\left(\sup_{0\leq\lambda\leq 1}\left(BB(\lambda) + c_W\frac{1}{\sigma\cdot 2\sqrt{\pi}}\delta_\tau(\lambda)\right) \geq q_\alpha\right) \tag{4.46}$$

where $\sigma^2 = 1/12$ and we assume that

$$h_W(n) = \frac{c_W}{\sqrt{n}}, \qquad h_D(n) = \frac{c_D}{\sqrt{n}}.$$

Thus if we want both tests to have identical power, we must ensure that $c_D = c_W/(\sigma \cdot 2\sqrt{\pi})$, in other words

$$\frac{h_D(n)}{h_W(n)} = \frac{c_D}{c_W} = \frac{1}{\sigma \cdot 2\sqrt{\pi}}.$$

Now we define, quite similar to the proof for LRD observations, the probability

$$\psi(t) := P\left(\sup_{0\le\lambda\le 1} (BB(\lambda) + t\,\delta_\tau(\lambda)) \ge q_\alpha\right),$$

for whose generalized inverse ψ^- holds

$$P\left(\sup_{0\le\lambda\le 1} \left(BB(\lambda) + \psi^-(\beta)\,\delta_\tau(\lambda)\right) \ge q_\alpha\right) \ge \beta. \tag{4.47}$$

Now we compare (4.47) and (4.45) conclude: We can detect a level shift of size h at time $[n\tau]$ with the "difference-of-means" test of (asymptotic) level α and power β based on n observations, if $h_D(n) = \frac{c_D}{\sqrt{n}}$ and where c_D satisfies $c_D = \psi^-(\beta)$; hence we obtain that $h_D(n)$ has to satisfy

$$h_D(n) = \frac{1}{\sqrt{n}}\psi^-(\beta).$$

In the same manner, we get for the Wilcoxon test the conditions $h_W(n) = \frac{c_W}{\sqrt{n}}$ and $c_W/(\sigma 2\sqrt{\pi}) = \psi^-(\beta)$ and thus

$$h_W(n) = \frac{\sigma 2\sqrt{\pi}}{\sqrt{n}}\psi^-(\beta).$$

Solving these two equations for n again and denoting the resulting numbers of observations by n_D and n_W, respectively, we obtain

$$n_D = \left(\frac{1}{h_D}\psi^-(\beta)\right)^2$$
$$n_W = \left(\frac{2\sigma\sqrt{\pi}}{h_W}\psi^-(\beta)\right)^2,$$

and thus we have established the following theorem.

Theorem 4.6. *Let $X_1, \ldots, X_n$ be i.i.d. random variables with $X_i \sim \mathcal{N}(0,1)$. Then*

$$\boxed{ARE(W,D) = \lim_{h\to 0}\frac{n_D}{n_W} = (2\sigma\sqrt{\pi})^{-2} = \frac{3}{\pi},} \tag{4.48}$$

where D, W denote the one-sided "difference-of-means"-test, respectively the one-sided "Wilcoxon-type" test, for the test problem (H, A_{τ,h_n}).

4.5 ARE of Wilcoxon and Gauß test for the two-sample problem

Previously, we did not specify the position of the level shift, and we observed that the ARE of the "Wilcoxon type" test and the "difference-of-means" test for a change point equals 1 in the case of LRD Gaussian data and that it is $\frac{3}{\pi}$ for i.i.d. Gaussian data. Now we show that this effect is already present in the 2-sample problem, where we assume that the position of the level shift is known. In this situation, the "Wilcoxon type" test and the "difference-of-means" test are known as the *Wilcoxon test* and the *Gauß test*. We show that for LRD observations, the ARE of these two tests equals one, in contrast to i.i.d. observations where the ARE of both tests equals $\frac{3}{\pi}$, i.e. where the Wilcoxon test is less efficient than the Gauß test – which is common knowledge in statistics.

For simplicity, we consider equal (and even) sample size for both samples. Our observations are

$$X_i = \begin{cases} \xi_i, & \text{for } i = 1, \dots, [n/2] \\ \xi_i + h, & \text{for } i = [n/2]+1, \dots, n, \end{cases} \tag{4.49}$$

where $(\xi_i)_{i\geq 1}$ is fGn with Hurst coefficient H. We consider two statistics for testing the hypothesis $h = 0$ against the alternative $h > 0$,

$$D_n^{(\mathrm{LRD})} = \frac{1}{n^{H+1}\sqrt{\frac{1}{2^{2H}} - \frac{1}{4}}} \sum_{i=1}^{[n/2]} \sum_{j=[n/2]+1}^{n} (X_j - X_i) \tag{4.50}$$

$$W_n^{(\mathrm{LRD})} = \frac{1}{n^{H+1}\sqrt{\frac{1}{2^{2H}} - \frac{1}{4}} \int_{\mathbb{R}} J(x)dF(x)} \sum_{i=1}^{[n/2]} \sum_{j=[n/2]+1}^{n} \left(I_{\{X_i \leq X_j\}} - \frac{1}{2} \right), \tag{4.51}$$

compare to (4.28) and (4.29). Here, in the case of fGn, which has the covariance structure (1.5), the appropriate scaling is $n\, d_n = n^{2-D/2} = n^{H+1}$, because $D = 2 - 2H$. The normalization will prove to be the right one.

Proposition 4.7. *Under the above assumptions (i.e. fGn with a shift in the mean), $D_n^{(LRD)}$ is normally distributed:*

$$\boxed{D_n^{(LRD)} \sim \mathcal{N}\left(\frac{1/4 \cdot n^2 h}{n^{H+1}\sqrt{\frac{1}{2^{2H}} - \frac{1}{4}}}, 1 \right).}$$

Moreover, $W_n^{(LRD)}$ is asymptotically normally distributed:

$$\boxed{W_n^{(LRD)} \approx \mathcal{N}\left(\frac{1/4 \cdot n^2 h \int f^2(x)dx}{n^{H+1} \int J(x)dF(x)\sqrt{\frac{1}{2^{2H}} - \frac{1}{4}}}, 1 \right),}$$

where $f(x) = \varphi(x)$ is the standard normal p.d.f. and $F(x) = \Phi(x)$ is the standard normal c.d.f..

Remark. Note that the results of Proposition 4.7 hold in similar form in greater generality, not just for the Gaussian case, as long as $m = 1$. Then, by the previous Theorems 4.1 and 4.2, $D_n^{(\mathrm{LRD})}$ and $W_n^{(\mathrm{LRD})}$ are asymptotically normally distributed, one only may obtain different parameters: For $W_n^{(\mathrm{LRD})}$, the first Hermite coefficient a_1 of $G(\cdot)$ may be differernt, and for $W_n^{(\mathrm{LRD})}$, one may obtain different f and F (the respective p.d.f. and c.d.f. of the observations X_i) and a different first Hermite coefficient $J(x)$ of the class of functions $\{I_{\{G(\cdot)\le x\}}, x \in \mathbb{R}\}$. We give the proof only for fGn which is a simple model for Gaussian data. For other Gaussian or non-Gaussian data, which still staisfy $m = 1$, the proof can easily be adapted.

Proof. $D_n^{(\mathrm{LRD})}$ is a linear function of Gaussian variables and thus normally distributed itself. The variance $\frac{1}{2^{2H}} - \frac{1}{4}$ of the two-sample Gauß test statistic can be calculated quite easily; this is carried out in Section 6.3.1. The mean is also clear by some easy calculation.

The second part involves going through the proof of Theorem 4.2. In virtue of (4.12) and Theorem 3.1, $W_n^{(\mathrm{LRD})}$ is asymptotically normally distributed, and by (4.14) we see that for any choice of λ (here we have $\lambda = \tau = 1/2$)

$$\frac{1}{n^{H+1}} \sum_{i=1}^{[n\lambda]} \sum_{j=[n\tau]+1}^{n} I_{\{G(\xi_j)<G(\xi_i)\le G(\xi_j)+h_n\}}$$

is asymptotically close to

$$\begin{aligned}
&\frac{n^2}{n^{H+1}}\lambda(1-\tau)\int_{\mathbb{R}}(F(x+h)-F(x))\,dF(x)\\
&=\frac{n^2h}{n^{H+1}}\lambda(1-\tau)\int_{\mathbb{R}}\frac{F(x+h)-F(x)}{h}\,dF(x)\\
&\approx\frac{n^2h}{n^{H+1}}\frac{1}{4}\int_{\mathbb{R}}f^2(x)\,dx
\end{aligned}$$

like in the remark following Theorem 4.2 with $\lambda(1-\tau) = \delta_\tau(\lambda) = \delta_{1/2}(1/2) = 1/4$. For the standardization of the Wilcoxon test statistic, note that in the case of a strictly monotone transformation G, the limit distribution of the "difference-of-means" test is – up to a norming constant – the same as for the test based on Wilcoxon's rank statistic, or see e.g. for Gaussian data Section 6.3.2, such that we finally obtain

$$\begin{aligned}
E\left[W_n^{(\mathrm{LRD})}\right] &= E\left[\frac{\sum_{i=1}^{[n/2]}\sum_{j=[n/2]+1}^{n}\left(I_{\{X_i\le X_j\}}-\frac{1}{2}\right)}{n^{H+1}\sqrt{\frac{1}{2^{2H}}-\frac{1}{4}}\int_{\mathbb{R}}J(x)dF(x)}\right]\\
&\approx\frac{E\left[\frac{1}{n^{H+1}}\sum_{i=1}^{[n/2]}\sum_{j=[n/2]+1}^{n}I_{\{G(\xi_j)<G(\xi_i)\le G(\xi_j)+h_n\}}\right]}{\sqrt{\frac{1}{2^{2H}}-\frac{1}{4}}\int_{\mathbb{R}}J(x)dF(x)}\\
&\approx\frac{\frac{n^2h}{n^{H+1}}\frac{1}{4}\int_{\mathbb{R}}f^2(x)\,dx}{\sqrt{\frac{1}{2^{2H}}-\frac{1}{4}}\int_{\mathbb{R}}J(x)dF(x)}.
\end{aligned}$$

□

So we clearly see that the ARE of the "Wilcoxon-type" test over the "difference-of-means" of 1 in the case of LRD data can already be observed in the two-sample situation with the Wilcoxon test and the Gauß test.

4.6 Simulations

We have rigorously proven that for Gaussian data, the "difference-of-means" test and the "Wilcoxon-type" test show asymptotically the same performance: Their ARE is 1. For Pareto(3,1) distributed data, we obtained an ARE of approximately $(2.68)^{2/D}$. Now we illustrate these findings by three simulation studies.

- In Section 4.6.1, we repeat the power simulations from Section 3.6.1 for fGn, but with sample size $n = 2,000$ instead of $n = 500$. These new simulations were computationally intensive because the Wilcoxon test takes a long time due to the comparison/ordering of the data.
- Second, we show in Section 4.6.3 that the asymptotic relative efficiency of 1 can already be observed in two-sample problems: We compare the "difference-of-means" two-sample test, also known as the Gauß test, and the Wilcoxon two-sample test in situations with fGn which possesses a level shift of size $h_n = cn^{-D/2}$, according to our theoretical power considerations above.
- Finally, we consider in Section 4.6.4 both two-sample tests also under Pareto(3,1) distributed data in order to illustrate the ARE of around $(2.68)^{2/D}$.

4.6.1 Power of change-point tests

We consider realizations $\xi_1, \ldots, \xi_n$ of a fGn process with Hurst parameter $H = 0.7$ ($D = 0.6$) and observations

$$X_i = \begin{cases} G(\xi_i) & \text{for } i = 1, \ldots, [n\lambda] \\ G(\xi_i) + h & \text{for } i = [n\lambda] + 1, \ldots, n \end{cases},$$

for a function $G \in \mathcal{G}^2$, the class of (with respect to the standard normal measure) normalized and square-integrable functions. Here, we choose $G(t) = t$ in order to obtain Gaussian observations $X_1, \ldots, X_n$. In other words, we consider data which follow the alternative

$$A_{\lambda,h} : \begin{cases} E[X_i] = 0 & \text{for } i = 1, \ldots, [n\lambda] \\ E[X_i] = h & \text{for } i = [n\lambda] + 1, \ldots, n. \end{cases}$$

We let both the break point $k = [\lambda n]$ and the level shift $h := \mu_{k+1} - \mu_k$ vary; specifically, we choose $k = 100, 200, 600, 1000$ and $h = 0.5, 1, 2$. For each of these situations, we compare the power of the "difference-of-means" test and the "Wilcoxon-type" test in the

$h \setminus \lambda$	0.05	0.1	0.3	0.5
0.5	1.024	1.041	1.006	0.999
1	1.165	1.377	1.005	1.000
2	2.199	1.265	1.000	1.000

$h \setminus \lambda$	0.05	0.1	0.3	0.5
0.5	1.019	1.069	1.003	0.997
1	1.201	1.153	1.000	1.000
2	2.579	1.000	1.000	1.000

Table 4.1: Power of the "difference-of-means" test relative to the power of the "Wilcoxon-type" test, β_D/β_W, for $n = 500$ (left) and $n = 2,000$ (right) observations of fGn with LRD parameter $H = 0.7$, different break points $[\lambda n]$ and different level shifts h. Both tests have asymptotically level 5%. The calculations are based on 10,000 simulation runs.

test problem (H, A), see (3.1) and (3.2). In contrast to the simulations in Section 3.6.1, we have here a sample size of $n = 2,000$ instead of $n = 500$. We have repeated each simulation $10,000$ times.

Since our theoretical considerations yield an ARE of 1, we expect that both tests detect jumps equally well – that means that both tests, set on the same level, detect jumps of the same height and at the same position in the same number of observations with the same relative frequency. And indeed, we can clearly see in Figure 4.1 that the power of both tests approximately coincides at many points; differences can be spot only when the break is large or occurs early in the data. Table 4.1 renders this impression more precisely: Here, the quotient of the power of the "difference-of-means" test and the "Wilcoxon-type" test is given, and indeed, it is only in the lower left quarter not close to 1.

Figure 4.2 illustrates how an increased sample size influences the power of the "difference-of-means" test and the "Wilcoxon-type" test. Here the relative change in the power $(\beta_{2000} - \beta_{500})/\beta_{500}$ is displayed for both tests, where β_n denotes the power in a simulation with n observations. As one can expect, there is a significant increase in the power for small and for early level shifts which traditionally are hard to detect. For level shifts that are easy to detect, like big jumps or jumps in the middle, there is no big difference in the test performance, when the sample size is increased from $n = 500$ to $n = 2,000$.

4.6.2 Power of two-sample tests, setting

Now we consider data $X_1, \ldots, X_n$ of the form

$$X_i = \begin{cases} G(\xi_i) & \text{for } i = 1, \ldots, [\frac{n}{2}] \\ G(\xi_i) + h_n & \text{for } i = [\frac{n}{2}] + 1, \ldots, n, \end{cases}$$

where $(\xi_i)_{i \geq 1}$ is standard fGn, $G \in \mathcal{G}^2$ and $h_n > 0$ is a certain positive jump, i.e. in the language of change-point tests, we assume to know the location $\tau = 1/2$ of the change-point. At first we consider $G(t) = t$ which produces fGn as observations, afterwards

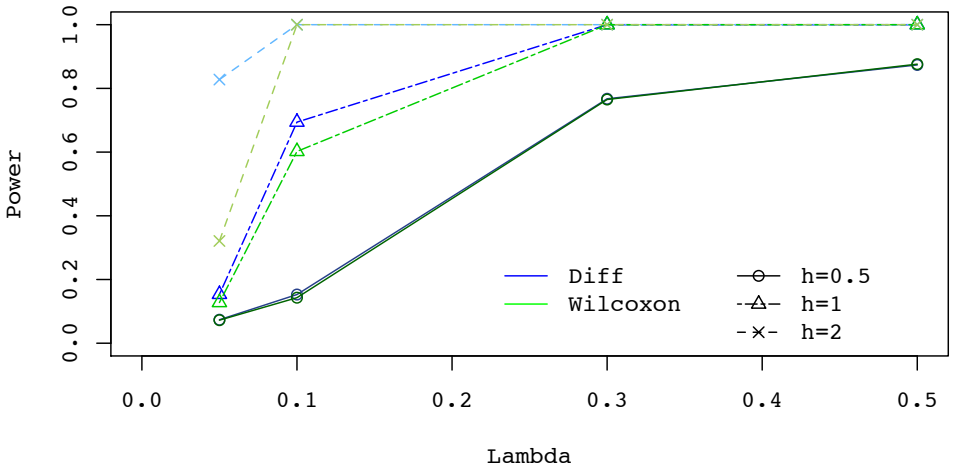


Figure 4.1: Power of "difference-of-means" test (blue) and power of "Wilcoxon-type" test (green) for $n = 2,000$ observations of fGn with LRD parameter $H = 0.7$, different break points $[\lambda n]$ and different level shifts h. Both tests have asymptotically level 5%. The calculations are based on 10,000 simulation runs.

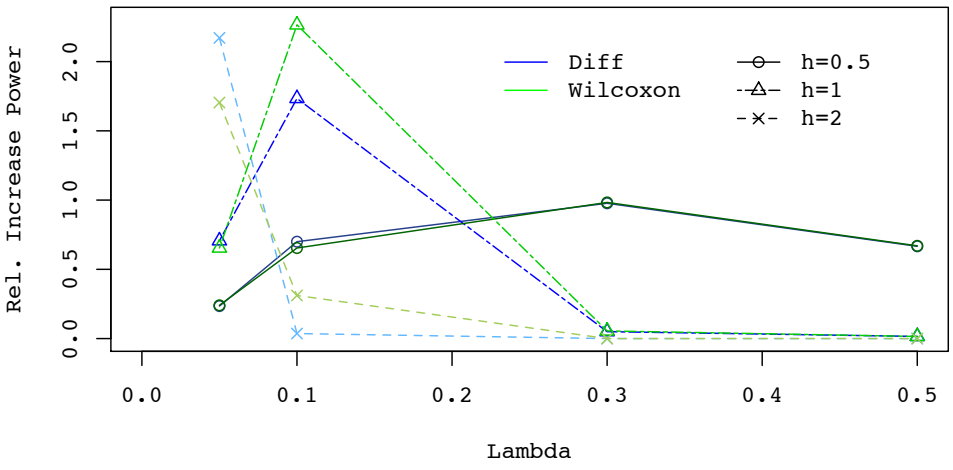


Figure 4.2: Relative change of the power of the "difference-of-means" test (blue) and of the "Wilcoxon-type" test (green) if the sample size increases from $n = 500$ to $n = 2,000$. The simulation is based on fGn with LRD parameter $H = 0.7$, different break points $[\lambda n]$, different level shifts h and each 10,000 simulation runs. Both tests have asymptotically level 5%.

we look at a transformation G which yields standardised Pareto(3,1) distributed data. As jump height we choose

$$h_n = c\frac{d_n}{n} = cn^{-D/2}$$

and analyse how well this jump can be detected, both by the two-sample Gauß test and by the two-sample Wilcoxon test: We test

$$H : h_n = 0 \quad \text{against} \quad A_{1/2,h_n} : h_n > 0.$$

For the Gauß test, we use the statistic

$$D_{1/2,n} = \frac{1}{n\,d_n} \sum_{i=1}^{[n/2]} \sum_{j=[n/2]+1}^{n} (X_j - X_i).$$

Here we have $m = 1$, since G is strictly monotone, thus we obtain by Theorem 3.5

$$D_{1/2,n} \xrightarrow{\mathcal{D}} -a_1 \left(Z_1(1/2) - \frac{1}{2} Z_1(1) \right) \stackrel{\mathcal{D}}{=} -a_1 \mathcal{N}(0, \sigma_{0.5}^2)$$

with

$$\sigma_{0.5}^2 := \mathrm{Var} \left[Z_1(1/2) - \frac{1}{2} Z_1(1) \right] = \frac{1}{2^{2H}} - \frac{1}{4}, \tag{4.52}$$

because Z_1 is fBm. So the two-sample Gauß test rejects H if

$$\frac{D_{1/2,n}}{|a_1|\sigma_{0.5}} > z_\alpha, \tag{4.53}$$

where z_α denotes the upper α-quantile of the standard normal distribution. By Theorem 4.1, we obtain under the alternative the convergence

$$\frac{D_{1/2,n}}{a_1\sigma_{0.5}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) + \frac{c}{4a_1\sigma_{0.5}}.$$

For the Wilcoxon test, we compute the statistic

$$W_{1/2,n} = \frac{1}{n\,d_n} \sum_{i=1}^{[n/2]} \sum_{j=[n/2]+1}^{n} \left(I_{\{X_i \le X_j\}} - \frac{1}{2} \right).$$

By Theorem 3.4, it holds under the null hypothesis of no change

$$\frac{2\sqrt{\pi}\, W_{1/2,n}}{\sigma_{0.5}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

with $\sigma_{0.5}$ as in (4.52), and the two-sample Wilcoxon test rejects H if

$$\frac{2\sqrt{\pi}\, W_{1/2,n}}{\sigma_{0.5}} > z_\alpha, \tag{4.54}$$

where z_α still denotes the upper α-quantile of the standard normal distribution. In Theorem 4.2 we have shown that under the alternative

$$\frac{2\sqrt{\pi}\, W_{1/2,n}}{\sigma_{0.5}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) + \frac{c\,2\sqrt{\pi}}{4\sigma_{0.5}} \int_{\mathbb{R}} f^2(x)\, dx,$$

where f denotes the p.d.f. of the observations X_i.

In what follows, we will compare these two tests based on $D_{1/2,n}$ and $W_{1/2,n}$, as set in (4.53) and (4.54), in two different data situations.

4.6.3 Power of two-sample tests, Gaussian observations

Using the `fArma` package in `R`, we have simulated $n = 50, 100, 500, 1000, 2000$ realizations $\xi_1, \ldots, \xi_n$ of fGn with Hurst parameter $H = 0.7$ (respectively $D = 0.6$). At first, we have considered observations $X_i = G(\xi_i)$ with $G(t) = t$, i.e. fGn. To the second sample of observations, $X_{[n/2]+1}, \ldots, X_n$, we have added a constant, depending on n,

$$h_n = c\frac{d_n}{n} = cn^{-D/2} = \begin{cases} 0.102c & \text{if } n = 2,000 \\ 0.126c & \text{if } n = 1,000 \\ 0.155c & \text{if } n = 500 \\ 0.251c & \text{if } n = 100 \\ 0.309c & \text{if } n = 50. \end{cases} \tag{4.55}$$

To these data, we have applied the two-sample Gauß test and the 2-sample Wilcoxon test, as set in (4.53) and (4.54). Here, the first Hermite coefficient of G is $a_1 = 1$ and $\int_{\mathbb{R}} f^2(x)\, dx = (2\sqrt{\pi})^{-1}$ since f is the standard normal p.d.f.. Under 10,000 simulation runs, we have counted the number of true rejections as an estimate of the power of the respective test; the results are shown in Table 4.2. In Table 4.3, their quotient is given; it is close to 1. Our theoretical findings state that

$$ARE = \lim_{n\to\infty} \frac{n_D}{n_W} = \left(\frac{a_1 \int_{\mathbb{R}} f^2(x)\, dx}{\int_{\mathbb{R}} J_1(x) f(x)\, dx} \right)^{2/D} = 1,$$

which means that both tests, set on the same level, need on large scale the same number of observations in order to detect the same jump with a given probability. And indeed, for the same number of observations and on the same level both tests detect the same jump nearly equally often. So our theoretical findings that the ARE is 1 for Gaussian data can already be seen in two sample situations.

Note that the simulations also illustrate that the special form of h_n (which is defined in (4.55) and provided by our limit theorem) is just right: According to our theoretical findings, a jump(-sequence) of height $c\,h_n$ leads to a non-degenerate limit of the test statistic. And indeed, for a fixed c, the observed power of each test is nearly the same for all sample sizes n which is a plausible evidence that the test statistic under data with jump $c\,h_n$ approaches a certain non-degenerate limit.

4.6.4 Power of two-sample tests, Pareto(3,1) observations

To n realizations $\xi_1, \ldots, \xi_n$ of fGn we have now applied the transformation

$$G(t) = \frac{1}{\sqrt{3/4}} \left((\Phi(t))^{-1/3} - \frac{3}{2} \right),$$

which produces observations $X_i = G(\xi_i)$ which follow a standardised Pareto(3,1) distribution. To the second sample of observations, $X_{[n/2]+1}, \ldots, X_n$, we have again added a

$c \setminus n$	50	100	500	1000	2000
0.5	0.097	0.097	0.093	0.102	0.094
1	0.174	0.168	0.168	0.178	0.169
2	0.402	0.403	0.396	0.404	0.403
4	0.875	0.867	0.872	0.874	0.872

$c \setminus n$	50	100	500	1000	2000
0.5	0.110	0.104	0.096	0.103	0.093
1	0.192	0.180	0.173	0.182	0.171
2	0.414	0.411	0.403	0.407	0.404
4	0.876	0.870	0.872	0.875	0.870

Table 4.2: Power of the two-sample Gauß test (upper table) and the two-sample Wilcoxon test (lower table), based on $10,000$ repetitions of a fGn series of length n with Hurst parameter $H = 0.7$ (D=0.6) and with a jump of height $h_n = cn^{-D/2}$ at the half.

$c \setminus n$	50	100	500	1000	2000
0.5	0.886	0.933	0.967	0.991	1.013
1	0.907	0.931	0.972	0.981	0.986
2	0.969	0.979	0.983	0.992	0.998
4	0.998	0.997	0.999	1.000	1.002

Table 4.3: The power of the two-sample Gauß test relative to the power of the two-sample Wilcoxon test, β_D/β_W, as given in Table 4.2.

constant h_n, as defined in (4.55). To these data, we have applied the two-sample Gauß test and the 2-sample Wilcoxon test, as set in (4.53) and (4.54). Here, the first Hermite coefficient of G is (obtained by numerical integration) $a_1 \approx -0.678$.

Now our theoretical considerations predict for this situation

$$ARE = \lim_{n\to\infty} \frac{n_D}{n_W} = \left(\frac{a_1 \int_{\mathbb{R}} f^2(x)\, dx}{\int_{\mathbb{R}} J_1(x) f(x)\, dx} \right)^{2/D} \approx (2.67754)^{2/0.6} \approx 26.655.$$

This means that the two-sample Gauß test needs approximately 26.655 times as many observations as the two-sample Wilcoxon test to detect the same jump on the same level with the same probability. In order to reveal this behaviour, we have therefore analysed the power of the Wilcoxon test for $n = 10, 50, 100, 200$ observations and the power of the Gauß test for $n = 266, 1332, 2666, 5330$ observations.

As above, we want to check if the height of the jump(-sequence) $c\, h_n$, as provided by our theory, is appropriate to lead to a non-degenerate limit, so again we add a jump

of height $c\,h_n$. But in order to check if the calculated ARE can be seen in simulations, we need to apply both tests to the same kind of jumps. Observe that

$$h_{n_W} = c\,n_W^{-D/2} = c\,\left(\frac{n_D}{ARE}\right)^{-D/2} = c\,n_D^{-D/2} 2.67754,$$

so in order to have the same jump heights for both tests, we choose for the Wilcoxon two-sample test $c = c'/2.67754$ whenever we choose for the two-sample Gauß test $c = c'$ for any c'.

The simulation results are shown in Table 4.4. We observe at first that for each test the power stays for all n in the same order of magnitude for a fixed c. This confirms that the jump height $c\,h_n$ leads to a non-degenerate limit. However, in contrast to the simulations under Gaussian data, we observe some gradual changes which may indicate the underlying convergence to the limit. This convergence may also explain why the power of both tests is not fully equal, as expected. But one can regognize an approximation, the tendency is clear: Indeed, the Gauß test needs quite a number of observations more to detect the same jump on the same level with the same probability – as predicted by our calculation around 25 times as many.

$c \setminus n$	266	1332	2666	5330
0.5	0.144	0.144	0.137	0.141
1	0.273	0.276	0.273	0.273
2	0.664	0.655	0.658	0.657
4	0.981	0.987	0.988	0.990

$c \setminus n$	10	50	100	200
0.5/2.67754	0.189	0.131	0.136	0.130
1/2.67754	0.289	0.242	0.248	0.246
2/2.67754	0.466	0.486	0.506	0.533
4/2.67754	0.684	0.805	0.855	0.895

Table 4.4: Power of the two-sample Gauß test (top) and the two-sample Wilcoxon test (bottom), based on $10,000$ repetitions of a Pareto(3,1) transformed fGn series of length n with Hurst parameter $H = 0.7$ (D=0.6) and with a jump of height $h_n = cn^{-D/2}$ at the half.

Chapter 5

Change-point processes based on U-statistics

There are several asymptotic results for U-statistics of LRD data (Dehling and Taqqu, 1991; Hsing and Wu, 2004; Beutner and Zähle, 2011; Beutner, Wu and Zähle, 2012, e.g.). In Chapter 3 we developed a non-parametric change-point test for LRD data which is based on Wilcoxon's two-sample test statistic which can be represented as a U-statistic. We determined its asymptotic distribution under the null hpothesis that no change occured by representing the test statistic as a functional of the two-parameter empirical process of the data for which there are limit theorems (Dehling and Taqqu, 1989, 1991). Since our test is based on the Wilcoxon two-sample rank statistic $W_{k,n} = (n\, d_n)^{-1} \sum_{i=1}^{k} \sum_{j=k+1}^{n} I_{\{X_i \leq X_j\}}$ and since $W_{k,n}$ in its above representation is a U-statistic, the natural question arises if the technique can universally be extended to U-statistics

$$U_{k,n} = \sum_{i=1}^{k} \sum_{j=k+1}^{n} h(X_i, X_j)$$

with a general kernel h. Of course, such a generalization will require more technical and formal work because a general kernel does not possess the conducive form of the kernel $h(x,y) = I_{\{x \leq y\}}$, which leads to the Wilcoxon statistic and has substantial formal similarity to the empirical distribution function. The idea of the proof shall nevertheless still be to express the statistic as a functional of the empirical process for which we have an asymptotic theory: As shown in Chapter 3, by the Dudley-Wichura version of Skorohod's representation theorem (Shorack and Wellner, 1986, Th. 2.3.4) it follows from the empirical process non-central limit theorem of Dehling and Taqqu (1989) that

$$\sup_{\lambda,x} \left| d_n^{-1} [n\lambda](F_{[\lambda n]}(x) - F(x)) - J(x)Z(\lambda) \right| \longrightarrow 0 \quad a.s. \tag{5.1}$$

and

$$\sup_{\lambda,x} \left| d_n^{-1} (n - [n\lambda])(F_{[n\lambda]+1,n}(x) - F(x)) - J(x)(Z(1) - Z(\lambda)) \right| \longrightarrow 0 \quad a.s., \tag{5.2}$$

where Z denotes a certain Hermite process. In this chapter, I will present generalizations to the work by Dehling, Rooch and Taqqu (2012) and develop limit theorems for the general U-statistic $U_{k,n}$ under the null hypothesis of no change in the mean. In Chapter 6, we present a different approach to the limit behaviour of $U_{k,n}$.

5.1 Special kernels

We start with general kernels of a special form: Factorizable kernels $h(x,y) = a(x)b(y)$ and additive kernels $h(x,y) = a(x)+b(y)$. Note that both types allow a direct approach, see the remark on page 114. Nevertheless, for illustration of our method, we choose a technique based on a representation via the e.d.f..

Theorem 5.1 (NCLT for two-sample U-statistics with factorizable kernels)**.** *Suppose that $(\xi_i)_{i\geq 1}$ is a stationary Gaussian process with mean zero, variance 1 and autocovariance function as in (1.1) with $0 < D < \frac{1}{m}$. Moreover let $G : \mathbb{R} \to \mathbb{R}$ be a measurable function with $E[G(\xi_i)] = 0$ and define*

$$X_k = G(\xi_k).$$

Assume that X_k has a continuous distribution function F. Let m denote the Hermite rank of the class of functions $I_{\{G(\xi_i)\leq x\}} - F(x)$, $x \in \mathbb{R}$. Assume that $h(x,y) = a(x)b(y)$ satisfies the following conditions:

(i) $\int |a(x)|\, dF(x) < \infty$ *and* $\int |b(x)|\, dF(x) < \infty$

(ii) $a(x)$ *and* $b(y)$ *have both bounded total variation.*

(iii) $\int J(x)\, da(x)$ *and* $\int J(x)\, db(x)$ *exist and are finite.*

Then with the notation of Chapter 3,

$$\boxed{\frac{1}{n\, d_n} \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} \left(h(X_i, X_j) - \iint h(x_1, x_2)\, dF(x_1) dF(x_2) \right), \quad 0 \leq \lambda \leq 1,} \tag{5.3}$$

converges in distribution towards the process

$$\boxed{\begin{aligned} &- (1-\lambda) Z(\lambda) \int J(x)\, da(x) \int b(x)\, dF(x) \\ &\quad - \lambda (Z(1) - Z(\lambda)) \int J(x)\, db(x) \int a(x)\, dF(x), \quad 0 \leq \lambda \leq 1. \end{aligned}} \tag{5.4}$$

We set $Z(\lambda) = Z_m(\lambda)/m!$, where $Z_m(\lambda)$ denotes the m-th order Hermite process.

Remark. (a) Using the special form of the kernel, the statistic (5.3) can be written as

$$\frac{1}{n\, d_n} \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} \left(a(X_i)b(X_j) - E[a(X)]\, E[b(X)] \right), \quad 0 \leq \lambda \leq 1,$$

where $X \sim F$ has the same distribution as the single X_i's.

(b) Condition (i) is equivalent to $E|h(X,Y)| < \infty$ for independent $X, Y \sim F$.

(c) If h has bounded variation in the sense of Hardy-Krause (see Appendix B.3), then condition (ii) is fulfilled.

Proof. In order to use (5.1) and (5.2), we write sums as integrals with respect to the e.d.f. F_n and then enforce expressions of type "$F_n - F$" in our test statistic.

$$\begin{aligned}
&\sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} \left(h(X_i, X_j) - \iint h(x_1, x_2)\, dF(x_1) dF(x_2) \right) \\
&= (n - [n\lambda]) \sum_{i=1}^{[\lambda n]} \int h(X_i, x_2)\, dF_{[\lambda n]+1,n}(x_2) - [\lambda n](n - [\lambda n]) \iint h(x_1, x_2)\, dF(x_1) dF(x_2) \\
&= [\lambda n](n - [n\lambda]) \Bigg\{ \int \left(\int h(x_1, x_2)\, dF_{[\lambda n]+1,n}(x_2) \right) dF_{[\lambda n]}(x_1) \\
&\qquad - \iint h(x_1, x_2)\, dF(x_1) dF(x_2) \Bigg\} \\
&= [\lambda n](n - [n\lambda]) \Bigg\{ \int \left(\int h(x_1, x_2)\, d(F_{[\lambda n]+1,n} - F)(x_2) \right) dF_{[\lambda n]}(x_1) \\
&\qquad + \int \left(\int h(x_1, x_2)\, dF(x_2) \right) d(F_{[\lambda n]} - F)(x_1) \Bigg\} \qquad (5.5)
\end{aligned}$$

and with $h(x, y) = a(x)b(y)$

$$\begin{aligned}
&= [\lambda n](n - [n\lambda]) \Bigg\{ \int a(x)\, dF_{[\lambda n]}(x) \int b(x)\, d(F_{[\lambda n]+1,n} - F)(x_2) \\
&\qquad + \int a(x)\, d(F_{[\lambda n]} - F)(x) \int b(x)\, dF(x) \Bigg\} \\
&= [\lambda n](n - [n\lambda]) \Bigg\{ \left(\int a(x)\, d(F_{[\lambda n]} - F)(x) + \int a(x)\, dF(x) \right) \int b(x)\, d(F_{[\lambda n]+1,n} - F)(x) \\
&\qquad + \int a(x)\, d(F_{[\lambda n]} - F)(x) \int b(x)\, dF(x) \Bigg\}
\end{aligned}$$

Now we integrate by parts in order to have the "$F_n - F$" terms as integrands and the deterministic terms as integrators.

$$\begin{aligned}
\int a(x)\, d(F_{[\lambda n]} - F)(x) &= \big[a(x)(F_{[\lambda n]} - F)(x) \big]_{-\infty}^{\infty} - \int (F_{[\lambda n]} - F)(x)\, da(x) \\
\int b(x)\, d(F_{[\lambda n]+1,n} - F)(x) &= \big[b(x)(F_{[\lambda n]+1,n} - F)(x) \big]_{-\infty}^{\infty} - \int (F_{[\lambda n]+1,n} - F)(x)\, db(x)
\end{aligned}$$

We assumed by (ii) that $a(x)$ and $b(x)$ have bounded total variation. This ensures that the integrals on the right-hand side exist, moreover it follows that $a(x)$ and $b(x)$ are bounded and the boundary terms on the right-hand side vanish. So we obtain

$$\begin{aligned}
&\frac{1}{n\,d_n}\sum_{i=1}^{[\lambda n]}\sum_{j=[\lambda n]+1}^{n}\left(h(X_i,X_j)-\iint h(x_1,x_2)\;dF(x_1)dF(x_2)\right)\\
&=\frac{[\lambda n](n-[n\lambda])}{n\,d_n}\Bigg\{\left(\int(F_{[\lambda n]}-F)(x)\;da(x)-\int a(x)\;dF(x)\right)\int(F_{[\lambda n]+1,n}-F)(x)\;db(x)\\
&\quad-\int(F_{[\lambda n]}-F)(x)\;da(x)\int b(x)\;dF(x)\Bigg\}. \qquad (5.6)
\end{aligned}$$

We will now show that this converges to the process given in (5.4). We look into the single terms.

$$\begin{aligned}
&\frac{[\lambda n](n-[n\lambda])}{n\,d_n}\int(F_{[\lambda n]+1,n}-F)(x)\;db(x)-\lambda\int J(x)(Z(1)-Z(\lambda))\;db(x)\\
&\quad=\frac{[\lambda n]}{n}\int d_n^{-1}(n-[n\lambda])\left((F_{[\lambda n]+1,n}-F)(x)-J(x)(Z(1)-Z(\lambda))\right)\;db(x)\\
&\qquad+\left(\frac{[\lambda n]}{n}-\lambda\right)\int J(x)(Z(1)-Z(\lambda))\;db(x)
\end{aligned}$$

The first term on the right-hand side converges to 0 due to the bounded total vartiation of $b(x)$ and (5.2). The second term converges to 0 because we assumed in (iii) that the integral is finite, and $[\lambda n]/n\to\lambda$ uniformly. Since $\|F_{[\lambda n]}-F\|_\infty\to 0$ almost surely,

$$\int(F_{[\lambda n]}-F)(x)\;da(x)\longrightarrow 0.$$

Finally

$$\begin{aligned}
&-\frac{[\lambda n](n-[n\lambda])}{n\,d_n}\int(F_{[\lambda n]}-F)(x)\;da(x)+(1-\lambda)\int J(x)Z(\lambda)\;da(x)\\
&\quad=-\frac{n-[\lambda n]}{n}\int\left(d_n^{-1}[\lambda n](F_{[\lambda n]}-F)(x)-J(x)Z(\lambda)\right)da(x)\\
&\qquad-\left(\frac{n-[\lambda n]}{n}-(1-\lambda)\right)\int J(x)Z(\lambda)\;da(x)
\end{aligned}$$

converges to 0 by the same arguments as above: $a(x)$ is of bounded variation and because of (5.1), the first integral on the right-hand side vanishes; the second integral is finite by assumtion (iii), and $(n-[\lambda n])/n\to(1-\lambda)$ uniformly. For putting everything together in (5.6), note that $\int J(x)\,da(x)$ and $\int J(x)\,db(x)$ are finite due to condition (iii). □

For additive kernels, we obtain a quite similar result.

Theorem 5.2 (NCLT for two-sample U-statistics with additive kernels)**.** *Suppose that $(\xi_i)_{i\geq 1}$ is a stationary Gaussian process with mean zero, variance* 1 *and auto-covariance*

function as in (1.1) *with* $0 < D < \frac{1}{m}$. *Moreover let* $G : \mathbb{R} \to \mathbb{R}$ *be a measurable function with* $E[G(\xi_i)] = 0$ *and define*

$$X_k = G(\xi_k).$$

Assume that X_k *has a continuous distribution function* F. *Let* m *denote the Hermite rank of the class of functions* $I_{\{G(\xi_i)\le x\}} - F(x)$, $x \in \mathbb{R}$. *Assume that* $h(x, y) = a(x)+b(y)$ *satisfies the following conditions:*

(i) $\int |a(x)|\, dF(x) < \infty$ *and* $\int |b(x)|\, dF(x) < \infty$

(ii) $a(x)$ *and* $b(y)$ *have both bounded total variation.*

(iii) $\int J(x)\, da(x)$ *and* $\int J(x)\, db(x)$ *exist and are finite.*

Then with the notation of Chapter 3,

$$\boxed{\frac{1}{n\, d_n} \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} \left(h(X_i, X_j) - \iint h(x_1, x_2)\, dF(x_1) dF(x_2) \right), \quad 0 \le \lambda \le 1,}$$

converges in distribution towards the process

$$\boxed{-(1-\lambda) Z(\lambda) \int J(x)\, da(x) - \lambda (Z(1) - Z(\lambda)) \int J(x)\, db(x), \quad 0 \le \lambda \le 1.} \tag{5.7}$$

Again, we set $Z(\lambda) = Z_m(\lambda)/m!$, *where* $Z_m(\lambda)$ *denotes the* m*-th order Hermite process.*

Remark. (a) Due to the special form of the kernel, the statistic can also be written as

$$\frac{1}{n\, d_n} \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} \left((a(X_i) - E[a(X)]) + (b(X_j) - E[b(X)]) \right), \quad 0 \le \lambda \le 1,$$

or

$$\frac{n - [\lambda n]}{n\, d_n} \sum_{i=1}^{[\lambda n]} (a(X_i) - E[a(X)]) + \frac{[\lambda n]}{n\, d_n} \sum_{j=[\lambda n]+1}^{n} (b(X_j) - E[b(X)]), \quad 0 \le \lambda \le 1,$$

where $X \sim F$ has the same distribution as the single X_i's.

(b) The conditions under which Theorem 5.2 holds are the same as for Theorem 5.1.

Proof. The idea behind the proof stays the same: We aim to express the statistic as a functional of the empirical process such that we can use the convergence theorems

(5.1) and (5.2). We start as in the proof before, and from (5.5) we obtain, using that $h(x,y) = a(x) + b(y)$,

$$\begin{aligned}
&\sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} \left(h(X_i, X_j) - \iint h(x_1, x_2)\, dF(x_1) dF(x_2) \right) \\
&\quad = [\lambda n](n - [n\lambda]) \Bigg\{ \int \left(\int a(x_1)\, d(F_{[\lambda n]+1,n} - F)(x_2) \right) dF_{[\lambda n]}(x_1) \\
&\qquad + \int \left(\int a(x_1)\, dF(x_2) \right) d(F_{[\lambda n]} - F)(x_1) \\
&\qquad + \int \left(\int b(x_2)\, d(F_{[\lambda n]+1,n} - F)(x_2) \right) dF_{[\lambda n]}(x_1) \\
&\qquad + \int \left(\int b(x_2)\, dF(x_2) \right) d(F_{[\lambda n]} - F)(x_1) \Bigg\} \\
&\quad = [\lambda n](n - [n\lambda]) \Bigg\{ \int a(x)\, d(F_{[\lambda n]} - F)(x) + \int b(x)\, d(F_{[\lambda n]+1,n} - F)(x) \Bigg\},
\end{aligned}$$

where we have used that F and F_n are distribution functions and that for any pair of c.d.f.'s F, G it holds $\int dF = 1$ and $\int d(F - G) = 0$. Here it also contributes condition (i). Observe that by integration by parts

$$\begin{aligned}
\left| \int a(x)\, dF_{[\lambda n]}(x) \right| &= \left| \int a(x)\, d(F_{[\lambda n]} - F)(x) + \int a(x)\, dF(x) \right| \\
&\le \left| \int (F_{[\lambda n]} - F)(x)\, da(x) \right| + \left| \int a(x)\, dF(x) \right| < \infty.
\end{aligned}$$

Now we integrate by parts in order to have the "$F_n - F$" terms as integrands and the deterministic terms as integrators, exactly as in the proof before. This yields

$$\begin{aligned}
&\frac{1}{n\, d_n} \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} \left(h(X_i, X_j) - \iint h(x_1, x_2)\, dF(x_1) dF(x_2) \right) \\
&\quad = [\lambda n](n - [n\lambda]) \Bigg\{ - \int (F_{[\lambda n]} - F)(x)\, da(x) - \int (F_{[\lambda n]+1,n} - F)(x)\, db(x) \Bigg\}. \qquad (5.8)
\end{aligned}$$

As in the proof before, due to the bounded total vartiation of $a(x)$, $b(x)$ and in virtue of (5.1) and (5.2), this converges to the process given in (5.7). □

Remark. The just treated cases of special kernels, $h(x,y) = a(x)b(y)$ and $h(x,y) = a(x)+b(y)$, allow both a direct approach: one can trace them back to the limit theorems for single partial sums, as presented in Theorem (1.1). For example for factorizable kernels one has

$$\frac{1}{d_n(m_a) d_n(m_b)} \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} a(X_i) b(X_j) = \left(\frac{1}{d_n(m_a)} \sum_{i=1}^{[\lambda n]} a(X_i) \right) \left(\frac{1}{d_n(m_b)} \sum_{j=[\lambda n]+1}^{n} b(X_j) \right),$$

where m_a, m_b are the Hermite rank of the functions $a(G(\cdot))$, $b(G(\cdot))$ respectively. For such an approach, one needs the joint convergence. Compare this to the approach in Chapter 6.

5.2 General kernels

Theorem 5.3 (General NCLT for two-sample U-statistics)**.** *Suppose that $(\xi_i)_{i\geq 1}$ is a stationary Gaussian process with mean zero, variance 1 and auto-covariance function as in (1.1) with $0 < D < \frac{1}{m}$. Moreover let $G : \mathbb{R} \to \mathbb{R}$ be a measurable function with $E[G(\xi_i)] = 0$ and define*

$$X_k = G(\xi_k).$$

Assume that X_k has a continuous distribution function F. Let m denote the Hermite rank of the class of functions $I_{\{G(\xi_i)\leq x\}} - F(x)$, $x \in \mathbb{R}$. Define

$$\tilde{h}(x_1) := \int h(x_1, x_2)\, dF(x_2)$$

and assume that the following conditions hold:

(i) The kernel $h(x, y)$ is of bounded total variation in each variable (each with the other one fixed).

(ii) $h(x, y)$ satisfies the following growth conditions:

$$\sup_{x\in\mathbb{R}} \lim_{y\to\infty} \big(h(x, y)(1 - F(y)) - h(x, -y)F(-y)\big) = 0$$
$$\lim_{x\to\infty} \big(\tilde{h}(x)(1 - F(x)) - \tilde{h}(-x)F(-x)\big) = 0$$

(iii) $\tilde{h}(x_1)$ is of bounded total variation.

(iv) $\int d|h(x_1, x_2)(x_2)|_{TV} \leq c < \infty$

(v) $\int J(x_1)\, d\tilde{h}(x_1)$ and $\int \big(\int J(x_2)\, dh(x_1, x_2)(x_2)\big)\, dF(x_1)$ exist and are finite.

Then, with the notation of Chapter 3,

$$\boxed{\frac{1}{n\, d_n} \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} \left(h(X_i, X_j) - \iint h(x_1, x_2)\, dF(x_1) dF(x_2) \right), \quad 0 \leq \lambda \leq 1,}$$

converges in distribution towards the process

$$\boxed{\begin{aligned} &-(1-\lambda) Z(\lambda) \int J(x_1)\, d\tilde{h}(x_1) \\ &\quad -\lambda (Z(1) - Z(\lambda)) \int \left(\int J(x_2)\, dh(x_1, x_2)(x_2) \right) dF(x_1), \quad 0 \leq \lambda \leq 1. \end{aligned}} \tag{5.9}$$

Again, we set $Z(\lambda) = Z_m(\lambda)/m!$, where $Z_m(\lambda)$ denotes the m-th order Hermite process.

Remark. (a) Conditions (i)–(iv) are sufficient, but not neccessary. During the proof, we will have to ensure that certain integrals are finite; the conditions guarantee upper bounds. Of course in particular cases, the conditions may be violated, but nevertheless the respective integrals may be finite. This is also the case for condition (ii) of Theorem 5.1 and Theorem 5.2: It may be too restrictive.

(b) Let us spend a moment to discuss the conditions and some possible implications:

- Condition (i) means that the kernel function does not fluctuate too much. It ensures that one can one-dimensionally integrate with h as integrator (with one variable fixed). The condition is met for any bounded function which is monotone in each variable, like $h(x,y) = I_{\{x \le y\}}$ (which, by the way, is not of bounded total variation, see Section B.3.2), while it is not met for any unbounded function like $h(x,y) = xy$.

- Condition (ii) is also needed for the integration by parts: If the tails of the kernel h grow to fast, the boundary terms may not vanish. If h is bounded, (ii) is always fulfilled.

- Condition (iii) is a sufficient condition that $\int f(x) d\tilde{h}(x)$ is finite for bounded f. If h is factorizable, (i) implies (iii).

- Condition (iv) demands that $h(x_1, x_2)$ is bounded in x_1 (which is also required by condition (i)). If $h(x_1, x_2)$ is monotone increasing in x_2, (iv) reduces to $\left|\int dh(x_1, x_2)(x_2)\right| \le c < \infty$.

- If J is bounded, (iii) and (iv) imply (v). By the remark following the proof, one can express $\int J(x_1)\, d\tilde{h}(x_1)$ as a Hermite coefficient of the function $\tilde{h}(G(\cdot))$. Thus if $\tilde{h}(G(\cdot)) = \int h(G(\cdot), y)\, dF(y) \in L^2(\mathbb{R}, \mathcal{N})$, then $\int J(x_1)\, d\tilde{h}(x_1)$ exists and is finite.

Proof. As before, we express the statistic as a functional of the empirical process. For this purpose, we write sums as integrals with respect to the e.d.f. F_n and enforce expressions of type "$F_n - F$". By (5.5) it is

$$\begin{aligned}
&\sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} \left(h(X_i, X_j) - \iint h(x_1, x_2)\, dF(x_1) dF(x_2) \right) \\
&\quad = [\lambda n](n - [n\lambda]) \Bigg\{ \int \left(\int h(x_1, x_2)\, d(F_{[\lambda n]+1,n} - F)(x_2) \right) dF_{[\lambda n]}(x_1) \\
&\qquad + \int \left(\int h(x_1, x_2)\, dF(x_2) \right) d(F_{[\lambda n]} - F)(x_1) \Bigg\},
\end{aligned}$$

and now we integrate by parts in order to get the "$F_n - F$" terms as integrands and the deterministic terms as integrators.

$$\begin{aligned}
&\int h(x_1,x_2)\; d(F_{[\lambda n]+1,n} - F)(x_2) \\
&\quad = \big[h(x_1,x_2)\cdot(F_{[\lambda n]+1,n} - F)(x_2)\big]_{x_2=-\infty}^{\infty} - \int (F_{[\lambda n]+1,n} - F)(x_2)\; dh(x_1,x_2)(x_2)
\end{aligned} \tag{5.10}$$

$$\begin{aligned}
&\int \left(\int h(x_1,x_2)\; dF(x_2)\right)\; d(F_{[\lambda n]} - F)(x_1) \\
&\quad = \Big[\tilde{h}(x_1)\cdot(F_{[\lambda n]} - F)(x_1)\Big]_{x_1=-\infty}^{\infty} - \int (F_{[\lambda n]} - F)(x_1)\; d\tilde{h}(x_1)
\end{aligned} \tag{5.11}$$

The boundary terms vanish due to assumption (ii) and the remaining integrals are defined because of assumptions (i) and (iii). It is important that the boundary term in (5.10) does not only vanish for any fixed x_1, but uniformly, due to assumption (ii), because the integration by parts takes place in an integrand of an integral with respect to $F_{[\lambda n]}(x_1)$. (Another approach is possible to handle the term on the right-hand side of (5.11); see the remark following this proof.) After integration by parts we thus obtain

$$\begin{aligned}
&\frac{1}{n\,d_n}\sum_{i=1}^{[\lambda n]}\sum_{j=[\lambda n]+1}^{n}\left(h(X_i,X_j) - \iint h(x_1,x_2)\; dF(x_1)dF(x_2)\right) \\
&\quad = \frac{[\lambda n](n-[n\lambda])}{n\,d_n}\Bigg\{\int\left(-\int (F_{[\lambda n]+1,n} - F)(x_2)\; dh(x_1,x_2)(x_2)\right)\; dF_{[\lambda n]}(x_1) \\
&\qquad - \int (F_{[\lambda n]} - F)(x_1)\; d\tilde{h}(x_1)\Bigg\}.
\end{aligned} \tag{5.12}$$

We now go into both terms and show that they converge to the limit given in (5.9).

$$\begin{aligned}
&\frac{[\lambda n](n-[n\lambda])}{n\,d_n}\int (F_{[\lambda n]} - F)(x_1)\; d\tilde{h}(x_1) - (1-\lambda)\int J(x_1)Z(\lambda)\; d\tilde{h}(x_1) \\
&\quad = \frac{n-[n\lambda]}{n}\int \big(d_n^{-1}[\lambda n](F_{[\lambda n]} - F)(x_1) - J(x_1)Z(\lambda)\big)\; d\tilde{h}(x_1) \\
&\qquad + \left(\frac{n-[\lambda n]}{n} - (1-\lambda)\right)\int J(x_1)Z(\lambda)\; d\tilde{h}(x_1)
\end{aligned} \tag{5.13}$$

The first summand on the right-hand side converges to zero because of (5.1) and the bounded total variation of $\tilde{h}$ due to condition (iii), the second summand since $\sup_{0\le\lambda\le 1}|(n-[n\lambda])/n - (1-\lambda)| \to 0$ as above and since $\int J(x_1)\, d\tilde{h}(x_1) < \infty$ by assumption (v).

$$
\begin{aligned}
&\frac{[\lambda n](n-[n\lambda])}{n\,d_n}\int\left(\int (F_{[\lambda n]+1,n}-F)(x_2)\;dh(x_1,x_2)(x_2)\right)\;dF_{[\lambda n]}(x_1)\\
&\qquad-\lambda\int\left(\int J(x_2)(Z(1)-Z(\lambda))\;dh(x_1,x_2)(x_2)\right)\;dF(x_1)\\
&=\frac{[n\lambda]}{n}\int\Bigg\{\int d_n^{-1}(n-[n\lambda])(F_{[n\lambda]+1,n}-F)(x_2)\\
&\qquad-J(x_2)(Z(1)-Z(\lambda))\;dh(x_1,x_2)(x_2)\Bigg\}\;dF_{[n\lambda]}(x_1)\\
&\qquad+\frac{[n\lambda]}{n}(Z(1)-Z(\lambda))\int\left(\int J(x_2)\;dh(x_1,x_2)(x_2)\right)\;d(F_{[n\lambda]}-F)(x_1)\\
&\qquad+\left(\frac{[n\lambda]}{n}-\lambda\right)(Z(1)-Z(\lambda))\int\left(\int J(x_2)\;dh(x_1,x_2)(x_2)\right)\;dF(x_1)\qquad(5.14)
\end{aligned}
$$

All three terms on the right-hand side converge to zero. For the last term this is a consequence of $\sup_{0\le\lambda\le1}|[n\lambda]/n-\lambda|\to0$ as above and of assumption (v). The convergence of the first term follows from (5.2) and from assumption (iv): With the abbreviation $K_{n,\lambda}(x):=d_n^{-1}(n-[n\lambda])(F_{[n\lambda]+1,n}(x)-F(x))-J(x)(Z(1)-Z(\lambda))$, the first term on the right-hand side of (5.14) is

$$
\begin{aligned}
&\iint K_{n,\lambda}(x_2)\;dh(x_1,x_2)(x_2)\;dF_{[n\lambda]}(x_1)\\
&\qquad\le\sup_{\lambda,x}|K_{n,\lambda}(x)|\iint d|h(x_1,x_2)(x_2)|_{TV}\;dF_{[n\lambda]}(x_1)\\
&\qquad\le\sup_{\lambda,x}|K_{n,\lambda}(x)|\,c\|F_{[\lambda n]}\|_{TV},
\end{aligned}
$$

due to assumption (iv). This goes to 0 because $\sup_{\lambda,x}|K_{n,\lambda}(x)|\to0\quad a.s.$ by (5.2). The second term in (5.14) can be written as

$$
\begin{aligned}
&\frac{[n\lambda]}{n}\int\left(\int J(x_2)\;dh(x_1,x_2)(x_2)\right)\;d(F_{[n\lambda]}-F)(x_1)\\
&\qquad=\frac{1}{n}\sum_{i=1}^{[\lambda n]}\int J(x_2)\;dh(X_i,x_2)(x_2)-E\left[\int J(x_2)\;dh(X_1,x_2)(x_2)\right].
\end{aligned}
$$

The ergodic theorem states that $k^{-1}\sum_{i=1}^{k}(f(X_i)-Ef(X_i))\to0$ a.s., and thus can be applied here with $f(x)=\int J(t)\;dh(x,t)(t)$; the neccessary requirement $E|f(X_i)|=\int|f(x)|dF(x)<\infty$ is guaranteed by assumption (v).

So (5.13) and (5.14) converge to zero, and bearing this in mind, (5.12) proves the statement. □

Remark. The preceding proof started with writing

$$\sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} \left(h(X_i, X_j) - \iint h(x_1, x_2)\, dF(x_1) dF(x_2) \right)$$
$$= [\lambda n](n - [n\lambda]) \Bigg\{ \int \left(\int h(x_1, x_2)\, d(F_{[\lambda n]+1,n} - F)(x_2) \right) dF_{[\lambda n]}(x_1)$$
$$+ \int \left(\int h(x_1, x_2)\, dF(x_2) \right) d(F_{[\lambda n]} - F)(x_1) \Bigg\}. \tag{5.15}$$

Using integration by parts, we showed that the second term on the right-hand side of (5.15) converges to

$$-(1-\lambda) \int J_1(x_1) Z(\lambda) d\tilde{h}(x_1),$$

where $Z(\lambda) = Z_m(\lambda)/m!$ and m is the Hermite rank of the class of functions $I_{\{G(\xi)\le x\}} - F(x)$. Note that another approach is possible since the second term on the right-hand side of (5.15) can be written as

$$\frac{[\lambda n](n-[n\lambda])}{n\, d_n} \frac{1}{[\lambda n]} \sum_{i=1}^{[\lambda n]} \left(\tilde{h}(X_i) - E\tilde{h}(X_i) \right) \xrightarrow{\mathcal{D}} (1-\lambda) \frac{a_p}{p!} Z_p(\lambda)$$

by direct application of Theorem 1.1, where p denotes the Hermite rank of $\tilde{h}(G(\cdot))$ and a_p is the associated Hermite coefficient.

This means that

$$-(1-\lambda) \int J_1(x_1) Z(\lambda) d\tilde{h}(x_1) \stackrel{\mathcal{D}}{=} (1-\lambda) \frac{a_p}{p!} Z_p(\lambda) \tag{5.16}$$

and as a consequence, it follows that first both Hermite ranks are the same, $m = p$. In other words, m is the smallest non-zero integer which satisfies

$$E\left[\tilde{h}(G(\xi)) H_m(\xi) \right] \neq 0$$
$$E\left[(I_{\{G(\xi)\le x\}} - F(x)) H_m(\xi) \right] \neq 0 \quad \text{for some } x \in \mathbb{R}$$

(and it satisfies both conditions if it satisfies one of them). Second, relation (5.16) may help to evaluate the constants on either side. As an example, we consider the case $G(t) = t$ (which entails $m = 1$ and $H_m(x) = x$). On the left hand side, we obtain

$$\begin{aligned}
-\int J_1(x)\, d\tilde{h}(x) &= -\iint I_{\{s\le x\}} s\, d\Phi(s)\, d\tilde{h}(x) \\
&= \int \varphi(x)\, d\tilde{h}(x) && \text{since } \int I_{\{s\le x\}} s\, d\Phi(s) = -\varphi(x) \\
&= -\int \tilde{h}(x)\, d\varphi(x) && \text{by integration by parts} \\
&= -\iint h(x,y)\, d\Phi(y)\, d\varphi(x) && \text{by definition of } \tilde{h} \\
&= \iint h(x,y) x\, d\Phi(x)\, d\Phi(y) && \text{with } \frac{d\varphi(x)}{dx} = -x\varphi(x).
\end{aligned}$$

For the expression on the right-hand side of (5.16), we obtain in our example

$$\begin{aligned} a_1 &= E\left[\tilde{h}(\xi)H_m(\xi)\right] \\ &= \int \tilde{h}(x)\, x\, d\Phi(x) \\ &= \iint h(x,y)\, x\, d\Phi(x)\, d\Phi(y), \end{aligned}$$

which is obviously the same.

5.3 Examples

The kernel $h(x,y) = I_{\{x\le y\}}$

In Chapter 3, we have shown that the technique of Theorem 5.3 works for the kernel $h(x,y) = I_{\{x\le y\}}$ which yields the *Mann-Whitney-Wilcoxon statistic*

$$W_{[\lambda n],n} = \frac{1}{n\, d_n} \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} \left(I_{\{X_i \le X_j\}} - \frac{1}{2} \right),$$

but of course we can get the same result as well from the general approach.

General case

We consider data of the general type $X_i = G(\xi_i)$ and check the consitions of Theorem 5.3:

(i) Observe that $h(x,y) = I_{\{x\le y\}} = H(y-x)$, where H denotes the Heaviside step function. So for a fixed x or a fixed y, the kernel $h(x,y) = H(y-x)$ is of bounded variation (it has each only one jump of height 1).

(ii) h is bounded, so the growth conditions are fulfilled.

(iii) We have

$$\tilde{h}(x_1) = \int h(x_1,x_2)\, dF(x_2) = \int_{x_1}^{\infty} dF(x_2) = 1 - F(x_1),$$

and this is of bounded total variation since F is monotone and bounded.

(iv) The kernel h can be represented as a Heaviside step function H, see above, and H is the integral of the Dirac delta function. Thus we have

$$dh(x_1,x_2)(x_2) = d(I_{\{x_1\le x_2\}})(x_2) = \delta_{x_1}(x_2)dx_2,$$

where $\delta_a(x)$ is the Dirac delta function with mass in point a. We obtain

$$\left| \int_{\mathbb{R}} dh(x_1,x_2)(x_2) \right| = \left| \int_{\mathbb{R}} \delta_{x_1}(x_2)dx_2 \right| = 1 < \infty.$$

(v) By the above calculations it holds

$$\int J(x_1)\, d\tilde{h}(x_1) = -\int J(x_1)\, dF(x_1)$$

and

$$\iint J(x_2)\, dh(x_1, x_2)(x_2)\, dF(x_1) = \int J(x_1)\, dF(x_1).$$

$|\int J(x)\, dF(x)|$ is finite, because with $J(x) = J_m(x) = E\left[H_m(\xi) I_{\{G(\xi)\le x\}}\right]$ it equals

$$\left|\iint_{\mathbb{R}^2} H_m(s) I_{\{G(s)\le x\}}\, d\Phi(s)\, dF(x)\right| \le \iint_{\mathbb{R}^2} |H_m(s)|\, d\Phi(s)\, dF(x),$$

and H_m is a polynomial and thus integrable with respect to the standard normal density, moreover F is a c.d.f..

So Theorem 5.3 reproduces the result of Chapter 3: $W_{[\lambda n],n}$ converges in distribution to the process

$$(Z(\lambda) - \lambda Z(1)) \int J(x)\, dF(x).$$

In fact, we have just shown a little (but essential) bit more, namely that the integrals in the limit are always finite such that the limit process is not ∞.

Gaussian observations

Let us also consider the Gaussian case, i.e. the special case $G(t) = t$ such that the process $(X_i)_{i\ge1} = (\xi_i)_{1\ge1}$ is standard normally distributed with auto-covariance function $\gamma(k) = k^{-D}L(k)$. F is then the c.d.f. of a Gaussian variable, denoted by Φ. We consider the expansion

$$I_{\{X\le x\}} - F(x) = \sum_{q=m}^{\infty} \frac{J_q(x)}{q!} H_q(X)$$

with $J_q(x) = E\left[I_{\{X\le x\}} H_q(X)\right]$. The Hermite rank is $m = 1$, because

$$J_m(x) = E\left[H_m(\xi) I_{\{G(\xi)\le x\}}\right] = E\left[\xi I_{\{\xi\le x\}}\right] = -\varphi(x),$$

is the negative of the p.d.f. of a Gaussian variable and non-zero. Above we have verified almost all conditions of Theorem 5.3, we only need to check that the limit is finite, but this is the case because $\int_{\mathbb{R}} \varphi(x)\, d\Phi(x) = (2\sqrt{\pi})^{-1} < \infty$, so Theorem 5.3 can be applied. It states that $W_{[\lambda n],n}$ converges in distribution to

$$\left(-\frac{1}{2\sqrt{\pi}}\right)(Z(\lambda) - \lambda Z(1)),$$

where $Z(\lambda) = B_{1-D/2}(\lambda)$ denotes the standard fBm with parameter $H = 1 - D/2$.

If we fix λ, we obtain a result on the two-sample Wilcoxon test under LRD data:

$$U_{W,[\lambda n],n} = \frac{n^{-2+D/2}L(n)^{-1/2}}{\sigma_W} \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} \left(I_{\{X_i \le X_j\}} - \frac{1}{2} \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

with

$$\sigma_W^2 := \frac{c_1}{4\pi} \operatorname{Var}[Z(\lambda) - \lambda Z(1)] = \frac{2}{4\pi(1-D)(2-D)} \left(\lambda^2 - \lambda + (1-\lambda)\lambda^{2-D} + \lambda(1-\lambda)^{2-D} \right).$$

The kernel $h(x,y) = x - y$

This kernel leads to the *difference of means statistic*

$$\begin{aligned} \frac{1}{n\, d_n} U_{\text{diff},[\lambda n],n} = D_{[\lambda n],n} &= \frac{1}{n\, d_n} \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} (X_i - X_j) \\ &= \frac{[\lambda n](n - [\lambda n])}{n\, d_n} \left(\bar{X}_{[\lambda n]} - \bar{X}_{[\lambda n]+1,n} \right), \end{aligned}$$

where $\bar{X}_{[\lambda n]}$ denotes the arithmetic mean of the first $[\lambda n]$ observations and $\bar{X}_{[\lambda n]+1,n}$ denotes the arithmetic mean of the last $n - [\lambda n]$ observations. Here, the conditions of Theorem 5.3 are not met; nevertheless, if we ignore this incommodious fact for reasons of inquisitiveness, carfreely applying the theorem yields the correct limit behaviour. In order to handle the conditions that involve F, we consider the Gaussian case: $G(t) = t$ such that $m = 1$, $J(x) = -\varphi(x)$, $F(x) = \Phi(x)$.

(i) The kernel $h(x,y) = x - y$ is not bounded in both variables, thus it does not have bounded total variation in its single variables. Nevertheless, it has locally bounded total variation since it is continuous.

(ii) We have

$$\tilde{h}(x_1) = \int_{\mathbb{R}} h(x_1, x_2)\, dF(x_2) = \int_{\mathbb{R}} (x_1 - x_2)\, d\Phi(x_2) = x_1.$$

$\Phi(x)$ converges to its limits, as $x \to \pm\infty$, fast enough, so that the growth conditions are fulfilled.

(iii) As the kernel itself, $\tilde{h}(x_1) = x_1$ is only of locally bounded variation, not of bounded total variation at all, since it is unbounded.

(iv) We have

$$dh(x_1, x_2)(x_2) = d(x_1 - x_2)(x_2) = -dx_2,$$

and of course, $\int_{\mathbb{R}} dh(x_1, x_2)(x_2)$ is not bounded. Nevertheless, the integrator has locally bounded variation again.

(v) Both integrals have the same absolute value, namely

$$\left| \int_{\mathbb{R}} J(t)\, dh(x,t)(t) \right| = \int_{\mathbb{R}} \varphi(t)\, dt = 1 < \infty,$$

thus the integrals in the limit exist and are finite.

So, if one ignored that some of the conditions are violated, Theorem 5.3 would state that $U_{\text{diff},[\lambda n],n}$ converges in distribution to the process

$$\begin{aligned}
&(1-\lambda)Z(\lambda) \int \varphi(x_1)\, dx_1 - \lambda(Z(1) - Z(\lambda)) \int \left(\int \varphi(x_2)\, dx_2 \right) d\Phi(x_1) \\
&\quad = (1-\lambda)Z(\lambda) - \lambda(Z(1) - Z(\lambda)) \\
&\quad = Z(\lambda) - \lambda Z(1),
\end{aligned}$$

where $Z(\lambda) = B_{1-D/2}(\lambda)$ denotes the standard fBm with parameter $H = 1 - D/2$. This "result" is confirmed by the calculations from Section 3.4.2 and by Horváth and Kokoszka (1997).

If we fix the parameter λ and change some notation, we reproduce the results from the introductory consideration of $\bar{X} - \bar{Y}$. Set $Y_j := X_{[\lambda n]+j}$, and Theorem 5.3 states

$$\begin{aligned}
\frac{1}{n\, d_n} U_{\text{diff},[\lambda n],n} = D_{[\lambda n],n} &= \frac{[\lambda n](n - [\lambda n])}{n\, d_n} \left(\bar{X}_{[\lambda n]} - \bar{Y}_{n-[\lambda n]} \right) \\
&\xrightarrow{\mathcal{D}} Z(\lambda) - \lambda Z(1).
\end{aligned}$$

It follows that

$$\frac{n^{D/2} L(n)^{-1/2} \left(\bar{X}_{[\lambda n]} - \bar{Y}_{n-[\lambda n]} \right)}{\sqrt{c_1 \operatorname{Var}[Z(\lambda) - \lambda Z(1)]}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

and since the auto-covariace of fBm is known so that $\operatorname{Var}[Z(\lambda) - \lambda Z(1)]$ can be calculated, after some algebra we find that

$$c_1 \operatorname{Var}[Z(\lambda) - \lambda Z(1)] = \frac{2}{(1-D)(2-D)} \frac{\lambda^{1-D} + (1-\lambda)^{1-D} - 1}{\lambda(1-\lambda)},$$

which is nothing else than the statement of Theorem 2.3.

5.4 Simulations

I have simulated time series $\xi_1, \ldots, \xi_n$ of fGn for different Hurst parameters H (respectively $D = 2 - 2H$). In this model, the auto-covariances are

$$\gamma_k \sim H(2H-1)k^{2H-2} = \left(1 - \frac{D}{2}\right)(1-D)k^{-D}.$$

For each a set of n observations (for varying sample sizes n) and different cutting points λ I have calculated the two-sample Wilcoxon test statistic

$$\frac{1}{\sigma_W\, n^{2-\frac{D}{2}}} U_{W,[\lambda n],n} = \frac{1}{\sigma_W} \left(n^{-2+\frac{D}{2}} \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda N]+1}^{n} \left(I_{\{\xi_i \le \xi_j\}} - \frac{1}{2} \right) \right)$$

with

$$\sigma_W^2 := \frac{1}{4\pi}\left(\lambda^2 - \lambda + (1-\lambda)\lambda^{2-D} + \lambda(1-\lambda)^{2-D}\right).$$

I have repeated this $10,000$ times for each choice of parameters n, λ, H. By the above theory, $(\sigma_W\, n^{2-\frac{D}{2}})^{-1} U_{W,[\lambda n],n}$ converges to a standard normal distribution. In Figure 5.1, the density of the $10,000$ simulated values for $U_{W,[\lambda n],n}$ is shown; their sample variances are given in Appendix D.6. Figure 5.1 indeed conveys some impression that U_W is asymptotically normally distributed.

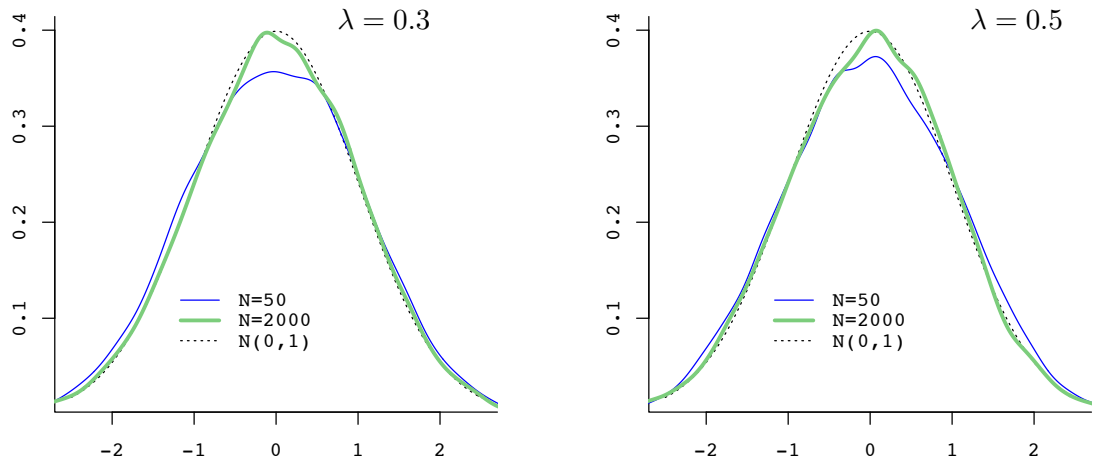


Figure 5.1: Density of the scaled and normalized Wilcoxon statistic $U_{W,[\lambda n],n}$. fGn with $k = 10,000$ and $H = 0.7$ ($D = 0.6$).

Chapter 6

Change-point processes based on U-statistics, a direct approach

In Chapter 5 we analysed the limit behaviour of the general change-point statistic

$$\sum_{i=1}^{k} \sum_{j=k+1}^{n} h(X_i, X_j)$$

with kernel h. In this chapter we make a different approach to this two-sample U-statistic. As a start, we consider a stationary Gaussian process $(\xi_i)_{i\geq 1}$ with mean zero, variance 1 and auto-covariance function (1.1) and a function $h \in L^2(\mathbb{R}, \mathcal{N})$. As described in Section 1.4.2, in order to analyse the limit behaviour of a partial sums of observations $h(\xi_1), \ldots, h(\xi_n)$, it is useful to expand h in Hermite polynomials. Because under some assumptions, only the first term in this expansion contributes to the limit, and one obtains the elegant Theorem 1.1,

$$\frac{1}{d_n} \sum_{i=1}^{[\lambda n]} h(\xi_i) \xrightarrow{\mathcal{D}} \frac{a_m}{m!} Z_m(\lambda),$$

where $d_n = c_m n^{2-Dm} L^m(n)$, $c_m = 2m!((1-Dm)(2-Dm))^{-1}$, m denotes the Hermite rank of h, a_m the associated Hermite coefficient and where Z_m is the m-th order Hermite process. In order to analyse the limit behaviour of two-sample U-statistics

$$U_{\lambda,n} = \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} h(\xi_i, \xi_j), \tag{6.1}$$

we will follow a similar route: We will expand h into Hermite polynomials and deduce the limit behaviour of $U_{\lambda,n}$ from this expansion. Due to the dependencies of the $(\xi_i)_{i\geq 1}$, some problems will arise.

In the one-sample case, each summand of a partial sum $\sum_{i=1}^{[\lambda n]} h(\xi_i)$ contains only one single observation $\xi_i \sim \mathcal{N}(0,1)$. Now in order to handle the partial sum, we use

that Hermite polynomials form an orthogonal basis of $L^2(\mathbb{R}, \mathcal{N})$, the space of square-integrable functions on $\mathbb{R}$ with respect to the standard normal measure[1] $\varphi(x)dx$: We expand h into Hermite polynomials, and this expansion is the same for each summand in the partial sum $\sum_{i=1}^{[\lambda n]} h(\xi_i)$. In the two-sample case, the situation is different: Each summand in (6.1) contains now a pair of observations $(\xi_i, \xi_j) \sim \mathcal{N}(0, \gamma_{|i-j|})$ and thus lives each in another space $L^2(\mathbb{R}^2, \mathcal{N}_{\gamma_k})$, the space of square-integrable functions on $\mathbb{R}^2$ with respect to the normal measure $\varphi_k(x, y)dx\,dy$ where

$$\varphi_k(x, y) = \frac{1}{2\pi|\Sigma_{\gamma_k}|^{1/2}} \exp\left(-\frac{1}{2}(x, y)^T \Sigma_{\gamma_k}^{-1}(x, y)\right)$$

is the p.d.f. of a (two-dimensional) $\mathcal{N}_2(0, \Sigma_{\gamma_k})$-distributed random vector with $k = |i-j|$ and

$$\Sigma_{\gamma_k} = \begin{pmatrix} 1 & \gamma_k \\ \gamma_k & 1 \end{pmatrix}.$$

Thus we would have to expand each summand of $U_{\lambda,n}$ into a different set of polynomials, which is far away from handy, not to mention the question which would be the "Hermite polynomials" in the space $L^2(\mathbb{R}^2, \mathcal{N}_{\gamma_k})$, i.e. appropriate orthogonal polynomials under a normal measure with non-trivial covariance matrix – and if the limit theory applies for them.

Thus we will expand $U_{\lambda,n}$ in usual Hermite polynomials as if (ξ_i, ξ_j) were independent, such that these problems do not arise. Instead, we have to find conditions that guarantee that this formal expansion converges and really represents $U_{\lambda,n}$.

This approach, which we restricted for a start to Gaussian data $\xi_1, \ldots, \xi_n$, can of course be extended to general data $G(\xi_1), \ldots, G(\xi_n)$ by considering the kernel $h(G(x), G(y))$ instead of $h(x, y)$.

6.1 The limit distribution under the null hypothesis

Let $(\xi_i)_{i\geq 1}$ be a stationary Gaussian processes with mean zero, variance 1 and auto-covariance function (1.1). We assume that we observe n data and that this series is cut into two pieces at point $[\lambda n]$, $\lambda \in (0, 1)$, so that we have artificially two samples:

$$\xi_1, \xi_2, \ldots, \xi_{[\lambda n]} \quad \text{and} \quad \xi_{[\lambda n]+1}, \xi_{[\lambda n]+2}, \ldots, \xi_n$$

We now want to study the asymptotic behaviour of the two-sample U-statistic

$$U_{\lambda,n} = \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} h(\xi_i, \xi_j),$$

[1] Sometimes we denote this measure simply by $\mathcal{N}$, i.e. like the normal distribution.

which is based on these observations, with a measurable kernel function $h : \mathbb{R}^2 \to \mathbb{R}$, $h \in L^2(\mathbb{R}^2, \mathcal{N})$. We can assume without loss of generality that $E[h(\xi, \eta)] = 0$ (otherwise just substract the mean). Then we have the Hermite expansion

$$h(x,y) = \sum_{k,l=0}^{\infty} \frac{a_{kl}}{k!\,l!} H_{kl}(x,y) = \sum_{k,l=0}^{\infty} \frac{a_{kl}}{k!\,l!} H_k(x) H_l(y). \tag{6.2}$$

This is an L^2 expansion with respect to the bivariate standard normal density. In this expansion, the coefficients are

$$a_{kl} = E\left[h(\xi,\eta) H_k(\xi) H_l(\eta)\right] = \iint_{\mathbb{R}^2} h(x,y) H_k(x) H_l(y)\, \varphi(x)\varphi(y)\, dx\, dy.$$

Note that because of the assumption that h is centralised, $a_{0,0} = 0$ always. We would like to order the terms in this expansion (6.2) according to the size of $k + l$:

$$h(x,y) = \sum_{q=m}^{\infty} \sum_{\substack{k,l: \\ k+l=q}} \frac{a_{kl}}{k!\,l!} H_k(x) H_l(y), \tag{6.3}$$

where m is the smallest integer for which there exists a non-zero Hermite coefficient $a_{kl} \neq 0$ with $k + l = m$. m is called the *Hermite rank* of $h(x, y)$.

Definition 6.1 (Hermite rank for two-dimensional functions)**.** The *Hermite rank* of a function $h(x, y)$ is defined as

$$m = \inf\{k + l \,|\, k, l \geq 0, a_{kl} \neq 0\},$$

where a_{kl} is the coefficient in the Hermite expansion (6.3).

In our setting, $a_{0,0} = 0$ always, as stated above, so the kernel h has always Hermite rank $m \geq 1$. In order to show the convergence of $U_{\lambda,n}$ with h represented by its Hermite expansion (6.3), we need a multi-dimensional version of Theorem 1.1. This result can be obtained using the same techniques that yield the one-dimensional NCLT. Details for the two-dimensional case will appear in a forthcoming paper by Taqqu (2012).

Theorem 6.1 (Multi-dimensional NCLT for LRD processes)**.** *For any* $1 \leq m \leq 1/D$

$$\boxed{\left(\frac{1}{d_n(1)} \sum_{i=1}^{[\lambda_1 n]} H_1(\xi_i), \frac{1}{d_n(2)} \sum_{i=1}^{[\lambda_2 n]} H_2(\xi_i), \ldots, \frac{1}{d_n(m)} \sum_{i=1}^{[\lambda_m n]} H_m(\xi_i) \right)}$$

converges in distribution to the m*-dimensional process*

$$\boxed{\left(\frac{Z_1(\lambda_1)}{1!}, \frac{Z_2(\lambda_2)}{2!}, \ldots, \frac{Z_m(\lambda_m)}{m!} \right)}$$

in $D[0,1]^m$*, where*

$$d_n^2(k) = c_k n^{2-Dk} L^k(n)$$

is the usual scaling for the partial sum of the k*-th Hermite polynomial and* Z_k *denotes the* k*-th order Hermite process, as defined in* (1.11)*.*

Definition 6.2 (Centralized/normalized $L^p(\mathbb{R}^2)$-functions)**.** Let $\xi, \eta \sim \mathcal{N}(0,1)$ be two independent standard normal random variables. We define

$$\mathcal{G}^1(\mathbb{R}^2,\mathcal{N}) := \{G : \mathbb{R}^2 \to \mathbb{R} \text{ integrable} \mid E[G(\xi,\eta)] = 0\} \subset L^1(\mathbb{R}^2,\mathcal{N}),$$

the class of (with respect to the standard normal measure) centralized and integrable functions on $\mathbb{R}^2$, and

$$\mathcal{G}^2(\mathbb{R}^2,\mathcal{N}) := \{G : \mathbb{R}^2 \to \mathbb{R} \text{ measurable} \mid E[G(\xi,\eta)] = 0, E[G^2(\xi,\eta)] = 1\} \subset L^2(\mathbb{R}^2,\mathcal{N}),$$

the class of (with respect to the standard normal measure) normalized and square-integrable functions on $\mathbb{R}^2$.

Any function $G : \mathbb{R}^2 \to \mathbb{R}$ which is measurable with mean zero and finite variance under standard normal measure can be normalized by dividing the standard deviation, so it can be considered as a function in $\mathcal{G}^2 = \mathcal{G}^2(\mathbb{R}^2,\mathcal{N})$.

Theorem 6.2. *Let $(\xi_i)_{i\geq 1}$ be a stationary Gaussian process with mean 0, variance 1 and covariances* (1.1)*. Let $Dm < 1$ and let $h \in \mathcal{G}^2(\mathbb{R}^2,\mathcal{N})$ be a function with Hermite rank m whose Hermite coefficients satisfy*

$$\sum_{k,l} \frac{|a_{kl}|}{\sqrt{k!\,l!}} < \infty. \tag{6.4}$$

Then as $n \to \infty$

$$\frac{1}{d'_n\, n}\left|\sum_{i=1}^{[\lambda n]}\sum_{j=[\lambda n]+1}^{n} h(\xi_i,\xi_j) - \sum_{\substack{k,l:\\ k+l=m}} \frac{a_{kl}}{k!\,l!}\sum_{i=1}^{[\lambda n]}\sum_{j=[\lambda n]+1}^{n} H_k(\xi_i)H_l(\xi_j)\right| \xrightarrow{L^1} 0$$

uniformly in $\lambda \in [0,1]$ and

$$\boxed{\frac{1}{d'_n\, n}\sum_{i=1}^{[\lambda n]}\sum_{j=[\lambda n]+1}^{n} h(\xi_i,\xi_j) \xrightarrow{\mathcal{D}} \sum_{\substack{k,l:\\ k+l=m}} \frac{a_{kl}}{(k!\,l!)^2}(c_k c_l)^{1/2} Z_k(\lambda)(Z_l(1) - Z_l(\lambda))} \tag{6.5}$$

in $D[0,1]$, $\lambda \in [0,1]$, where

$$d'^2_n = n^{2-mD} L(n)^m \tag{6.6}$$

and the $Z_k(\lambda)$, $k = 0,\dots,m$, are dependent processes which can be expressed as k-fold Wiener-Itō-Integrals, see (1.11)*.*

Remark. (a) The scaling factor (6.6) differs slightly from the usual scaling (1.12): It does not include the normalizing constant c_m. This is caused by the fact that the limit now is a linear combination of two possibly different Hermite processes Z_k, Z_l and thus the associated factors c_k, c_l cannot be divided out and must remain inside the sum of the right-hand side of (6.5).

(b) For the most interesting and simple case, we can give a handy explicit representation of the limit (6.5) (this is what makes the case interesting), because then $Z_1(\lambda)$ is fBm $B_H(\lambda)$ with $H = 1 - D/2$. We will do this hereafter.

Proof. The two-dimensional Hermite polynomials $H_{kl}(x,y) = H_k(x)H_l(y)$ form an orthogonal basis of the space $L^2(\mathbb{R}^2,\mathcal{N})$ (and $H_k(x)H_l(y)/\sqrt{k!l!}$ form an orthonormal basis), so the expansion (6.3) of $h(x,y)$ in Hermite polynomials converges to h in $L^2(\mathbb{R}^2,\mathcal{N})$. But in order to handle $h(\xi_i,\xi_j)$, this expansion is not suitable for now, since any pair (ξ_i,ξ_j) is dependent and underlies a joint normal distribution with non-standard covariance matrix. So first we ensure that the expansion (6.3) is nevertheless applicable in our situation. We show at first that under condition (6.4), (6.3) converges almost surely to $h(x,y)$.

Observe that

$$\begin{aligned} E\left[\sum_{k,l}\Big|\frac{a_{kl}}{k!\,l!}H_k(\xi_i)H_l(\xi_j)\Big|\right] &\le \sum_{k,l}\frac{|a_{kl}|}{k!\,l!}E\,|H_k(\xi_i)H_l(\xi_j)| \\ &\le \sum_{k,l}\frac{|a_{kl}|}{k!\,l!}\sqrt{E\left[H_k^2(\xi_i)\right]E\left[H_l^2(\xi_j)\right]} \\ &= \sum_{k,l}\frac{|a_{kl}|}{\sqrt{k!\,l!}}, \end{aligned}$$

and this is finite by our assumption. Thus

$$\sum_{k,l}\frac{a_{kl}}{k!\,l!}H_k(\xi_i)H_l(\xi_j)$$

is almost surely absolutely convergent. So the Hermite expansion converges almost surely, and since it converges to h in $L^2(\mathbb{R}^2,\mathcal{N})$, the same holds almost surely (since the measures are equivalent).

Thus we have

$$h(\xi_i,\xi_j) = \sum_{\substack{k,l:\\ k+l\ge m}}\frac{a_{kl}}{k!\,l!}H_k(\xi_i)H_l(\xi_j)$$

and hence

$$\sum_{i=1}^{[\lambda n]}\sum_{j=[\lambda n]+1}^{n} h(\xi_i,\xi_j) - \sum_{\substack{k,l:\\ k+l\ge m}}\sum_{i=1}^{[\lambda n]}\sum_{j=[\lambda n]+1}^{n}\frac{a_{kl}}{k!\,l!}H_k(\xi_i)H_l(\xi_j) = 0$$

and so

$$\begin{aligned} &\sup_{0\le\lambda\le 1}\frac{1}{d_n' n}\left(\sum_{i=1}^{[\lambda n]}\sum_{j=[\lambda n]+1}^{n} h(\xi_i,\xi_j) - \sum_{\substack{k,l:\\ k+l=m}}\sum_{i=1}^{[\lambda n]}\sum_{j=[\lambda n]+1}^{n}\frac{a_{kl}}{k!\,l!}H_k(\xi_i)H_l(\xi_j)\right) \\ &\quad = \sup_{0\le\lambda\le 1}\frac{1}{d_n' n}\sum_{\substack{k,l:\\ k+l\ge m+1}}\sum_{i=1}^{[\lambda n]}\sum_{j=[\lambda n]+1}^{n}\frac{a_{kl}}{k!\,l!}H_k(\xi_i)H_l(\xi_j). \end{aligned}$$

We will show that the lower term converges in L^1 to 0, and so will the upper one. Note at first that the supremum here is in fact a maximum, since λ appears only in terms

of the integer $[\lambda n]$, thus by setting $b := [\lambda n]$ and by the fact that we need to have $[\lambda n] \geq 1$ in order to have a two-sample statistic, we can replace $\sup_{0\leq\lambda\leq 1}$ by $\max_{1\leq b\leq n}$. Using $\max_b |f(b)g(b)| \leq \max_b |f(b)| \max_b |g(b)|$ and the Cauchy–Bunyakovsky–Schwarz inequality, we obtain

$$
\begin{aligned}
&E\left|\sup_{0\leq\lambda\leq 1} \sum_{\substack{k,l:\\ k+l\geq m+1}} \frac{a_{kl}}{k!\,l!} \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} H_k(\xi_i) H_l(\xi_j)\right| \\
&\leq E\left[\frac{1}{d'_n\, n} \sum_{\substack{k,l:\\ k+l\geq m+1}} \frac{|a_{kl}|}{k!\,l!} \max_{1\leq b\leq n}\left|\sum_{i=1}^{b} H_k(\xi_i)\right| \max_{1\leq b\leq n}\left|\sum_{j=b+1}^{n} H_l(\xi_j)\right|\right] \\
&\leq \frac{1}{d'_n\, n} \sum_{\substack{k,l:\\ k+l\geq m+1}} \frac{|a_{kl}|}{k!\,l!} \left(E\left[\max_{1\leq b\leq n}\left|\sum_{i=1}^{b} H_k(\xi_i)\right|\right]^2 E\left[\max_{1\leq b\leq n}\left|\sum_{j=b+1}^{n} H_l(\xi_j)\right|\right]^2\right)^{1/2}. \qquad (6.7)
\end{aligned}
$$

In order to show that this term converges to 0, we need bounds for the expectations of the squared maxima.

The growth of the partial sum $\sum_{i=1}^{b} H_k(\xi_i)$ is determined by the degree k of the Hermite polynomial and the size of the LRD parameter $D \in (0,1)$: For $Dk > 1$ we observe usual SRD behaviour, while for $Dk < 1$ we observe a faster rate of growth, remember (1.8). First we consider the SRD case, that is $Dk > 1$. Here we have by (1.8)

$$E\left[\sum_{i=1}^{b} H_k(\xi_i)\right]^2 \leq Ck! \cdot b,$$

and thus by stationarity and an inequality by Móricz (1976, Theorem 3)

$$E\left[\max_{1\leq b\leq n}\left|\sum_{i=1}^{b} H_k(\xi_i)\right|\right]^2 \leq 4Ck! \cdot n(\log_2 n)^2. \qquad (6.8)$$

Here we used the estimate $\log_2(2n) \leq 2\log_2 n$ for $n \geq 2$. Now we turn to the LRD case, that is $Dk < 1$. Here we have by (1.8) and the simple estimate $b^{2-Dk} \leq bn^{1-Dk}$ for all $b \leq n$ (and we do not consider any other b)

$$E\left[\sum_{i=1}^{b} H_k(\xi_i)\right]^2 \leq \tilde{C}(k)k! \cdot n^{1-Dk} \max_{1\leq b\leq n} L^k(b) \cdot b,$$

and thus by the same inequality of Móricz (1976, Theorem 3)

$$E\left[\max_{1\leq b\leq n}\left|\sum_{i=1}^{b} H_k(\xi_i)\right|\right]^2 \leq 4\tilde{C}(k)k! \cdot n^{2-Dk} \max_{1\leq b\leq n} L^k(b) \cdot (\log_2 n)^2. \qquad (6.9)$$

Note that the same estimates hold in (6.7) for the sum that starts at $b+1$ because for some $b' \in \{1, \dots, n\}$

$$\max_{1\le b\le n}\left|\sum_{j=b+1}^{n} H_l(\xi_j)\right| \overset{\mathcal{D}}{=} \max_{1\le b\le n}\left|\sum_{j=1}^{n-b} H_l(\xi_j)\right| = \max_{1\le b'\le n}\left|\sum_{j=1}^{b'} H_l(\xi_j)\right|$$

where $\overset{\mathcal{D}}{=}$ denotes equality in distribution since the $(\xi_i)_{i\ge 1}$ are stationary.

Now depending on the size of k and l, both sums in (6.7) can be SRD or LRD – and thus they can be bounded by (6.8) or by (6.9) –, such that we have to descriminate four cases.

- When $k, l < 1/D$, such that both sums are LRD, (6.7) is bounded by

$$\begin{aligned}
&\frac{1}{n^{2-Dm/2}L^{m/2}(n)} \sum_{\substack{k+l\ge m+1\\ k,l<1/D}} \frac{|a_{kl}|}{k!\,l!}\Bigg(\tilde{C}(k)\sqrt{k!}n^{1-Dk/2}\max_{1\le b\le n} L^{k/2}(b)\log_2 n \\
&\qquad \cdot \tilde{C}(l)\sqrt{l!}n^{1-Dl/2}\max_{1\le b\le n} L^{l/2}(b)\log_2 n\Bigg) \\
&\le \sum_{\substack{k+l\ge m+1\\ k,l<1/D}} \frac{|a_{kl}|}{\sqrt{k!\,l!}}\Bigg(\tilde{C}(k)\tilde{C}(l) \\
&\qquad \cdot n^{\frac{D}{2}(m-(k+l))}\max_{1\le b\le n} L^{k/2}(b)\max_{1\le b\le n} L^{l/2}(b)L^{-m/2}(n)(\log_2 n)^2\Bigg)
\end{aligned}$$

  Now $n^{\frac{D}{2}(m-(k+l))} = n^{-\varepsilon}$ for some $\varepsilon > 0$, and $L^{-m/2}(n)$ and $\log_2^2 n$ are $o(n^{\varepsilon})$ for any $\varepsilon > 0$. We will immediately show that also $\max_{1\le b\le n} L^{k/2}(b)$ is $o(n^{\varepsilon})$ for any $\varepsilon > 0$ and $k \in \mathbb{N}$. Because the summation over k, l is only finite, the sum on the right-hand side is finite, and thus the right-hand side converges to 0.

  Now we show that $\max_{1\le b\le n} L^{k/2}(b)$ is $o(n^{\varepsilon})$ for any $\varepsilon > 0$ and $k \in \mathbb{N}$. When L is slowly varying, $L^{k/2}(x)$ is it as well. So we need to consider

$$\begin{aligned}
\max_{1\le b\le n}\frac{L(b)}{n^{\varepsilon}} &\le \max_{1\le b\le\sqrt{n}}\frac{L(b)}{\sqrt{n}^{\varepsilon}\sqrt{n}^{\varepsilon}} + \max_{\sqrt{n}\le b\le n}\frac{L(b)}{n^{\varepsilon}} \\
&\le \frac{1}{\sqrt{n}^{\varepsilon}}\max_{1\le b\le\sqrt{n}}\frac{L(b)}{b^{\varepsilon}} + \max_{\sqrt{n}\le b\le n}\frac{L(b)}{b^{\varepsilon}},
\end{aligned}$$

  and since $L(b)/b^{\varepsilon} \to 0$ as $b \to \infty$, $\max_{1\le b\le\sqrt{n}} \frac{L(b)}{b^{\varepsilon}}$ is bounded and $\max_{\sqrt{n}\le b\le n} \frac{L(b)}{b^{\varepsilon}}$ converges to 0.

- When $k < 1/D$ and $l > 1/D$, such that the sum over i is LRD and the sum over j is SRD, (6.7) is bounded by

$$\begin{aligned}
&\frac{1}{n^{2-Dm/2}L^{m/2}(n)} \sum_{\substack{k+l\geq m+1\\ k<1/D,\, l>1/D}} \frac{|a_{kl}|}{k!\,l!}\Bigg(C(k)\sqrt{k!}n^{1-Dk/2}\max_{1\leq b\leq n} L^{k/2}(b)\log_2 n \\
&\qquad \cdot \sqrt{l!}\sqrt{n}\log_2 n\Bigg) \\
&\leq \sum_{\substack{k+l\geq m+1\\ k<1/D,\, l>1/D}} \frac{|a_{kl}|}{\sqrt{k!\,l!}}\left(C(k)n^{-\frac{1}{2}+\frac{Dm}{2}-\frac{Dk}{2}}\max_{1\leq b\leq n} L^{k/2}(b)L^{-m/2}(n)(\log_2 n)^2\right)
\end{aligned}$$

  Here, we have summed up some constants in order to keep the expression simple. Now $n^{-\frac{1}{2}+\frac{Dm}{2}-\frac{Dk}{2}} = n^{-\varepsilon}$ for a $\varepsilon > 0$, because $\frac{Dm}{2}, \frac{Dk}{2} \in (0, \frac{1}{2})$. $\max_{1\leq b\leq n} L^{l/2}(b)$, $L^{-m/2}(n)$ and $\log_2^2 n$ are $o(n^{\varepsilon})$ for any $\varepsilon > 0$ as above, and the sum on the right hand side is finite, because of (6.4) and since the summation over k is only finite.

- When $k > 1/D$ and $l < 1/D$, such that the sum over i is SRD and the sum over j is LRD, (6.7) converges to 0 by the same arguments.

- When $k, l > 1/D$, such that both sums are SRD, (6.7) is bounded by

$$\begin{aligned}
&\frac{1}{n^{2-Dm/2}L^{m/2}(n)} \sum_{\substack{k+l\geq m+1\\ k,l>1/D}} \frac{|a_{kl}|}{k!\,l!}\left(C\sqrt{k!}\sqrt{n}\log_2 n\cdot\sqrt{l!}\sqrt{n}\log_2 n\right) \\
&\leq C\sum_{\substack{k+l\geq m+1\\ k,l>1/D}} \frac{|a_{kl}|}{\sqrt{k!\,l!}}\left(n^{-1+Dm/2}L^{-m/2}(n)\log_2^2 n\right)
\end{aligned}$$

  Now $n^{-1+Dm/2} = n^{-\varepsilon}$ for a $\varepsilon > 0$, because $\frac{Dm}{2} \in (0, \frac{1}{2})$. $L^{-m/2}(n)$ and $\log_2^2 n$ are $o(n^{\varepsilon})$ for any $\varepsilon > 0$ as above, and the sum on the right hand side is finite, because of (6.4).

So all in all, (6.7) converges to 0, and the first statement of the Theorem is proved.

For the second statement we consider the cases where $k + l = m$.

$$\begin{aligned}
&\frac{1}{d_n' n}\sum_{i=1}^{[\lambda n]}\sum_{j=[\lambda n]+1}^{n} H_k(\xi_i)H_l(\xi_j) \\
&\quad = n^{-2+Dm/2}L(n)^{-m/2}\sum_{i=1}^{[\lambda n]} H_k(\xi_i)\left(\sum_{j=1}^{n} H_l(\xi_j) - \sum_{j=1}^{[\lambda n]} H_l(\xi_j)\right) \\
&\quad \xrightarrow{\mathcal{D}} c_k^{1/2}\frac{Z_k(\lambda)}{k!}\cdot c_l^{1/2}\frac{(Z_l(1) - Z_l(\lambda))}{l!}
\end{aligned}$$

uniformly in $\lambda \in [0, 1]$ by Theorem 6.1 and the continuous mapping theorem. □

6.2 The limit distribution in special situations

6.2.1 Hermite rank $m = 1$

Corollary. *If the Hermite rank of $h(x, y)$ is $m = 1$, the statement of Theorem 6.2 simplifies to*

$$\boxed{\frac{1}{d'_n\, n}\sum_{i=1}^{[\lambda n]}\sum_{j=[\lambda n]+1}^{n} h(\xi_i,\xi_j) \xrightarrow{\mathcal{D}} \sqrt{c_1}\left(a_{1,0}(1-\lambda)B_H(\lambda) + a_{0,1}\lambda(B_H(1)-B_H(\lambda))\right),} \tag{6.10}$$

where $B_H(\lambda)$ is fBm with parameter $H = 1 - D/2$. If $\lambda \in [0,1]$ is fixed, it holds

$$\boxed{\frac{1}{d'_n\, n}\sum_{i=1}^{[\lambda n]}\sum_{j=[\lambda n]+1}^{n} h(\xi_i,\xi_j) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma^2)} \tag{6.11}$$

with

$$\begin{aligned}\sigma^2 = \frac{2}{(1-D)(2-D)}\big(&-(1-\lambda)\lambda^{2-D}a_{1,0}\left(\lambda a_{0,1} - (1-\lambda)a_{1,0}\right)\\ &+\lambda a_{0,1}\left(a_{1,0} - \lambda a_{1,0} + (1-\lambda)^{2-D}\left(\lambda a_{0,1} - (1-\lambda)a_{1,0}\right)\right)\big).\end{aligned}$$

Proof. By Theorem 6.2, the limit is

$$\begin{aligned}&\sum_{k+l=1}\frac{a_{kl}}{k!\,l!}\sqrt{c_k c_l}Z_k(\lambda)(Z_l(1)-Z_l(\lambda))\\ &\quad= a_{1,0}\sqrt{c_1}Z_1(\lambda)\left(Z_0(1)-Z_0(\lambda)\right) + a_{0,1}\sqrt{c_1}Z_0(\lambda)\left(Z_1(1)-Z_1(\lambda)\right)\\ &\quad= a_{1,0}\sqrt{c_1}(1-\lambda)B_H(\lambda) + a_{0,1}\sqrt{c_1}\lambda(B_H(1)-B_H(\lambda))\end{aligned}$$

with $c_1 = 2/((1-D)(2-D))$, because $Z_0(t) = t$ and $Z_1(t) = B_H(t)$; this proves the first statement.

For the second statement we see that

$$\begin{aligned}\operatorname{Var}\left[B_H(t)\right] &= \operatorname{Cov}\left[B_H(t), B_H(t)\right]\\ &= t^{2H}\\ \operatorname{Var}\left[B_H(1)-B_H(t)\right] &= \operatorname{Var}[B_H(1)] - 2\operatorname{Cov}[B_H(1),B_H(t)] + \operatorname{Var}[B_H(t)]\\ &= 1-(1+t^{2H}-(1-t)^{2H})+t^{2H}\\ &= (1-t)^{2H}\\ \operatorname{Cov}\left[B_H(t), B_H(1)-B_H(t)\right] &= \operatorname{Cov}[B_H(1),B_H(t)] - \operatorname{Var}[B_H(t)]\\ &= \frac{1}{2}(1+t^{2H}-(1-t)^{2H}) - t^{2H},\end{aligned}$$

so that $(B_H(\lambda), B_H(1)-B_H(\lambda)) \sim \mathcal{N}_2(0,\Sigma)$ with

$$\Sigma = \begin{pmatrix} \lambda^{2H} & \frac{1}{2}\left(\lambda^{2H}+1-(1-\lambda)^{2H}\right) - \lambda^{2H}\\ \frac{1}{2}\left(\lambda^{2H}+1-(1-\lambda)^{2H}\right) - \lambda^{2H} & (1-\lambda)^{2H}\end{pmatrix}.$$

A linear combination of two bivariate normally distributed random variables is univariate normally distributed: For $b = (b_1, b_2) \in \mathbb{R}^2$ and $Z = (Z_1, Z_2)^t \sim \mathcal{N}_2(0, \Sigma)$ it holds $b \cdot Z \sim \mathcal{N}(0, b\Sigma b^t)$. Here, $b = \sqrt{c_1}(a_{1,0}(1-\lambda), a_{0,1}\lambda)$, and so

$$\begin{aligned} b^t \Sigma b = c_1 \big(&- (1-\lambda)\lambda^{2H} a_{1,0} \left(\lambda a_{0,1} - (1-\lambda) a_{1,0}\right) \\ &+ \lambda a_{0,1} \left(a_{1,0} - \lambda a_{1,0} + (1-\lambda)^{2H} \left(\lambda a_{0,1} - (1-\lambda) a_{1,0}\right)\right) \big). \end{aligned}$$

□

6.2.2 Two independent samples

Now we consider the behaviour of $U_{\lambda,n}$ in a situation with two independent samples

$$\xi_1, \xi_2, \ldots, \xi_{[\lambda n]} \quad \text{and} \quad \eta_1, \eta_2, \ldots, \eta_{n-[\lambda n]}$$

where $\lambda \in (0,1)$. In this situation, we can show in Theorem 6.2 convergence in L^2.

Theorem 6.3. *Let $(\xi_i)_{i\geq 1}$ and $(\eta_j)_{j\geq 1}$ be two stationary Gaussian processes, independent of each other, each with mean 0, variance 1 and covariances* (1.1)*. Let h be a function in $\mathcal{G}^2(\mathbb{R}^2, \mathcal{N})$ with Hermite rank m whose Hermite coefficients satisfy* (6.4)*. Let $Dm < 1$. Then as $n \to \infty$*

$$\frac{1}{d'_n\, n} \left| \sum_{i=1}^{[\lambda n]} \sum_{j=1}^{n-[\lambda n]} h(\xi_i, \eta_j) - \sum_{\substack{k,l:\\ k+l=m}} \frac{a_{kl}}{k!\, l!} \sum_{i=1}^{[\lambda n]} \sum_{j=1}^{n-[\lambda n]} H_k(\xi_i) H_l(\eta_j) \right| \xrightarrow{L^2} 0$$

uniformly in $\lambda \in [0,1]$ and

$$\boxed{\frac{1}{d'_n\, n} \sum_{i=1}^{[\lambda n]} \sum_{j=1}^{n-[\lambda n]} h(\xi_i, \eta_j) \xrightarrow{\mathcal{D}} \sum_{\substack{k,l:\\ k+l=m}} \frac{a_{kl}}{(k!\, l!)^2} \sqrt{c_k c_l}\, Z_k(\lambda) Z'_l(1-\lambda)} \tag{6.12}$$

in $D[0,1]$, where d'^2_n is as in (6.6) *and $Z_k(\lambda), Z'_l(\lambda)$, $k, l = 0, \ldots, m$, $\lambda \in [0,1]$ are independent Hermite processes, see* (1.11)*.*

Proof. Set temporarily $S_k := \sum_{i=1}^{[\lambda n]} H_k(\xi_i)$ and $S_l := \sum_{j=1}^{n-[\lambda n]} H_l(\eta_j)$ ($S_{k'}$ and $S_{l'}$ respectively). Like in the proof of Theorem 6.2, it holds

$$\sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} h(\xi_i, \eta_j) - \sum_{\substack{k,l:\\ k+l=m}} \frac{a_{kl}}{k!\, l!} S_k S_l = \sum_{\substack{k,l:\\ k+l\geq m+1}} \frac{a_{kl}}{k!\, l!} S_k S_l,$$

and thus with $b = [\lambda n] > 1$

$$E\left|\sup_{0\le\lambda\le 1}\frac{1}{n\,d_n'}\left(\sum_{i=1}^{[\lambda n]}\sum_{j=[\lambda n]+1}^{n}h(\xi_i,\eta_j)-\sum_{\substack{k,l:\\ k+l=m}}\frac{a_{kl}}{k!\,l!}S_kS_l\right)\right|^2$$
$$=E\left|\max_{1\le b\le n}\frac{1}{n\,d_n'}\sum_{\substack{k,l:\\ k+l\ge m+1}}\frac{a_{kl}}{k!\,l!}S_kS_l\right|^2.$$

We will show that the term on the right-hand side converges to 0, and so will the term on the left-hand side. Like in the proof of Theorem 6.2 and since ξ_i and η_j are independent, the right-hand side is bounded by

$$E\left[\frac{1}{n\,d_n'}\sum_{\substack{k,l:\\ k+l\ge m+1}}\frac{|a_{kl}|}{k!\,l!}\max_{1\le b\le n}|S_k|\max_{1\le b\le n}|S_l|\right]^2$$
$$=\frac{1}{(n\,d_n')^2}\sum_{\substack{k,l:\\ k+l\ge m+1}}\sum_{\substack{k',l':\\ k'+l'\ge m+1}}\left(\frac{|a_{kl}|}{k!\,l!}\frac{|a_{k'l'}|}{k'!\,l'!}\right.$$
$$\left.\cdot E\left[\max_{1\le b\le n}|S_k|\max_{1\le b\le n}|S_{k'}|\right]E\left[\max_{1\le b\le n}|S_l|\max_{1\le b\le n}|S_{l'}|\right]\right)$$
$$\le\frac{1}{(n\,d_n')^2}\sum_{\substack{k,l:\\ k+l\ge m+1}}\sum_{\substack{k',l':\\ k'+l'\ge m+1}}\left(\frac{|a_{kl}|}{k!\,l!}\frac{|a_{k'l'}|}{k'!\,l'!}\right.$$
$$\left.\cdot\sqrt{E\left[\max_{1\le b\le n}|S_k|\right]^2E\left[\max_{1\le b\le n}|S_{k'}|\right]^2}\sqrt{E\left[\max_{1\le b\le n}|S_l|\right]^2E\left[\max_{1\le b\le n}|S_{l'}|\right]^2}\right)$$

by the Cauchy–Bunyakovsky–Schwarz inequality, and finally by sorting

$$=\frac{1}{n\,d_n'}\sum_{\substack{k,l:\\ k+l\ge m+1}}\left(\frac{|a_{kl}|}{k!\,l!}\sqrt{E\left[\max_{1\le b\le n}|S_k|\right]^2E\left[\max_{1\le b\le n}|S_l|\right]^2}\right)$$
$$\cdot\frac{1}{n\,d_n'}\sum_{\substack{k',l':\\ k'+'l\ge m+1}}\left(\frac{|a_{k'l'}|}{k'!\,l'!}\sqrt{E\left[\max_{1\le b\le n}|S_{k'}|\right]^2E\left[\max_{1\le b\le n}|S_{l'}|\right]^2}\right).$$

Now we have two expressions like in the right-hand side of (6.7) which converge to 0, as we have shown in the proof of Theorem 6.2. Thus, the first statement of the Theorem is proved.

For the second statement we consider the cases where $k + l = m$.

$$\begin{aligned}
&\frac{1}{d'_n\, n} \sum_{i=1}^{[\lambda n]} \sum_{j=1}^{n-[\lambda n]} H_k(\xi_i) H_l(\eta_j) \\
&\quad = \frac{1}{n^{1-kD/2} L^{k/2}(n)} \sum_{i=1}^{[\lambda n]} H_k(\xi_i) \cdot \frac{1}{n^{1-lD/2} L^{l/2}(n)} \sum_{j=1}^{n-[\lambda n]} H_l(\eta_j) \\
&\quad \xrightarrow{\mathcal{D}} \sqrt{c_k} \frac{Z_k(\lambda)}{k!} \cdot \sqrt{c_l} \frac{Z'_l(1-\lambda)}{l!}
\end{aligned}$$

by Theorem 1.1. □

Remark. Observe that the limits in Theorem 6.2 (one sample $\xi_1, \ldots, \xi_n$ which is divided in $\xi_1, \ldots, \xi_{[\lambda n]}$ and $\xi_{[\lambda n]+1}, \ldots, \xi_n$) and in Theorem 6.3 (two independent samples $\xi_1, \ldots, \xi_{[\lambda n]}$ and $\eta_1, \eta_2, \ldots, \eta_{n-[\lambda n]}$) differ. This is in contrast to the weak dependent case.

6.3 Examples

6.3.1 "Differences-of-means" test

The kernel $h(x, y) = x - y$ leads to the *difference-of-means statistic*

$$\boxed{U_{\mathrm{diff},\lambda,n} = \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} (\xi_i - \xi_j) = [\lambda n](n - [\lambda n]) \left(\bar{X}_1^{[\lambda n]} - \bar{X}_{[\lambda n]+1}^{n} \right).}$$

This kernel is of course in $L^2(\mathbb{R}^2, \mathcal{N})$ and its Hermite expansion can be read off without calculating:

$$h(x, y) = x - y = \frac{a_{1,0}}{1!0!} H_1(x) + \frac{a_{0,1}}{0!1!} H_1(y),$$

so its Hermite coefficients are

$$a_{kl} = \begin{cases} 1 & k = 1, l = 0 \\ -1 & k = 0, l = 1 \\ 0 & \text{else} \end{cases},$$

and condition (6.4) is trivially fulfilled.

"Differences-of-means" test for one divided sample

The Corrolary to Theorem 6.2 states

$$\frac{1}{d'_n\, n} U_{\mathrm{diff},\lambda,n} \xrightarrow{\mathcal{D}} \sqrt{c_1} \left((1-\lambda) B_H(\lambda) - \lambda (B_H(1) - B_H(\lambda)) \right),$$

and this is exactly the result of Section 3.4.2 and also confirmed by Horváth and Kokoszka (1997).

For a *fixed* λ, we obtain the asymptotic behaviour of the two-sample Gauß test statistic:

$$\frac{1}{d'_n\, n} U_{\mathrm{diff},\lambda,n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, s^2)$$

with $s^2 = \frac{2}{(1-D)(2-D)} \left((1-\lambda)\lambda^{2-D} - \lambda\left(1-\lambda-(1-\lambda)^{2-D}\right)\right)$ or

$$n^{D/2} L(n)^{-1/2} \frac{\left(\bar{X}_1^{[\lambda n]} - \bar{X}_{[\lambda n]+1}^{n}\right)}{\sigma_{\mathrm{diff}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

with

$$\sigma^2_{\mathrm{diff}} := \frac{2}{(1-D)(2-D)} \frac{\lambda^{1-D} - 1 + (1-\lambda)^{1-D}}{\lambda(1-\lambda)},$$

and this is exactly the same as we have calculated manually in Theorem 2.3.

"Differences-of-means" test for two independent samples

We consider two samples of independent observations $\xi_1, \ldots, \xi_{n_1}$ and $\eta_1, \ldots, \eta_{n_2}$ which are independent of each other. A common test statistic to detect differences in the location of these two samples, is the simple *difference of means statistic*, also known as the *Gauß test statistic*, based on the kernel $h(x,y) = x - y$:

$$U'_{\mathrm{diff},n_1,n_2} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} h(\xi_i, \eta_j) = n_1 n_2 (\bar{\xi}_{n_1} - \bar{\eta}_{n_2})$$

Theorem 6.3 states

$$\begin{aligned} \frac{1}{d'_n\, n} U'_{\mathrm{diff},n_1,n_2} &\xrightarrow{\mathcal{D}} \sum_{\substack{k,l:\\ k+l=1}} \frac{a_{kl}}{k!\, l!} \sqrt{c_k c_l} Z_k(\lambda) Z_l'(1-\lambda) \\ &= \sqrt{c_1}\left(Z_1(\lambda)(1-\lambda) - \lambda Z_1'(1-\lambda)\right). \end{aligned}$$

A short check with basic theory, i.e. applying Theorem 1.1 directly, yields the same:

$$\begin{aligned} \frac{1}{d'_n\, n} U'_{\mathrm{diff},n_1,n_2} &= (1-\lambda)\frac{1}{d_n} \sum_{i=1}^{[\lambda n]} \xi_i - \lambda \frac{1}{d_n} \sum_{j=1}^{n-[\lambda n]} \eta_i \\ &\xrightarrow{\mathcal{D}} Z_1(\lambda)(1-\lambda) - \lambda Z_1'(1-\lambda). \end{aligned}$$

For $\lambda \in [0,1]$ *fixed*, we obtain

$$\frac{1}{d'_n\, n} U'_{\mathrm{diff},n_1,n_2} \xrightarrow{\mathcal{D}} \mathcal{N}(0, s'^2_{\mathrm{diff}})$$

with

$$\begin{aligned} s'^2_{\mathrm{diff}} &= \mathrm{Var}\left[\sqrt{c_1}\left(Z_1(\lambda)(1-\lambda) - \lambda Z_1'(1-\lambda)\right)\right] \\ &= c_1\left(\lambda^{2-D}(1-\lambda)^2 - \lambda(1-\lambda)^{2-D}\right) \end{aligned}$$

or

$$\sqrt{\frac{n_1 n_2}{n^{2-D} L(n)\, \sigma^2_{\text{diff},}}}(\bar{\xi}_{n_1} - \bar{\eta}_{n_2}) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

with $n = n_1 + n_2$ and $\sigma^2_{\text{diff},} = \frac{2}{(1-D)(2-D)}(\lambda^{1-D}(1-\lambda) + \lambda(1-\lambda)^{1-D})$. This is exactly the same as we have calculated manually in Theorem 2.4.

"Differences-of-means" test for non-Gaussin data

So far, we have considered Gaussian observations $\xi_1, \dots, \xi_n$. As mentioned in the introduction to this chapter, all results can be extended to general data $G(\xi_1), \dots, G(\xi_n)$, where G is a transformation, by considering the kernel $h(G(x), G(y))$ instead of $h(x,y)$. This is what we will do now exemplarily.

Consider a function $G : \mathbb{R} \to \mathbb{R}$, $G \in \mathcal{G}^2(\mathbb{R}, \mathcal{N}) \subset L^2(\mathbb{R}, \mathcal{N})$, like the quantile transformations from Section 3.6. The Hermite coefficients of the function $h(G(x), G(y))$ are

$$\begin{aligned} a_{kl} &= \iint_{\mathbb{R}^2} (G(x) - G(y)) H_k(x) H_l(y)\, d\Phi(x)\, d\Phi(y) \\ &= \int_{\mathbb{R}} G(x) H_k(x)\, d\Phi(x) \cdot \int_{\mathbb{R}} H_l(y)\, d\Phi(y) - \int_{\mathbb{R}} G(y) H_l(y)\, d\Phi(y) \cdot \int_{\mathbb{R}} H_k(x)\, d\Phi(x) \\ &= \begin{cases} 0 & \text{if } k, l \neq 0 \\ -a_l & \text{if } k = 0, l \neq 0 \\ a_k & \text{if } k \neq 0, l = 0 \end{cases}, \end{aligned}$$

where $a_p = E[G(\xi) H_p(\xi)]$ is the p-th Hermite coefficient of G. Thus for such G and $h(x,y) = x - y$, the summability condition (6.4) is satisfied:

$$\sum_{k,l} \frac{a_{kl}}{\sqrt{k!\, l!}} = \sum_{k=1}^{\infty} \frac{a_k}{\sqrt{k!}} - \sum_{l=1}^{\infty} \frac{a_l}{\sqrt{l!}} = 0$$

6.3.2 "Wilcoxon-type" test

Using $h(x,y) = I_{\{x \leq y\}}$ yields the famous *Mann-Whitney-Wilcoxon statistic*

$$\boxed{U_{W,\lambda,n} = \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} I_{\{\xi_i \leq \xi_j\}}.}$$

We will now show that this kernel does not fulfil the summability condition (6.4), but that – if we ignore this – the above theorems nevertheless reproduce our results from the last chapters. This confirms that condition (6.4) may be too strong and suggests that Theorem 6.2 holds under milder assumptions.

Summability condition not fulfilled

We will now demonstrate that $h(x,y) = I_{\{x\le y\}}$ does not satisfy (6.4), which is neither intuitively visible nor enjoyable to verify. The Hermite coefficients are

$$\begin{aligned} a_{kl} &= \iint_{\{x\le y\}} H_k(x)H_l(y)\varphi(x)\varphi(y)\,dx\,dy \\ &= \int_{\mathbb{R}} H_l(y)\varphi(y)\left(\int_{-\infty}^{y} H_k(x)\varphi(x)\,dx\right)dy \end{aligned}$$

and by the definition of the Hermite polynomials and for $k \ge 1$

$$= \int_{\mathbb{R}} \frac{(-1)^{k+l}}{2\pi}\left(\frac{d^l}{dy^l}e^{-y^2/2}\right)\left(\frac{d^{k-1}}{dy^{k-1}}e^{-y^2/2}\right)\,dy.$$

If we now integrate by parts, the boundary terms vanish. Therefore $k-1$ times integration by parts yields

$$\begin{aligned} a_{k,l} &= (-1)^{k-1}\frac{(-1)^{k+l}}{2\pi}\int_{\mathbb{R}}\left(\frac{d^{l+k-1}}{dy^{l+k-1}}e^{-y^2/2}\right)e^{-y^2/2}\,dy \\ &= \frac{(-1)^{2k+l-1}}{2\pi}\int_{\mathbb{R}} H_{l+k-1}(y)e^{-y^2}\,dy. \end{aligned}$$

From the symmetry of H_{l+k-1} we conclude that $a_{k,l} = 0$ if $l+k-1$ is odd. But if $l+k-1$ is even, we are in the unpleasant situation that we need to evaluate the integral, but the Hermite polynomials H_{l+k-1} do not form an orthogonal family in L^2 with respect to the weight e^{-y^2}. But we can transform them into physicists' Hermite polynomials $H^{(phy)}_{l+k-1}$, so that we can apply the following formula (Bateman, 1953, p. 195):

$$\int_{\mathbb{R}} e^{-y^2}H^{(phy)}_{2m}(ay)\,dy = \sqrt{\pi}\frac{(2m)!}{m!}(a^2-1)^m$$

In doing so we obtain for even $l+k-1$

$$\begin{aligned} a_{k,l} &= \frac{(-1)^{2k+l-1}}{2\pi}2^{-\frac{l+k-1}{2}}\int_{\mathbb{R}} H^{(phy)}_{l+k-1}(2^{-\frac{1}{2}}y)e^{-y^2}\,dy \\ &= \frac{(-1)^k}{\sqrt{\pi}}2^{-\frac{l+k-1}{2}-1}\frac{(l+k-1)!}{\left(\frac{l+k-1}{2}\right)!}\left(-\frac{1}{2}\right)^{\frac{l+k-1}{2}} \\ &= \frac{(-1)^{\frac{l+k-1}{2}+k}}{\sqrt{\pi}}\frac{\Gamma(l+k)\Gamma\left(\frac{l+k}{2}\right)}{2^{l+k}\Gamma\left(\frac{l+k}{2}+\frac{1}{2}\right)\Gamma\left(\frac{l+k}{2}\right)}. \end{aligned}$$

We have expanded the fraction with $\Gamma\left(\frac{l+k}{2}\right)$ in order to use the Legendre duplication formula $\Gamma(z)\Gamma(z+\frac{1}{2}) = 2^{1-2z}\sqrt{\pi}\Gamma(2z)$ in the denominator. So finally we have found an explicit expression for the Hermite coefficients of $h(x,y) = I_{\{x\le y\}}$:

$$a_{k,l} = \begin{cases} \frac{(-1)^{\frac{l+3k-1}{2}}}{2\pi}\Gamma\left(\frac{l+k}{2}\right) & l+k \text{ odd and positive} \\ 0 & l+k \text{ even and positive} \\ \frac{1}{2} & l=k=0 \end{cases} \tag{6.13}$$

Now we show that $\sum_{k,l=1}^{\infty} |a_{k,l}|/\sqrt{k!\,l!}$ diverges. It is enough to consider the first odd diagonal where $l = k+1$, because there we have already with Sterling's approximation

$$\frac{|a_{k,l}|}{\sqrt{k!\,l!}} \sim \frac{(2k-1)^k e}{2^k (k+1)^{k/2+3/4} k^{k/2+1/4}} = \frac{\left(1-\frac{1}{2k}\right)^k \frac{1}{k} e}{\left(1+\frac{1}{k}\right)^{k/2} \left(1+\frac{1}{k}\right)^{3/4}} \sim \frac{1}{k}.$$

"Wilcoxon-type" test for one divided sample

Let us for a moment ignore that the Wilcoxon kernel does not fulfill the summability condition (6.4), which may be too rigorous anyway, and apply Theorem 6.2. To this end, we use (6.13) or calculate the first Hermite coefficients manually:

$$\begin{aligned}
a_{0,0} &= \iint_{\{x\le y\}} H_0(x)H_0(y)\varphi(x)\varphi(y)\,dx\,dy = \iint_{\{x\le y\}} \varphi(x)\varphi(y)\,dx\,dy = \frac{1}{2} \\
a_{1,0} &= \iint_{\{x\le y\}} x\varphi(x)\varphi(y)\,dx\,dy = -\frac{1}{2\sqrt{\pi}} \\
a_{0,1} &= \iint_{\{x\le y\}} y\varphi(x)\varphi(y)\,dx\,dy = \frac{1}{2\sqrt{\pi}}
\end{aligned}$$

Since we formulated the theorem for centralized kernels, we consider $h(x,y) - E[h(\xi,\eta)] = I_{\{x\le y\}} - 1/2$, which has Hermite rank $m = 1$. So the Corrolary to Theorem 6.2 states that

$$\begin{aligned}
\frac{1}{n\,d_n'} \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} I_{\{\xi_i\le\xi_j\}} &\xrightarrow{\mathcal{D}} \sqrt{c_1}\,(a_{1,0}(1-\lambda)B_H(\lambda) + a_{0,1}\lambda(B_H(1) - B_H(\lambda))) \\
&= \frac{\sqrt{c_1}}{2\sqrt{\pi}}\,(\lambda B_H(1) - B_H(\lambda))\,.
\end{aligned}$$

And for *fixed* $\lambda \in [0,1]$ we obtain

$$\frac{1}{n\,d_n'} \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} I_{\{\xi_i\le\xi_j\}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma_W^2)$$

with

$$\sigma_W^2 = \frac{c_1}{4\pi}\left(\lambda^2 - \lambda + (1-\lambda)\lambda^{2-D} + \lambda(1-\lambda)^{2-D}\right).$$

Bearing in mind that $\int_{\mathbb{R}} J_1(x)\,d\Phi(x) = -(2\sqrt{\pi})^{-1}$, we have just reproduced Theorem 3.4 for the Gaussian case and the findings of Section 5.3.

The Wilcoxon two-sample test for independent samples

We ignore that the Wilcoxon kernel does not fulfill the summability condition (6.4) and apply Theorem 6.3 for the case of two independent LRD samples. For two samples of independent observations $\xi_1,\dots,\xi_{n_1}$ and $\eta_1,\dots,\eta_{n_2}$ which are independent of

each other and have the same distribution but may differ in their location, the kernel $h(x,y) = I_{\{x \le y\}}$ yields the famous *Mann-Whitney-Wilcoxon statistic*

$$U'_{W,n_1,n_2} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I_{\{\xi_i \le \eta_j\}},$$

also known as *Mann-Whitney U* or *Wilcoxon rank-sum test* (since it can be written as a statistic of ranks). It is well known that U'_{W,n_1,n_2} is asymptotically normally distributed:

$$\frac{U'_{W,n_1,n_2} - \frac{n_1 n_2}{2}}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

respectively written in a more convenient notation

$$n^{-3/2} \frac{U'_{W,n_1,n_2} - m_W}{s_{W'}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

with $m_W = n_1 n_2/2$ and $s_{W'} = \sqrt{\lambda(1-\lambda)/12}$. In our situation here, with two independent samples of LRD observations, one expects at least a different normalizing and a stronger scaling, and indeed, this is true.

By the above calculation, the Hermite rank of $h(x,y) = I_{\{x \le y\}}$ is $m = 1$, and Theorem 6.3 states

$$\frac{1}{n\, d'_n} (U'_{W,n_1,n_2} - m_W) \xrightarrow{\mathcal{D}} \frac{\sqrt{c_1}}{2\sqrt{\pi}} \left(-Z_1(\lambda)(1-\lambda) + \lambda Z'_1(1-\lambda) \right),$$

and this is the result of Theorem 3.2 for the Gaussian case.

When we consider a *fixed* $\lambda \in [0,1]$, then we obtain

$$n^{-2+\frac{D}{2}} L(n)^{-1/2} \frac{U_W - m_W}{\sigma_{W'}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

with $m_W = n_1 n_2/2$ and

$$\sigma^2_{W'} = \mathrm{Var}\left[\frac{\sqrt{c_1}}{2\sqrt{\pi}} \left(-Z_1(\lambda)(1-\lambda) + \lambda Z'_1(1-\lambda) \right) \right] = \frac{\lambda^{2-D}(1-\lambda)^2 + (1-\lambda)^{2-D}\lambda^2}{2\pi(1-D)(2-D)}.$$

And indeed, the normalizing is different and the scaling is stronger than for mutually independent observations: $n^{3/2} < n^{2-\frac{D}{2}} < n^2$.

"Wilcoxon-type' test for non-Gaussin data

Like for the "difference-of-means" test in the previous section, we will shortly show that we can easily pass over from Gaussian observations $\xi_1, \ldots, \xi_n$ to non-Gaussian observations $G(\xi_1), \ldots, G(\xi_n)$, where G is a transformation – irrespective of the fact, that the kernel $h(x,y) = I_{\{x \le y\}}$ does not fulfill the summability condition (6.4). Although in general, this changeover produces even more difficult integrals when calculating Hermite coefficients, the indicator kernel, like the differences kernel $h(x,y) = x - y$, makes it easy.

Consider a strictly monotone function $G : \mathbb{R} \to \mathbb{R}$, $G \in \mathcal{G}^2(\mathbb{R}, \mathcal{N}) \subset L^2(\mathbb{R}, \mathcal{N})$. The Hermite coefficients of the function $h(G(x), G(y))$ are

$$\begin{aligned} a_{kl} &= \iint_{\mathbb{R}^2} I_{\{G(x)\le G(y)\}} H_k(x) H_l(y)\, d\Phi(x)\, d\Phi(y) \\ &= \iint_{\mathbb{R}^2} I_{\{x\le y\}} H_k(x) H_l(y)\, d\Phi(x)\, d\Phi(y), \end{aligned}$$

and this is nothing else than the Hermite coefficients of $h(x, y) = I_{\{x\le y\}}$ itself. Here we have essentially used that G is strictly monotone. Note that this result is in concordance with the invariance of the "Wilcoxon-type' test statistic which we proved in Lemma 3.3.

6.4 In search of handy criteria

There is no denying that the summability condition (6.4) is somewhat unhandy. If we are not in the unlikely situation that h is just a polynomial and has therefore a finite Hermite expansion, most kernels h will not do us the favour that one look at them tells us if (6.4) is fulfilled. So next I will investigate if there are criteria which are easier to verify.

Lemma 6.4. *Let h be four times differentiable and let h and its four derivatives be in $L^1(\mathbb{R}^2, \lambda)$. Then condition* (6.4) *is fulfilled.*

Proof. By the Plancherel theorem we can write Hermite coefficients in the following way:

$$\begin{aligned} a_{kl} &= \frac{1}{2\pi} \iint_{\mathbb{R}^2} h(x, y) H_k(x) H_l(y) e^{-(x^2+y^2)/2}\, dx\, dy \\ &= \frac{1}{2\pi} \iint_{\mathbb{R}^2} \hat{h}(s, t) g(s, t)\, ds\, dt, \end{aligned}$$

where $\hat{h} = \mathcal{F}(h)$ and $g(s, t) = \mathcal{F}(H_k(x)H_l(y)e^{-(x^2+y^2)/2})$ denote the Fourier transform of h and $H_k(x)H_l(y)e^{-(x^2+y^2)/2}$, as defined in (6.14). Now we want to give an explicit representation of $g(s, t)$, and we will use the following properties of Fourier transforms and Hermite polynomials:

- $\mathcal{F}\left(e^{-(x^2+y^2)/2}\right) = e^{-(s^2+t^2)/2}$
  This is a standard result from elementary Fourier analysis.
- $\mathcal{F}\left(\frac{\partial^{k+l}}{\partial x^k \partial x^l} f(x, y)\right) = i^{k+l} s^k t^l \hat{f}(s, t)$
  This as well.
- $H_k(x)H_l(y)e^{-(x^2+y^2)/2} = \frac{\partial^{k+l}}{\partial x^k \partial x^l} e^{-(x^2+y^2)/2} (-1)^{k+l}$
  This follows easily from the definition of Hermite polynomials.

With this formulae we can write

$$a_{kl} = \frac{1}{2\pi} \iint_{\mathbb{R}^2} \hat{h}(s, t)(-i)^{k+l} s^k t^l e^{-(s^2+t^2)/2}\, ds\, dt.$$

Now we will bound this expression, using factorials and exponential functions. For reasons of simplicity, we write $(k/2)! := \Gamma(k/2+1)$. Stirling's approximation still holds, i.e. $(k/2)! \sim \sqrt{2\pi(k/2)}(k/(2e))^{k/2}$, and from this it follows that $(k/2)!/\sqrt{k!} \approx C2^{-k/2}k^{1/4}$ and we obtain

$$\begin{aligned}\sum_{k,l}\frac{|a_{kl}|}{\sqrt{k!\,l!}} &\le C\sum_{k,l}\iint_{\mathbb{R}^2}|\hat{h}(s,t)|\frac{|s|^k|t|^l}{\sqrt{k!\,l!}}e^{-(s^2+t^2)/2}\,ds\,dt\\ &\approx C\sum_{k,l}\iint_{\mathbb{R}^2}|\hat{h}(s,t)|\frac{|s|^k|t|^l}{(k/2)!\,(l/2)!}(kl)^{1/4}2^{-(k+l)/2}e^{-(s^2+t^2)/2}\,ds\,dt\\ &= C\sum_{k,l}\iint_{\mathbb{R}^2}|\hat{h}(s,t)|e^{-(s^2+t^2)/2}\frac{\left(\frac{s^2}{2}\right)^{k/2}\left(\frac{t^2}{2}\right)^{l/2}}{(k/2)!\,(l/2)!}(kl)^{1/4}\,ds\,dt.\end{aligned}$$

Note that $k^{1/4} \le k/2$ for $k \ge 3$ and

$$\sum_{k=0}^{\infty}\frac{x^{k/2}k/2}{(k/2)!} = \frac{\sqrt{x}}{\sqrt{\pi}} + xe^x(1+\operatorname{erf}(\sqrt{x})) \le \frac{\sqrt{x}}{\sqrt{\pi}} + 2xe^x,$$

where $\operatorname{erf}(x)$ denotes the Gaussian error function $2/\sqrt{\pi}\int_0^x e^{-t^2}\,dt$ which is bounded by 1, such that we obtain

$$\sum_{k,l}\frac{|a_{kl}|}{\sqrt{k!\,l!}} \le C\iint_{\mathbb{R}^2}|\hat{h}(s,t)|\left(\sqrt{\frac{s^2}{2}}e^{-s^2/2}+s^2\right)\left(\sqrt{\frac{t^2}{2}}e^{-t^2/2}+t^2\right)\,ds\,dt.$$

Now we know for any $f \in L^1(\mathbb{R}^2,\lambda)$ that $\hat{f}(\xi) \to 0$ as $|\xi| \to \infty$, and we know for any k-times differentiable f that $\mathcal{F}(\frac{d^k}{dx^k}f(x)) = \xi^k\hat{f}(\xi)$. By our assumptions, h and its first four derivatives are in $L^1(\mathbb{R}^2,\lambda)$, so we receive $\xi^4\hat{f}(\xi) \to 0$, respectively $\hat{f}(\xi) = o(\xi^{-4})$ as $|\xi| \to \infty$. So the integrand is bounded around 0 and on large scale it decreases at least like $(st)^{-2}$, and thus the integral is finite. □

Example. Any function $h \in L^1(\mathbb{R}^2,\lambda)$ with four integrable derivatives satisfies the summability condition (6.4), for instance:

(i) a (normalized) Hermite function

$$\begin{aligned}\tilde{h}_{kl}(x,y) &= \frac{1}{\sqrt{2^{k+l}k!\,l!\pi}}H_{kl}^{(phy)}(x,y)e^{-(x^2+y^2)/2}\\ &= \frac{1}{\sqrt{2^{k+l}k!\,l!\pi}}(-1)^{k+l}e^{(x^2+y^2)/2}\frac{d^k}{dx^k}e^{-x^2}\frac{d^l}{dy^l}e^{-y^2}\end{aligned}$$

(ii) a Gaussian function

$$g(x,y) = a\exp\left\{-b\cdot\left(\begin{pmatrix}x\\y\end{pmatrix}-\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix}\right)^t\Sigma^{-1}\left(\begin{pmatrix}x\\y\end{pmatrix}-\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix}\right)\right\}$$

with $a, b, \mu_1, \mu_2 \in \mathbb{R}$ and $\Sigma \in \mathbb{R}^{2\times 2}$ a symmetric positive-definite matrix

(iii) a smooth function with bounded support like the bump function

$$f(x,y) = \begin{cases} e^{\frac{-1}{1-x^2}} e^{\frac{-1}{1-y^2}} & |x|, |y| < 1 \\ 0 & \text{else} \end{cases}$$

Now we turn to a even more tricky criterion involving Fourier transforms. (6.4) is a condition on the summability of the Hermite coefficients. Vemuri (2008) has shown that the Hermite coefficients of a function $f : \mathbb{R} \to \mathbb{R}$ decay exponentially under some growth and smoothness conditions, namely that both the function and its Fourier transform are dominated by a Gaussian of large variance. Such an exponential deacy would be sufficient to imply (6.4), so we will try to adapt this technique for our porposes and extend Vemuri's results to the multi-dimensional case. Since the method uses techniques from complex analysis – which I dare say are not common among probabilists –, I will work out the proof in details. We will use the following multi-dimensional notation:

- We write $x \cdot y = \sum_{i=1}^{d} x_i y_i$ for the inner product on $\mathbb{R}^d$ and abbreviate $x \cdot x = x^2$.
- At times we use multi-index notation when we refer to a general result: A d-dimensional multi-index is a vector $\nu = (\nu_1, \ldots, \nu_d)$ of non-negative integers. For two multi-indices $\nu, \beta \in \mathbb{N}_0^d$ and $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ one defines

$$\begin{aligned} \nu \pm \beta &= (\nu_1 \pm \beta_1,\, \nu_2 \pm \beta_2, \ldots,\, \nu_d \pm \beta_d) \\ |\nu| &= \nu_1 + \nu_2 + \cdots + \nu_d \\ \nu! &= \nu_1! \cdot \nu_2! \cdots \nu_d! \\ x^\nu &= x_1^{\nu_1} x_2^{\nu_2} \ldots x_d^{\nu_d} \\ \partial^\nu &= \partial_1^{\nu_1} \partial_2^{\nu_2} \ldots \partial_d^{\nu_d} = \frac{\partial^{\nu_1}}{\partial x_1^{\nu_1}} \cdots \frac{\partial^{\nu_d}}{\partial x_d^{\nu_d}} \end{aligned}$$

The idea of the proof is as follows: When calculating the k,l-th Hermite coefficient of a kernel $h \in L^2(\mathbb{R}^d, \mathcal{N})$, we have to evaluate the inner product of h with the k,l-th Hermite polynomial with respect to the bivariate standard normal measure

$$\langle h, H_{k,l} \rangle_{\mathcal{N}} = \int_{\mathbb{R}^2} h(x,y) H_{k,l}(x,y)\; d\Phi(x) d\Phi(y).$$

We can represent this expression as the inner product of a slightly modified kernel f and the k,l-th Hermite function with respect to the Lebesgue measure:

$$\langle h, H_{k,l} \rangle_{\mathcal{N}} = \langle f, h_{k,l} \rangle_{\lambda}$$

In order to bound this, we will use a transformation into another space which preserves this inner product. This is the so called Bargmann transformation which translates Hermite functions into monomials. Moreover, the new space in which we operate then

contains entire functions on $\mathbb{C}^d$, i.e. functions that are holomorphic on the whole $\mathbb{C}^d$, and so we can apply tools from complex analysis to find bounds corresponding to the monomials, most notably the Phragmén-Lindelöf principle.

The Phragmén-Lindelöf principle is an extension of the well-known maximum modulus principle, which states that an analytic function f on a bounded region in $\mathbb{C}$ with $|f(z)| \leq 1$ on the boundary is bounded by 1 in the interior as well. This does not apply to unbounded regions, but under certain growth conditions, a bound on the edges still implies a bound in the interior.

Theorem 6.5 (Phragmén-Lindelöf principle)**.** *Let $f(z)$ be an analytic function of a complex variable $z = re^{i\vartheta}$ defined on the region D between two straight rays making an angle π/b at the origin and on the lines themselves, i.e.*

$$D = \left\{ z = re^{i\vartheta} \,\middle|\, \vartheta \in [\vartheta_0, \vartheta_0 + \pi/b] \right\}$$

for $\vartheta_0, b \in \mathbb{R}$. Suppose that for some constant M

$$|f(z)| \leq M$$

on the lines, and that for every $\delta > 0$

$$f(z) = O(e^{\delta r^b})$$

uniformly in the angle as $r \to \infty$. Then $|f(z)| \leq M$ throughout the whole region D.

For a proof, some different formulations of the principle (it applies to strips as well as to cones) and a general discussion of the concept, see Titchmarsh (1964, p. 176 –187)[2].

Definition 6.3 (Bargmann transform)**.** The *Bargmann transform* of a function f on $\mathbb{R}^d$ is defined by

$$\boxed{Bf(z) = \frac{e^{-z^2/4}}{2^{d/4}\pi^{d/2}} \int_{\mathbb{R}^d} f(x) e^{xz} e^{-x^2/2}\, dx,}$$

where $z^2 = z_1^2 + \ldots + z_d^2$. $Bf(z)$ is an element of the so-called Bargmann-Fock space $\mathcal{F}^2(\mathbb{C}^d)$, the Hilbert space of all entire functions F on $\mathbb{C}^d$ with finite norm

$$\|F\|_{\mathcal{F}}^2 = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{C}^d} |F(z)|^2 e^{-|z|^2/2}\, dz.$$

[2]Titchmarsh requires the function to be regular. This emanates from his definition of functions which is obsolete, but still can be found in complex analysis at times: A function f is not an unique mapping, but a formal expression whose values $f(z)$ are obtained by all possible limits approaching z. In this case, $f(z)$ can have several values, consider for example $\sqrt{z}$, which reaches a different value (namely a value rotated by π) if one approaches z going on a circle around 0. A function is said to be regular, if it is one-valued (Titchmarsh, 1964, p. 142–143). By our modern convention, each function is one-valued and therefore regular; a function like $\sqrt{z}$ is either not a function in our sense or its domain has to be adapted (for example by cutting out the negative real axis.)

The inner product on $\mathcal{F}^2(\mathbb{C}^d)$ is

$$\langle F, G\rangle_{\mathcal{F}} = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{C}^d} F(z)\overline{G(z)} e^{-|z|^2/2}\, dz.$$

The Bargmann transform $B : L^2(\mathbb{R}^d) \to \mathcal{F}^2(\mathbb{C}^d)$ is an isometry and its range is dense: it is a unitary operator. For a discussion see Gröchenig (2001, Chap. 3.4)[3]. The Bargmann transform has two properties which are very useful for our purposes. Let $\hat{f}$ denote the Fourier transform

$$\boxed{\hat{f}(\xi) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(x) e^{-ix\cdot\xi}\, dx} \tag{6.14}$$

(keep in mind that this expression is only applicable without problems for $f \in L^1(\mathbb{R}^d)$, but it extends to an unitary operator on $L^2(\mathbb{R}^d)$ by approximation procedures), then

$$B\hat{f}(z) = Bf(-iz)$$

for all $z \in \mathbb{C}^d$ which can easily be verified:

$$\begin{aligned}
B\hat{f}(z) &= \frac{e^{-z^2/4}}{2^{d/4}\pi^{d/2}} \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(t) e^{itx} e^{xz} e^{-x^2/2}\, dt\, dx \\
&= \frac{e^{-z^2/4}}{2^{d/4}\pi^{d/2}} \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(t) e^{(z-it)x - x^2/2}\, dx\, dt \\
&= \frac{e^{-z^2/4}}{2^{d/4}\pi^{d/2}} \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(t) e^{b^2/2} \int_{\mathbb{R}^d} e^{-\frac{1}{2}(x-b)^2}\, dx\, dt \qquad \text{with } b := (z - it) \\
&= \frac{e^{-z^2/4}}{2^{d/4}\pi^{d/2}} \int_{\mathbb{R}^d} f(t) e^{z^2/2 - izt - t^2/2} dt = Bf(-iz)
\end{aligned}$$

The second useful property is: If we set temporarily

$$h_\nu(x) = \frac{1}{\sqrt{\nu!\, 2^{|\nu - 1/2|}}} H_\nu^{(\mathrm{phy})}(x) e^{-x^2/2},$$

with ν a multi-index, then we have

$$Bh_\nu(z) = \frac{z^\nu}{\sqrt{2^{|\nu|}\nu!}},$$

see Gröchenig (2001, Th. 3.4.2 and p. 57). h_ν is a normalized Hermite function. Such functions form an orthonormal basis of $L^2(\mathbb{R}^d)$, and – this is what we will use later – they can be used to calculate the coefficients $a_\nu = \langle f, H_\nu\rangle_{\mathcal{N}}$ in a Hermite expansion of a function f. Here, we have normalized with respect to the stretched Lebesgue measure $\lambda/(2\pi)^{d/2}$.

[3] Gröchenig gives a different definition of the Bargmann transform. This is due to his definition of the Fourier transform, because both are related in a certain way. See the remark on page 147 for details.

Technical remark. a) The cause for the stretching factor in the measure is the definition of the L^2 space in the paper of Vemuri (2008). Probably for aesthetical reasons, he defines the norm on $L^2(\mathbb{R}^d)$ as

$$\langle f, g\rangle = (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(x)g(x)\, dx.$$

We have seen above that Hermite functions that are normal with respect to this measure pass over to rather elegant monomials. We will follow this definition as long as we go along Vemuri's work. This is not consistent with the standard definitions used here, but probably less confusing as if we would struggle with additional roots and powers of π. One short example: The normalized Hermite functions in $L^2(\mathbb{R}^d, \lambda)$ are

$$h_\nu(x) = \frac{1}{\sqrt{\nu!\, 2^\nu \pi^{d/2}}}\, H_\nu^{(\mathrm{phy})}(x) e^{-x^2/2},$$

and we have

$$Bh_\nu(z) = \frac{z^\nu}{\sqrt{2^{|\nu|}\nu!(2\pi)^{d/2}}}.$$

Since we are going to deduce bounds for Hermite coefficients to get an idea of their rate of growth, we can ignore constants, so we can temporarily work with a compressed measure without any problems.

b) The formal expression of a Hermite function is not only affected by the particular measure on the space, but also on the way of defining the Fourier transform: It is per definition an eigenfunction of the Fourier transform, so there are as many ways of defining a Hermite function as there are ways of notating a Fourier transform. And if one wants to keep the property that Hermite functions are linked with orthonormal monomials in $\mathcal{F}^2(\mathbb{C}^d)$ via the unitary Bargmann transform – which is an adequate way of defining the transform –, the particular look of the Fourier transform has an impact on the formal expression of the Bargmann transform as well.

In Gröchenig (2001), the Fourier transform is – as often in signal processing – a function of the ordinary frequency ξ, given in hertz:

$$\hat{f}_1(\xi) = \int_{\mathbb{R}^d} f(x) e^{-2\pi i x\cdot\xi}\, dx$$

This definition and the one used here (which workes in real life with angular frequency in rad/s) are related by $\hat{f}(\omega) = (2\pi)^{-d/2} \hat{f}_1\left(\omega/(2\pi)\right)$.

Starting with this definition, the Bargmann transform is defined by

$$Bf(z) = 2^{d/4} \int_{\mathbb{R}^d} f(x) e^{2\pi x\cdot z - \pi x^2 - \frac{\pi}{2} z^2}\, dx$$

and the Bargmann-Fock space $\mathcal{F}^2(\mathbb{C}^d)$ is the Hilbert space of all entire functions F on $\mathbb{C}^d$ with inner product

$$\langle F, G\rangle_{\mathcal{F}} = \int_{\mathbb{C}^d} F(z)\overline{G(z)} e^{-\pi|z|^2}\, dz.$$

The Bargmann transform is then still a unitary operator $B : L^2(\mathbb{R}^d) \to \mathcal{F}^2(\mathbb{C}^d)$ (Gröchenig, 2001, Th. 3.4.3), and it keeps the property $B\hat{f}(z) = Bf(-iz)$. The normalized Hermite functions are in this setting

$$\psi_\nu(x) = \frac{1}{\sqrt{2^{|\nu-1/2|}\nu!}} H_\nu^{(\mathrm{phy})}(\sqrt{2\pi}x) e^{-\pi x^2},$$

with ν a multi-index, and they are still the counterpart of orthonormal monomials in $\mathcal{F}^2(\mathbb{C}^d)$ under the Bargmann transformation:

$$B\psi_\nu(z) = \left(\frac{\pi^{|\nu|}}{\nu!}\right)^{1/2} z^\nu$$

One can pass over from one notation to the other by a chain of isometries. A function $f(x)$ in our $L^2(\mathbb{R}^d)$ corresponds to a function $\tilde{f}(x) = u(f)(x) = f((2\pi)^{d/2}x)$ in Gröchenig's $L^2(\mathbb{R}^d)$ and a function

$\tilde{F}(z)$ in Gröchenig's $\mathcal{F}^2(\mathbb{C}^d)$ or $L^2(\mathbb{R}^d)$ corresponds to a function $F(x) = u^{-1}(\tilde{F})(x) = \tilde{F}((2\pi)^{-d/2}x)$ in our spaces.

$$\begin{array}{lccc} \text{Grö.'s situation} & L^2(\mathbb{R}^d,\lambda) & \xrightarrow{B} & \mathcal{F}^2(\mathbb{C}^d) \\ & \uparrow u & & \downarrow u^{-1} \\ \text{situation here} & L^2(\mathbb{R}^d, \frac{\lambda}{(2\pi)^{d/2}}) & \xrightarrow[B]{} & \mathcal{F}^2(\mathbb{C}^d) \end{array} \qquad \begin{array}{ccc} \tilde{f} & \longmapsto & 2^{d/4}\int_{\mathbb{R}^d} \tilde{f}(x) e^{2\pi x\cdot z - \pi x^2 - \frac{\pi}{2}z^2}dx \\ \uparrow & & \downarrow \\ f & \longmapsto & \frac{e^{-z^2/4}}{2^{d/4}\pi^{d/2}}\int_{\mathbb{R}^d} f(x) e^{xz} e^{-x^2/2}dx. \end{array}$$

The basis for our criterion for summability condition (6.4) will be the following generalization of the theorem of Vemuri (2008).

Lemma 6.6. *Let $a \in (0,1)$ and $C \in \mathbb{R}_+$ be constant and let $g_a(x) = e^{-ax^2/2}$ be a Gaussian function of $x \in \mathbb{R}^d$ with variance $1/\sqrt{a}$. If*

$$|f(x)| \le C g_a(x) \qquad \text{and as well} \qquad |\hat{f}(\xi)| \le C g_a(\xi), \tag{6.15}$$

then

$$\boxed{|\langle f, h_\nu\rangle| \le C\left(\frac{2\pi}{1+a}\right)^{d/2} \sqrt{\nu!}\left(\frac{e}{\nu}\right)^{\nu/2}\left(\frac{1-a}{1+a}\right)^{|\nu|/4}}$$

for any multi-index $\nu \in \mathbb{N}_+^d$.

We will formulate the proof in several single propositions. For reasons of simplicity we will consider only the two-dimensional case. During the proof it becomes clear what has to be done in higher dimensions.

Proposition 6.7. *Under the conditions of Lemma 6.6, the Bargmann transform of a function $f : \mathbb{R}^2 \to \mathbb{R}$ can be bounded as follows:*

$$\boxed{|Bf(z)| \le C\frac{2\pi}{1+a}\exp\left\{\frac{r_1^2\left(\mu + (1-\mu)\sin^2\vartheta_1\right) + r_2^2\left(\mu + (1-\mu)\sin^2\vartheta_2\right)}{4}\right\}} \tag{6.16}$$

and

$$\boxed{|Bf(z)| \le C\frac{2\pi}{1+a}\exp\left\{\frac{r_1^2\left(\mu + (1-\mu)\cos^2\vartheta_1\right) + r_2^2\left(\mu + (1-\mu)\cos^2\vartheta_2\right)}{4}\right\},} \tag{6.17}$$

where $\mu = (1-a)/(1+a) \in \mathbb{R}$.

Proof. We represent $z \in \mathbb{C}^2$ as

$$z = u + iv = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} + i\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} r_1 e^{i\vartheta_1} \\ r_2 e^{i\vartheta_2} \end{pmatrix}$$

with $u_j, v_j, r_j, \vartheta_j \in \mathbb{R}$ for $j = 1, 2$. By the first hypothesis in (6.15) and completing the square in the exponent we obtain

$$\begin{aligned}\left|\iint_{\mathbb{R}^2} e^{xz} e^{-x^2/2} f(x_1, x_2)\, dx_1 dx_2\right| &\le C \iint_{\mathbb{R}^2} \left|e^{xz - \frac{1}{2}(1+a)x^2}\right| dx \\ &= C e^{\frac{u^2}{2(1+a)}} \iint_{\mathbb{R}^2} e^{-\frac{1+a}{2}\left(x - \frac{u}{1+a}\right)^2} dx \\ &= C e^{\frac{u^2}{2(1+a)}} \frac{2\pi}{1+a}.\end{aligned}$$

Thus a simple bound for the Bargmann transformation is

$$\begin{aligned}|Bf(z)| &= \left|\frac{e^{-z^2/4}}{2^{1/2}\pi} \iint_{\mathbb{R}^2} e^{xz} e^{-x^2/2} f(x_1, x_2)\, dx_1 dx_2\right| \\ &\le C \frac{2\pi}{1+a} \exp\left\{\frac{u^2}{2(1+a)} + \frac{v^2 - u^2}{4}\right\} \\ &= C \frac{2\pi}{1+a} \exp\left\{\frac{v^2 + \mu u^2}{4}\right\}\end{aligned}$$

with $\mu = \frac{1-a}{1+a}$, respectively by writing it in real polar coordinates

$$\begin{aligned}|Bf(z)| &\le C \frac{2\pi}{1+a} \exp\left\{\frac{v_1^2 + v_2^2 + \mu\left(u_1^2 + u_2^2\right)}{4}\right\} \\ &= C \frac{2\pi}{1+a} \exp\left\{\frac{r_1^2\left(\mu + (1-\mu)\sin^2\vartheta_1\right) + r_2^2\left(\mu + (1-\mu)\sin^2\vartheta_2\right)}{4}\right\},\end{aligned}$$

which proves the first statement.

Now we make the same calculation for $B\hat{f}$. $f \in L^2(\mathbb{R}^2)$ implies $\hat{f} \in L^2(\mathbb{R}^2)$, and so $B\hat{f}$ is defined, and by the second hypothesis in (6.15), $\hat{f}$ has the same bound as f. We obtain

$$\begin{aligned}|Bf(z)| &= \left|B\hat{f}(iz)\right| \\ &\le C \frac{2\pi}{1+a} \exp\left\{\frac{r_1^2\left(\mu + (1-\mu)\sin^2(\vartheta_1 + \frac{\pi}{2})\right) + r_2^2\left(\mu + (1-\mu)\sin^2(\vartheta_2 + \frac{\pi}{2})\right)}{4}\right\} \\ &= C \frac{2\pi}{1+a} \exp\left\{\frac{r_1^2\left(\mu + (1-\mu)\cos^2\vartheta_1\right) + r_2^2\left(\mu + (1-\mu)\cos^2\vartheta_2\right)}{4}\right\}.\end{aligned}$$

□

We will now improve these estimates with the Phragmén-Lindelöf principle.

Proposition 6.8. *Under the conditions of Lemma 6.6, the upper bounds* (6.16) *and* (6.17) *for the Bargmann transform of a function $f : \mathbb{R}^2 \to \mathbb{R}$ can be refined to*

$$\boxed{|Bf(z)| \le C \frac{2\pi}{1+a} \exp\left\{\frac{\sqrt{\mu}}{4} r_1^2\right\} \exp\left\{\frac{\sqrt{\mu}}{4} r_2^2\right\}} \tag{6.18}$$

for z_1 and z_2 each in the first quadrant.

Proof. Note that Bf is a holomorphic function, since the Bargmann transform maps functions into the Bargmann-Fock space $\mathcal{F}^2(\mathbb{C}^2)$ of entire functions. Consider

$$F(z) = \exp\left\{i\frac{\sqrt{\mu}}{4}z_1^2\right\} Bf(z)$$

which is entire, too, because both factors are entire: Bf is entire and

$$\frac{\partial}{\partial \bar{z}} \exp\left\{i\frac{\sqrt{\mu}}{4}z_1^2\right\} = 0.$$

We fix the second variable and regard F as a function of one complex variable z_1. It stays entire. Now we show that F (with fixed second argument) is bounded by an exponential everywhere and by a constant on certain rays. The Phragmén-Lindelöf principle then guarantees that F is bounded by this constant on the whole cone between the rays.

$$\begin{aligned}\left|\exp\left\{i\frac{\sqrt{\mu}}{4}z_1^2\right\}\right| &= \exp\left\{-\frac{\sqrt{\mu}}{4}2u_1v_1\right\} \\ &= \exp\left\{-\frac{\sqrt{\mu}}{4}2r_1^2\cos\vartheta_1\sin\vartheta_1\right\} \le \exp\left\{\frac{\sqrt{\mu}}{4}r_1^2\right\}\end{aligned}$$

Therefore and by (6.16)

$$\begin{aligned}|F(z_1)| &\le C(z_2)\exp\left\{\frac{\sqrt{\mu}}{4}r_1^2\right\}\exp\left\{\frac{r_1^2\left(\mu+(1-\mu)\sin^2\vartheta_1\right)}{4}\right\} \\ &\le C(z_2)e^{r_1^2/4}e^{r_1^2/4} \\ &\le C(z_2)e^{|z_1|^2}\end{aligned}$$

with $C(z_2) = C\frac{2\pi}{1+a}\exp\left\{r_2^2\left(\mu+(1-\mu)\sin^2\vartheta_2\right)/4\right\}$. Keep in mind that we could have just as well $C(z_2) = C\frac{2\pi}{1+a}\exp\left\{r_2^2\left(\mu+(1-\mu)\cos^2\vartheta_2\right)/4\right\}$, if we had taken (6.17) for the estimate.

Let now

$$\vartheta_{c_1} = \frac{1}{2}\arctan\left(\frac{2\sqrt{\mu}}{1-\mu}\right) \quad \text{and} \quad \vartheta_{c_2} = \frac{\pi}{2} - \vartheta_{c_1}.$$

Observe that $\vartheta_{c_1} \in (0, \frac{\pi}{4})$ and $\vartheta_{c_2} - \vartheta_{c_1} < \frac{\pi}{2}$, so that the cone, which is delimited by the rays $\vartheta_1 = \vartheta_{c_1}$ and $\vartheta_1 = \vartheta_{c_2}$ lies in the first quadrant, see Figure 6.1. On these rays, $F(z_1)$ is bounded. First look at $\vartheta_1 = \vartheta_{c_1}$.

$$\left|\exp\left\{i\frac{\sqrt{\mu}}{4}z_1^2\right\}\right| = \exp\left\{-\frac{\sqrt{\mu}}{4}\Im(z_1^2)\right\} = \exp\left\{-\frac{\sqrt{\mu}}{4}r_1^2\sin(2\vartheta_{c_1})\right\}, \tag{6.19}$$

where $\Im(z_1^2)$ is the imaginary part of z_1^2. In addition, (6.16) and the trigonometric identities

$$\sin(\arctan x) = \frac{x}{\sqrt{1+x^2}}, \quad \cos(\arctan x) = \frac{1}{\sqrt{1+x^2}}, \quad \sin\frac{x}{2} = \pm\sqrt{\frac{1-\cos x}{2}}$$

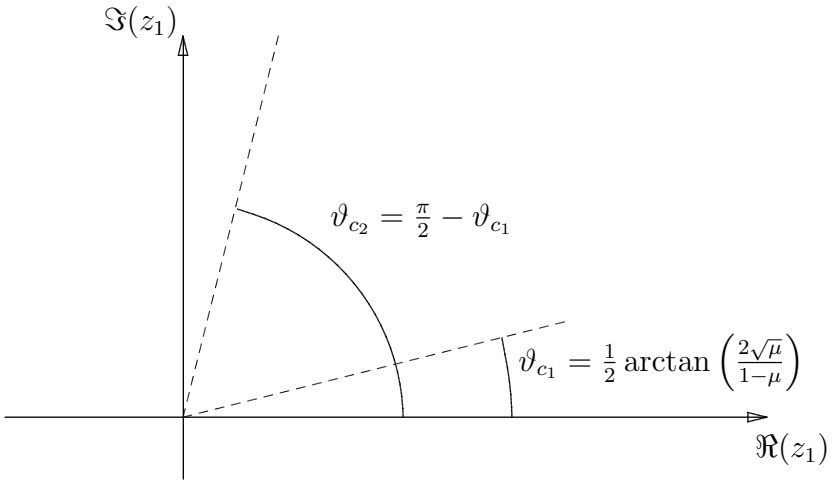


Figure 6.1: On the rays, $F(z_1)$ is bounded, and the Phragmén-Lindelöf principle extends the bound on the interior $\vartheta_{c_1} \le \vartheta_1 \le \vartheta_{c_2}$.

yield

$$|F(z_1)| \le C(z_2)\exp\left\{-\frac{\sqrt{\mu}}{4}r_1^2\sin(2\vartheta_{c_1})\right\}\exp\left\{\frac{r_1^2\left(\mu+(1-\mu)\sin^2\vartheta_{c_1}\right)}{4}\right\}$$
$$= C(z_2)\exp\left\{-\frac{\sqrt{\mu}}{4}r_1^2\frac{2\sqrt{\mu}}{1+\mu}\right\}\exp\left\{\frac{r_1^2}{4}\frac{2\mu}{1+\mu}\right\} = C(z_2).$$

Since $z_{c_1}^2 = (r_1 e^{i\vartheta_{c_1}})^2 = r_1^2 e^{i2\vartheta_{c_1}}$ and $z_{c_2}^2 = (r_1 e^{i\vartheta_{c_2}})^2 = r_1^2 e^{i(\pi-2\vartheta_{c_1})}$ have the same imaginary part and since $\cos(\frac{\pi}{2} - x) = \sin x$, we can easily deduce the same bound for $F(z_1)$ on the ray $\vartheta = \vartheta_{c_2}$ via (6.17).

Now it follows from the Phragmén-Lindelöf principle[4] that

$$|F(z_1)| \le C(z_2)$$

on the whole cone $\vartheta_{c_1} \le \vartheta_1 \le \vartheta_{c_2}$. So we obtain

$$|Bf(z)| \le C(z_2)\exp\left\{\frac{\sqrt{\mu}\sin 2\vartheta_1}{4}r_1^2\right\}$$

for $\vartheta_{c_1} \le \vartheta_1 \le \vartheta_{c_2}$. Trivially, we can bound this more roughly:

$$|Bf(z)| \le C(z_2)\exp\left\{\frac{\sqrt{\mu}}{4}r_1^2\right\} \tag{6.20}$$

We will now verify, that this estimate holds even for the whole frist quadrant. (6.16) as a function of z_1 only is monotone increasing in ϑ_1 in the first quadrant and therefore

[4]Let us qickly check the conditions, as we have stated them in Theorem 6.5: Here we have $b > 2$ and the bound $|F(z_1)| \le Ce^{r_1^2}$. So

$$\frac{|F(z_1)|}{e^{\delta r_1^b}} \le Ce^{r_1^2-\delta r_1^b} \le Ce^{r_1^2-\delta r_1^{2+\varepsilon}} = Ce^{r_1^2(1-\delta r_1^\varepsilon)}$$

for a certain $\varepsilon > 0$, and this goes to 0 as $r_1 \to \infty$ for any fixed $\delta > 0$. The convergence is uniform in the angle, because it only depends on the absolut value of z_1.

maximal on $[0, \vartheta_{c_1}]$ for $\vartheta_1 = \vartheta_{c_1}$. With the same trigonometric indentities as before and a simple convexity argument one can easily show that

$$\mu + (1-\mu)\sin^2\left(\frac{1}{2}\arctan\left(\frac{2\sqrt{\mu}}{1-\mu}\right)\right) = \frac{2\mu}{1+\mu} \leq \sqrt{\mu}$$

for $\mu \in (0,1)$. Analogously, (6.17) is monotone decreasing in ϑ_1 in the first quadrant and therefore maximal on $[\vartheta_{c_2}, \frac{\pi}{2}]$ for $\vartheta_1 = \vartheta_{c_2} = \frac{\pi}{2} - \vartheta_{c_1}$. With $\cos(\frac{\pi}{2} - x) = \sin x$, we obtain exactly the same upper bound as on the first cone $[0, \vartheta_{c_1}]$, so in the whole first quadrant of z_1, estimate (6.20) is valid.

By the previous calculation we have achieved the bounds

$$|Bf(z)| \leq C\frac{2\pi}{1+a}\exp\left\{\frac{\sqrt{\mu}}{4}r_1^2\right\}\exp\left\{\frac{r_2^2\left(\mu + (1-\mu)\sin^2\vartheta_2\right)}{4}\right\} \tag{6.21}$$

and

$$|Bf(z)| \leq C\frac{2\pi}{1+a}\exp\left\{\frac{\sqrt{\mu}}{4}r_1^2\right\}\exp\left\{\frac{r_2^2\left(\mu + (1-\mu)\cos^2\vartheta_2\right)}{4}\right\}, \tag{6.22}$$

and now we take the second argument into account and fix the first argument (of which only the absolute value r_1 is left). We will repeat the calculation that we have done for z_1. For that purpose, consider

$$F(z) = \exp\left\{i\frac{\sqrt{\mu}}{4}z_2^2\right\}Bf(z).$$

This F as a function of z_2 is entire, too. Now use (6.21) and (6.22) instead of (6.16) and (6.17) to show

$$|F(z_2)| \leq C(z_1)e^{|z_2|^2}$$

with $C(z_1) = C\frac{2\pi}{1+a}\exp\left\{\sqrt{\mu}r_1^2/4\right\}$ and

$$|F(z_2)| \leq C(z_1)$$

on the rays $\vartheta_2 = \vartheta_{c_1}$ and $\vartheta_2 = \vartheta_{c_2}$, and so via Phragmén-Lindelöf on the domain between the rays. With the same arguments as before we extend the bound to the whole first quadrant of z_2, and so we obtain after all the bound

$$|Bf(z)| \leq C\frac{2\pi}{1+a}\exp\left\{\frac{\sqrt{\mu}}{4}r_1^2\right\}\exp\left\{\frac{\sqrt{\mu}}{4}r_2^2\right\}$$

for z_1 and z_2 each in the first quadrant. □

Proposition 6.9. *The upper bound* (6.18) *holds not only for the first quadrant, but everywhere.*

Proof. We have just established (6.18) for the first quadrant, starting from a cone with rays ϑ_{c_1}, $\vartheta_{c_2} = \pi/2 - \vartheta_{c_1}$. At first consider the third quadrant. We rotate the cone by an angle of π so that we have a domain between the rays $\vartheta_{c_5} = \pi + \vartheta_{c_1}$ and $\vartheta_{c_6} = \pi + \vartheta_{c_2}$. On these rays, z_1 has the same imaginary part as on the rays in the first quadrant, compare to (6.19):

$$\begin{aligned}\Im(z_1^2) &= r_1^2 \sin(2\vartheta_{c_5}) = r_1^2 \sin(2\pi + 2\vartheta_{c_1}) = r_1^2 \sin(2\vartheta_{c_1}) \\ \Im(z_1^2) &= r_1^2 \sin(2\vartheta_{c_6}) = r_1^2 \sin(3\pi - 2\vartheta_{c_1}) = r_1^2 \sin(2\vartheta_{c_1})\end{aligned}$$

And the functions $\sin^2\vartheta$ and $\cos^2\vartheta$ are π-periodic, so that we can do exactly the same calculation as before and obtain the bound (6.18) for the third quadrant.

For the second quadrant, we rotate the cone in the first quadrant by an angle of $\pi/2$ and operate on the rays $\vartheta_{c_3} = \frac{\pi}{2} + \vartheta_{c_1}$ and $\vartheta_{c_4} = \frac{\pi}{2} + \vartheta_{c_2}$. Now the imaginary party of z_1 has changed its sign:

$$\begin{aligned}\Im(z_1^2) &= r_1^2 \sin(2\vartheta_{c_3}) = r_1^2 \sin(\pi + 2\vartheta_{c_1}) = -r_1^2 \sin(2\vartheta_{c_1}) \\ \Im(z_1^2) &= r_1^2 \sin(2\vartheta_{c_4}) = r_1^2 \sin(2\pi - 2\vartheta_{c_1}) = -r_1^2 \sin(2\vartheta_{c_1})\end{aligned}$$

Therefore we make use of a slightly different entire function F, namely

$$F(z) = \exp\left\{-i\frac{\sqrt{\mu}}{4} z_1^2\right\} Bf(z).$$

Then we can deduce the same exponential growth bound for this F as for the former version, and since rotation by $\pi/2$ only swaps sine and cosine, the bounds (6.16) and (6.17) are still available and only have to be applied vice versa.

Finally, the fourth qudadrant can be put down to this last case by rotating it by an angle of π, since $\sin^2\vartheta$ and $\cos^2\vartheta$ are π-periodic and again $\Im(z_1^2) = -r_1^2 \sin(2\vartheta_{c_1})$ on the rays $\vartheta_{c_7} = \pi + \frac{\pi}{2} + \vartheta_{c_1}$ and $\vartheta_{c_8} = \pi + \frac{\pi}{2} + \vartheta_{c_2}$. So (6.18) holds on whole $\mathbb{C}^2$. □

Finally, we turn to the

Proof of Lemma 6.6. The statement to proove is

$$|\langle f, h_k h_l\rangle| \leq C \frac{2\pi}{1+a} \sqrt{k!\, l!} \left(\frac{e}{k}\right)^{k/2} \left(\frac{e}{l}\right)^{l/2} \left(\frac{1-a}{1+a}\right)^{(k+l)/4}$$

for all $k, l \in \mathbb{N}_+$, if

$$|f(x_1, x_2)| \leq C g_a(x_1, x_2) \qquad \text{and} \qquad |\hat{f}(\xi_1, \xi_2)| \leq C g_a(\xi_1, \xi_2).$$

Consider the Bargmann transform Bf of f. By Proposition 6.9,

$$|Bf(z)| \leq C \frac{2\pi}{1+a} \exp\left\{\frac{\sqrt{\mu}}{4} r_1^2\right\} \exp\left\{\frac{\sqrt{\mu}}{4} r_2^2\right\},$$

and since Bf is analytic (it is an element of the Bargmann-Fock space of entire functions), we can write it as a power series

$$Bf(z) = \sum_{m,n=0}^{\infty} c_{m,n} z_1^m z_2^n$$

with Taylor coefficients

$$c_{mn} = \frac{\partial^{m+n}}{\partial z_1^m \partial z_2^n} \frac{Bf(z)}{m!n!}.$$

In complex analysis there is a famous estimate for derivatives, Cauchy's inequality[5]. Applying this to the polydisk $B(0,r_1) \times B(0,r_2)$ with arbitrary $r_1, r_2 > 0$ and with (6.18), we obtain

$$\frac{\partial^{m+n} Bf(z)}{(\partial z_1)^m (\partial z_2)^n} \leq \frac{m!n!}{r_1^m r_2^n} C \frac{2\pi}{1+a} \exp\left\{\frac{\sqrt{\mu}}{4} r_1^2\right\} \exp\left\{\frac{\sqrt{\mu}}{4} r_2^2\right\}$$

and so

$$|c_{m,n}| \leq C \frac{2\pi}{1+a} \exp\left\{\frac{\sqrt{\mu}}{4} r_1^2\right\} \exp\left\{\frac{\sqrt{\mu}}{4} r_2^2\right\} r_1^{-m} r_2^{-n}.$$

This bound is valid for any $r_1, r_2 > 0$, so we can choose such r_1, r_2 that it becomes minimal. Note that the function $g_{k,\mu}(r) := \exp(\sqrt{\mu}/4 \cdot r^2) r^{-k}$ has derivative $g'_{k,\mu}(r) = \exp(\sqrt{\mu}/4 \cdot r^2) r^{-k-1} (\sqrt{\mu}/2 \cdot r^2 - k)$, and this has a null in $r > 0$ for $\sqrt{\mu}/2 \cdot r^2 - k = 0$. Thus $|c_{mn}|$ gets minimal for $r_1 = \sqrt{2m/\sqrt{\mu}}$, $r_2 = \sqrt{2n/\sqrt{\mu}}$, and this is

$$|c_{m,n}| \leq C \frac{2\pi}{1+a} \left(\frac{e\sqrt{\mu}}{2m}\right)^{m/2} \left(\frac{e\sqrt{\mu}}{2n}\right)^{n/2}. \tag{6.23}$$

Now the stage is set to bound the Hermite coefficients of f. Remember that B is an isometry and that we temporarily work with the measure $dz/(2\pi)$.

$$\begin{aligned} |\langle f, h_{kl} \rangle| &= |\langle Bf, Bh_{kl} \rangle| \\ &= \left| \int_{\mathbb{C}^2} \left(\sum_{m,n=0}^{\infty} c_{m,n} z_1^m z_2^n \right) \overline{\left(\frac{z_1^k}{\sqrt{2^k k!}} \frac{z_2^l}{\sqrt{2^l l!}} \right)} \frac{e^{-r_1^2/2} e^{-r_2^2/2}}{4\pi^2} \, dz \right| \end{aligned}$$

[5]Let $D \in \mathbb{C}^d$ be a domain, $f : D \to \mathbb{C}$ analytic, $z_0 \in D$ a point and $P^d(z_0, r) \subset\subset D$ a polydisk with distinguished boundary T. Then

$$|D^\nu f(z_0)| \leq \frac{\nu!}{r^\nu} \sup_T |f|.$$

Here, ν is a multi-index, $U \subset\subset V$ means that U lies relatively compact in V (the closure of U is compact in V), a polydisk is the Cartesian product of d usual (complex-)one-dimensional disks and the distinguished boundary the Cartesian product of their boundary, i.e. of d circles (Fritzsche and Grauert, 2002, Chap. 1.4.)

In $\mathcal{F}^2(\mathbb{C}^d)$, monomials form an orthogonal family[6]; in the one-dimensional case we have for instance $z = re^{i\vartheta}$ and

$$\begin{aligned}\int_{\mathbb{C}} z^n \overline{z^k} e^{-r^2/2}\, dz &= \int_{\mathbb{C}} r^{n+k} e^{i\vartheta(n-k)} e^{-r^2/2}\, dz \\ &= \int_0^\infty r^{n+k+1} e^{-r^2/2} \int_0^{2\pi} e^{i\vartheta(n-k)}\, d\vartheta\, dr = 0\end{aligned}$$

since $\int_0^{2\pi} e^{i\vartheta(n-k)}\, d\vartheta = 0$ for $n \neq k$. So only one term remains of the sum:

$$\begin{aligned}|\langle f, h_{kl}\rangle| &= \left| \frac{c_{kl}}{\sqrt{2^{k+l} k!\, l!}} \int_{\mathbb{C}^2} r_1^{2k} r_2^{2l} \frac{e^{-r_1^2/2} e^{-r_2^2/2}}{4\pi^2}\, dz \right| \\ &= \frac{|c_{kl}|}{\sqrt{2^{k+l} k!\, l!}} \int_0^\infty \int_0^\infty r_1^{2k+1} r_2^{2l+1} e^{-r_1^2/2} e^{-r_2^2/2}\, dr_1\, dr_2 \\ &= |c_{kl}| \sqrt{2^{k+l} k!\, l!} \\ &\leq C \frac{2\pi \sqrt{k!\, l!}}{1+a} \left(\frac{e}{k}\right)^{k/2} \left(\frac{e}{l}\right)^{l/2} \left(\frac{1-a}{1+a}\right)^{(k+l)/4}\end{aligned}$$

For the last step, we have used (6.23), for the step before $\int_0^\infty r^{2k+1} e^{-r^2/2}\, dr = 2^k k!$ for all $k \in \mathbb{N}$. □

After this groundwork, we can now give a criterion for the summability condition (6.4).

Lemma 6.10. *Let the function $h : \mathbb{R}^2 \to \mathbb{R}$ fulfill the smoothness and growing conditions*

$$|h(\sqrt{2}x, \sqrt{2}y) e^{-(x^2+y^2)/2}| \leq C g_a(x, y) \tag{6.24}$$

and

$$|\mathcal{F}\left(h(\sqrt{2}x, \sqrt{2}y) e^{-(x^2+y^2)/2}\right)(\xi_1, \xi_2)| \leq C g_a(\xi_1, \xi_2) \tag{6.25}$$

where $a \in (0, 1)$ and $C \in \mathbb{R}_+$ are constant, $g_a(x, y) = e^{-a(x^2+y^2)/2}$ is a Gaussian function with variance $1/\sqrt{a}$ and $\mathcal{F}(f)$ denotes the Fourier transform of f, see (6.14)*. Then* (6.4) *is fulfilled.*

[6]Even more holds: They form an orthogonal basis, see Gröchenig (2001), Theorem 3.4.2. But this should not be surprising: The normalized Hermite polynomials are an orthonormal basis for $L^2(\mathbb{R}^d)$, and the Bargmann transform, which translates them into monomials in $\mathcal{F}^2(\mathbb{C}^d)$, is an unitary mapping. Gröchenig uses this relation just the other way round: He proves that the monomials form an orthonormal basis for $\mathcal{F}^2(\mathbb{C}^d)$ to conclude that the normalized Hermite polynomials are an orthonormal basis of L^2 (incidentally, without giving an explicit expression for them).

Proof. In order to keep track of how the above theory applies here, we will keep the d-dimensional multi-index notation. We can write Hermite coefficients in the following way:

$$\begin{aligned}
a_\nu &= \langle h, H_\nu \rangle_{\mathcal{N}} \\
&= (2\pi)^{-d/2} \int_{\mathbb{R}^d} h(x) H_\nu(x) e^{-x^2/2}\, dx \\
&= (2\pi)^{-d/2} \int_{\mathbb{R}^d} h(x) H_\nu^{(phy)}(x/\sqrt{2}) 2^{-|\nu|/2} e^{-x^2/2}\, dx \\
&= \frac{1}{\sqrt{2^{|\nu|}\,\nu!\,\pi^{d/2}}} \frac{\sqrt{\nu!\,\pi^{d/2}}}{\pi^{d/2}} \int_{\mathbb{R}^d} h(\sqrt{2}x) H_\nu^{(phy)}(x) e^{-x^2}\, dx \\
&= \frac{\sqrt{\nu!}}{\pi^{d/4}} \int_{\mathbb{R}^d} h(\sqrt{2}x) e^{-x^2/2} h_\nu(x)\, dx \\
&= \frac{\sqrt{\nu!}}{\pi^{d/4}}\, \langle h(\sqrt{2}x) e^{-x^2/2}, h_\nu(x) \rangle_\lambda
\end{aligned}$$

Now we apply Lemma 6.6 with $f(x) = h(\sqrt{2}x)e^{-x^2/2}$ and obtain

$$|a_\nu| \le C\nu! \left(\frac{e}{\nu}\right)^{\nu/2} \mu^{|\nu|/4}$$

with $\mu = (1-a)/(1+a) \in (0,1)$ and $C \in \mathbb{R}_+$ a constant. Thus we have

$$\sum_{k,l} \frac{|a_{kl}|}{\sqrt{k!\,l!}} \le C \left(\sum_k \sqrt{k!} \left(\frac{e}{k}\right)^{k/2} \mu^{k/4} \right) \left(\sum_l \sqrt{l!} \left(\frac{e}{l}\right)^{l/2} \mu^{l/4} \right),$$

and we make a ratio test to see that the sums converge:

$$\begin{aligned}
&\sqrt{\frac{(k+1)!}{k!}} \left(\frac{e}{k+1}\right)^{(k+1)/2} \left(\frac{k}{e}\right)^{k/2} \mu^{(k+1)/4} \mu^{-k/4} \\
&\qquad = \sqrt{e}\mu^{1/4} \left(1 + \frac{1}{k}\right)^{-k/2} \to \mu^{1/4},
\end{aligned}$$

and so

$$\sum_{k,l} \frac{|a_{kl}|}{\sqrt{k!\,l!}} < \infty.$$

□

Example. Lemma 6.10 enables us to handle for instance a Gaussian function like $g(x,y) = e^{-(x^2+y^2)/2}$, because it has itself as Fourier transformation, to be more precisely:

$$\begin{aligned}
|g(\sqrt{2}x, \sqrt{2}y) e^{-(x^2+y^2)/2}| &= e^{-\frac{3}{2}(x^2+y^2)} \le C g_{\frac{1}{3}}(x,y) \\
|\mathcal{F}\left(g(\sqrt{2}x, \sqrt{2}y) e^{-(x^2+y^2)/2}\right)(\xi_1,\xi_2)| &= \frac{1}{3} e^{-(\xi_1^2+\xi_2^2)/6} \le C g_{\frac{1}{3}}(\xi_1,\xi_2).
\end{aligned}$$

The question is if this is "easier to verify" as promised. But the answer is positive: Fourier transforms can be determined somehow in many cases while a general expression for Hermite coefficients (to check the summability condition directly) may be hard to achieve.

Chapter 7

Solutions for estimation problems

In Chapter 3, we have developed a non-parametric change-point test for changes in the mean of certain LRD processes, which was based on the Wilcoxon two-sample test, and compared it to a test based on the differences of means. Along the way, some important questions arose with which we will deal now.

We still consider a stochastic process $(X_i)_{i\geq 1}$ which is an instantaneous functional of a stationary and LRD Gaussian process:

$$X_i = G(\xi_i), \quad i \geq 1$$

where $(\xi_i)_{i\geq 1}$ is a stationary mean zero Gaussian process with $E[\xi_i^2] = 1$ and autocovariance function (1.1) and $G \in \mathcal{G}^1$ or $G \in \mathcal{G}^2$. In other words, we consider the situation under the null hypothesis (3.1) where there is no change in the mean, and without loss of generality we can assume that the common mean of the observations is 0. For strictly monotone G and if $(\xi_i)_{i\geq 1}$ is fGn, the tests from Chapter 3 are as follows: For the "Wilcoxon-type" test, reject the null hypothesis of no change if

$$W_n = \frac{1}{n\, d_n} \max_{1\leq k\leq n-1} \left| \sum_{i=1}^{k} \sum_{j=k+1}^{n} \left(I_{\{X_i \leq X_j\}} - \frac{1}{2} \right) \right|$$

is large, i.e. if it is greater or equal the upper 5%-quantile $q_{0.05}$ of its asymptotic distribution $(2\sqrt{\pi})^{-1} \sup_{0\leq\lambda\leq 1} |Z_1(\lambda) - \lambda Z_1(1)|$; for the "difference-of-means" test, reject the null hypothesis if

$$D_n = \frac{1}{n\, d_n} \max_{1\leq k\leq n-1} \left| \sum_{i=1}^{k} \sum_{j=k+1}^{n} (X_i - X_j) \right|$$

is large, i.e. if it is greater or equal the upper 5%-quantile $q_{0.05}$ of its asymptotic distribution $|a_1| \sup_{0\leq\lambda\leq 1} |Z_1(\lambda) - \lambda Z_1(1)|$, where $|a_1|$ is the first Hermite coefficient of G.

If one wants to apply these tests in real situations, one must face two essential road-blocks:

- First, the above described tests depend heavily on the LRD parameter D of the data, respectively the Hurst parameter $H = 1 - D/2$, which is not known in practice and must be estimated. In Section 7.1, we compare in a simulation study different estimators for H and investigate how they influence the above tests. In doing so, we confirm what has been observed by different authors in the last decade: that change-points lead to false estimates of H and thus adulterate also change-point tests itself.
  In Section 7.2, we propose new methods to estimate H even under a jump in the mean. We will show in a further simulation study that these new methods considerably improve the estimation of H if there is a change in the mean, and moreover that, if there is no change in the mean, our methods do not affect the usual estimation.

- Second, we demonstrated in the Corollary to Theorem 3.5 that if G is strictly monotone, it has always Hermite rank 1; thus in order to apply the "difference-of-means" test, one only needs to know the first Hermite coeficient a_1 of G. But in real situations, one does not know the function G which generated the data from fGn in the model. In Section 7.3, we will propose an estimator for a_1.

7.1 The influence of an estimated Hurst parameter

In the simulation study in Chapter 3, where we compared the performance of the "difference-of-means" test and the "Wilcoxon-type" test, we assumed the LRD parameter of the data to be known in order to assess the practicability of the test procedures; but in real applications of course, one has to estimate it from the data. Since such an estimation is error prone and may heavily influence the test statistic, it is interesting to analyse the performance of the change-point test in such a situation. Thus, we will now repeat the simulations, but this time with estimated LRD parameters. (To tell the truth, one actually would have to estimate even more: In practice, also L is unknown. But to begin with, estimating L seems to be hardly possible, second, the influence of H is more important, so we will concentrate on estimating the Hurst parameter H here.)

In what follows, we will repeat parts of the simulation study in Chapter 3, but we will estimate $nd_n = n^{2-D/2}$ by $n^{2-\hat{D}/2}$, where $\hat{D}$ is an appropriate estimator for the LRD parameter D (respectively $H = 1 - D/2$) of the data. (Note that c_1 and $L(n)$ cancel each other out if $(\xi_i)_{i\geq 1}$ is fGn.) First of all, we will choose such an estimator by a comparing simulation study.

7.1.1 Methods of estimating in comparison

The competing estimators

Taqqu, Teverovsky and Willinger (1995) study various techniques for estimating LRD; they list nine different methods and analyse their performance in a simulation study

based on 50 fGn and FARIMA time series with each 10,000 observations; Taqqu and Teverovsky (1998) take this up under different assumptions. We have chosen five of these nine estimators to investigate in a more realistic finite sample setting; they all are based on different techniques, and among similar estimators they exhibit the least mean squared error in the study by Taqqu, Teverovsky and Willinger (1995):

- Absolute values of the aggregated series
  The original time series is divided into blocks of size M in which the observations are avagared, so one considers the variables

$$X_k^{(M)} = \frac{1}{M} \sum_{i=(k-1)M+1}^{kM} X_i, \quad k = 1, 2, \ldots$$

  and calculates the mean of their absolute values, $\frac{1}{n/M} \sum_{k=1}^{n/M} |X_k^{(M)}|$. If the original data $(X_i)_{i \geq 1}$ have LRD parameter H and the logarithm of this statistic is plotted versus $\log M$, the result should be a line with slope $H - 1$.

- Periodogram method
  The periodogram

$$I(\lambda) = \frac{2}{2\pi n} \left| \sum_{k=1}^{n} X_k e^{ik\lambda} \right|^2$$

  is an estimator for the spectral density of the variables, which should be proportional to $|\lambda|^{1-2H}$ at the origin. A regression of $\log I(\lambda)$ on $\log \lambda$ should give a coefficient of $1 - 2H$.

- Boxed periodogram method
  The just described method of estimating H can be modified in order to compensate that in a log-log plot, most frequencies fall on the far right (which skews the regression); this is done by averaging the periodogram values over logarithmically equally spaced blocks.

- Peng's method / Residuals of regression
  The original series is split up into blocks of size M. Within each of these blocks, the partial sums Y_i, $i = 1, \ldots, M$ are calculated. Take the partial sums of the first block, fit a least-squares line to the Y_i and calculate the sample variances of the residuals. Repeat this with the other $(n/M) - 1$ blocks. Finally, take the average of all n/M sample variances. For large M, this value is proportional to M^{2H} for fGn.

- Whittle estimator
  The Whittle estimator is also based on the periodogram. It is the only estimator in this survey which does not use graphical methods, but which estimates H by minimizing a certain function. It is rather inflexible since it assumes that the parametric form of the spectral density is known.

For details on these methods, see Taqqu, Teverovsky and Willinger (1995) and the references therein.

Simulation results

We have simulated 10,000 repetitions of a time series $G(\xi_1), \ldots, G(\xi_n)$, where $(\xi_i)_{i\geq 1}$ is fGn with Hurst parameter H and G is a certain function, and we let both the sample size n and the Hurst parameter H vary: We simulated the time series for $n = 100, 500, 1000$ and for $H = 0.6, H = 0.9$ to cover low strong dependence and high strong dependence as well as time series of different realistic lengths.

In a first set of simulations, we chose G to be the identity $G(t) = t$ so that the observations are simple fGn. In a second set, we chose G to be

$$G(t) = \frac{1}{\sqrt{3/4}} \left((\Phi(t))^{-1/3} - \frac{3}{2} \right),$$

where Φ denotes the standard Gaussian c.d.f., so that the observations are Pareto(3,1) distributed. This corresponds to two extrem cases: Well-behaved Gaussian data on the one hand and heavy tailed data on the other hand.

In Figure 7.1, boxplots of the simulations are shown. For each estimation method (shortly referenced as `absval`, `boxper`, `peng`, `per` and `whittle`), we have three boxplots: for $n = 100$, $n = 500$ and $n = 1000$. We see that all estimators work better with increasing sample size and that the Whittle estimator and the Peng estimator seem to perform rather well.

In Figure 7.2, the square root of the mean squared error,

$$\sqrt{\frac{1}{k} \sum_{i=1}^{k} (\hat{H}_i - H)^2},$$

is given. Here, $\hat{H}_i$ denotes the estimated value of H in the i-th of the $k = 10,000$ simulations. Even though there is always an inherent bias and the estimators underestimate the true value if the data is not fGn, the Whittle estimator and the Peng estimator show the least MSE. Since the Peng estimator is very slow (which of course is no problem in real life application where only a few time series have to analysed, but which seriously extends the simulation time, due to the 10,000 repetitions), we decided to use the Whittle estimator in the following.

7.1.2 Change-point tests with estimated Hurst parameter

Since the estimation of H only changes the scaling in the tests of Chapter 3, and not the test procedures itself, we just took the simulation results and rescaled them by multiplying the 10,000 values of W_n and D_n for each set of data and parameters by $n^{2-D/2}/n^{2-\hat{D}/2}$. After that we counted the number of rejections, following the same rejection rules as were used there.

We see in Figure 7.1 that the Whittle estimator tends to underestimate the true value of H in our setting, thus we expect $\hat{H} < H$, respectively $\hat{D} > D$. As a consequence, $n^{2-D/2} > n^{2-\hat{D}/2}$; in other words, we rather scale down the test statistic too slowly if we estimate H. This is supported by Figure 7.3, where the new scaling $n^{2-D/2}/n^{2-\hat{D}/2}$ is shown under fGn and Pareto data, without jump: The rescaling is a bit higher than 1, which enlarges the data and will probably cause that the tests reject more often than they should. We will see that this is really the case.

In contrast, if there is a change-point in the time series, the Hurst parameter is easily overestimated (in Section 7.2, we propose methods to avoid this misjudging). Figure 7.4 shows the rescaling under different alternatives (a jump of height 0.5, 1 and 2 at different positions) and we clearly see that higher jumps cause a higher estimate. For fGn, where the Whittle estimator works well under the null hypothesis, this has the effect that data are scaled down when the jump gets big. For Pareto(3,1) distributed data, the overestimation caused by high jumps is compensated to some extent by the underestimation due to the non-normal distribution. So we expect the tests to detect jumps less often than in the situation when H is known. This foreboding is also supported by the following simulations.

Either way, note that the asymptotic level of the tests is probably not 5% any more, because we know neither the asymptotic distribution of $\hat{H}$ nor the asymptotic distribution of the test statistic when H is estimated by $\hat{H}$, so the critical value which we use is not adapted to the estimation of H.

Normally distributed data

As the Whittle estimator works well under Gaussian data, the influence of estimation errors is not big here.

In Figure 7.5, the relative frequency of false rejections under nearly $10,000$ simulation runs[1] is shown. Especially for very strong dependence and small sample sizes (i.e. $H = 0.9$ and $n = 50$ or $n = 100$) the influence of the Whittle estimator is visible: The test statistic is scaled down too slowly, so that it is bigger and exceeds the asymptotic critical value more often than usual.

Figure 7.6 shows the the relative frequency of true rejections[2]. Here we cannot spot any remarkable difference to the original simulation results where H was assumed to be known; with an estimated H, both tests detect jumps close to the borders of the time series a bit more often (i.e. after 5% and 10% of the data, respectively after 90% and 95%).

The exact simulation results are presented in Table D.20 and Table D.21 in Appendix D.

[1] For numerical reasons, sometimes H could not be estimated. In this case, the simulation run was skipped, so that we analyse a little less than $10,000$ repetitions here.

[2] Again under nearly $10,000$ simulation runs, because a few of the $10,000$ simulations were skipped.

Heavy-tailed data

For Pareto(3,1) distrbuted data, the situation is different: Here, the impact of underestimating the LRD parameter comes to light.

Figure 7.7 shows the level of the tests; obviously, it is far away from 5%, and it veers away as the sample size n increases. This can be explained by Figure 7.1: For small sample sizes, the estimation has a big variance and includes values near the true value of H; as n increases, $\hat{H}$ gets more and more concentrated below the true parameter.

Estimating H boosts as well the power of the tests. Figure 7.8 gives the observed power of the "difference-of-means" test and the "Wilcoxon-type" test, for sample size n = 500 and various positions and heights of the level shift. As in the original situation with a known H, the Wilcoxon-type test has larger power than the "difference-of-means" test for small level shifts h, but the "difference-of-means" test outperforms the "Wilcoxon type" test for larger level shifts.

The exact simulation results can be found in Table D.22 and Table D.23 in Appendix D.

7.1.3 Summary

We have seen that detecting change-points in LRD time series becomes difficult when the LRD parameter of the data is estimated, because the behaviour of the test statistic gets unpredictable and incalculable to some extent: When the LRD parameter is underestimated (which happens easily in Pareto distributed data, e.g.), the value of the test statistic is increased, while a break in a time series causes a too high estimate of H which entails that the test statistic is scaled down too heavily. So when there is no level shift, the tests falsely reject too often, while they detect existing breaks too rarely.

Actually, this even happens in the simple setting that the data are an instantaneous functional of fGn, where only the LRD parameter has to be estimated. In a more general situation, one would even have to estimate the auto-covariance function.

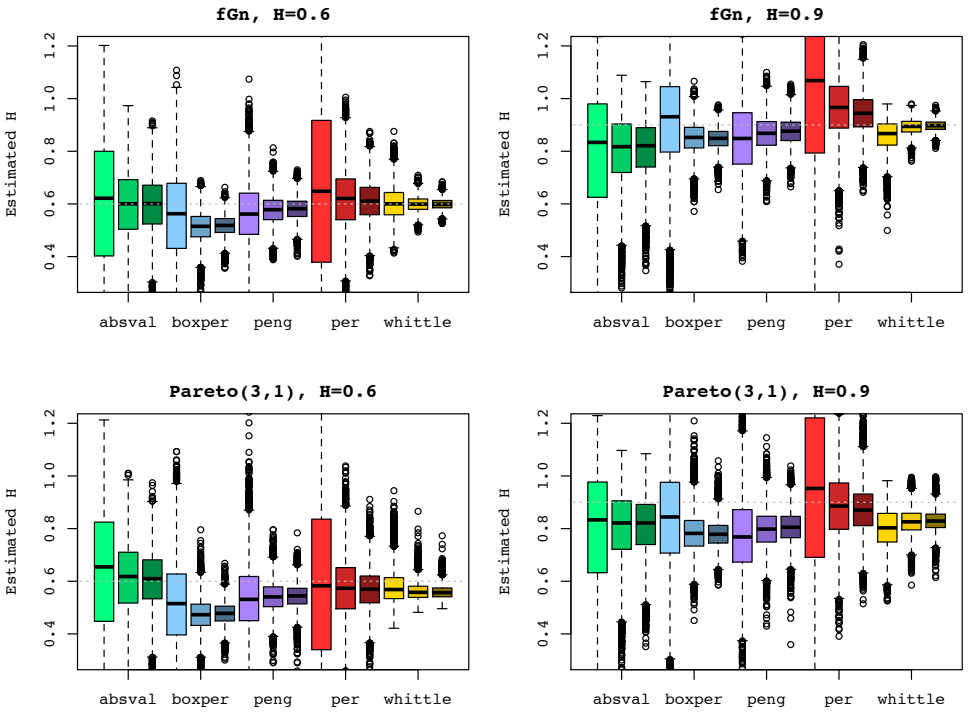


Figure 7.1: Boxplots of simulated Hurst parameters. Each method (`absval`, `boxper`, `peng`, `per`, `whittle`) was applied to time series of sample size $n = 100, 500, 1000$ (from left to right, from light to dark tones), each plot is based on $10,000$ estimations. The dotted line indicates the real value of H.

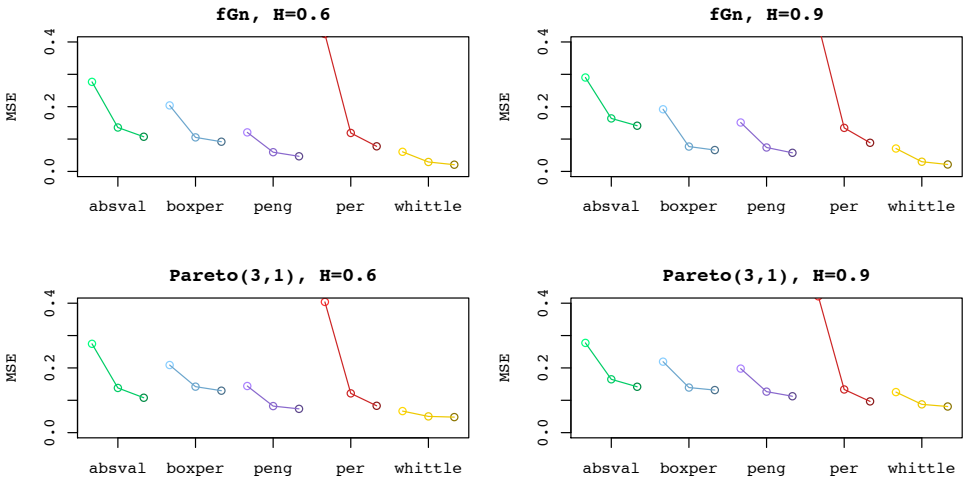


Figure 7.2: Square root of the MSE of different Hurst parameter estimates. Each method (`absval`, `boxper`, `peng`, `per`, `whittle`) was applied to time series of sample size $n = 100, 500, 1000$ (from left to right, from light to dark tones), each MSE is based on $10,000$ estimations.

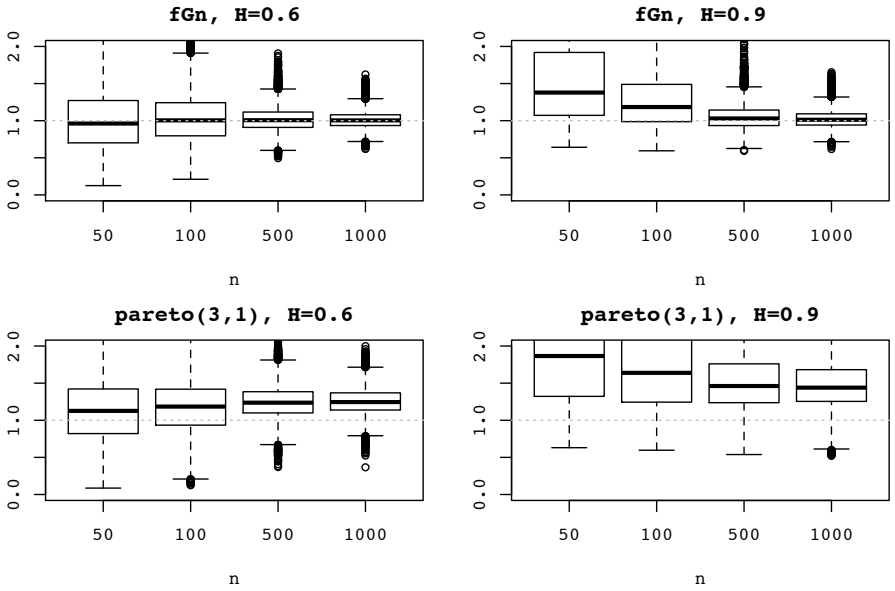


Figure 7.3: Boxplots of the rescaling $n^{2-D/2}/n^{2-\hat{D}/2}$, based on 10,000 samples of fGn and Pareto(3,1) distributed data, sample size $n = 50, 100, 500, 1000$, without break. The dotted line indicates the optimal value of 1.

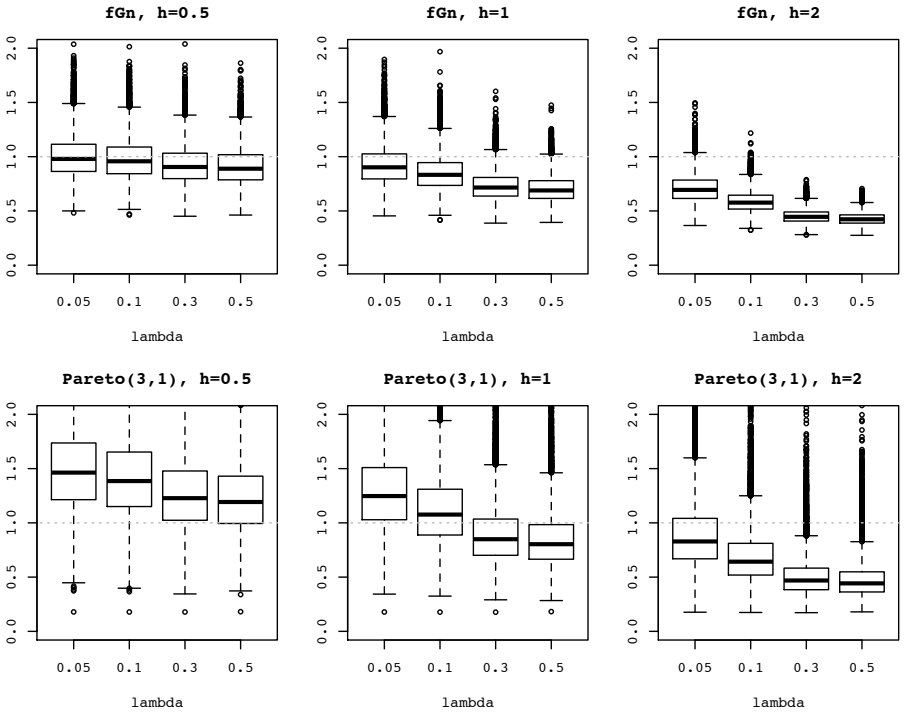


Figure 7.4: Boxplots of the rescaling $n^{2-D/2}/n^{2-\hat{D}/2}$, based on 10,000 samples of 500 observations of fGn and Pareto(3,1) distributed data with $H = 0.7$, with jumps of height h after the $[\lambda n]$-th observation. The dotted line indicates the optimal value of 1.

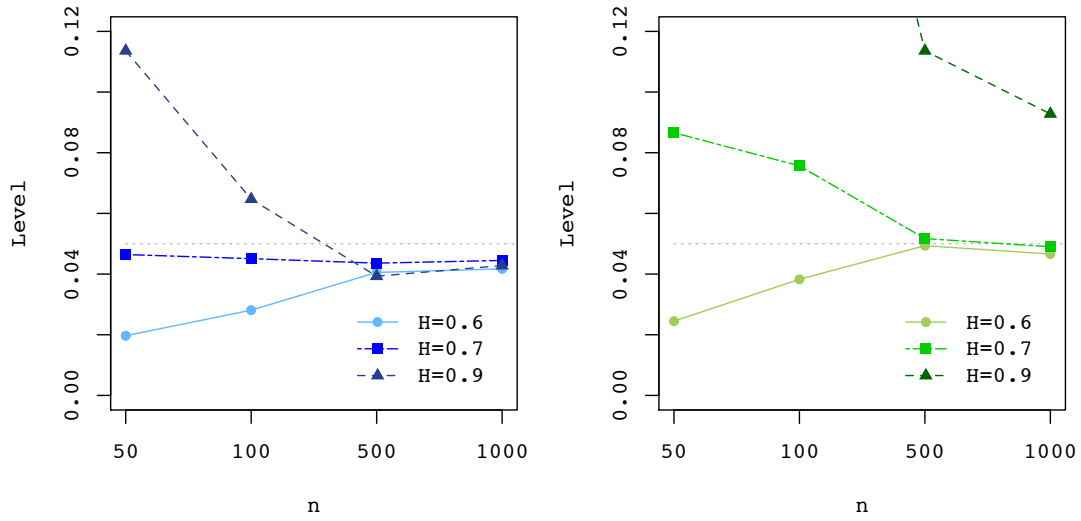


Figure 7.5: Level of "difference-of-means" test (left) and level of "Wilcoxon-type" test (right) for fGn time series with LRD parameter H, estimated by the Whittle estimator; 10,000 simulation runs.

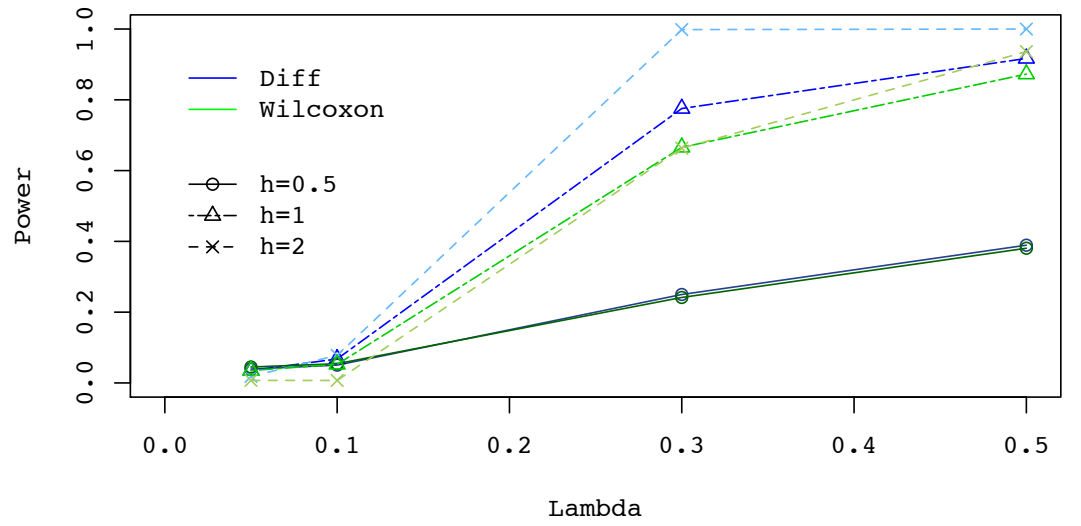


Figure 7.6: Power of "difference-of-means" test (left) and power of "Wilcoxon-type" test (right) for $n = 500$ observations of fGn with LRD parameter $H = 0.7$, estimated by the Whittle estimator; different break points $[\lambda n]$ and different level shifts h. The calculations are based on 10,000 simulation runs.

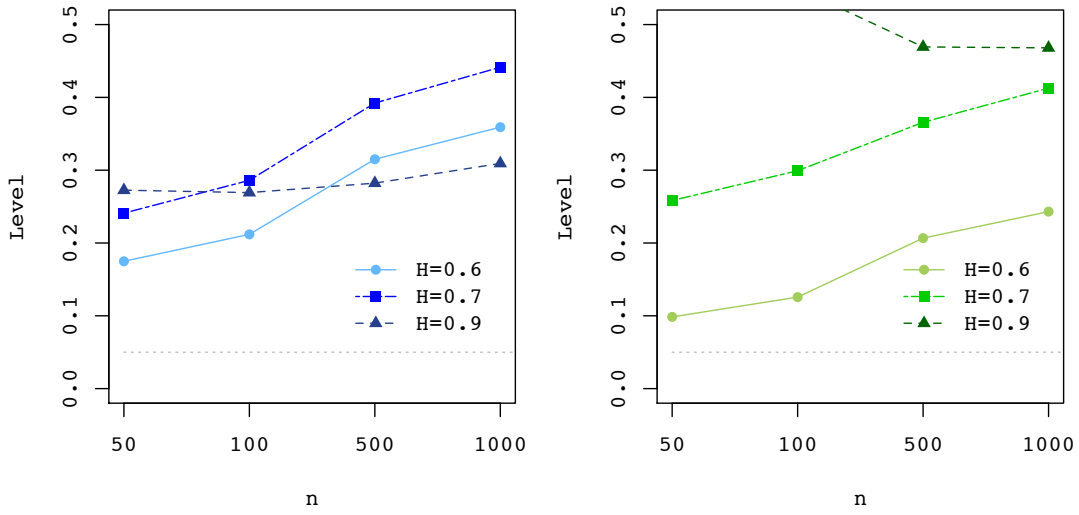


Figure 7.7: Level of "difference-of-means" test (left) and level of "Wilcoxon-type" test (right) for standardised Pareto(3,1)-transformed fGn with LRD parameter H, estimated by the Whittle estimator; 10,000 simulation runs.

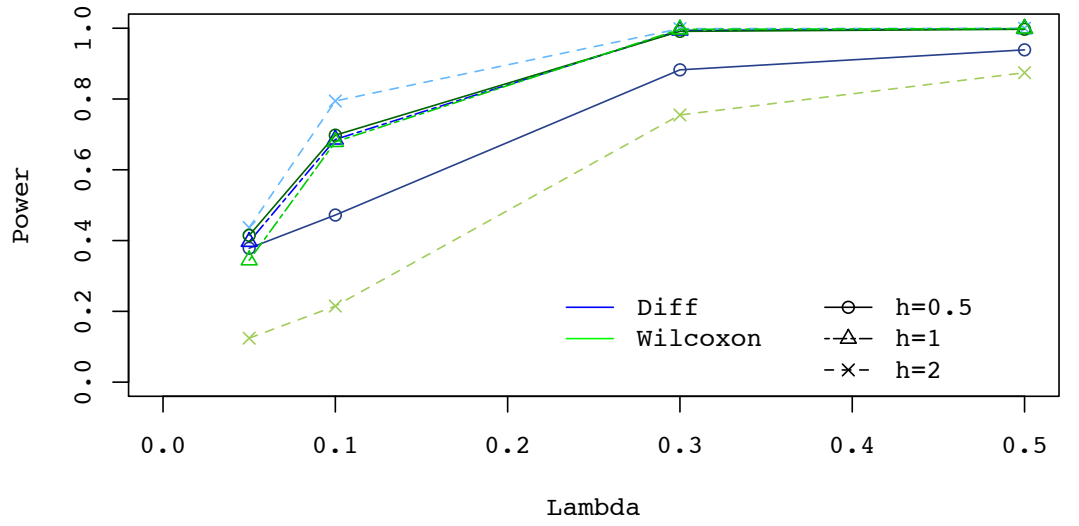


Figure 7.8: Power of "difference-of-means" test (left) and power of "Wilcoxon-type" test (right) for $n = 500$ observations of standardised Pareto(3,1)-transformed fGn with LRD parameter $H = 0.7$, estimated by the Whittle estimator; different break points $[\lambda n]$ and different level shifts h. The calculations are based on 10,000 simulation runs.

7.2 Estimating the LRD parameter under a change in the mean

There are several methods to estimate the LRD parameter. For an overview see the articles by Taqqu, Teverovsky and Willinger (1995) and Taqqu and Teverovsky (1998). But unfortunately, when one estimates H in order to apply a change-point test, one maybe has to do with a time series that includes a jump (why else should one test for it otherwise?), and such a jump intrinsically gets in the way of estimating the LRD parameter since the usual estimators do not consider a change-point and likely missspecify the resulting structure of the data as a more intense LRD. As a consequence, the power of a change-point test likely decreases, because the jump may cause an overestimation of H which may bring the test to interpret the jump as LRD behaviour. We have just seen this effect in Section 7.1. Sibbertsen and Willert (2010) report that tests for a break in persistence, i.e. a break in the long range-dependence structure, are neither robust against a shift in the mean and must be adapted in this case. For an overview about how structural breaks and trends lead to missspecification of LRD and methods how to distinguish both effects, see the survey of Sibbertsen (2004) who also points out that change-point estimators can not distinguish between LRD and break points. Krämer, Sibbertsen and Kleiber (2002) illustrate this with data from the German stock market, and Krämer and Sibbertsen (2002) show that in linear regression models disturbances that exhibit LRD can be confused with structural changes.

It is thus a challenging problem to distinguish between behaviour that originates from long memory and behaviour that originates from change-points. CUSUM-type tests which discriminate between both, an LRD time series and a short-range dependent time series with changes in the mean, have been proposed in the articles by Berkes et al. (2006) and Shao (2011) which both contain a lot of references concerning LRD and structural breaks. We focus on the situation where a LRD time series may possess a change-point. In this situation, especially for change-point tests which usually require the true, but unknown LRD parameter, there is a strong need for methods which estimate the LRD parameter without being confused by the possible mean shift. Our goal is thus to develop estimators for H which are on the one hand not heavily biased by a jump in the mean of the data, but which on the other hand still work well if there is no jump.

Künsch (1987) and Hsu (2005) proposed usage of a local estimator to achieve this. Hassler and Olivares (2007) applied such a method to German stock market data. Kuswanto (2009) also analysed German stock market returns and used the invariance of the LRD parameter under aggregation to develop an improved estimator of H by looking at combinations of paired of aggregated series.

We propose a broader approach. We will develop and analyse three general methods to estimate the LRD parameter in a time series which may include a jump that work with any usual estimation technique. These adapted methods will prove to work better

than the standard estimations which ignore jumps. All our methods have in common that they base on the initial insight that, if it includes a jump, the whole LRD time series is not appropriate for estimating the LRD parameter and that any estimation approach must thus at first segregate the jump or concentrate on local environments in the sample.

7.2.1 Adaption techniques

Seperating in two blocks

The first technique divides the time series into two blocks

$$X_1, \ldots, X_k \quad \text{and} \quad X_{k+1}, \ldots, X_n$$

and estimates H on each block seperately. This is done for all possible cutting points. Because the estimation of LRD intrinsically needs many observations, it is not sensible if one allows early or late cutting points which produce a small block on which the estimation of H is useless. We thus let the cutting point k take values in the set

$$K = \{k_{\text{low}}, k_{\text{low}} + 1, \ldots, k_{\text{up}}\}$$

with

$$\begin{aligned} k_{\text{low}} &= \max\{[n/10], 10\} + 1 \\ k_{\text{up}} &= n - k_{\text{low}} \end{aligned}$$

and to end up to $|K| = k_{\text{up}} - k_{\text{low}} + 1$ pairs of estimations of H: for each cutting point $k \in K$, we obtain an estimate $\hat{H}_k^{(1)}$ of H on the first block and an estimation $\hat{H}_k^{(2)}$ of H based on the second block.

As an estimate for the Hurst parameter of the whole sample, two functions of the $\hat{H}_k^{(1)}, \hat{H}_k^{(2)}, k \in K$ are suitable: At first,

$$\boxed{\hat{H}_{\text{mean}} = \frac{1}{|K|} \sum_{k \in K} \frac{\hat{H}_k^{(1)} + \hat{H}_k^{(2)}}{2}}$$

is the mean value of the estimate on the first and the second block, averaged over all cutting points. Then,

$$\boxed{\hat{H}_{\text{mindiff}} = \frac{\hat{H}_{k^*}^{(1)} + \hat{H}_{k^*}^{(2)}}{2}}$$

with

$$k^* = \arg\min_{k \in K} \left| \hat{H}_k^{(2)} - \hat{H}_k^{(1)} \right|$$

is the mean value of the estimate on the first and the second block, diveded by the special cutting point k^* where both estimates differ least, measured by the absolute difference.

Estimating on a moving window

The second technique estimates H only on a part of the observations, on a certain window

$$X_{m-w}, X_{m-w+1}, \ldots, X_m, \ldots, X_{m+w}$$

of length $2w+1$ around the center X_m. Now this window is moved through the time series $X_1, \ldots, X_n$: For each $m \in M = \{w+1, w+2, \ldots, n-w\}$, we estimate H on the window around the midpoint X_m, and in doing so, we obtain finally $|M| = n-2w$ estimates $\hat{H}_{w,m}$ for H.

We expect that for a fixed (and not too big) window size w, these estimates $\hat{H}_{w,m}$, $m \in M$, show a high variablilty because each estimation relies on a small window which may include the jump or wich cover a rather steady part of the observations (both resulting in too large estimation) or which may not include the jump or cover a rather fluctuating section of the observations (both resulting in too low estimations) – or a mixture of all these scenarios. Thus it seems reasonable to average all estimations and consider

$$\boxed{\hat{H}_{MV,w} = \frac{1}{|M|} \sum_{m \in M} \hat{H}_{w,m}}$$

as an estimate for the true Hust parameter H.

In the context of blocks and windows for statistical inference, the choice of the length of the block or window is traditionally an important issue. As flank length $w = w(n)$ as a function of the overall sample size n (which results in a window of size $2w+1$) we choose

$$w_1 = w_1(n) = \max\left\{[\sqrt{n}], 10\right\}$$

since the square root has often proven to be a good choice and since we do not want the window to be too small as the estimation of an LRD parameter inherently gets bad for a short sample of observations. For a comparison, we also analysed

$$w_2 = w_2(n) = \max\left\{[n/10], 10\right\}$$

and

$$w_3 = w_3(n) = \max\left\{[n/5], 10\right\}.$$

Whenever moving window techniques are employed, it is much discussed if the windows should be overlapping or non-overlapping. In our situation, the latter could have the advantage that only one window of data is affected with the jump which results in only one most likely erroneous estimate. So it is interesting (and a contribution to the above discussion) to analyse also a non-overlapping moving window approach. Here, H is estimated on blocks

$$X_1, \ldots, X_w, \qquad X_{w+1}, \ldots, X_{2w}, \qquad X_{2w+1}, \ldots, X_{3w}, \quad \ldots \quad X_{([N/w]-1)w+1}, \ldots, X_{[N/w]w}$$

of length w, resulting in $[N/w]$ estimates $\check{H}_{w,k}$, $k = 1, \ldots, [N/w]$, for H. Again it is reasonable to average and consider

$$\boxed{\hat{H}_{MVnl,w} = \frac{1}{[N/w]} \sum_{k=1}^{[N/w]} \check{H}_{w,k}}$$

as an estimate for the true Hust parameter H. As window sizes, it is reasonable to choose the same w_1, w_2, w_3 as above. Since this method does not yield satisfying results, for reasons of comparison, we only concentrated on window size w_1 in the following simulation study (see Section 7.2.2).

Pre-estimating the jump

Our third method is the most natural approach: Since the jump disturbes the estimation of H, we try to remove it. For this purpose, we estimate the position and the height of the jump, ignoring that we do not know the LRD parameter H yet, and then either eliminate the jump and estimate H on the whole (now) jumpfree time series or we estimate H on the observations before and on the observations after the jump and take the mean value as an estimate for H on the overall sample.

We get down to details: Given the observations $X_1, \ldots, X_n$, we apply a change-point test. In principle, we may take any change-point test we like. Here, we take the "Wilcoxon-type" change-point test from Chapter 3 which rejects the null hypothesis that there is no change in the mean for large values of the test statistic

$$\max_{1 \le k \le n-1} \left| \frac{1}{n\, d_n} \sum_{i=1}^{k} \sum_{j=k+1}^{n} \left(I_{\{X_i \le X_j\}} - \frac{1}{2} \right) \right|. \tag{7.1}$$

The change-point is supposed to take place after the observation X_{k^*} for the $k^* \in \{1, \ldots, n-1\}$, for which the test statistic (7.1) takes its maximum value. Since the scaling factor $(n\, d_n)^{-1} = n^{-2+D/2}(c_1 L(n))^{-1/2}$ does not depend on k, it does not influence k^*. Knowledge about the LRD parameter H is therefore only essential for a test decision *if* there is a change-point; the possible location k^* is unaffected. This allows us to pre-estimate the jump location without knowing H by

$$k^* = \operatorname*{arg\,max}_{1 \le k \le n-1} \left| \sum_{i=1}^{k} \sum_{j=k+1}^{n} \left(I_{\{X_i \le X_j\}} - \frac{1}{2} \right) \right|. \tag{7.2}$$

If there is a jump after $X_{[\tau n]}$ for a $\tau \in (0,1)$, k^* estimates its location $[\tau n]$; if there is no jump in the time series, k^* is some more or less meaningsless index in $\{1, \ldots, n\}$, but it will not affect the following estimation procedure.

Given the observations $X_1, \ldots, X_n$ and an estimate k^* for the jump location as described in (7.2), we divide the observation into two parts

$$X_1, \ldots, X_{k^*} \quad \text{and} \quad X_{k^*+1}, \ldots, X_n.$$

The jump height can be estimated by

$$\hat{h} = \frac{1}{n-k^*}(X_{k^*+1} + \ldots + X_n) - \frac{1}{k^*}(X_1 + \ldots + X_{k^*}),$$

and so we can remove the jump by considering the time series

$$X_1, \ldots, X_{k^*}, X_{k^*+1} - \hat{h}, \ldots, X_n - \hat{h}.$$

On this new time series which is regarded as jump-free, we can estimate H; we denote this estimate, which is based on pre-estimating and on one long time series, by

$$\boxed{\hat{H}_{\mathrm{pre},1} = \text{estimation of } H \text{ on sample } X_1, \ldots, X_{k^*}, X_{k^*+1} - \hat{h}, \ldots, X_n - \hat{h}.}$$

Alternatively, we can estimate H on both arising blocks seperately, which yields two estimates $\hat{H}^{(1)}_{k^*}$ and $\hat{H}^{(2)}_{k^*}$, and take their mean value as an estimate for H on the whole sample:

$$\boxed{\hat{H}_{\mathrm{pre},2} = \frac{\hat{H}^{(1)}_{k^*} + \hat{H}^{(2)}_{k^*}}{2}}$$

Remark. (i) Other change-point tests which can be used to pre-detect the jump (instead of the "Wilcoxon-type" test based on (7.1) and (7.2) above) may need the LRD parameter H – which is just what the method itself is aiming at. A way out of this vicious circle could be an iterative procedure: Starting with an estimate of H on the whole sample, one could apply the change-point test which yields a new estimate for H, which can then be used to execute the change-point test again, and so on. One can cherish hopes that such a procedure converges or yields, when stopped by a certain rule, an useful estimate for H.
(ii) Such an iterative procedure can not be used to refine the estimator $\hat{H}_{\mathrm{pre},1}$ although it seems to be reasonable: The removal of the jump is error prone since both the location and the height of the jump are estimated, so after removing the jump, one could again apply the change-point test, remove the resulting jump and estimate H on this new (double) jump-free sample, and so on; one could stop the iterations when the estimated value of H does not change significantly (for example when the difference between the new estimate and the estimate from the previous iteration is ≤ 0.01). But this does not lead to an useful estimate since the technique removes too much of the structure of the data and H is heavily underestimated.
(iii) An iterative procedure like that to refine $\hat{H}_{\mathrm{pre},2}$ is impossible: As described, the estimation of the jump location by k^* is not affected by the value of H – we would obtain the same estimate $\hat{H}_{\mathrm{pre},2}$ in any iteration step.

7.2.2 Simulations

We have simulated $n = 500$ realizations $\xi_1, \ldots, \xi_{500}$ of fGn with Hurst parameter $H = 0.7$ (i.e. $D = 0.6$), using the `fArma` package in `R`. We have choosen $G(t) = t$ and

obtained observations $X_1, \ldots, X_{500}$ of fGn with $H = 0.7$. To this time series, we have added a jump of height $h = 0.5, 1, 2$ after a proportion of $\lambda = 0.1, 0.5$ (i.e. after 50 and after 250 observations). To these resulting seven time series (one without jump, six with jump at different positions and of different heights) we have applied each of the above defined estimators, using exemplarily the Whittle Estimator (Section 7.2.2) and the Box-Periodogram Estimator (Section 7.2.2) to estimate the value of the Hurst parameter $H = 1 - D/2$, see Taqqu, Teverovsky and Willinger (1995) and Giraitis and Taqqu (1999). We have repeated these simulations $1,000$ times, so after all we obtained $1,000$ estimates $\tilde{H}_1 \ldots, \tilde{H}_{1000}$ for each estimator $\tilde{H}$ and each data situation.

Simulation results for the Whittle Estimator

In a first set of simulations, we have choosen the Whittle Estimator as generic estimation method for H. In Table 7.1 the relative difference between the average of the above described estimates $\tilde{H}_1 \ldots, \tilde{H}_{1000}$ and the true parameter is given for each situation, i.e. the value

$$\frac{\text{mean}\{\tilde{H}_1 \ldots, \tilde{H}_{1000}\}}{H} - 1$$

where H denotes the true parameter and $\tilde{H}$ the respective estimate. For example, for the usual estimator $\hat{H}$ (here: the Whittle estimator) and a jump of height 2 after 10% of the data ($h = 2$, $\lambda = 0.1$), we have tabulated 0.126 which means that in this situation the usual estimator overestimates the true value on average by 12.6% (and yields $\hat{H} = 0.788$ as average estimate for the true parameter $H = 0.7$).

At a glance, we see that there is no uniformly best estimator; instead we observe the following:

- For early or late jumps ($\lambda = 0.1$)
  - and for small jump heights, the estimators $\hat{H}_{\text{mindiff}}$, $\hat{H}_{\text{pre},1}$, $\hat{H}_{MV,w_2}$ and $\hat{H}_{MV,w_3}$ yield the best results, while
  - for high jumps, $\hat{H}_{MV,w_1}$ is the best.
- For jumps in the middle ($\lambda = 0.5$)
  - and for small jump heights, we obtain the best estimates with $\hat{H}_{\text{mean}}$, $\hat{H}_{\text{pre},2}$ and $\hat{H}_{MV,w_2}$, while
  - for high jumps, $\hat{H}_{\text{pre},2}$ and again $\hat{H}_{MV,w_1}$ are the best.
- The non-overlapping moving window technique is worse than the regular (overlapping) moving window technique. We conjecture that this arises from underestimating H on small samples without jump and overestimating H on samples with jump: In the non-overlapping technique, H is rather underestimated on most of the few small windows, while in the regular overlapping technique, this is compensated by rather overestimating H on all those many windows which include the jump.

		$\hat{H}$	$\hat{H}_{\text{mean}}$	$\hat{H}_{\text{mindiff}}$	$\hat{H}_{\text{pre,1}}$	$\hat{H}_{\text{pre,2}}$
no jump		-0.002	-0.006	-0.004	-0.022	-0.020
$\lambda = 0.1$	$h = 0.5$	0.010	0.005	0.003	-0.014	-0.012
	$h = 1$	0.042	0.031	0.020	0.000	0.006
	$h = 2$	0.126	0.091	0.068	0.031	0.051
$\lambda = 0.5$	$h = 0.5$	0.027	0.005	0.011	-0.017	-0.013
	$h = 1$	0.086	0.031	0.040	-0.013	-0.009
	$h = 2$	0.198	0.084	0.090	-0.010	-0.006

		$\hat{H}$	$\hat{H}_{MV,w_1}$	$\hat{H}_{MV,w_2}$	$\hat{H}_{MV,w_3}$	$\hat{H}_{MVnl,w_1}$
no jump		-0.002	-0.020	-0.010	-0.003	-0.040
$\lambda = 0.1$	$h = 0.5$	0.010	-0.017	-0.007	-0.001	-0.039
	$h = 1$	0.042	-0.011	0.001	0.005	-0.036
	$h = 2$	0.126	0.002	0.017	0.021	-0.030
$\lambda = 0.5$	$h = 0.5$	0.027	-0.017	-0.003	0.013	-0.038
	$h = 1$	0.086	-0.011	0.012	0.048	-0.035
	$h = 2$	0.198	0.001	0.043	0.125	-0.028

Table 7.1: Estimators for the Hurst parameter H in time series without ('no jump') and with change-point (jump of height h after a proportion of λ), relative difference between average estimate and the true parameter, each based on $1,000$ simulation runs with $n = 500$ realizations of fGn with $H = 0.7$, each based on the Whittle Estimator

- Our methods yield very reliable results even if there is no shift in the sample. While the usual estimator which is designed for exactly this situation yields a relative error of 0.2% on average, our methods with the best average relative error of 0.3% and the worst one of only 2.2% can easily compete with this.

In order to get an overview of the performance of the different estimators, we looked in Table 7.1 only at the average values. But estimating the Hurst parameter is always precarious when the sample of observations is not very large because LRD intrinsically manifests oneself on large scales, so the sucess of any estimation from a small sample depends more or less on fortunate circumstances – the sample has to be appropriate to reveal the dependence structure –, so as a consequence, the estimation from small samples is error-prone and is likely subject to fluctuations; thus, we seletced the most promising estimators from Table 7.1 for both scenarios (early/late jumps and jumps in the middle) and looked at their variation in our 1,000 simulation runs which is shown in Figure 7.9.

And indeed, we observe the following:

- The estimators, which all yield very good results on average, have all an interquartile range (IQR, the difference between the upper and the lower quartile

of a sample) of around 0.04, in other words: half of all estimates (of the true parameter $H = 0.7$) fluctuate between 0.66 and 0.74, which is pleasing.

- But the whiskers and outliers indicate that there are situations in which the estimators miss the true value by up to around 0.1 – or even worse –, in other words in which the estimation yields 0.6 or 0.8 instead of the true parameter value $H = 0.7$.

- Taking into account the variability of the estimator, the moving window estimator with the $\sqrt{n}$-window size, $\hat{H}_{MV,w_1}$, is uniformly the best in our survey: on average, it yields good estimates, and the variation is not only rather constant over jump heights and jump positions, but also one of the smallest among the analysed estimators.

Simulation results for the Box-Periodogram Estimator

In a second set of simulations, we have choosen the Box-Periodogram Estimator as underlying estimation method for H. In our simulation study, this estimator has proven to be biased: it tends to underestimate the true value of H in time series without jump; moreover, it has a bigger variance. In Table 7.2, again the relative difference between the average of the above described estimates $\tilde{H}_1 \ldots, \tilde{H}_{1000}$ and the true parameter is given for each situation.

We observe clear differences to the simulations with the Whittle Estimator:

- In the high number of underestimations, even in time series with jump, the bias of the underlying estimator becomes apparent.

- In combination with overestimating H when there is a jump in the mean, which even slightly occurs when using our new jump-adapted estimation procedures, this original bias causes that we sometimes obtain better results than with the Whittle Estimator. This occurs here and there, but most likely for high jumps ($h = 1, h = 2$) in the middle ($\lambda = 0.5$).

- Combined with the non-overlapping moving window approach $\hat{H}_{MVnl}$, the Box-Periodogramm Estimator reveals a remarkable behaviour: it yields drasticly worse results than the Box-Periodogramm Estimator without any modification. The reason may be that, the larger the sample is, the more changes the bias of the Box-Periodogramm Estimator from positive to negative as we have seen in our inital simulation study. In the non-overlapping moving window approach, H is estimated on only a few small windows, and on these, the Box-Periodogramm Estimator yields impressive overestimations.

Again, we selected the most promising estimators from Table 7.2 for both scenarios (early/late jumps and jumps in the middle) and looked at their variation; these are

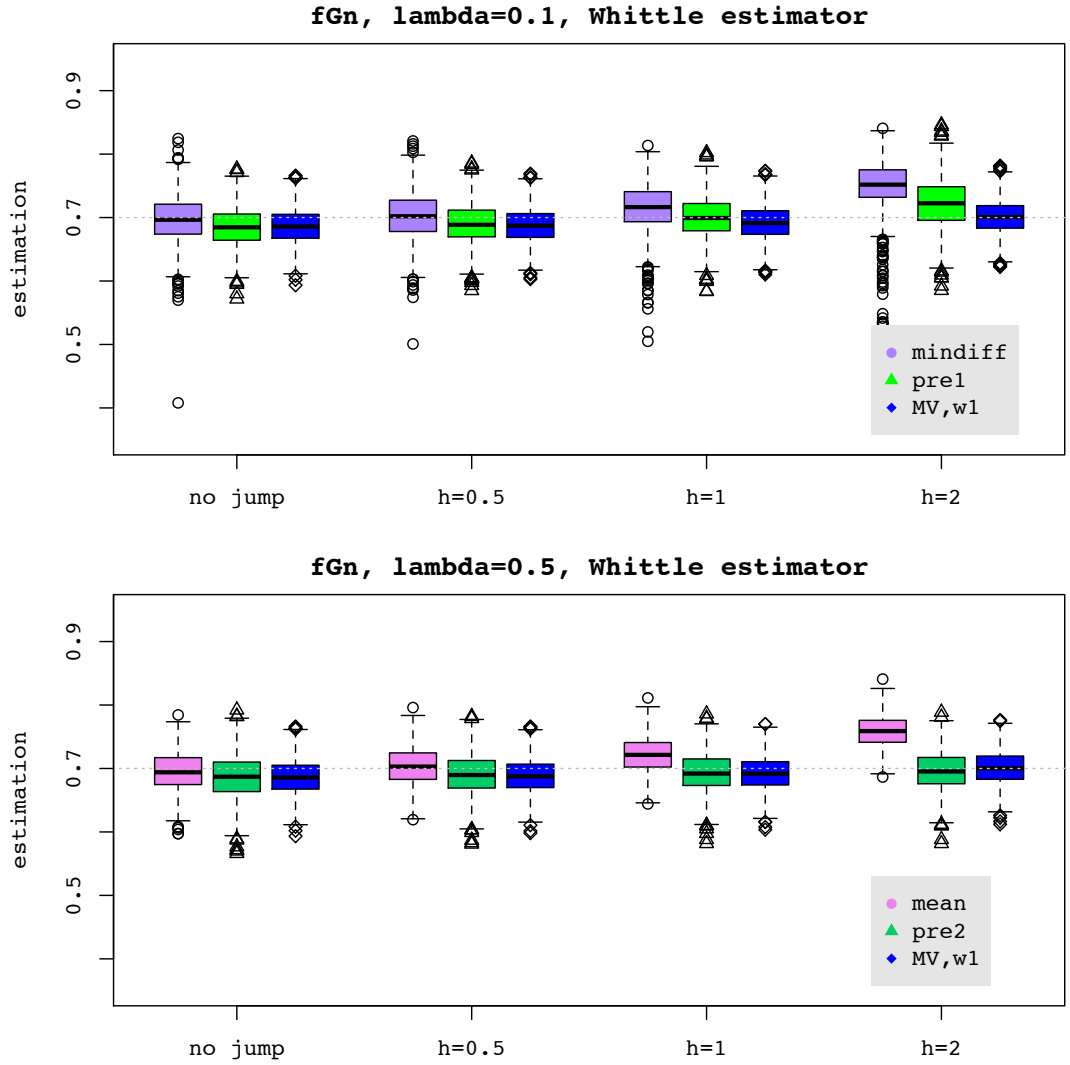


Figure 7.9: Estimators for the Hurst parameter $H = 0.7$ in time series with change-point (jump of height h after a proportion of λ), based on each 1,000 simulation runs with $n = 500$ realizations of fGn with $H = 0.7$.

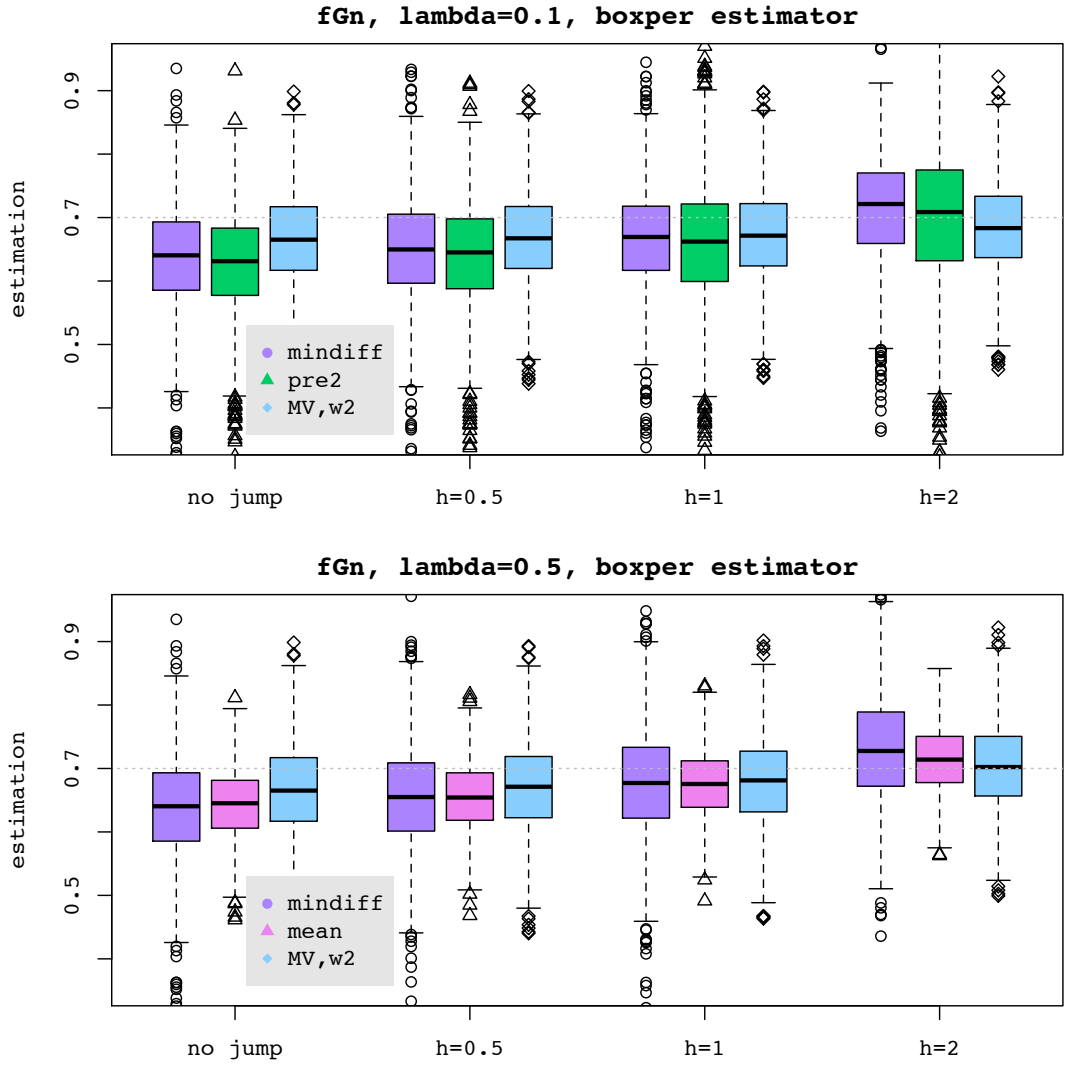


Figure 7.10: Estimators for the Hurst parameter $H = 0.7$ in time series with change-point (jump of height h after a proportion of λ), based on each 1,000 simulation runs with $n = 500$ realizations of fGn with $H = 0.7$.

		$\hat{H}$	$\hat{H}_{\text{mean}}$	$\hat{H}_{\text{mindiff}}$	$\hat{H}_{\text{pre,1}}$	$\hat{H}_{\text{pre,2}}$
no jump		-0.104	-0.083	-0.090	-0.148	-0.105
$\lambda = 0.1$	$h = 0.5$	-0.082	-0.070	-0.074	-0.129	-0.087
	$h = 1$	-0.030	-0.039	-0.049	-0.099	-0.063
	$h = 2$	0.096	0.0310	0.015	-0.052	-0.001
$\lambda = 0.5$	$h = 0.5$	-0.064	-0.067	-0.066	-0.139	-0.106
	$h = 1$	-0.010	-0.039	-0.033	-0.131	-0.101
	$h = 2$	0.081	0.017	0.042	-0.126	-0.097

		$\hat{H}$	$\hat{H}_{MV,w_1}$	$\hat{H}_{MV,w_2}$	$\hat{H}_{MV,w_3}$	$\hat{H}_{MVnl,w_1}$
no jump		-0.104	0.047	-0.050	-0.093	0.130
$\lambda = 0.1$	$h = 0.5$	-0.082	0.051	-0.047	-0.090	0.166
	$h = 1$	-0.030	0.059	-0.041	-0.083	0.219
	$h = 2$	0.096	0.081	-0.024	-0.064	0.281
$\lambda = 0.5$	$h = 0.5$	-0.064	0.051	-0.044	-0.076	0.153
	$h = 1$	-0.010	0.059	-0.031	-0.042	0.187
	$h = 2$	0.081	0.080	0.003	0.034	0.231

Table 7.2: Estimators for the Hurst parameter H in time series without ('no jump') and with change-point (jump of height h after a proportion of λ), relative difference between average estimate and the true parameter, each based on $1,000$ simulation runs with $n = 500$ realizations of fGn with $H = 0.7$, based on the Box-Periodogram Estimator.

in parts different estimators than in the setting based on the Whittle estimator. The results are shown in Figure 7.10 and reveal the following:

- Without any exception, the estimations have a much higher variance and a fatally large range.
- Compared to the estimations based on the Whittle Estimator, choosing the Box-Periodogram Estimator as underlying estimation method is not recommendable.
- Taking into account the variance of the estimator, the moving window estimator with $n/10$-window size, $\hat{H}_{MV,w_2}$, is uniformly the best (of the bad) in our survey: its variation is rather constant over jump heights and jump positions, and it is one of the smallest.

7.2.3 Conclusion and outlook

In time series which are supposed to exhibit LRD, it is important for any statistical inferece to estimate the LRD parameter like the Hurst parameter $H \in (0.5, 1)$, respectively $D = 2H - 2$ in (1.1). There exist several different approaches for this problem,

but they all suffer from the defect that they are confused by structural changes in the data: If there is a shift in the mean, the usual estimation methods easily missjudge the structure of the data and may detect spurious or too heavy LRD. This drawback is a fatal double bind when it comes to test for change-points in LRD time series: Change-point tests naturally require knowledge of the Hurst parameter H in order to discriminate between a change-point and behaviour which originates from the long memory, but in practival situations, H must be estimated. Thus a jump in the data may lead to overestimating H which may lead to overlooking the jump.

We have proposed three types of methods how estimation procedures can be adapted in order to make allowance for even time series with a change in the mean. In a simulation study we have compared these methods with different parameters and in different jump-contaminated and also jump-free situations. Our research shows that estimation of H can considerably be improved by our methods which yield better estimates when there is a jump in the time series and which do not affect the estimation when there is none.

It would be interesting to see how the proposed methods perform under different data scenarios, e.g. heavy-tailed observations, which may occur if one chooses another transormation G in our model, or a FARIMA model. Moreover, one could explore if one could reach improvements by e.g. choosing other window sizes for the moving window methods or by introducing weighted means for the methods which seperate the sample in two blocks (instead of keeping the cutting point k away from the borders, one could down-weight estimations where one of the blocks has a small size). It would also be interesting to investigate how the methods can be adapted to allow for multiple breaks.

7.3 Estimating the first Hermite coefficient

The observations $X_i = G(\xi_i)$ considered here exhibit LRD, and thus it is well-known by now, but still astonishing, that the limit behaviour of the process

$$d_n^{-1} \sum_{i=1}^{[\lambda n]} X_i = d_n^{-1} \sum_{i=1}^{[\lambda n]} G(\xi_i), \quad 0 \leq \lambda \leq 1,$$

only depends on very little properties of G, namely only on the Hermite rank of G and the associated Hermite coefficient a_m (see Theorem 1.1): m determines the scaling factor d_n and the kind of limit distribution; a_m is a multiplicative factor in this limit distribution. But this means: If one wants to do statistics based on the observations X_i, e.g. the "difference-of-means" test from Section 3.4.2, one has to know the function G, or at least its Hermite rank and the belonging coefficient, otherwise the limit distribution is unknown. Even in the comforting situation that G is strictly monotone in which we already know by the Corollary to Theorem 3.5 that the "difference-of-means" test statistic has always Hermite rank $m = 1$, we still need to know the first Hermite coefficient a_1. In what follows, we propose a method to estimate this Hermite coefficient

if one only observes $X_i = G(\xi_i)$, $i = 1, \ldots, n$, for the broad class of strictly monotone functions G.

7.3.1 The sort and replace method

For a start, we restrict ourselves to strictly monotonely increasing G. We want to estimate the first Hermite coefficient

$$a_1 := E[\xi\, G(\xi)],$$

where $\xi \sim \mathcal{N}(0,1)$. A natural way to estimate this expectation is the mean

$$\hat{a}_1 = \frac{1}{n}\sum_{i=1}^{n} \xi_i G(\xi_i) = \frac{1}{n}\sum_{i=1}^{n} \xi_i X_i,$$

but this does not help since we do not observe the ξ_i. But here the monotonicity of G comes in handy: We know that small X_i originate from small ξ_i and big X_i originate from big ξ_i, so we sort the summands:

$$\hat{a}_1 = \frac{1}{n}\sum_{i=1}^{n} \xi_i G(\xi_i) = \frac{1}{n}\sum_{i=1}^{n} \xi_{(i)} G(\xi_{(i)}) = \frac{1}{n}\sum_{i=1}^{n} \xi_{(i)} X_{(i)},$$

where in the last step we have essentially used that a strictly monotonely increasing G does not change the order of the data. In order to estimate a_1, we can now replace the unknown ξ_i by new independent random variables ξ_i' with the same distribution.

Theorem 7.1. *Suppose that $(\xi_i)_{i\geq 1}$ is a stationary Gaussian process with mean zero, variance 1 and auto-covariance function (1.1) with $0 < D < 1$. For $G \in \mathcal{G}^2$ define*

$$X_k = G(\xi_k).$$

Let the X_k have a continuous c.d.f. F. Let $\xi_i' \sim \mathcal{N}(0,1)$, $i = 1, \ldots, n$, be i.i.d. random variables, independent of the ξ_i that generate the observations $X_i = G(\xi_i)$, as defined above.
(i) If G is strictly monotonely increasing,

$$\boxed{\widetilde{a_1} := \frac{1}{n}\sum_{i=1}^{n} \xi_{(i)}' X_{(i)} \xrightarrow{P} a_1.} \tag{7.3}$$

(ii) If G is strictly monotonely decreasing,

$$\boxed{\widetilde{a_1} := \frac{1}{n}\sum_{i=1}^{n} \xi_{(i)}' X_{(n-i)} \xrightarrow{P} a_1.} \tag{7.4}$$

Proof. (i) G is strictly increasing. $G \in L^2(\mathbb{R}, \mathcal{N})$ entails $f(t) = t\,G(t) \in L^1(\mathbb{R}, \mathcal{N})$, thus by the ergodic theorem $\frac{1}{n}\sum_{i=1}^{n}(\xi_i X_i - a_1) \to 0$, almost surely. So it is to show that

$$\frac{1}{n}\sum_{i=1}^{n}(\xi_{(i)} - \xi'_{(i)})X_{(i)} \xrightarrow{P} 0.$$

By the Cauchy-Bunyakovsky-Schwarz inequality

$$\left|\frac{1}{n}\sum_{i=1}^{n}(\xi_{(i)} - \xi'_{(i)})X_{(i)}\right| \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\xi_{(i)} - \xi'_{(i)})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2},$$

and the second factor converges to a constant, since $G \in L^2(\mathbb{R}, \mathcal{N})$. In order to show that the first factor converges to zero, we truncate; set an arbitrary constant $c > 0$ and define

$$\bar{\xi}_i := \begin{cases} -c & \text{if } \xi_i < -c \\ c & \text{if } \xi_i > c \\ \xi_i & \text{else} \end{cases}$$

and $\bar{\xi}'_i$ analogously. With Minkowski's inequality and since truncation does not change the order of the data, so that $\sum_{i=1}^{n}(\xi_{(i)} - \bar{\xi}_{(i)})^2 = \sum_{i=1}^{n}(\xi_i - \bar{\xi}_i)^2$, we obtain

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\xi_{(i)} - \xi'_{(i)})^2} \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\xi_i - \bar{\xi}_i)^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\xi'_i - \bar{\xi}'_i)^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\bar{\xi}_{(i)} - \bar{\xi}'_{(i)})^2}. \tag{7.5}$$

The first term on the right-hand side of (7.5) vanishes asymptotically for $n, c \to \infty$ because

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\xi_i - \bar{\xi}_i)^2} \overset{n\to\infty}{\longrightarrow} \sqrt{E\left[(\xi_i - \bar{\xi}_i)^2\right]}$$

and $E\left[(\xi_i - \bar{\xi}_i)^2\right] \to 0$ as $c \to \infty$ due to the dominated convergence theorem, because $\xi_i - \bar{\xi}_i \to 0$ and $|\xi_i - \bar{\xi}_i| < |\xi_i|$. This holds of course for the second term in (7.5) as well. We well now rigorously prove the intuition correct that for two samples $\xi_1, \ldots, \xi_n$ and $\xi'_1, \ldots, \xi'_n$ of identically distributed random variables $\xi_{(i)}$ gets close to $\xi'_{(i)}$, when n increases. With $|\bar{\xi}_{(i)} - \bar{\xi}'_{(i)}| \leq 2c$, we obtain for the last term on the right-hand side of (7.5)

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\bar{\xi}_{(i)} - \bar{\xi}'_{(i)})^2 &\leq 2c \max_{i=1,\ldots,n}|\bar{\xi}_{(i)} - \bar{\xi}'_{(i)}| \\ &\leq 2c \max_{i=1,\ldots,n}\left(\left|\bar{\xi}_{(i)} - \bar{\Phi}^{-1}\left(\frac{i}{n}\right)\right| + \left|\bar{\xi}'_{(i)} - \bar{\Phi}^{-1}\left(\frac{i}{n}\right)\right|\right),\end{aligned}$$

where $\bar{\Phi}$ denotes the c.d.f of the single $\bar{\xi}_i$. Note that by definition of $\bar{\xi}_i$, $\bar{\Phi}$ equals the standard normal c.d.f. Φ on the interval $(-c, c)$.

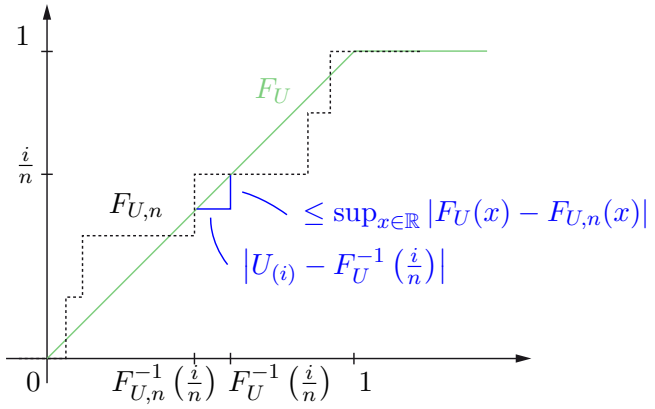


Figure 7.11: A lower bound for the difference of the empirical and the true distribution function for uniformly distributed variables: $\sup_{x\in\mathbb{R}} |F_U(x) - F_{U,n}(x)|$ is at least as big as $\left|U_{(i)} - F_U^{-1}\left(\frac{i}{n}\right)\right|$. Note that $F_{U,n}^{-1}\left(\frac{i}{n}\right) = \inf\{t \,|\, F_{U,n}(t) \geq \frac{i}{n}\} = U_{(i)}$.

For a moment consider i.i.d. random variables $U_i \sim U[0,1]$. Here the c.d.f. F_U is basically the bisecting line, and thus one can graphically show (see Figure 7.11) that

$$\left|U_{(i)} - F_U^{-1}\left(\frac{i}{n}\right)\right| \leq \sup_{x\in\mathbb{R}} |F_U(x) - F_{U,n}(x)|\,, \tag{7.6}$$

where $F_{U,n}$ is the empirical distribution function of the U_i. Such an inequality for non-uniformly distributed random variables can be traced back on (7.6) by using the mean value theorem:

$$(b-a) \inf_{t\in(a,b)} f'(t) \leq f(b) - f(a),$$

if f is continuous on $[a,b]$ and differentiable in the interior. With $f = \bar{\Phi}$ and $[a,b] = [-c,c]$ we obtain

$$\begin{aligned}
\left|\bar{\xi}_{(i)} - \bar{\Phi}^{-1}\left(\frac{i}{n}\right)\right| &\leq \frac{1}{\inf_{t\in(-c,c)} \varphi(t)} \left|\bar{\Phi}\left(\bar{\xi}_{(i)}\right) - \bar{\Phi}\left(\bar{\Phi}^{-1}\right)\left(\frac{i}{n}\right)\right| \\
&\leq \frac{1}{\inf_{t\in(-c,c)} \varphi(t)} \left|U_{(i)} - \frac{i}{n}\right| \\
&\leq \frac{1}{\inf_{t\in(-c,c)} \varphi(t)} \sup_{x\in\mathbb{R}} \left|F_U(x) - \bar{F}_{U,n}(x)\right|,
\end{aligned}$$

where $\bar{F}_{U,n}(x)$ denotes the e.d.f. of the (uniformly distributed) $\bar{\Phi}(\bar{\xi}_i)$.

So we obtain for the last term on the right-hand side of (7.5)

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} (\bar{\xi}_{(i)} - \bar{\xi}'_{(i)})^2 &\leq \frac{2c}{\inf_{t\in(-c,c)} \varphi(t)} \left(\sup_{x\in\mathbb{R}} \left|F_U(x) - \bar{F}_{U,n}(x)\right| + \sup_{x\in\mathbb{R}} \left|F_U(x) - \bar{F}'_{U,n}(x)\right|\right) \\
&= \frac{2c}{\inf_{t\in(-c,c)} \varphi(t)} \left(\sup_{x\in\mathbb{R}} \left|\bar{\Phi}(x) - \bar{\Phi}_n(x)\right| + \sup_{x\in\mathbb{R}} \left|\bar{\Phi}(x) - \bar{\Phi}'_n(x)\right|\right),
\end{aligned} \tag{7.7}$$

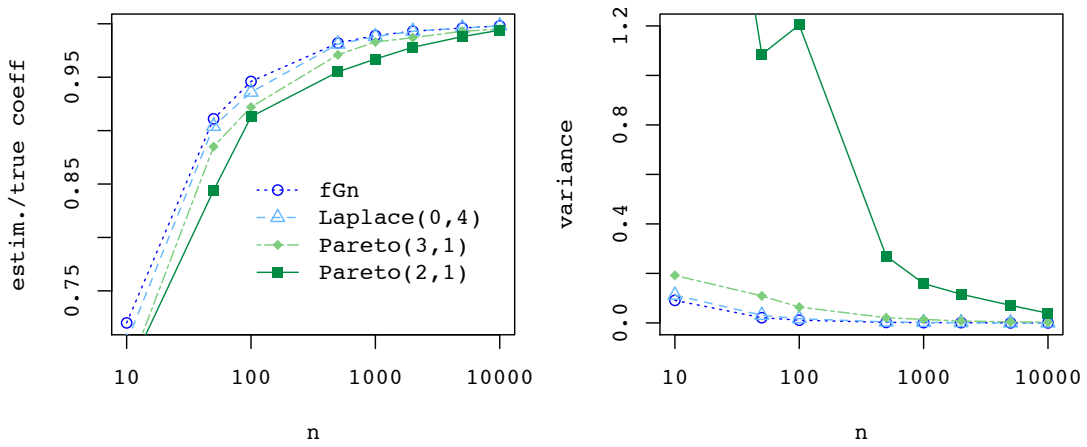


Figure 7.12: Estimated Hermite coefficients $\widetilde{a_1}$, the mean relative to the true value a_1 (left) and the variance (right), based on 10,000 repetitions, for different G.

since the transformation to the uniform distribution only rescales the x-axis. (To see this, just plug in $\bar{\Phi}(x)$ instead of x, it makes no difference if we evaluate the supremum over all $x \in \mathbb{R}$ or all $\bar{\Phi}(x) \in \bar{\Phi}(\mathbb{R})$. Recall: $\bar{F}_{U,n}(x)$ denotes the e.d.f. of the $\bar{\Phi}(\bar{\xi}_i)$, $\bar{F}'_{U,n}(x)$ denotes the e.d.f. of the $\bar{\Phi}(\bar{\xi}'_i)$, $\bar{\Phi}_n(x)$ denotes the e.d.f. of the $\bar{\xi}$, $\bar{\Phi}(x)$ denotes the true c.d.f. of the $\bar{\xi}$ and $\bar{\Phi'}_n(x)$ denotes the e.d.f. of the $\bar{\xi}'$.) Now with the ergodic version of the theorem of Glivenko-Cantelli, (7.7) converges to 0 for all fixed c, as $n \to \infty$, thus the right-hand side of (7.5) converges to zero, and the statement is proved.

(ii) If G is strictly decreasing, only small changes have to be made. Under a decreasing transformation $X_i = G(\xi_i)$, small X_i originate from big ξ_i and big X_i originate from small ξ_i, so we now sort the summands

$$\hat{a}_1 = \frac{1}{n}\sum_{i=1}^{n} \xi_i G(\xi_i) = \frac{1}{n}\sum_{i=1}^{n} \xi_{(i)} G(\xi_{(i)}) = \frac{1}{n}\sum_{i=1}^{n} \xi_{(i)} X_{(n-i+1)}.$$

So just replace $X_{(i)}$ by $X_{(n-i+1)}$ in the preceding proof. Since $(\xi_i)_{i\geq 1}$ is stationary and thus $(X_i)_{i\geq 1}$ is it, this does not change anything. □

7.3.2 Simulations

In this section, we will analyse the behaviour of the estimator $\widetilde{a_1}$ in (7.3) in finite sample settings (with sample size ranging from $n = 10$ up to $n = 10,000$). We have simulated n realisations $\xi_1, \ldots, \xi_n$ of fractional Gaussian noise (fGn) with Hurst parameter $H = 0.7$ and generated the observations $X_i = G(\xi_i)$ by applying different functions G on the fGn. We have repeated each simulation 10,000 times. To see how good the estimation is, we have divided the sample mean of these 10,000 estimates (for each set of simulations) by the respective true Hermite coefficient $a_1 = E[\xi G(\xi)]$ which has been determined by numerical integration. We have also calculated the sample variance of the 10,000

estimates. The simulation results are presented in Figure 7.12; the exact simulation results are given in Table D.24 in Appendix D.

- Gaussian data.
  With the increasing transformation $G(t) = t$ we obtain standard fGn as obervations. This is the most simple non-trivial case in this model since G does not change the underlying ξ_i. As one can expect, the estimation is very good in this case.

- Symmetric data.
  The function
  $$G(t) = -(2^{-1/2}) \operatorname{sgn}\left(\Phi(t) - \frac{1}{2}\right) \log\left(1 - 2\left|\Phi(t) - \frac{1}{2}\right|\right).$$
  first transforms the data to a $U[-\frac{1}{2}, \frac{1}{2}]$ distribution, then applies a quantile transformation and finally centralises the data. This G is increasing and yields standardised Laplace(0,4) distributed data with p.d.f.
  $$f_{\mathrm{st}}(x) = \frac{1}{\sqrt{2}} \exp\left(-|\sqrt{2}x|\right),$$
  i.e. normal-tailed, symmetric data. G is a well-behaved transformation, so the data are not too wild: The estimation is close to the Gaussian case.

- Not-so-heavy-tailed data.
  A function G that provides heavy tails but a finite variance, so that it is coverd by our techniques, is $G(t) = (3/4)^{-1/2}\left((\Phi(t))^{-1/3} - \frac{3}{2}\right)$. G is decreasing and yields standardised Pareto(3,1) distribution with p.d.f.
  $$\sqrt{\frac{3}{4}} \cdot f_{3,1}\left(\sqrt{\frac{3}{4}}\, x + \frac{3}{2}\right) = \begin{cases} 3\sqrt{\frac{3}{4}}\left(\sqrt{\frac{3}{4}}\, x + \frac{3}{2}\right)^{-4} & \text{if } x \geq -\sqrt{\frac{1}{3}} \\ 0 & \text{else} \end{cases}.$$
  The estimated value is comparable to the estimated value under Pareto(2,1) data (the case which we will treat next), but the variance of the estimation is fortunately smaller.

- Heavy-tailed data.
  $G(t) = (\Phi(t))^{-1/2} - 2$ is decreasing and transforms the data to a centralised Pareto(2,1)-distribution with p.d.f.
  $$f_{2,1}\left(x + \frac{3}{2}\right) = \begin{cases} 2\left(x + \frac{3}{2}\right)^{-3} & \text{if } x \geq -\frac{1}{2} \\ 0 & \text{else} \end{cases}.$$
  So the X_i have heavy tails and and infinite variance. Note that such transformation G are actually not covered by Theorem 7.1, but it is interesting to study

the performance of the estimator in such a case. Unsurprisingly, the estimation is bad: For for small sample sizes, the mean is far away from the true value (although it relies on 10,000 simulation runs), and the estimation is afflicted with a huge variance.

Appendix A

A short introduction into stochastic integration

In Section 1.3, I have mentioned (and we have seen it over and over throughout this work) that the theory of LRD processes is strongly related to stochastic integration, integration with respect to a stocastic process (instead of a deterministic function). Now I will explain the idea behind these objects and their relation to LRD.

A.1 Wiener integral

The simplest stochastic integral is the *Wiener integral*

$$\boxed{I_W(f) = \int_a^b f(t)\, dB(t,\omega),}$$

where f is a deterministic function (i.e. it does not depend on ω) and B is a Brownian motion. It can be constructed for arbitrary square-integrable functions $f \in L^2([a,b],\lambda)$ by approximation procedures using step functions. To this end, divide the interval $[a,b]$ into n pieces: $a = \tau_0 < \tau_1 < \ldots < \tau_{n-1} < \tau_n = b$. For step functions $f(t) = \sum_{i=1}^n a_i I_{[\tau_{i-1},\tau_i)}(t)$, $a_i \in \mathbb{R}$, the integral is defined as

$$I_W(f) = \sum_{i=1}^n a_i \left(B(\tau_i) - B(\tau_{i-1})\right).$$

Clearly, I_W is a linear map on the space of step functions: $I_W(\alpha f + \beta g) = \alpha I_W(f) + \beta I_W(g)$ for any $\alpha, \beta \in \mathbb{R}$ and any step functions f, g. Using that the increments $B(t) - B(s)$ of a Brownian motion are independent $\mathcal{N}(0, t-s)$-distributed, one can easily show that

$$I_W(f) \sim \mathcal{N}(0, \sigma^2)$$

with

$$\sigma^2 = E\left[I_W(f)\right]^2 = \int_a^b f^2(t)\, dt.$$

Now consider an arbitrary function $f \in L^2([a,b],\lambda)$ and choose a sequence $(f_n)_{n\geq 1}$ of approximating step functions (i.e. $f_n \to f$). The last identity ensures that $(I_W(f_n))_{n\geq 1}$ is a Cauchy sequence in L^2 and thus converging. We define its limit as the Wiener integral of f.

Definition A.1 (Wiener integral)**.** For any $f \in L^2([a,b],\lambda)$, the *Wiener integral* of f is

$$I_W(f) = \lim_{n\to\infty} I_W(f_n),$$

where the convergence is in L^2.

A Wiener integral has the following properties:

- The limit in the definition above is independent of the chosen approximating sequence of step functions, such that $I_W(f)$ is well-defined.
- $I_W(f) = \int_a^b f(t)\ dB(t,\omega)$ is a Gaussian random variable with mean 0 and variance $\|f\|^2 = \int_a^b f^2(t)\ dt$.
- From this it follows by some calculations that

  $$E\left[I_W(f) I_W(g)\right] = \int_a^b f(t)g(t)\ dt,$$

  and so, $I_W(F)$ and $I_W(g)$ are independent, if f and g are orthogonal in $L^2([a,b],\lambda)$.
- The stochastic process

  $$M_t = \int_a^t f(s)\ dB(s)$$

  is a martingale.

Verifying these properties is not excessively difficult (Kuo, 2006, p. 9–21, gives detailed proofs).

A.2 Itō integral

A natural question, at least for mathematicians, is now if one can define an integral with respect to a Brownian motion even for stochastic processes as integrands. The answer is positive; a so called *Itō integral*

$$\boxed{I_I(f) = \int_a^b f(t,\omega)\ dB(t,\omega)}$$

arises. Now f is a stochastic process, and this extension yields some severe problems. If we want to keep the computationally advantageous martingale property, we need to place some demands on f.

Definition A.2 (Adapted L^2 space)**.** Consider a probability space $(\Omega, \mathcal{F}, P)$. Let $L^2_{\text{ad}}([a,b] \times \Omega, \lambda \times P)$ be the class of functions

$$f(t,\omega) : [a,b] \times \Omega \to \mathbb{R}$$

such that

(i) $(t,\omega) \mapsto f(t,\omega)$ is $\mathcal{B} \times \mathcal{F}$-measurable, where $\mathcal{B}$ denotes the Borel σ-algebra on $[a,b]$,

(ii) $f(t,\omega)$ is adapted to the filtration $\{\mathcal{F}_t\}$, where $\mathcal{F}_t = \sigma(B_s, s \le t)$ is the σ-algebra generated by the random variables B_s with $s \le t$ (one can think of $\mathcal{F}_t$ as being the history of B_s up to time t),

(iii) $E\left[\int_a^b f(t,\omega)^2 \, dt\right] < \infty$.

For such processes f we can define a nice stochastic integral, and we will do this once more by approximation procedures. At first, divide the interval $[a,b]$ into n pieces $a = \tau_0 < \tau_1 < \ldots < \tau_{n-1} < \tau_n = b$ and consider a step process

$$f(t,\omega) = \sum_{i=1}^{n} a_i(\omega) I_{[\tau_{i-1},\tau_i)}(t),$$

where a_i is now a $\mathcal{F}_{\tau_{i-1}}$-measurable, quadratically integrable random variable. For such processes, the integral is defined as

$$I_I(f) = \sum_{i=1}^{n} a_i(\omega) \left(B(\tau_i) - B(\tau_{i-1})\right)(\omega).$$

Clearly, this is again linear: $I_I(\alpha f + \beta g) = \alpha I_I(f) + \beta I_I(g)$ for any $\alpha, \beta \in \mathbb{R}$ and any step processes f, g. Not as easily as for Wiener integrals, one can show that

$$E\left[I_I(f)\right] = 0$$
$$E\left[I_I(f)\right]^2 = \int_a^b E\left[f(t,\omega)\right]^2 \, dt,$$

but of course $I_I(f)$ is in general not normal distributed. This is a first raw version of the so called *Itō isometry*.

With the aid of this isometry, one can extend the definition from step processes to processes in $L^2_{\text{ad}}([a,b] \times \Omega)$. At first, choose an arbitrary process $g \in L^2_{\text{ad}}$ which is bounded and continuous in t for each ω. Then we can find an approximating sequence $(f_n)_{n \ge 1}$ of step processes such that

$$E\left[\int_a^b (g - f_n)^2 \, dt\right] \to 0 \qquad \text{as } n \to \infty.$$

Next, one can approximate any bounded $h \in L^2_{\text{ad}}$ by such bounded functions $g_n \in L^2_{\text{ad}}$ which are continuous in t for each fixed ω:

$$E\left[\int_a^b (h - g_n)^2 \, dt\right] \to 0 \qquad \text{as } n \to \infty$$

Finally, we can approximate any arbitray $f \in L^2_{\text{ad}}$ by bounded functions $h_n \in L^2_{\text{ad}}$:

$$E\left[\int_a^b (f - h_n)^2 \, dt\right] \to 0 \qquad \text{as } n \to \infty$$

The technical details of these approximations are worked out for example in the books by Kuo (2006, Chap. 4.3) or Øksendal (1998, Chap. 3.1). These three steps together mean that, in order to define an Itō integral, we can approximate any $f \in L^2_{\text{ad}}$ by step functions for which we have defined $I_I(f)$.

Definition A.3 (Itō integral)**.** For any $f \in L^2_{\text{ad}}([a,b] \times \Omega, \lambda \times P)$, the *Itō integral* is defined by

$$I_I(f) = \int_a^b f(t,\omega) \, dB_t(\omega) = \lim_{n\to\infty} I_I(f_n),$$

where $(f_n)_{n\geq 1}$ is a sequence of step processes approximating f in $L^2([a,b] \times \Omega)$:

$$E\left[\int_a^b (f(t,\omega) - f_n(t,\omega))^2 \, dt\right] \to 0 \qquad \text{as } n \to \infty.$$

An Itō integral has the following properties:

- The limit in the definition above is independent of the chosen approximating sequence of step processes, such that $I_I(f)$ is well-defined.
- $I_I(f) = \int_a^b f(t,\omega) \, dB(t,\omega)$ is a random variable with mean 0 and variance

  $$E\left[\int_a^b f(t,\omega) \, dB_t(\omega)\right]^2 = \int_a^b E[f^2(t,\omega)] \, dt$$

  (this is called *Itō isometry*).
- For a sequence $(f_n)_{n\geq 1} \in L^2_{\text{ad}}$ with $E\left[\int_a^b (f(t,\omega) - f_n(t,\omega))^2 \, dt\right] \to 0$ it holds

  $$\int_a^b f_n(t,\omega) \, dB_t(\omega) \to \int_a^b f(t,\omega) \, dB_t(\omega)$$

  in $L^2(P)$ as $n \to \infty$.
- For all $f, g \in L^2_{\text{ad}}$

  $$E\left[I_I(f) I_I(g)\right] = \int_a^b E\left[f(t,\omega) g(t,\omega)\right] \, dt.$$

- $\int_a^b f(t,\omega) \, dB_t(\omega)$ is $\mathcal{F}_b$-measurable.
- The stochastic process

  $$M_t = \int_a^t f(s,\omega) \, dB_s(\omega)$$

  is a martingale with respect to the filtration $\{\mathcal{F}_t\} = \sigma(B(s), s \leq t)$.

- M_t has a t-continuous version: There exists a t-continuous stochastic process $\tilde{M}_t$ on $(\Omega, \mathcal{F}, P)$ with $P(\tilde{M}_t = M_t) = 1$ for all $t \in [a, b]$.

One major difference between stochastic and deterministic integrals is the following: As usual, one approximates $f(t, \omega)$ by a step process $\sum_j f(t_j^*, \omega) I_{[t_{j-1}, t_j)}(t)$ with some $t_j^* \in [t_{j-1}, t_j)$, and the integral $\int_a^b f(t, \omega)\ dB_t(\omega)$ is defined by the limit of $\sum_j f(t_j^*, \omega)(B_{t_j} - B_{t_{j-1}})$. But what is now a bit surprising is that, in contrast to Riemann integrals, it makes a difference which point $t_j^* \in [t_{j-1}, t_j)$ we choose, even though the intervals become infinitely small.

Two choices have become common, for they have been proven useful: Choosing the left endpoint of the interval $t_j^* = t_{j-1}$ leads to the Itō integral above, choosing the mid-point $t_j^* = (t_{j-1} + t_j)/2$ leads to the so called *Stratonovic integral*. While the Itō integral is a martingale and thus features some computational advantage, the Stratonovich integral is not, however it behaves more pleasingly under transformations: It allows for a chain rule without second order terms (which are inevitable for Itō integrals, see section A.3 below). Øksendal (1998, p. 36–37) gives a short discussion of both concepts; Kuo (2006, Chap. 8.3) discusses the Stratonovich integral in the context of Itō processes and the Itō formula (which we will shortly introduce).

It is possible to extend the definition of an Itō integral to a wider class of integrands f. Condition (iii) in Definition A.2, $E[\int_a^b f(t, \omega)^2\ dt] < \infty$, can be relaxed to $\int_a^b f(t, \omega)^2\ dt < \infty$ a.s., which may be not integrable.

With the original definition, $I_I(f)$ belongs to $L^2(\Omega, P)$, and since P is a probability measure and therefore finite, it follows that $I_I(f) \in L^1(\Omega, P)$ as well, that means that the Itō integral is an integrable random variable, and $\int_a^t f(s, \omega) dB_s(\omega)$ is a martingale. A function f that fulfills only the new condition $\int_a^b f(t, \omega)^2\ dt < \infty$ a.s. does not necessarily lead to an integrable $I_I(f)$, and if it is not integrable, it cannot be a martingale. But it is at least a so called *local martingale* which has as well a continuous version (Kuo, 2006, Chap. 5).

In Definition A.2, condition (ii) can also be weakened. This allows to define multidimensional Itō integrals and, as a special case, Itō integrals with respect to one single coordinate of n-dimensional Brownian motion while the integrand is a function of other of its coordinates. This is carried out by Øksendal (1998, Chap. 3.3) where some bibliographical references are given as well. It is also possible to extend the definition of an Itō integral to a wider class of integrators, namely to martingales. For $f, g \in L^2_{\text{ad}}([a, b] \times \Omega)$ and a martingale $M_t = \int_a^t g(s, \omega)\ dB_s(\omega)$ one defines

$$\int_a^b f(t, \omega)\ dM_t(\omega) = \int_a^b f(t, \omega) g(t, \omega)\ dB_t(\omega).$$

To obtain a useful definition, some further conditions on f are necessary; the construction can be found in the book of Kuo (2006, Chap. 6), the essential tool is the

Doob-Meyer decomposition.

As for ordinary random variables, we will ommit the argument ω when we deal with stochastic processes if there is no danger of confusing something.

A.3 Itō process and Itō formula

As in ordinary calculus, one does not want to evaluate stochastic integrals by their definition; it is laborious. But while ordinary integrals can be evaluated with the aid of the fundamental theorem of calculus, with antiderivatives, such useful combination of integration and differentiation is not on hand here, simply because in the world of stochastic processes there is no differentiation. For example a chain rule like $(f(g(t)))' = f'(g(t))g'(t)$ in the context of processes,

$$(f(B(t)))' = f'(g(t))B'(t),$$

is meaningless since almost all sample paths of $B(t)$ are nowhere differentiable. Nevertheless, there is an analogue to the chain rule which lends us a hand at stochastic integration.

Theorem A.1 (Itō formula I)**.** *For any $f \in C^2(\mathbb{R})$*

$$\boxed{f(B(t)) - f(B(a)) = \int_a^t f'(B(s))\, dB(s) + \frac{1}{2}\int_a^t f''(B(s))\, ds,}$$

where the first summand on the right side is an Itō integral and the second one is an ordinary Riemann integral for any sample path of $B(s)$.

Obviously, the second term on the right side shows (or even constitutes) the difference between ordinary calculus in Newton/Leibnz sense and Itō calculus. It originates in the Brownian motion which has a non zero quadratic variation. Using this simple version of the Itō formula (or as well by calculating manually which is not too difficult in this simple case), one obtains for example

$$\begin{aligned}\frac{1}{2}B_t^2 &= \frac{1}{2}B_t^2 - \frac{1}{2}B_0^2 \\ &= \int_0^t B_s\, dB(s) + \frac{1}{2}\int_0^t ds \\ &= \int_0^t B_s\, dB(s) + \frac{1}{2}t,\end{aligned}$$

and this shows that a harmless transformation like $g(x) = x^2/2$ converts an Itō integral like $B_t = \int_0^t dB_s$ in something which is not an Itō integral any more. Mathematicians dislike such inconsistent behaviour, and so the need for another class of processes that is stable under smooth transformations is obvious.

Definition A.4 (Itō process)**.** A stochastic process of the form

$$\boxed{X_t(\omega) = X_a(\omega) + \int_a^t f(s,\omega)\, dB_s(\omega) + \int_a^t g(s,\omega)\, ds,} \tag{A.1}$$

where X_a is $\mathcal{F}_a$-measurable, $a \le t \le b$ and f, g are in $L^2_{\mathrm{ad}}([a,b] \times \Omega)$ (or in the generalisation of this space mentioned above) is called an *Itō process*. Often, such processes are written in the convenient shorthand notation

$$dX_t = f(t)\, dB_t + g(t)\, dt, \tag{A.2}$$

which is of course only a symbolic expression since the differential of B is not defined because B is nowhere differentiable.

$B_t = \int_0^t dB_s$ in the example above is an Itō process, and its transformation under the map $x \mapsto x^2/2$ is it as well: It is the sum of a dB_s- and a ds-integral. Its short differential notation is

$$d\left(\frac{1}{2}B_t^2\right) = B_t\, dB_t + \frac{1}{2}\, dt.$$

Theorem A.2 (Itō formula II)**.** *For an Itō process X_t, as in* (A.1) *or* (A.2)*, and a function $\theta(t,x) \in C^2([0,\infty) \times \mathbb{R})$,*

$$Y_t = \theta(t, X_t)$$

is again an Itō process, and

$$\begin{aligned}\theta(t, X_t) = \theta(a, X_a) &+ \int_a^t \frac{\partial\theta}{\partial x}(s, X_s) f(s)\, dB_s \\ &+ \int_a^t \left(\frac{\partial\theta}{\partial t}(s, X_s) + \frac{\partial\theta}{\partial x}(s, X_s) g(s) + \frac{1}{2}\frac{\partial^2\theta}{\partial x^2}(s, X_s) f^2(s)\right) ds\end{aligned}$$

or in shorthand notation

$$\boxed{dY_t = \frac{\partial\theta}{\partial x}(t, X_t)\, dX_t + \frac{\partial\theta}{\partial t}(t, X_t)\, dt + \frac{1}{2}\frac{\partial^2\theta}{\partial x^2}(t, X_t)\, (dX_t)^2,}$$

where $(dX_t)^2 = (dX_t) \cdot (dX_t)$ is computed by (A.2) *and the symbolic multiplication table*[1]

$\times$	dB_t	dt
dB_t	dt	0
dt	0	0

.

Some important properties, extensions and applications are:

- With the aid of the Itō formula, a lot of stochastic integrals can be evaluated using symbolic notation.

[1] Even not in full generality, Kuo (2006, p. 103) gives a nicely simple instruction how to use the short notation and the multiplication table.

- Itō processes and the Itō formula can be defined in the multi-dimensional case.
- The (multi-dimensional) Itō formula applies as well to processes with respect to continuous and square integrable martingales (and not only with respect to the Brownian motion).
- By the Itō formula, an integration by parts formula can be derived: If f is deterministic (that means $f(s,\omega) = f(s)$ only depends on s), continuous and of bounded variation on $[0,t]$, then $\int_0^t f(s)\, dB_s = f(t)B_t - \int_0^t B_s\, df(s)$.
- As stated above, an Itō integral is a martingale, but the converse is also true. With the Itō formula, one can proove the important *Martingale Representation Theorem*: Any martingale (adapted to $\mathcal{F}_t$, with respect to the probability measure P) can be represented as an Itō integral.

A.4 Multiple Wiener-Itō integrals

As we have seen, we can find analogies between ordinary integration and stochastic integration in many fields, but one obvious question remains: Can one define multiple stochstic integrals? The answer is positive, but the construction of multiple stochstic integrals is not self-evident. We restrict ourselfs to multiple Wiener integrals, i.e. to deterministic integrands. The usual approach is to approximate such a deterministic function $f(t,s)$ by step functions

$$\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij} I_{[\tau_{i-1},\tau_i)}(t) I_{[s_{i-1},s_i)}(s),$$

where $a = \tau_0 < \tau_1 < \ldots < \tau_{n-1} < \tau_n = b$ and $a = s_0 < s_1 < \ldots < s_{m-1} < s_m = b$ are partitions of the interval $[a,b]$ and $a_{ij} \in \mathbb{R}$. This construction yields for example

$$\int_0^1 \int_0^1 1\, dB_t\, dB_s = B_1^2,$$

and this has the major drawback that it is unfortunately not orthogonal to constant functions. In general, integrals of different degrees are by this approach, first introduced by Wiener (1938), not orthogonal to each other.

Itō (1951) found a remedy (and this is why the construction is called *multiple Wiener-Itō integral*): f has to be approximated by step functions which spare out the diagonal of the domain of integration. For a start, we consider the two-dimensional example above. Let $\Delta_n = (\tau_0, \ldots, \tau_n)$ be a partition of the intervall $[a,b] = [0,1]$ into n pieces. It naturally extends to a partition of the unit square

$$[0,1)^2 = \bigcup_{i,j=1}^{n} [\tau_{i-1},\tau_i) \times [\tau_{j-1},\tau_j).$$

Defining step functions on this partition, where the interval in both dimensions is split up in the same way, yields of course the same result as above because it is only a special case of a general partition into rectangles: For the integrad $f \equiv 1$, we obtain the Riemann sum

$$\sum_{i,j=1}^{n} (B_{\tau_i} - B_{\tau_{i-1}})(B_{\tau_j} - B_{\tau_{j-1}}) = \left(\sum_{i=1}^{n} (B_{\tau_i} - B_{\tau_{i-1}}) \right)^2 = B_1^2.$$

But now we remove the diagonal elements and evaluate the increments of the Brownian motion over the domain

$$[0,1)^2 \setminus \bigcup_{i=1}^{n} [\tau_{i-1}, \tau_i)^2 = \bigcup_{1 \le i \ne j \le n} [\tau_{i-1}, \tau_i) \times [\tau_{j-1}, \tau_j),$$

i.e. the remaining off-diagonal squares:

$$\begin{aligned} S_n &= \sum_{1 \le i \ne j \le n} (B_{\tau_i} - B_{\tau_{i-1}})(B_{\tau_j} - B_{\tau_{j-1}}) \\ &= \sum_{i,j=1}^{n} (B_{\tau_i} - B_{\tau_{i-1}})(B_{\tau_j} - B_{\tau_{j-1}}) - \sum_{i=1}^{n} (B_{\tau_i} - B_{\tau_{i-1}})^2 \\ &\to B_1^2 - 1 \qquad \text{as } \|\Delta_n\| \to 0 \end{aligned}$$

where $\|\Delta_n\| = \max_{1 \le i \le n} \{\tau_i - \tau_{i-1}\}$ is the fineness of the partition and where we have used that the quadratic variation of the Brownian motion on an interval $[a, b]$ is

$$\lim_{\|\Delta_n\| \to 0} \sum_{i=1}^{n} (B_{\tau_i} - B_{\tau_{i-1}})^2 = b - a,$$

see Kuo (2006, Th. 4.1.2). Thus

$$\int_0^1 1 \, dB_t \, dB_s = B_1^2 - 1,$$

which is obviously orthogonal to a constant function.

This way of constructing multiple stochastic integrals has proved convenient. The general construction for a dimension $d \in \mathbb{N}$ is as follows. Let $[a, b] \subset \mathbb{R}$. A subset of $[a, b]^d$ of the form $[t_1^{(1)}, t_1^{(2)}) \times \ldots \times [t_d^{(1)}, t_d^{(2)})$ is called a *rectangle*. The *diagonal set* of $[a, b]^d$ is the set $D = \{(t_1, \ldots, t_d) \in [a, b]^d \mid \exists i \ne j : t_i = t_j\}$ of all points which have at least two identical coordinates.

Definition A.5 (Off-diagonal step function)**.** A function

$$f(t_1, \ldots, t_d) = \sum_{1 \le i_1, \ldots, i_d \le n} a_{i_1, \ldots, i_d} I_{[\tau_{i_1 - 1}, \tau_{i_1})}(t_1) \cdots I_{[\tau_{i_d - 1}, \tau_{i_d})}(t_d)$$

on the rectangle $[a, b]^d$ with a partition $a = \tau_0 < \tau_1 < \ldots < \tau_n = b$ in each dimension is called a *step function*. It is called an *off-diagonal step function* if it vanishes on D, that means if the coefficients satisfy

$$\boxed{a_{i_1, \ldots, i_d} = 0 \qquad \text{if } i_p = i_q \text{ for some } p \ne q.}$$

The class of off-diagonal step functions is a vector space. For an off-diagonal step function f define the d-dimensional multiple Wiener-Itō integral by

$$\boxed{\begin{aligned} I_d(f) &= \int_{[a,b]^d} f(t_1,\dots,t_d)\, dB_{t_1}\, dB_{t_2}\cdots dB_{t_d} \\ &= \sum_{1\le i_1,\dots,i_d\le n} a_{i_1,\dots,i_d}\, \xi_{i_1}\xi_{i_2}\cdots\xi_{i_d}, \end{aligned}}$$

where $\xi_{i_p} = B(\tau_{i_p-1}) - B(\tau_{i_p})$ is the increment of B over the p-th piece of the partition of $[a,b]$. Note that the above representation of an off-diagonal step function is not unique, but the multiple Wiener-Itō integral $I_d(f)$ is well-defined (it does not depend on the representation of f). Again, $I_d(f)$ is linear on the vector space of off-diagonal step functions.

Now note that we can write the diagonal set D as $D = \bigcup_{1\le i\ne j\le d}(D\cap\{t_i = t_j\})$, in other words: D is a finite union of intersections of D with $(d-1)$-dimensional hyperplanes. Thus, D is a set of the Lebesgue measure 0. This allows us to approximate a function on $[a,b]^d$ by step functions on $[a,b]^d\backslash D$, i.e. by off-diagonal step functions, because we can cover $[a,b]^d$ with rectangles that come arbitrarily close to the diagonal set without touching it. So we summarize: For each $f\in L^2([a,b]^d)$ we can find a sequence $(f_n)_n$ of off-diagonal step functions such that

$$\lim_{n\to\infty}\int_{[a,b]^d} |f(t_1,\dots,t_d) - f_n(t_1,\dots,t_d)|^2\, dt_1\, dt_2\cdots dt_d = 0.$$

Definition A.6 (Symmetrization)**.** Given a function $f(t_1,\dots,t_d)$, the *symmetrization* of f is

$$\hat f(t_1,\dots,t_d) = \frac{1}{d!}\sum_\sigma f(t_{\sigma(1)},\dots,t_{\sigma(d)}),$$

where the sum is to be taken over all permutations σ of $\{1,2,\dots,d\}$.

If f is an off-diagonal step function, $\hat f$ is it as well. Since the Lebesgue measure is symmetric, one can show that the multiple Wiener-Itō integrals based on f and $\hat f$ coincide: $I_d(f) = I_d(\hat f)$. Moreover,

$$E\,[I_d(f)]^2 = d!\int_{[a,b]^d} |\hat f(t_1,\dots,t_d)|^2\, dt_1\, dt_2\cdots dt_d.$$

With this, we can proove for the above sequence $(f_n)_n$ of off-diagonal step functions that $(I_d(f_n))_n$ is a Cauchy sequence in $L^2(\Omega)$ and therefore converging. The limit does not not depend on the coice of the approximating sequence, so we can give

Definition A.7 (Multiple Wiener-Itō integral)**.** For a function $f\in L^2([a,b]^d)$, the *multiple Wiener-Itō integral* $I_d(f)$ is defined by

$$I_d(f) = \int_{[a,b]^d} f(t_1,\dots,t_d)\, dB_{t_1}\, dB_{t_2}\cdots dB_{t_d} := \lim_{n\to\infty} I_d(f_n),$$

where $(f_n)_n$ is a sequence of off-diagonal step functions approximating f and the convergence is in $L^2(\Omega)$.

Multiple Wiener-Itō integrals have the following properties:

- It does not matter, if we integrate over f or its symmetrization $\hat{f}$: $I_d(f) = I_d(\hat{f})$.
- $I_d(f)$ is a random variable with expectation $E[I_d(f)] = 0$ and variance $E[I_d(f)]^2 = d!\|\hat{f}\|^2$, where $\|\cdot\|$ is the norm on $L^2([a,b]^d)$.
- $I_1(f)$ is the simple Wiener integral.
- Multiple Wiener-Itō integrals of different orders are orthogonal: For $n \neq m$ and for any $f \in L^2([a,b]^n)$, $g \in L^2([a,b]^m)$ it holds $E\left[I_n(f)I_m(g)\right] = 0$.
- To compute a multiple Wiener-Itō integral, it can be written as an iterated Itō integral:

$$\begin{aligned}&\int_{[a,b]^d} f(t_1,\ldots,t_d)\, dB_{t_1}\, dB_{t_2}\cdots dB_{t_d}\\ &= d!\int_a^b \cdots \int_a^{t_{d-2}} \left(\int_a^{t_{d-1}} \hat{f}(t_1,\ldots,t_d)\, dB_{t_d}\right) dB_{t_{d-1}}\cdots dB_{t_1}\end{aligned}$$

A.4.1 Relation between Hermite polynomials and multiple Wiener-Itō integrals

We will shortly illuminate the relation between Hermite polynomials and multiple Wiener-Itō integrals. Let $(\Omega, \mathcal{F}, P)$ be a probability space and $B(t)$ a Brownian motion with respect to P. Note that for a function $f \in L^2([a,b])$, the Wiener integral $\int_a^b f(t)\, dB(t)$ is measurable with respect to the sigma field

$$\mathcal{F}_B := \sigma\left\{B(t) \,|\, a \leq t \leq b\right\}$$

which is smaller than $\mathcal{F}$ and in general not equal. Now let $L^2_B(\Omega) \subset L^2(\Omega)$ denote the Hilbert space of P-square integrable functions on Ω which are mesurable with respect to $\mathcal{F}_B$.

Define the *tensor product* of functions $f_1,\ldots,f_k \in L^2([a,b])$ as

$$f_1 \otimes \cdots \otimes f_k(t_1,\ldots,t_k) := f_1(t_1)\cdots f_k(t_k).$$

The notation $g_1^{\otimes n_1} \otimes \cdots \otimes g_k^{\otimes n_k}$ means that g_j is repeated n_j times $(1 \leq j \leq k)$; it is a tensor product of $n_1 + \ldots + n_k$ factors. Now the Wiener-Itō integral of the tensor product of $f_1,\ldots,f_k$ can be calculated as the product of some Hermite polynomials of Wiener-Itō integrals of the single f_j.

Theorem A.3. *Let $f_1,\ldots,f_k$ be non-zero orthogonal functions in $L^2([a,b])$ and $n_1,\ldots,n_k \in \mathbb{N}$. Set $n := n_1 + \ldots + n_k$. Then*

$$\boxed{I_n\left(f_1^{\otimes n_1} \otimes \cdots \otimes f_k^{\otimes n_k}\right) = \prod_{j=1}^{k} H_{n_j}\left(I(f_j); \|f_j\|^2\right).}$$

And with this, one can show: Square integrable functions on Ω that are mesurable with respect to $\mathcal{F}_B$ can be represented as sum of multiple Wiener-Itō integrals:

Theorem A.4 (Wiener-Itō Theorem)**.** *The space $L_B^2(\Omega)$ can be decomposed into the orthogonal direct sum*

$$L_B^2(\Omega) = K_0 \oplus K_1 \oplus K_2 \oplus \dots,$$

where K_j consists of multiple Wiener-Itō integrals of order j. Each $f \in L_B^2(\Omega)$ has a unique representation

$$f = \sum_{j=0}^{\infty} I_j(f_j)$$

with certain $f_j \in L_{sym}^2([a,b]^j)$, the real Hilbert space of symmetric square integrable functions on $[a,b]^j$. Because of the orthogonality of Wiener-Itō integrals of different order, we have

$$\|f\|^2 = \sum_{j=0}^{\infty} j!\, \|f_j\|^2.$$

A proof and some comments on how to construct the before mentioned f_j are given by Kuo (2006, Chap. 9.7).

Appendix B

Additions

B.1 Proof of Theorem 2.3

As the X_i are jointly Gaussian with expectation 0, the numerator $\bar{X} - \bar{Y}$ is an affine linear transformation and therefore normally distributed with expectation 0 as well. We only need to investigate its asymptotic variance.

For this purpose, we decompose the variance into three parts – one containing only the variances of the first sample, one only containing these of the second sample and one third including the interdependencies between both:

$$\begin{aligned}
\operatorname{Var}[\bar{X} - \bar{Y}] &= \frac{1}{m^2} \operatorname{Var}\left[\sum_{i=1}^{m} X_i\right] + \frac{1}{n^2} \operatorname{Var}\left[\sum_{i=m+1}^{m+n} X_i\right] - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=m+1}^{m+n} \operatorname{Cov}[X_i, X_j] \\
&= \frac{1}{m^2} \operatorname{Var}\left[\sum_{i=1}^{m} X_i\right] + \frac{1}{n^2} \operatorname{Var}\left[\sum_{i=m+1}^{m+n} X_i\right] - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=m+1-i}^{m+n-i} \gamma_j \qquad \text{(B.1)} \\
&=: A + B + C
\end{aligned}$$

Now we look into the convergence of each summand, starting with the last and most difficult one.

Ad C. To detect the limiting behaviour we will replace some terms by other ones that are asymptotically equivalent and easier to handle. For keeping an overview we will at first show these replacements and give the detailed justification afterwards. During

the whole proof, all limits and asymptotic equivalences apply to an increasing overall sample size $N \to \infty$ (and thus $m, n \to \infty$ as well).

$$\begin{aligned}
\sum_{i=1}^{m} \sum_{j=m+1-i}^{m+n-i} \gamma_j &= \sum_{i=1}^{m} \left(\sum_{j=1}^{m+n-i} \gamma_j - \sum_{j=1}^{m-i} \gamma_j \right) \\
&\sim \sum_{i=1}^{m} \Big(\frac{c}{\Gamma(2-D)} (m+n-i)^{1-D} L(m+n-i) \qquad \text{(B.2)} \\
&\qquad - \frac{c}{\Gamma(2-D)} (m-i)^{1-D} L(m-i) \Big) \\
&= \sum_{k=1}^{m+n-1} \frac{c}{\Gamma(2-D)} k^{1-D} L(k) - \sum_{k=1}^{n-1} \frac{c}{\Gamma(2-D)} k^{1-D} L(k) \qquad \text{(B.3)} \\
&\qquad - \sum_{k=0}^{m-1} \frac{c}{\Gamma(2-D)} k^{1-D} L(k) \\
&\sim \frac{c}{\Gamma(3-D)} \Bigg((m+n-1)^{2-D} L(m+n-1) \qquad \text{(B.4)} \\
&\qquad - (n-1)^{2-D} L(n-1) - (m-1)^{2-D} L(m-1) \Bigg)
\end{aligned}$$

In (B.3) we have expanded the sum and changed the index of summation (in the first sum $m+n-i=k$, in the second sum $m-i=k$). Next we factor out $(m+n)^{2-D}$ and substitute $m = \lambda N$ and $n = (1-\lambda)N$.

$$\begin{aligned}
\sum_{i=1}^{m} \sum_{j=m+1-i}^{m+n-i} \gamma_j &\sim \frac{c}{\Gamma(3-D)} (m+n)^{2-D} \Bigg(\left(\frac{N-1}{N} \right)^{2-D} L(N-1) \\
&\qquad - \left(\frac{n-1}{N} \right)^{2-D} L(n-1) - \left(\frac{m-1}{N} \right)^{2-D} L(m-1) \Bigg) \\
&\sim \frac{c}{\Gamma(3-D)} N^{2-D} L(N) \left(1 - \lambda^{2-D} - (1-\lambda)^{2-D} \right) \qquad \text{(B.5)}
\end{aligned}$$

Now we give the promised justification of the asymptotic equivalences above. Look at step (B.2). We apply Lemma 2.1 (iii) with

$$\begin{aligned}
a_{i,m} &= \sum_{j=1}^{m+n-i} \gamma_j - \sum_{j=1}^{m-i} \gamma_j \\
\alpha_{i,m} &= \frac{c}{\Gamma(2-D)} \left((m+n-i)^{1-D} L(m+n-i) - (m-i)^{1-D} L(m-i) \right) \\
&= g_{m+n-i} - g_{m-i}.
\end{aligned}$$

We have

$$\sum_{i=1}^{m-1}(a_{i,m}-a_{i,m-1})+a_{m,m}=\sum_{i=1}^{m-1}\gamma_{m+n-i}-\sum_{i=1}^{m-1}\gamma_{m-i}+\sum_{i=1}^{n}\gamma_i$$
$$=\sum_{i=n+1}^{m+n-1}\gamma_i-\sum_{i=n+1}^{m-1}\gamma_i=\sum_{i=m}^{m+n-1}\gamma_i$$

and

$$\sum_{i=1}^{m-1}(\alpha_{i,m}-\alpha_{i,m-1})+\alpha_{m,m}=\sum_{i=1}^{m-1}(g_{m+n-i}-g_{m-i}-g_{m+n-i-1}-g_{m-i-1})+g_n$$
$$=\sum_{i=1}^{m-1}(g_{n+i}-g_i)-\sum_{i=0}^{m-2}(g_{n+i}-g_i)+g_n$$
$$=g_{m+n-1}-g_{m-1}.$$

We have to verify that these two expressions are asymptotically equivalent as m increases. By (2.6) in Lemma 2.2

$$\sum_{i=1}^{m+n-1}\gamma_i\sim\frac{c(m+n-1)^{1-D}}{\Gamma(2-D)}L(m+n-1)\quad\text{and}\quad\sum_{i=1}^{m-1}\gamma_i\sim\frac{c(m-1)^{1-D}}{\Gamma(2-D)}L(m-1),$$

and so by Lemma 2.1 (i),

$$\sum_{i=m}^{m+n-1}\gamma_i\sim\frac{c}{\Gamma(2-D)}\left((m+n-1)^{1-D}L(m+n-1)-(m-1)^{1-D}L(m-1)\right)$$
$$=g_{m+n-1}-g_{m-1},$$

because

$$\frac{(m+n-1)^{1-D}L(m+n-1)}{(m-1)^{1-D}L(m-1)}\sim\left(\frac{N}{\lambda N}\right)^{1-D}\frac{L(N)}{L(\lambda N)}\to\left(\frac{1}{\lambda}\right)^{1-D}>1.$$

It remains to verify that $\sum_{k=1}^{m}D_{k,m}$ is unbounded and strictly increasing. We shall see that from some point the increments stay positive.

$$\sum_{i=1}^{m}\alpha_{i,m}-\sum_{i=1}^{m-1}\alpha_{i,m-1}$$
$$=g_n+\sum_{i=1}^{m-1}(g_{m+n-i}-g_{m-i}-g_{m-1+n-i}+g_{m-1-i})$$
$$=g_{m+n-1}-g_{m-1}$$
$$=\frac{c}{\Gamma(2-D)}\left((m+n-1)^{1-D}L(m+n-1)-(m-1)^{1-D}L(m-1)\right)$$
$$=\frac{c}{\Gamma(2-D)}(N-1)^{1-D}L(N-1)\left(1-\left(\frac{\lambda N-1}{N-1}\right)^{1-D}\frac{L(\lambda N-1)}{L(N-1)}\right)$$

Trivially, the first three factors are positive. The factor in brackets converges to $1-\lambda^{1-D}>0$, so it is positive for large N. Thus the sum in (B.2) is monotonic from a certain index foreward. And we can see that the sum is unbounded: Its increments behave, at least from a certain index foreward and except a positive factor, like $N^{1-D}L(N)$, in other words they are increasing.

Now look at step (B.4). The asymptotic equivalence of each single term is nothing else than (2.6) with $D-1$ instead of D (as noted in the proof, Karamata's theorem still holds as long $D-1>-1$); the asymptotic equivalence of all differences follows similarly as before from Lemma 2.1 (i), since we have for the first difference

$$\frac{(N-1)^{2-D}L(N-1)}{((1-\lambda)N-1)^{2-D}L((1-\lambda)N-1)}\sim\left(\frac{1}{1-\lambda}\right)^{2-D}\frac{L(N)}{L((1-\lambda)N)}\to\left(\frac{1}{1-\lambda}\right)^{2-D}>1$$

and for the second difference

$$\frac{(N-1)^{2-D}L(N-1)-((1-\lambda)N-1)^{2-D}L((1-\lambda)N-1)}{(\lambda N-1)^{2-D}L(\lambda N-1)}\sim\frac{1-(1-\lambda)^{2-D}}{\lambda^{2-D}}$$

and recalling that $D,\lambda\in(0,1)$ we can bound this as follows:

$$\frac{1-(1-\lambda)^{2-D}}{\lambda^{2-D}}>\frac{1}{\lambda^2}-\left(\frac{1}{\lambda}-1\right)^2=\frac{2}{\lambda}-1>2-1=1.$$

Finally we deal with step (B.5). We have factored out $L(N)$ and we have replaced $L(N-1)/L(N)$, $L(n-1)/L(N)$ and $L(m-1)/L(N)$ by 1 (this is equivalent by Lemma 2.2 and since L is slowly varying). To ensure that we indeed are allowed to make these replacements in the differences, we bring once more lemma 2.1 (i) into action: $1/\lambda^{2-D}>1$ for the first difference is trivial, and for the second difference we have

$$\frac{1-\lambda^{2-D}}{(1-\lambda)^{2-D}}>\frac{1-\lambda^{2-D}}{1-\lambda}>\frac{1-\lambda}{1-\lambda}=1.$$

Ad A. Now that we coped with the covariances we confidently can tackle the variances of $\bar{X}$ and $\bar{Y}$.

$$\begin{aligned}\operatorname{Var}\left[\sum_{i=1}^m X_i\right]&=\sum_{i=1}^m\operatorname{Var}[X_i]+\sum_{1\le i\ne j\le m}\operatorname{Cov}[X_i,X_j]\\&=m\gamma_0+2\sum_{k=1}^{m-1}(m-k)\gamma_k\end{aligned}$$

Clearly, $\sum_{k=1}^{m-1}(m-k)\gamma_k=\sum_{k=1}^m m\gamma_k-\sum_{k=1}^m k\gamma_k$, and by (2.6) and by going through the proof of Lemma 2.2 (ii), we obtain

$$\sum_{k=1}^m m\gamma_k\sim cm^{2-D}\frac{L(m)}{\Gamma(2-D)},\qquad\sum_{k=1}^m k\gamma_k\sim cm^{2-D}\frac{L(m)}{(2-D)\Gamma(1-D)}.$$

Since

$$cm^{2-D}\frac{L(m)}{\Gamma(2-D)}\Big/cm^{2-D}\frac{L(m)}{(2-D)\Gamma(1-D)}=\frac{2-D}{1-D}>\frac{1}{1-D}>1,$$

we obtain with Lemma 2.1 (i)

$$\sum_{k=1}^{m-1}(m-k)\gamma_k\sim cm^{2-D}\frac{L(m)}{\Gamma(2-D)}\left(1-\frac{1-D}{2-D}\right)=cm^{2-D}\frac{L(m)}{\Gamma(3-D)}. \tag{B.6}$$

Now it is not a surprise that we apply Lemma 2.1 (i) one last time to detect finally the asymptotic equivalence of $m\gamma_0+2\sum_{k=1}^{m-1}(m-k)\gamma_k$. $\Gamma(3-D)\gamma_0\big/2cm^{1-D}L(m)\to 0\neq 1$ is self-evident, so in the end we receive

$$\operatorname{Var}\left[\sum_{i=1}^{m}X_i\right]\sim m\gamma_0+2cm^{2-D}\frac{L(m)}{\Gamma(3-D)}. \tag{B.7}$$

Ad B. Calculating the asymptotic variance of the sum of the Y's is exactly the same as we just have done, so just replace m by n in the result:

$$\begin{aligned}\operatorname{Var}\left[\sum_{i=m+1}^{m+n}X_i\right]=\operatorname{Var}\left[\sum_{i=1}^{n}X_{i+m}\right]&=n\gamma_0+2\sum_{k=1}^{n-1}(n-k)\gamma_k\\&\sim n\gamma_0+2cn^{2-D}\frac{L(n)}{\Gamma(3-D)}\end{aligned} \tag{B.8}$$

Now we look back on the decomposition (B.1) and put together the results on the single summands (B.5), (B.7) and (B.8) to establish the asymptotic (expediently scaled) variance of $\bar{X}-\bar{Y}$:

$$\begin{aligned}&\lim_{N\to\infty}\frac{mn}{(m+n)^{2-D}}\frac{1}{L(m+n)}\operatorname{Var}\left[\bar{X}-\bar{Y}\right]\\&\quad=\lim_{N\to\infty}\frac{mn}{(m+n)^{2-D}}\frac{1}{L(m+n)}\left[\frac{1}{m^2}m\gamma_0+\frac{1}{m^2}2cm^{2-D}\frac{L(m)}{\Gamma(3-D)}+\frac{1}{n^2}n\gamma_0\right.\\&\qquad\left.+\frac{1}{n^2}2cn^{2-D}\frac{L(n)}{\Gamma(3-D)}-\frac{2}{mn}\frac{c}{\Gamma(3-D)}N^{2-D}L(N)\left(1-\lambda^{2-D}-(1-\lambda)^{2-D}\right)\right]\\&\quad=\lim_{N\to\infty}\frac{\lambda(1-\lambda)}{N^{-D}L(N)}\left[\frac{\gamma_0}{\lambda N}+2c\lambda^{-D}N^{-D}\frac{L(\lambda N)}{\Gamma(3-D)}+\frac{\gamma_0}{(1-\lambda)N}\right.\\&\qquad+2c(1-\lambda)^{-D}N^{-D}\frac{L((1-\lambda)N)}{\Gamma(3-D)}\\&\qquad\left.-\frac{2c}{\lambda(1-\lambda)}\frac{1}{\Gamma(3-D)}N^{-D}L(N)\left(1-\lambda^{2-D}-(1-\lambda)^{2-D}\right)\right]\\&\quad=\frac{2c}{\Gamma(3-D)}\lambda(1-\lambda)\left(\lambda^{-D}+(1-\lambda)^{-D}-\frac{1-\lambda^{2-D}-(1-\lambda)^{2-D}}{\lambda(1-\lambda)}\right)\end{aligned}$$

This completes the proof.

B.2 An heuristic example illustrating Theorem 2.3

We now make a check on Theorem 2.3 with an heuristic example. We consider $\gamma_k = k^{-D}$ with $D \in (0,1)$ and disregard that this is actually not an auto-covariance function (see the technical remark on page 25). We will calculate the asymptotic variance of $\sqrt{\frac{mn}{(m+n)^{2-D}}}(\bar{X}-\bar{Y})$ elementarily[1]. $\gamma_k = \gamma(k)$ is monotonic decreasing and non-negative, and a well known result from introductory analysis for decreasing $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is

$$\boxed{\int_1^{n+1} f(x)\,dx \leq \sum_{i=1}^{n} f(i) \leq \int_1^n f(x)\,dx + f(1).}$$

We will use this to replace sums by integrals (afterwards we ensure that the remainders vanish asymptotically). As in the proof of Theorem 2.3, we analyse the single terms in the decomposition (B.1).

$$\begin{aligned}
\sum_{i=1}^{m} \sum_{j=m+1-i}^{m+n-i} \gamma_j &= \sum_{i=1}^{m} \sum_{j=m+1-i}^{m+n-i} j^{-D} && \text{(B.9)}\\
&\approx \sum_{i=1}^{m} \left(\int_{m+1-i}^{m+n-i} j^{-D}\,dj \right) && \text{(B.10)}\\
&\approx \frac{1}{1-D} \int_{i=1}^{m} \left((m+n-i)^{1-D} - (m+1-i)^{1-D} \right) di && \text{(B.11)}\\
&\approx \frac{1}{1-D} \int_{i=0}^{m} \left((m+n-i)^{1-D} - (m+1-i)^{1-D} \right) di && \text{(B.12)}\\
&= \frac{(m+n)^{2-D}}{(2-D)(1-D)} \left(1 - \left(\frac{m+1}{m+n} \right)^{2-D} - \left(\frac{n}{m+n} \right)^{2-D} + \left(\frac{1}{m+n} \right)^{2-D} \right)
\end{aligned}$$

It is easy to see that

$$\begin{aligned}
\gamma(m+1-i) &= (m+1-i)^{-D} = o\left((m+n)^{2-D} \right)\\
\int_{m+n-i}^{m+n-i+1} j^{-D}\,dj &= \frac{1}{1-D} \left((m+n-i+1)^{1-D} - (m+n-i)^{1-D} \right) = o\left((m+n)^{2-D} \right),
\end{aligned}$$

so in (B.10) really holds "=" for large n (that is what we denoted by $\approx$), and the same counts for (B.11). In (B.12) we have omitted the rest integral

$$\begin{aligned}
&\int_0^1 \left((m+n-i)^{1-D} - (m+1-i)^{1-D} \right) di\\
&\quad = \frac{(m+n)^{2-D}}{2-D} \left(1 + \left(\frac{m}{m+n} \right)^{2-D} - \left(\frac{m+1}{m+n} \right)^{2-D} - \left(\frac{m+n-1}{m+n} \right)^{2-D} \right)\\
&\quad = \frac{(m+n)^{2-D}}{2-D} o(1),
\end{aligned}$$

[1] We know by know that $\sqrt{nm/(n+m)^{2-D}}$ is the right scaling; but if we did not, we could sense it in the course of the calculation.

so we have for the covariance term in (B.1):

$$\begin{aligned}
&\lim_{n\to\infty} \frac{mn}{(m+n)^{2-D}} \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=m+1-i}^{m+n-i} \gamma_j \\
&\quad = \lim_{n\to\infty} \frac{1}{(1-\frac{D}{2})(1-D)} \left(1 - \left(\frac{m+1}{m+n}\right)^{2-D} - \left(\frac{n}{m+n}\right)^{2-D} + \left(\frac{1}{m+n}\right)^{2-D}\right) \\
&\quad = \frac{1}{(1-\frac{D}{2})(1-D)} \left(1 - \lambda^{2-D} - (1-\lambda)^{2-D}\right) \qquad \text{(B.13)}
\end{aligned}$$

To detect the limiting behaviour of the two variance terms in (B.1), we use[2] (1.8).

$$\begin{aligned}
\lim_{N\to\infty} \frac{mn}{(m+n)^{2-D}} \frac{1}{m^2} \operatorname{Var}\left[\sum_{i=1}^{m} X_i\right] &= \lim_{N\to\infty} \frac{\lambda(1-\lambda)}{N^{-D}} (\lambda N)^{-D} \frac{2}{(1-D)(2-D)} \\
&= \lambda^{1-D}(1-\lambda) \frac{2}{(1-D)(2-D)} \qquad \text{(B.14)}
\end{aligned}$$

$$\lim_{N\to\infty} \frac{mn}{(m+n)^{2-D}} \frac{1}{n^2} \operatorname{Var}\left[\sum_{i=1}^{n} X_{m+i}\right] = \lambda(1-\lambda)^{1-D} \frac{2}{(1-D)(2-D)} \qquad \text{(B.15)}$$

We now put together (B.13), (B.14), (B.15):

$$\begin{aligned}
&\frac{mn}{(m+n)^{2-D}} \operatorname{Var}\left[\bar{X} - \bar{Y}\right] \\
&\quad \overset{N\to\infty}{\longrightarrow} \frac{1}{(1-\frac{D}{2})(1-D)} \left(\lambda^{1-D}(1-\lambda) + \lambda(1-\lambda)^{1-D} - 1 + \lambda^{2-D} + (1-\lambda)^{2-D}\right)
\end{aligned}$$

Alternatively, we may apply Theorem 2.3 with $c = \Gamma(1-D)$ and $L(k) \equiv 1$. The result is obviously the same.

B.3 Bounded variation in higher dimensions

In Chapter 3 we derived the limit distribution of the Wilcoxon two-sample test statistic

$$W_{[n\lambda],n} = \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} I_{\{X_i \le X_j\}}, \quad 0 \le \lambda \le 1.$$

Since $W_{[n\lambda],n}$ is a U-statistic, one could try to adapt the technique of Dehling and Taqqu (1989) who treat one-sample U-statistics of LRD data in order to handle $W_{[n\lambda],n}$; but this approach fails since the kernel $h(x,y) = I_{\{x<y\}}$ does not have bounded variation – which is an essential technical requirement for the technique.

In this chapter, we will deal with the question what bounded variation is in higher dimensions since it is of general interest, even though it does not lead to the desired change-point test; we shall see that this is not trivial. At the end, we prove that $h(x,y) = I_{\{x<y\}}$ has (even locally) infinite variation.

[2] To draw on an existing limit theorem does of course not seem to be an "elementary" approach as promised, but we do this only for reasons of speed; we could deduce the results in the same elementary way as we have treated the rest of the calculation before.

B.3.1 Definition, properties and examples

Let f be a real valued function defined on a bounded interval $[a, b] \subset \mathbb{R}$. The total variation of f is defined by

$$\boxed{\|f\|_{V,[a,b]} := \sup \sum_{i=1}^{p} |f(a_i) - f(a_{i-1})|,}$$

where the supremum is taken over all finite sets of points $\{a_i \,|\, 0 \leq i \leq p < \infty\}$ which define a partition of the interval $[a, b]$, i.e. with $a = a_0 < \ldots < a_p = b$. If the underlying interval $[a, b]$ is clear, we shorty write $\|f\|_V$ instead of $\|f\|_{V,[a,b]}$. If $\|f\|_V < \infty$ then f is said to have bounded variation.

This one-dimensional variation has some nice properties (Kannan and Krueger (1996, Chap. 6) provide proofs and some more results):

- If f is monotone, then $\|f\|_V = |f(b) - f(a)| < \infty$.
- If f is Lipschitz continuous, then $\|f\|_V < \infty$.
- If f has a bounded derivative, then $\|f\|_V = \int_a^b |f'(x)|\, dx < \infty$.
- Sums, multiples and products of functions of bounded variation have itself bounded variation: For functions f, g on $[a, b]$ and a scalar α it holds $\|f + g\|_V \leq \|f\|_V + \|g\|_V$, $\|\alpha f\|_V = |\alpha| \|f\|_V$, and if $\|f\|_V, \|g\|_V < \infty$, then $\|fg\|_V < \infty$ (and the same is true for f/g as long as $|g(x)| \geq c > 0$ for all $x \in [a, b]$).
- The variation is additive: If $[a, c]$ is divided by $b \in (a, c)$, then $\|f\|_{V,[a,c]} = \|f\|_{V,[a,b]} + \|f\|_{V,[b,c]}$.
- f is of bounded variation, if and only if it is the difference of two increasing, bounded functions.
- If f has has bounded variation, then the left-hand limits $f(x_-)$ for $x \in (a, b]$ and the right-hand limits $f(x_+)$ for $x \in [a, b)$ exist, f can have at most a countable number of discontinuities, f' exists and is finite a.e..

The question is now how to define bounded variation for functions $f : [a, b] \subset \mathbb{R}^d \to \mathbb{R}$, $d \geq 2$, and which properties hold for this higher-dimensional variation. Unfortunately, there is no obvious way to extend the definition of bounded variation to several variables: Clarkson and Adams (1933) for example list six different definitions of bounded variation for $d = 2$. We will take a closer look at two of them which proved to be useful in the context of integration and measure; the first one is connected with the names of Vitali, Lebesgue, Fréchet and de la Vallée Poussin, the second one with the names of Hardy and Krause. Owen (2004) discusses both definitions in detail in a general setting (but with a notation that needs getting used to).

For $a = (a_1, \ldots, a_d), b = (b_1, \ldots, b_d) \in \mathbb{R}^d$ write $a \leq b$, if the inequality holds for all components (likewise define $a < b$). Define the hyperrectangle $[a, b] \subset \mathbb{R}^d$ as follows

$$[a, b] := \{x \in \mathbb{R}^d \,|\, a \leq x \leq b\}$$

(and likewise $(a, b]$, $[a, b)$ and (a, b)). For $f : [a, b] \subset \mathbb{R}^d \to \mathbb{R}$ define the d-increment of f over the rectangle $R := [a, b]$ as

$$\Delta_R f := \sum_{I \subset \{1,\ldots,d\}} (-1)^{|I|} f(x_I),$$

where $x_I \in \mathbb{R}^d$ has the components

$$(x_I)_i = \begin{cases} a_i & i \in I \\ b_i & i \notin I \end{cases}.$$

Note that $\emptyset \subset \{1, \ldots d\}$ always. In case $d = 1$ this definition reduces to

$$\Delta_R f = f(b) - f(a) = f(b_1) - f(a_1),$$

in case $d = 2$ to

$$\Delta_R f = f(b_1, b_2) - f(b_1, a_2) - f(a_1, b_2) + f(a_1, a_2).$$

There are several ways to note down d-increments. Young (1916) for example defines them recursively:

$$\begin{aligned} [f]_{x_1=a_1}^{b_1} &= f(b_1, x_{2,}, \ldots, x_d) - f(a_1, x_{2,}, \ldots, x_d) \\ [f]_{x_1=a_1, x_2=a_2}^{b_1, b_2} &= [f(b_1, x_{2,}, \ldots, x_d) - f(a_1, x_{2,}, \ldots, x_d)]_{x_2=a_2}^{b_2} \\ &\vdots \\ [f]_{a_1,\ldots,a_d}^{b_1,\ldots,b_d} &= \Delta_R f \end{aligned}$$

If it is clear over which rectangle the increments have to be computed, we will shortly write Δf instead of $\Delta_R f$.

Definition B.1 (Monotonicity)**.** A function $f : \mathbb{R}^d \to \mathbb{R}$ is called *monotone*, if $\Delta_R f \geq 0$ for all d-dimensional rectangles R.

Example. We consider some functions on the first quadrant $\mathbb{R}_+ \times \mathbb{R}_+$. By intuition we expect $f_1(x, y) = x^2 y$ to be monotone because the functions $x \mapsto x^2$ and $y \mapsto y$ have this property, and we expect $f_2(x, y) = -y \sin(x)$ not to be monotone because the sinus possesses hills and valleys – an educated guess. We compute the increments over the rectangle $[a_1, b_1] \times [a_2, b_2]$ in the first quadrant $\mathbb{R}_+ \times \mathbb{R}_+$ to verify this. The increment

$$\begin{aligned} \Delta f_1 &= f_1(b_1, b_2) - f_1(b_1, a_2) - f_1(a_1, b_2) + f_1(a_1, a_2) \\ &= (b_1 - a_1)(b_2 - a_2)(a_1 + b_1) \end{aligned}$$

is positive, so f_1 is in fact monotone. The (sign of the) increment

$$\Delta f_2 = (b_2 - a_2)(\sin(a_1) - \sin(b_1))$$

depends on the underlying rectangle (effectively, only the x-cordinates are crucial), $\sin(a_1) - \sin(b_1)$ can be negative, and so Δf_2 can be negative. f_2 is therefore not monotone.

But in general, more-dimensional monotonicity is not intuitively visible.

Example. We consider some monotone functions and take a closer look at their one- and two-dimensional increments[3] over the same rectangle as before. We will see that their behaviour is quite different.

a) For $f_1(x, y) = x^2 y$ we have

$$\begin{aligned}\Delta f_1 &= (b_1 - a_1)(b_2 - a_2)(a_1 + b_1) > 0\\ \Delta_x f_1 &= f_1(b_1, b_2) - f_1(a_1, b_2) = b_2(b_1^2 - a_1^2) > 0\\ \Delta_y f_1 &= f_1(b_1, b_2) - f_1(b_1, a_2) = b_1^2(b_2 - a_2) > 0.\end{aligned}$$

All increments in all dimensions are positive.

b) $f_3(x, y) = x^4 - y^5$ does not look monotone. But it is:

$$\begin{aligned}\Delta f_3 &= 0\\ \Delta_x f_3 &= b_1^4 - a_1^4 > 0\\ \Delta_y f_3 &= a_2^5 - b_2^5 < 0\end{aligned}$$

f_3 is increasing in x and decreasing in y. Curiously, the 2-dimensional increment is always 0. $f_4(x, y) = -y/x$ is an example of a function with the same 1-dimensional properties, but an always positive 2-dimensional increment.

c) For $f_5(x, y) = e^{-x}e^{-y}$ we have

$$\begin{aligned}\Delta f_5 &= (e^{b_1} - e^{a_1})(e^{b_2} - e^{a_2})e^{-(a_1+a_2+b_1+b_2)} > 0\\ \Delta_x f_5 &= e^{-b_2}(e^{-b_1} - e^{-a_1}) < 0\\ \Delta_y f_5 &= e^{-b_1}(e^{-b_2} - e^{-a_2}) < 0.\end{aligned}$$

Although both 1-dimensional increments are negative, the 2-dimensional increment is positive.

Definition B.2 (Vitali variation)**.** For a function $f : [a, b] \subset \mathbb{R}^d \to \mathbb{R}$ let $\mathcal{R}$ be a finite set of d-dimensional rectangles that exactly cover $[a, b]$, i.e. $\mathcal{R} = \{R_i \,|\, 1 \le i \le p < \infty\}$

[3]When we consider one-dimensional increment, we will fix one argument and regard $f(x, b_2)$ and $f(b_1, y)$ as functions of one variable. There is no deeper secret behind it that we fix the arguments just to the upper bounds b_1, b_2; this is a convention in the notation of Hardy-Krause variation and it does not matter which value in $[a_i, b_i]$, $i = 1, 2$ we take.

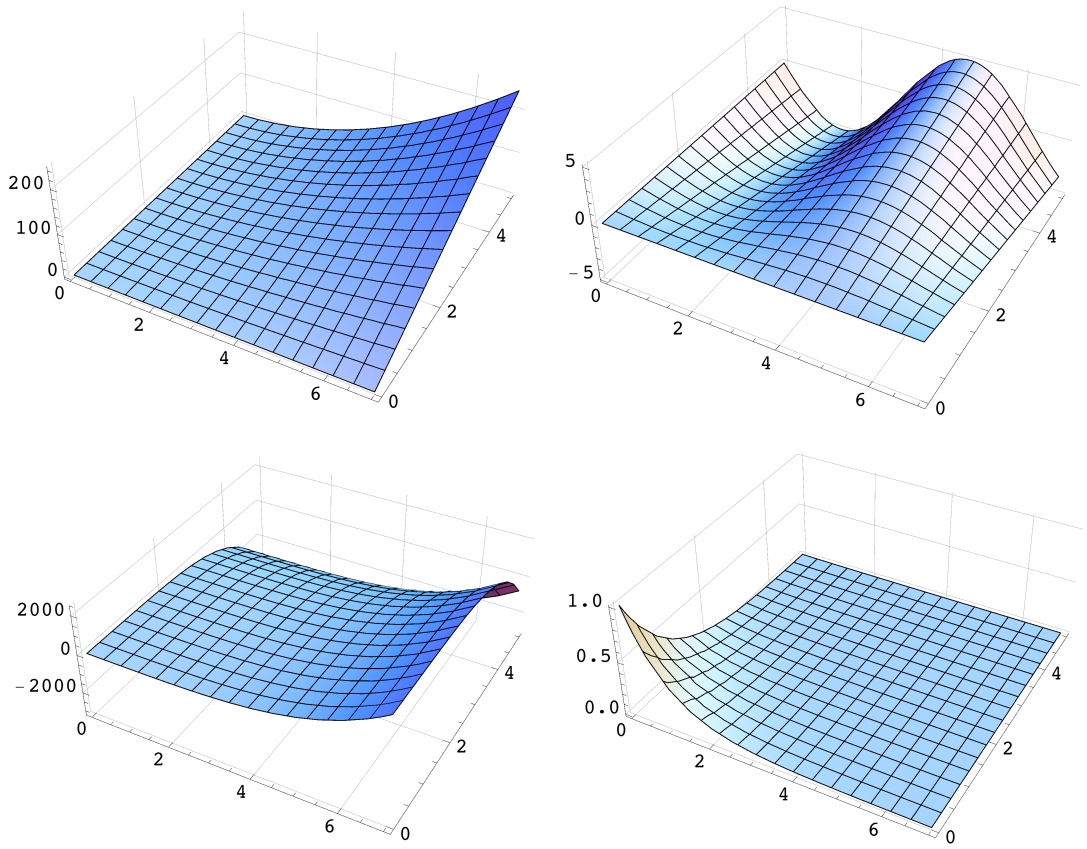


Figure B.1: Monotonicity in higher dimensions, here on the first quadrant $\mathbb{R}_+ \times \mathbb{R}_+$, is not intuitively visible. First row: $f_1(x, y) = x^2 y$ is monotone, because all increments in all dimensions are positive, but $f_2(x, y) = -y \sin(x)$ is clearly not. Second row: $f_3(x, y) = x^4 - y^5$ does not look monotone, but it is, and $f_5(x, y) = e^{-x} e^{-y}$ has negative one-dimensional increments, but is two-dimensional increasing.

such that $[a,b] = \bigcup_{i=1}^{p} R_i$ and the interiors of two different rectangles R_i, R_j are disjoint. The *Vitali variation* of f is

$$\boxed{\|f\|_{V,[a,b]} := \sup \sum_{R \in \mathcal{R}} |\Delta_R f|,} \tag{B.16}$$

where the supremum is taken over all such sets $\mathcal{R}$.

Clearly, for $d = 1$ this is the well-known standard definition of variation for real functions on the real line which we considered at the beginning.

Definition B.3 (Hardy-Krause variation)**.** Consider a function $f : [a,b] \subset \mathbb{R}^d \to \mathbb{R}$ and a non-empty subset $\emptyset \neq I \subset \{1,\dots,d\}$. Define a new function $f_I : \prod_{i \in I}[a_i, b_i] \to \mathbb{R}$ by setting the i-th argument of f equal to b_i for all $i \notin I$. The *Hardy-Krause variation* of f is

$$\boxed{\|f\|_{HK,[a,b]} := \sum_{\emptyset \neq I \subset \{1,\dots,d\}} \|f_I\|_{V,[a,b]}.} \tag{B.17}$$

For $d = 1$ the Hardy-Krause variation turns out to be the Vitali variation and therefore the standard variation for real functions on the real line. In higher dimensions both definitions differ, as we will shortly see.

Definition B.4 (Bounded variation)**.** (i) $f : [a,b] \subset \mathbb{R}^d \to \mathbb{R}$ has *bounded variation in the sense of Vitali* if

$$\|f\|_{V,[a,b]} < \infty.$$

The class of all such functions is denoted by $BV_V([a,b])$.
(ii) $f : \mathbb{R}^d \to \mathbb{R}$ has *bounded variation in the sense of Vitali*, if

$$\sup \left\{ \|f\|_{V,[a,b]} \,\middle|\, [a,b] \subset \mathbb{R}^d \right\} < \infty.$$

Analogously we define *bounded variation in Hardy-Krause sense*. Denote the class of functions with bounded Hardy-Krause variation by $BV_{HK}([a,b])$.

Remark. A function is of bounded variation in Hardy-Krause sense if the sum over all Vitali variations over $[a,b]$ and its upper faces (b_i is fixed as i-th argument) in all dimension is bounded. Owen (2004) refers to Young (1913, p. 142) who proved that this condition is equivalent to the original definition by Hardy (1905) that the sum is bounded regardless of which point $z_i \in [a_i, b_i]$ we fix as i-th argument. We will make use of this alternative definition later.

Now we collect some properties[4] of functions on $\mathbb{R}^2$ which have bounded variation in either Vitali or Hardy-Krause sense.

[4] A lot of results on bounded variation – in either way of defining it – does not hold in the same way for all dimensions d. Keep in mind that Adams and Clarkson (1934) and Clarkson and Adams (1933) only give proofs for $d = 2$ (it does not matter since we are interested just in two-dimensional results), while Owen (2004) provides general results.

Proposition B.1. *(i) BV_V and BV_{HK} are closed under $+$ and $-$.*
(ii) BV_{HK} is closed under $\cdot$ and $\div$ (as long as the quotient of two functions is defined), but BV_V is not.
(iii) If the rectangle R is split into subrectangles R_i, $i = 1, \dots k$, then

$$\boxed{f \in BV(R) \quad \Rightarrow \quad f \in BV(R_i) \quad \forall = 1, \dots k,}$$

and conversely

$$\boxed{f \in BV(R_i) \quad \forall = 1, \dots k \quad \Rightarrow \quad f \in BV\left(\bigcup_{i=1}^{k} R_i\right)}$$

as long as $\bigcup_{i=1}^{k} R_i$ is again a rectangle. This holds for both BV_V and BV_{HK}.

Proof. Owen (2004, Prop. 11) shows (i) and (ii), (iii) is demonstrated by Adams and Clarkson (1934, Th. 11 and 12). □

Proposition B.2. *Let f be defined on $[a, b] \subset \mathbb{R}^2$.*
(i)

$$\boxed{f \in BV_V \quad \Leftrightarrow \quad f = f_1 - f_2}$$

where f_1, f_2 are functions with all 2-increments non-negative: $\Delta_R f_i \geq 0$, $i = 1, 2$, for all 2-dimensional rectangles R.
(ii)

$$\boxed{f \in BV_{HK} \quad \Leftrightarrow \quad f = f_1 - f_2}$$

where f_1, f_2 are bounded functions with all increments non-negative: $\Delta_R f_i \geq 0$, $i = 1, 2$, for all rectangles R in all dimensions[5], in other words $\Delta(f_1)_I \geq 0$ and $\Delta(f_2)_I \geq 0$ for all $I \subset \{1, 2\}$. In particular

$$\boxed{f \in BV_{HK} \quad \Rightarrow \quad f \text{ is bounded.}}$$

Proof. See Adams and Clarkson (1934, Th. 5 and 6). □

These unequal representations of Vitali and Hardy-Krause variation reflect the differences between both definitions: We expect the function $f : [a_1, b_1] \times [a_2, b_2] \to \mathbb{R}$,

$$f(x, y) = \begin{cases} 1/(x - b_1)^2 & x < b_1 \\ 0 & \text{else} \end{cases},$$

not to have bounded variation, because it is not bounded. But surprisingly, it has bounded Vitali variation, since the 2-increments are always 0. By the last proposition, it can be written as $f = f - 0$, and f is not bounded. The notation of Hardy-Krause variation does not suffer from this defect, it is a stronger property.

[5] In case $d = 2$, Young (1916) calls such functions with all increments non-negative "monotonely monotone".

Proposition B.3.

$$\boxed{BV_{HK} \subset BV_V}$$

The other implication is false in general.

Proof. $f \in BV_{HK}$ can be written as difference of two bounded functions with all increments non-negative. This is of course a difference of two functions with all 2-increments non-negative, and therefore a function in BV_V.

To prove the other implication wrong consider the Dirichlet function $f(x, y) = I_{\mathbb{Q}}(x)$. Although f has unbounded variation in x (note that it is even worse: it is nowhere continuous in x), the 2-dimensional variation Δf is always 0. So $f \in BV_V$, but $f \notin BV_{HK}$. □

Such examples can easily be constructed, as the following proposition shows.

Proposition B.4. *If $f : \mathbb{R}^d \to \mathbb{R}$ does not depend on all d variables, then*

$$\boxed{\|f\|_V = 0.}$$

Proof. See Owen (2004, Prop. 8). □

Technical remark. While $BV_V \subset BV_{HK}$ is false in general, there are some further conditions we can impose on $f \in BV_V$ to be as well in BV_{HK} (Clarkson and Adams, 1933, p. 827, 841 and 846).

Proposition B.5. *$\|\cdot\|_V$ and $\|\cdot\|_{HK}$ are semi-norms on functions.*

Proof. This follows easily from the definitions. To see that they are not a norm, note that both variations vanish for constant but non-zero functions. □

B.3.2 The case $h(x, y) = I_{\{x \le y\}}$

One of the crucial points in the technique of Dehling and Taqqu (1989) is that h must be of bounded variation. This condition is necessary to ensure that some terms in the 2-dimensional integration by parts formula vanish at infinity, and above all it is necessary to define $\iint (F_m - F)(x)(G_n - G)(y)\, dh(x, y)$. Unfortunately, this requirement is not met for many examples. This is, e.g., just the case for the kernel $h(x, y) = I_{\{x \le y\}}$ which leads to the Mann-Whitney-Wilcoxon statistic. It has infinite variation (in Vitali sense, and so as well in Hardy-Krause sense), even on compact domains if they cross the diagonal $D = \{(x, y) \,|\, x = y\} \subset \mathbb{R}^2$.

To see this, consider $h(x, y) = I_{\{x \le y\}}$ on an arbitrary rectangle $R = [a, b] \subset \mathbb{R}^2$ which crosses D. It is always possible to find another rectangle $R' \subset R$ inside which touches the line D in such a way that three of its corners lie on the same side of D, see Figure B.2. The 2-increment over R' is 1, since the 2-increment of h over $[a, b]$ is

$$\Delta_R h = h(b_1, b_2) - h(b_1, a_2) - h(a_1, b_2) + h(a_1, a_2)$$

and h can take values 0 and 1. Now it is always possible to divide R' into three subrectangles with the same property. In so doing, we obtain a partition of $\tilde{R}$ with as many subrectangles R with $|\Delta_R h| = 1$ as we wish.

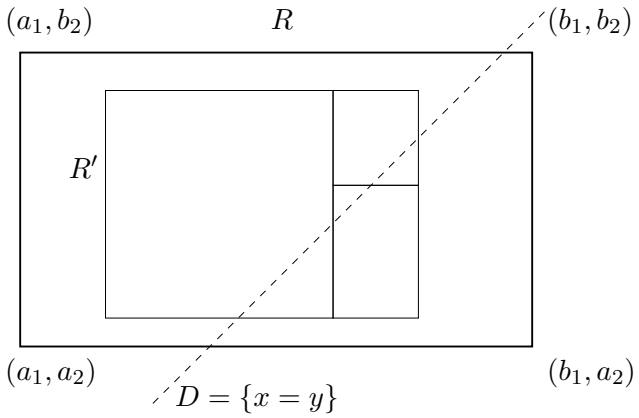


Figure B.2: $h(x, y) = I_{\{x \leq y\}}$ has infinite variation over any rectangle R crossing the diagonal, because one can easily construct a partition of R into arbitrarily many sub-rectangles on which h has a 2-increment of 1.

Appendix C

Source code of the simulations

C.1 Estimating the variance of $\bar{X}$ (section 2.4)

The routine `varBlockmeans(x,r)` estimates $\mathrm{Var}[\bar{X}]$ of a time series X (which is given as vector `x` and which has length `N`). The time series is divided in blocks of size `r`, the mean values over these blocks is calculated and as the desired estimator, the sample variance of these [`N`/`r`] block means is given out.

```
1  varBlockmeans <- function(x,r){
2     # x = given time series
3     # r = desired size of blocks
4     # r can be entered as absoulte size or relative to sample
          size N
5     N = length(x)
6
7     # stop, if r has unfeasible values
8     if(r<=0 || r>N) {
9        stop("Block␣size␣r␣must␣be␣positive␣and␣smaller␣than␣
             the␣sample␣size␣(0␣<␣r␣<␣N).")
10    }
11
12    # if block size r was entered as fractional amount of N,
         convert it
13    if(r < 1) {
14       r <- round(r*N)
15    }
16
17    # variance without Bessel's correction
18    varianz <- function(x) {n=length(x) ; var(x) * (n-1) / n}
19
20    if(r==1) {
21       warning("Block␣size␣r␣=␣1␣means:␣The␣variance␣of␣the␣
             single␣values␣in␣the␣time␣series␣is␣calculated␣(
             variance␣of␣means␣over␣bocks␣of␣length␣1).")
```

```
22        varianz(x)
23     }
24     else {
25     # divide x in blocks of length r
26     # Number of colums = integer part of N/r
27     nCols = N%/%r
28     # stop, if nCols < 2 (then variance of one single mean is
           0)
29     if(nCols < 2) {
30        stop("Only␣one␣block␣of␣the␣desired␣block␣size␣r␣could
              ␣be␣created␣from␣the␣time␣series␣x.␣The␣variance␣of
              ␣the␣block␣means␣(we␣have␣only␣one␣block!)␣is␣0.")
31     }
32
33     Y = matrix(x[1:(r * nCols)], byrow = FALSE, ncol = nCols)
34
35     # colMeans := vector of means of N%/%r columns
36     varianz(colMeans(Y))
37     }
38 }
```

C.2 Estimating the auto-covariance (section 2.5)

This is just the standard estimator for the auto-covariances of a vector `vect`, as defined in (2.18).

```
1  estimate.gamma  <- function(vect){
2     N = length(vect)
3     gamma.dach = rep(NA, N-1)
4
5     for(h in 1:(N-1)){
6        x.i = vect[1:(N-h)]
7        x.iplush = vect[(1:(N-h))+h]
8
9        # gamma.dach = 1/(N-h) sum_{i=1}^{N-h} X_{i}X_{i+h}
10       # i.e.\ sum(x.i*x.iplush)/(N-h)
11       gamma.dach[h] = mean(x.i*x.iplush)
12    }
13    gamma.dach
14 }
```

C.3 Estimating the variance of $\bar{X} - \bar{Y}$ (section 2.6)

This program puts together the two last routines, as described in (2.22). It needs as input the data sample `vect`, its Hurst parameter `HH`, the proportion `llambda` after

which the sample shall be divided, and the size `rr` of the blocks for the subroutine `varBlockmeans`.

```
estimate.varXqYq  <- function(vect, HH, llambda, rr){
   alpha = 2-2*HH
   D = alpha

   N = length(vect)
   m = floor(N*llambda)
   n = N-m

   # divide vect in m and N-m observation
   x = vect[1:m]
   y = vect[(m+1):N]

   gamma_dach = estimate.gamma(vect)

   # gamma_dach bad for large lags, so trim:
   # set to 0 after 10*log10(N) observations
   # this step may be commented out
   maxlag = min(floor(10*log10(N)), N-1)
   gamma_dach = c(gamma_dach[1:maxlag], rep(0, (N-1-maxlag))
      )

   # gamma_dach_sumoverj = sum_{j = m+1-i}^{N-i}, for i
      =1,...,m
   gamma_dach_sumoverj = rep(NA, m)

   for(i in 1:m){
      gamma_dach_sumoverj[i] = sum(gamma_dach[(m+1-i):(N-i)
         ])
   }

   estimator = (rr/m)^D * varBlockmeans(x, rr) + (rr/n)^D *
      varBlockmeans(y, rr) - 2/(m*n) * sum(gamma_dach_
      sumoverj)
   estimator
}
```

C.4 "Differences-of-means" test (section 3.6)

The following program code calculates `k` repetitions of the "difference-of-means" test statistic D_n, as defined in (3.16) for different sample lengths `N` and LRD parameter `H`. The input needed is a `N` $\times$ `k`-matrix with `k` LRD time series of length `N`. The routine is designed for time series with FGN; to apply it on other LRD series, one must adapt the constants in the scaling d_n.

```
# function divides vector vec after entry with index cut
# and gives mean(x)-mean(y)
divide.and.difference  <- function(vec, cut){
   l = length(vec)
   mean(vec[1:cut]) - mean(vec[(cut+1):l])
}

# function "calculate.Udiff" calculates U_diff =
# 1/(n dn) sum_{i=1}^{lambda n} sum_{j=lambda n + 1}^n (Xi-Xj)
# for two vectors X, Y (created by dividing one fGn series Z)
# returns vector with 10.000 values of Udiff
# needs divide.and.difference
calculate.Udiff <- function(Z, cutpt, alpha){
   # Z = given matrix of k fGn series (k columns of length N)
   # cutting point cutpt= 1,...,N-1

   # error if cutting point not an integer number
   if(cutpt - trunc(cutpt) != 0){
      stop("Cutpoint is not an integer.")
   }

   # read dimensions of Z
   N = nrow(Z)
   k = ncol(Z)

   # reserve vector U to store k values of Udiff
   U = rep(NA,k)

   # cutting point should be in {1,2, ..., N-1}
   if((cutpt >= 1) && (cutpt < N)){
      # apply "divide.and.difference" to columns of Z
      # and write k results in vector U
      U = apply(Z, 2, divide.and.difference, cut=cutpt)
   }

   # if cutpt out of range, T can't be calculated
   if((cutpt < 1) || (cutpt >= N)) {
      U = rep(NA,k)
   }

   lambd = cutpt/N

   # fGn has covariances (1-alpha)(2-alpha)/2 k^(-alpha)
```

```
44     Lk = (1-alpha)*(2-alpha)/2
45
46     # constant c in dn=c*n^(1-alpha/2) is here
47     # just: c= sqrt (1/Lk)
48
49     # U_diff = 1/(n dn) sum_{i=1}^{lambda n} sum_{j=lambda n
          + 1}^n (Xi-Xj)
50     #        = lambda (1-lambda) n^(alpha/2) c^{-1} L(n)^{-1/
          2} (mean(X)-mean(Y))
51     U = lambd*(1-lambd)*N^(alpha/2) * U
52 }
53
54 # function "make.k.times.DN" calculates
55 # k repetitions of DN = max_{1<=k<N} |Udiff|
56 # needs function calculate.Udiff
57 # output: table with k rows
58 make.k.times.DN <- function(H, N, Z){
59     alpha = 2-2*H
60
61     # read number of repetitions
62     k = ncol(Z)
63
64     # N observations => N-1 possible cutting points
65     l = N-1
66
67     # reserve marix to store Udiff
68     # l columns (for each cutting point), k=10.000 rows (
          repetitions)
69     Udiff.table = matrix(rep(NA, k*l), byrow = FALSE, ncol =
          l)
70
71     # reserve vector to store k repetitions of DN
72     k.times.DN = rep(NA, k)
73
74     # for each cutting point index i=1,...,N
75     for(i in 1:l){
76        # calculate Udiff
77        Udiff.table[,i] = calculate.Udiff(Z[1:N,], i, alpha)
78     }
79
80     # now we have an lxk matrix Udiff.table
81     # go through rows and find max(abs(..))
82
83     # function finds max(abs(vec)) of vector vec
84     # ignore "NA"-entries!
85     find.maxabs  <- function(vec){
```

```
86          max(abs(vec), na.rm=TRUE)
87      }
88
89      # apply "find.maxabs" to rows of Udiff.table
90      k.times.DN = apply(Udiff.table,1, find.maxabs)
91
92      # result: vector of k repetitions of DN = sup |Udiff|
93      return(k.times.DN)
94  }
```

The workflow to gain simulations for different parameters `H`, is easy: For `H=0.7` e.g., just load the file `fgn H0.7 N2000 k10000.txt` which contains 10,000 time series of fGn with `H=0.7` and length `N=2000`.

```
96  # call make.k.times.DN for different N and H
97  lets.go.niveau  <- function(H){
98      alpha = 2-2*H
99
100     # read k*N-matrix with fGn (k columns of length N)
101     Z.all = read.table(paste("fgn␣H", H , "␣N2000␣k10000.txt"
            , sep=""))
102     # for each length N
103     for(N in c(10,50,100,500,1000)){
104         ktimesDN = make.k.times.DN(H, N, Z.all)
105         write.table(ktimesDN, file = paste("Xquer-Yquer␣k.
                times.DN␣H", H, "␣N", N, ".txt", sep = ""))
106     }
107 }
```

Now executing the function `lets.go.niveau` yields files with `k=10000` values of D_n for sample sizes `N=10, 50, 100, 500, 1000`, e.g. `Xquer-Yquer k.times.DN H0.7 N100.txt`. In order to perform the "difference-of-means" test, one has to count how many of these `k=10000` values of D_n exceed the respective critical value.

```
109 # set quantiles manually
110 quantiles = c(1.1, 0.87, 0.44)
111 names(quantiles)=c("H=0.6", "H=0.7", "H=0.9")
112
113 make.niveau.table  <- function(quantiles){
114     # reserve marix to store frequencies
115     table.errorI = matrix(rep(NA,6*3), byrow = FALSE, ncol =
            3)
116
117     # give names to colums and rows
118     colnames(table.errorI) = c('H=0.6', 'H=0.7', 'H=0.9')
119     rownames(table.errorI) =  c('N=10', 'N=50', 'N=100', 'N
            =500', 'N=1000', 'N=2000')
120
```

```
121     # for each LRD parameter
122     for(H in c(0.6, 0.7, 0.9)){
123        alpha = 2-2*H
124        # and for each length N=50, ..., N=2000
125        for(N in c(10,50,100,500,1000)){
126           # read the simulation output Dn
127           dn = read.table(paste("Xquer-Yquer␣k.times.DN␣H", H
                  , "␣N", N, ".txt", sep=""))
128
129           # extract data (as vector) from data.frame T
130           dn = dn[,1]
131
132           table.errorI[paste("N=", N, sep=""),paste("H=", H,
                  sep="")] = length(which(dn > quantiles[   paste(
                  "H=", H, sep="")]))/length(dn)
133        }
134     }
135
136     table.errorI.rounded = round(table.errorI,3)
137     write.table(table.errorI,file="Tabelle␣Niveau␣Xq-Yq␣fGn.
           txt")
138     write.table(table.errorI.rounded,file="Tabelle␣Niveau␣Xq-
           Yq␣fGn.txt",append=TRUE)
139 }
140
141 # make.niveau.table(quantiles)
```

The function `make.niveau.table` takes three critical values for `H=0.6, 0.7, 0.9` and yields a table with the observed level for each H and each sample size `N`, i.e. the relative frequency of wrong rejections among the 10,000 simulation runs.

Now the routine to simulate the level of the test under various alternatives needs the above defined functions `calculate.Udiff` and `make.k.times.DN` and an additional function which adds a jump to the time series.

```
1  # function adds break-point of heigth hh
2  # at position lambda to a vector
3  add.break <- function(vec, lambd, hh){
4     ll = length(vec)
5     m = floor(lambd*ll)
6     vec.neu = c(vec[1:m], vec[(m+1):ll] + hh)
7     vec.neu
8  }
9
10 # now the workflow:
11 # call make.k.times.DN for different h and lambda
12 # to save simulation time: fix H and N
```

```
13 lets.go.alternative  <- function(h){
14    H=0.7
15    alpha = 2-2*H
16
17    N=500
18
19    # read k*N-matrix with FGN (k columns of length N)
20    Z.all = read.table(paste("fgn␣H", H , "␣N2000␣k10000.txt"
          , sep=""))
21
22    # cut time series to length N
23    Z.all = Z.all[1:N,]
24
25    # for different times...
26    for(lambda in c(0.05, 0.1, 0.3, 0.5)){
27       # ...add break of height h to time series
28       Z = apply(Z.all, 2, add.break, lambd=lambda, hh=h)
29       ktimesDN = make.k.times.DN(H, N, Z)
30       write.table(ktimesDN, file = paste("Xq-Yq␣k.DN␣H", H,
             "␣N", N, "␣h", h, "␣lambda", lambda, ".txt", sep =
             ""))
31    }
32 }
```

The function `lets.go.alternative` takes a value for the jump height `h` and yields a file with 10,000 values of the test statistic D_n, for the choosen `h` and for different values of `lambda`, the jump location, e.g. `Xq-Yq k.DN H0.7 N500 h1 lambda0.1.txt`.

```
34 make.power.table  <- function(quantiles){
35    # set LRD and N manually
36    H=0.7
37    alpha = 2-2*H
38    N=500
39
40    # reserve marix to store frequencies
41    table.power = matrix(rep(NA,3*4), byrow = FALSE, ncol =
          4)
42
43    # give names to colums and rows
44    colnames(table.power) = c(0.05, 0.1, 0.3, 0.5)
45    rownames(table.power) =  c('h=0.5', 'h=1', 'h=2')
46
47    for(lambda in c(0.05, 0.1, 0.3, 0.5)){
48       for(h in c(0.5, 1, 2)){
49          # read the simulation output Dn
50          dn = read.table(paste("Xq-Yq␣k.DN␣H", H, "␣N", N, "
                ␣h", h, "␣lambda", lambda, ".txt", sep=""))
```

```
51
52          # extract data (as vector) from data.frame
53          dn = dn[,1]
54
55          # count frequency of rejections
56          table.power[paste("h=", h, sep=""),paste(lambda)] =
                 length(which(dn > quantiles[ paste("H=", H, sep
                =="")]))/length(dn)
57       }
58    }
59
60    table.power.rounded = round(table.power,3)
61    write.table(table.power,file=paste("Tabelle␣Power␣Xq-Yq␣
         fGn␣N", N, ".txt", sep=""))
62    write.table(table.power.rounded,file=paste("Tabelle␣Power
         ␣Xq-Yq␣fGn␣N", N, "␣rounded.txt", sep=""),append=TRUE)
63 }
```

Like `make.niveau.table`, the function `make.power.table` takes three critical values for `H=0.6, 0.7, 0.9` and yields a table with the observed power for the various alternatives (jumps of height `h` after a proportion of `lambda` of the data; we restrict ourselves to `H=0.7` and sample size `N=500`), i.e. the relative frequency of true rejections among the 10,000 simulation runs.

C.5 $\bar{X} - \bar{Y}$ for one divided sample (section 2.2.2)

This is just a special application of the routine `calculate.Udiff` from Section C.4.

C.6 $\bar{X} - \bar{Y}$ for two independent samples (section 2.3.2)

This routine is almost the same as the one before; the only differences are first the input data (one must not divide one single time series, but now one has to consider two independent time series) and second the normalizing constant in Theorem 2.4. `calculate.Udiff` can easily be adapted to these changes.

An easy way to provide for this new situation without changing the original data set (e.g. `fgn H0.7 N2000 k10000.txt`) is the following: Instead of dividing a time series and calculate the difference of the means of the arising two samples (they would be dependent), take the first part of a time series as X-sample and the second part of the succeeding time series as Y-sample and the other way around – in this way, we obtain two pairs of independent samples X and Y. This can be implemented as follows.

```
1 # function "DiffOfMeans_indep" calculates mean(X)-mean(Y)
2 # for 2 vectors X, Y (read out of 1 matrix Z of FGN series)
3 # output: mean(X)-mean(Y) scaled and scaled+normalized
4 DiffOfMeans_indep <- function(Z,lambda){
```

```
5     # Z = matrix of k FGN series (k columns of length N)
6     # lambda = cutting point, 0<lambda<1
7
8     # error if not 0<lambda<1
9     if(lambda<=0 || lambda>=1) {
10       stop("lambda must be in (0,1).")
11    }
12
13    # divide Z into X and Y at point lambda*N
14    k = ncol(Z)
15    N = nrow(Z)
16    m = round(lambda*N)
17    n = round((1-lambda)*N)
18
19    # error if m=0 or n=0
20    if(m==0 || n==0) {
21       stop("One of both samples has size 0.")
22    }
23
24    # if k odd, take k-1 instead
25    # reason: see for-loop below
26    if(k %% 2) k <- k-1
27
28    # reserve vectors to store k values of T
29    T.unnormalized = rep(NA,k)
30    T = rep(NA,k)
31
32    # take all time series from Z, each time two at once
33    for(i in seq(1,k,2)){
34       # divide time series i and i+1 into pieces
35       # of length m and n=N-m
36       x1 = Z[1:m,i]
37       y1 = Z[(m+1):N,i]
38       x2 = Z[1:m,i+1]
39       y2 = Z[(m+1):N,i+1]
40       # calculate two times unnormalized and unscaled T-
            value
41       # and store in vector T.unnormalized
42       T.unnormalized[i] = mean(x1)-mean(y2)
43       T.unnormalized[i+1] = mean(x2)-mean(y1)
44    }
45
46    # calculate normalizing constant sigma_{Xbar-Ybar}^2
47    # if Z is not FGN, this constant is different!
48    sigma.squared = (lambda^(1-alpha)*(1-lambda) + lambda*(1-
         lambda)^(1-alpha))
```

```
49
50     # normalize and scale T.unnormalized
51     # to obtain limit distribution N(0,1) (predicted by
           theory)
52     # if Z is not FGN, scaling is different!
53     T = T.unnormalized*sqrt(m*n/N^(2-alpha))/sqrt(sigma.
           squared)
54
55     # scale T.unnormalized with overall sample size N (not
           normalizing)
56     # if Z is not FGN, scaling is different!
57     T.unnormalized = T.unnormalized*N^(alpha/2)
58
59     write.table(T.unnormalized, file = paste("Xquer-Yquer␣N",
           N, "␣lambda", lambda, "␣k", k, ".txt", sep = ""))
60     write.table(T, file = paste("Xquer-Yquer␣norm␣N", N, "␣
           lambda", lambda, "␣k", k, ".txt", sep = ""))
61 }
```

C.7 Quantiles of $\sup|Z(\lambda) - \lambda Z(1)|$ (section 3.4)

The distribution of $\sup_{0\leq\lambda\leq 1}|Z(\lambda) - \lambda Z(1)|$ was simulated, as described on page 61, with the following code.

```
1  simulate.distribution.Dn  <- function(H){
2     # fArma: long-range dependend time series
3     library(fArma)
4
5     alpha = 2-2*H
6
7     # from 0 to 1 in steps
8     # k repetitions
9     steps = seq(0,1,0.001)
10    l = length(steps)
11    k = 10000
12
13    # reserve vector Z for time series Z at l times
14    # reserve vectors for simulation result
15    Z = rep(NA,l)
16    result.different.lambdas = rep(NA,l)
17    result = rep(NA,k)
18
19    # j=1,...,k (repetitions)
20    for(j in 1:k){
21       # simulate fBm with time in [0,1] at l equidistant
             points 0, ..., 1
```

```
22        Z = fbmSim(l, H, doplot=F);
23
24        # for each time lambda=0,...,1
25        i = 1
26        for(lambda in steps){
27           # calculate Z(lambda)-lambda Z(lambda)
28           # remember that Z is a vector with entries Z[1], Z
                [2], ...
29           lambda2 = lambda*(l-1)+1
30           result.different.lambdas[i] = Z[lambda2]-lambda*Z[l
                ]
31        i=i+1
32        }
33
34        # store the max of the absolute value to obtain 1
             realization
35        result[j] = max(abs(result.different.lambdas))
36     }
37
38     write.table(result, file = paste("Xquer-Yquer␣Vtlg␣CP-
          test␣fGn␣H=", H, "␣steps=", l, ".txt", sep = ""))
39
40     detach("package:fArma")
41  }
```

The routine `simulate.distribution.Dn` produces a file, e.g. `Xquer-Yquer Vtlg CP-test fGn H=.7 steps=1001.txt`, with 10000 simulation results of $\sup_{0\leq\lambda\leq 1}|Z(\lambda)-\lambda Z(1)|$ from which properties like the quantiles of the distribution of $\sup_{0\leq\lambda\leq 1}|Z(\lambda)-\lambda Z(1)|$ can be approximated.

C.8 "Wilcoxon-type" test (section 3.6)

The following program code calculates `k` repetitions of the "Wilcox-type" test statistic W_n, as defined in (3.12) for different sample lengths `N` and LRD parameter `H`. The input needed, the design of the routine (for fGn; it must be adapted if th underlying Gaussian process is not fGn), the workflow needed and the output is like for the "diferences-of-means" test in section C.4.

```
1  # function divides vector vec after entry with index cut
2  # and gives Wilcoxon statistic
3  divide.and.Wilcox  <- function(vec, cut){
4     l = length(vec)
5
6     # function counts, how many entries of vector vv are >= n
7     which.vv.greater.n <- function(n,vv) length(which(n<=vv))
8
```

```
 9     # function compares all entries of a vector vec1
10     # with a vector vec2 via "which.vv.greater.n"
11     which.vec1.smaller.vec2  <- function(vec1, vec2) sum(
           sapply(vec1, which.vv.greater.n, vv=vec2))
12
13     # finally apply "which.vec1.smaller.vec2" to
14     # first and second part of vec
15     which.vec1.smaller.vec2(vec[1:cut], vec[(cut+1):l])
16 }
17
18 # function calculates
19 # U_W = 1/(n dn) sum_{i=1}^{lambda n} sum_{j=lambda n+1}^n (
       I{(Xi<=Xj} - 1/2)
20 # for two vectors X, Y (created by dividing one FGN series Z
       )
21 # returns vector with 10.000 values of U_W
22 calculate.U_W <- function(Z, cutpt, alpha){
23     # Z = given matrix of k FGN series (k columns of length N
           )
24     # cutting point cutpt= 1,...,N-1
25
26     # error if cutting point not an integer number
27     if(cutpt - trunc(cutpt) != 0){
28        stop("cutpoint␣is␣not␣an␣integer.")
29     }
30
31     # read dimensions of Z
32     k = ncol(Z)
33     N = nrow(Z)
34
35     # reserve vector U to store k values of U_W
36     U = rep(NA,k)
37
38     # cutting point should be in {1,2, ..., N-1}
39     if((cutpt >= 1) && (cutpt < N)){
40        # apply "divide.and.difference" to columns of Z
41        # and write k results in vector U
42        U = apply(Z, 2, divide.and.Wilcox, cut=cutpt)
43     }
44
45     # if cutpt out of range, T can't be calculated
46     if((cutpt < 1) || (cutpt >= N)) {
47        U = rep(NA,k)
48     }
49
50     lambd = cutpt/N
```

```
51
52      # FGN has covariances (1-alpha)(2-alpha)/2 k^(-alpha)
53      Lk = (1-alpha)*(2-alpha)/2
54
55      # constant c in dn=c*n^(1-alpha/2) is here
56      # just: c= sqrt (1/Lk)
57
58      # U_W = 1/(n dn) sum_{i=1}^{lambda n} sum_{j=lambda n +
            1}^n (I{(Xi<=Xj} -1/2)
59      #       = 1/(n^(2-alpha/2) c * sqrt(Lk)) * [#{Xi<=Xj} -
            lambda*(1-lambda) n^2/2 ]
60      U = 1/N^(2-alpha/2) * ( U - (lambd*(1-lambd)*N^2/2))
61  }
```

`calculate.U_W` calculates the "Wilcoxon-type" statistic $W_{k,n}$ as in (3.3) with cutpoint $k = [\lambda n]$. (We have tested some other ways to calculate the "Wilcoxon-type" statistic (one for example includes sorting in order to avoid time intensive componentwise comparing of long vectors), but they all proved to be slower than the one given here in the sub-routine `divide.and.Wilcox`.) The following routine `make.k.times.WN` calculates `k` repetitions of $W_n = \max_{1 \le k < N} |W_{k,n}|$.

```
63  make.k.times.WN <- function(H, N, Z){
64      alpha = 2-2*H
65
66      # read number of repetitions
67      k = ncol(Z)
68
69      # N observations => N-1 possible cutting points
70      l = N-1
71
72      # reserve marix to store U_W
73      # l columns (for each cutting point), k=10.000 rows (
            repetitions)
74      U_W.table = matrix(rep(NA, k*l), byrow = FALSE, ncol = l)
75
76      # reserve vector to store k repetitions of WN
77      k.times.WN = rep(NA, k)
78
79      # for each cutting point index i=1,...,N
80      for(i in 1:l){
81          # calculate U_W
82          U_W.table[,i] = calculate.U_W(as.matrix(Z[1:N,]), i,
                alpha)
83      }
84
85      # now we have an lxk matrix U_W.table
86      # go through rows and find max(abs(..))
```

```
87
88      # function finds max(abs(vec)) of vector vec
89      # ignore "NA"-entries!
90      find.maxabs  <- function(vec){
91         max(abs(vec), na.rm=TRUE)
92      }
93
94      # apply "find.maxabs" to rows of U_W.table
95      k.times.WN = apply(U_W.table, 1, find.maxabs)
96
97      # result: vector of k repetitions of DN = sup |U_W|
98      return(k.times.WN)
99  }
```

In order to apply the "Wilcoxon-type" test and to evaluate its level and power, this routine `make.k.times.WN` must be applied to data and its output has to be comapred with the respective quantiles. As mentioned before, this can be done like for the "diferences-of-means" test in section C.4.

The routine `make.k.times.WN` can easily be changed such that it does not store the values of the test statistic $W_n = \max |W_{k,n}|$, but the point $k \in \{1, N-1\}$ at which $W_{k,n}$ reaches it maximum, in other words, such that the routine stores an estimation of the change-point location. For this purpose, change the respective part like this:

```
1       # now we have an lxk matrix U_W.table
2       # go through rows and find argmax(abs(..))
3
4       # function finds argmax(abs(vec)) of vector vec
5       # ignore "NA"-entries!
6       find.argmaxabs  <- function(vec){
7          which.max(abs(vec))
8       }
9
10      # apply "find.argmaxabs" to rows of U_W.table
11      k.times.WNArg = apply(U_W.table, 1, find.argmaxabs)
```

C.9 Estimating the LRD parameter under a jump (section 7.2)

Given an arbitrary procedure to estimate the LRD parameter from a sample of observations, we presented in Section 7.2 different methods to adapt this procedure to time series which include a jump. Now we will give the `R` implementations of these methods.

The following subroutines are needed by all methods. `add.break(vec,lambd,hh)` adds a break-point of height `hh` to a time series `vec` after a proportion of `lambd`. In `estimate.H`, the procedure to estimate is specified. Here, we concentrate on the meth-

ods `whittleFit` and `boxperFit` from the `fArma` package. The routine `estimate.H.try` handles errors: If `estimate.H` produces an error, it gives out `NA`.

```
add.break <- function(vec, lambd, hh){
   ll = length(vec)
   m = floor(lambd*ll)
   vec.neu = c(vec[1:m], vec[(m+1):ll] + hh)
   vec.neu
}

estimate.H  <- function(vec){
   library(fArma)
   #y = whittleFit(vec)
   #y@hurst[[3]]
   y = boxperFit(vec)
   y@hurst[[1]]
}

estimate.H.try <- function(vec){
   return(tryCatch(
      estimate.H(vec),
      error=function(e) NA
   ))
}
```

With the following routines, the LRD parameter is estimated on each two blocks which arise from cutting the sample in two pieces. This is done for all cutting points (leaving out the margins of the sample). The third routine then calculates a function of these pairs of estimates – the means of all estimates or the mean of the two estimates at the cutting point where both estimates (left and right block) differ least.

```
# estimate H seperately on index set
# I1={1,...,k} and I2 = {k+1, ..., n}
estimate.H.piecewise  <- function(vec, cutpoint){
   N = length(vec)
   H.lower = estimate.H.try(vec[1:cutpoint])
   H.upper = estimate.H.try(vec[(cutpoint+1):N])
   c(H.lower, H.upper)
}

calculate.all.Hpiecewise  <- function(vec){
   N = length(vec)
   lower = max(c(floor(N*0.10), 10))+1
   upper = N-lower
   Hpiecewise = matrix(rep(NA, 2*(upper-lower+1)), ncol=2)
   for(i in lower:upper){
      Hpiecewise[i-lower+1,] = estimate.H.piecewise(vec, i)
   }
```

```
18    Hpiecewise
19 }
20
21 estimate.H.block  <- function(vec){
22    Hpiecewise = calculate.all.Hpiecewise(vec)
23    RMeans = rowMeans(Hpiecewise, na.rm=TRUE)
24
25    H.mean = mean(RMeans, na.rm=TRUE)
26    H.mindiff = mean(estimate.H.piecewise(vec, which.min(abs(
          Hpiecewise[,2]-Hpiecewise[,1]))))
27    c(H.mean, H.mindiff)
28 }
```

Now we give the code for estimating H on a moving (overlapping) window with flank size `w` (i.e. a windowf of size `2w+1`) from the data `vec`.

```
1  estimate.H.movwin <- function(vec, w){
2     N = length(vec)
3
4     window.indices <- function(w, midpoint){
5        # w = half window size
6        # if midpoint <= w, some indices <= 0, fill with "NA"
7        # should not happen, but to be on the safe side :-)
8        if(midpoint > w) indices = seq(midpoint-w, midpoint+w)
9        if(midpoint <= w) {
10          indices = seq(midpoint-w, midpoint+w)
11          indices[which(indices<=0)] = NA
12       }
13       indices
14    }
15
16    window.midpoints = seq(w+1, N-w)
17    H.movwin = rep(NA, length(window.midpoints))
18
19    # move index window [x-w, x+w] from x=w+1 to x=N-w
20    for(i in window.midpoints){
21       H.movwin[i] = estimate.H.try(vec[window.indices(w, i)
             ])
22    }
23    mean(H.movwin, na.rm=TRUE)
24 }
```

And this the code for a non-overlapping moving window based estimation:

```
26 estimate.H.movwin.nonlap <- function(vec, w){
27    N = length(vec)
28    # w: window size
29
30    window.startpoints = seq(1, N, w)
```

```
31     # remove last startpoint, if last window smaller than w
32     if(N%%w!=0){
33        window.startpoints = window.startpoints[-length(window
              .startpoints)]
34        }
35
36     H.movwin.nonlap = rep(NA, length(window.startpoints))
37
38     counter = 1
39     for(i in window.startpoints){
40        H.movwin.nonlap[counter] = estimate.H.try(vec[i:(i+w
              -1)])
41        counter=counter+1
42     }
43     mean(H.movwin.nonlap, na.rm=TRUE)
44  }
```

In what follows, we present the code for estimating H using pre-estimation of the possible change-point. This approach needs a method to estimate the change-point location, here this is the routine `make.k.times.WNArg`, a simple modification of `make.k.times.WN` as defined above at the end of Section C.8.

```
45  estimate.H.preestimate <- function(vec){
46     N = length(vec)
47
48     # load Wilcoxon change-point test statistic
49
50     # estimate H seperately on index set
51     # I1={1,...,k} and I2 = {k+1, ..., n}
52     estimate.H.piecewise  <- function(vec, cutpoint){
53        N = length(vec)
54        H.lower = estimate.H.try(vec[1:cutpoint])
55        H.upper = estimate.H.try(vec[(cutpoint+1):N])
56        c(H.lower, H.upper)
57     }
58
59     # start with arbitrary value for H, e.g. usual estimate
60     H.estim.usual = estimate.H.try(vec)
61
62     # under H=H.estim.usual, where detects Wilcoxon test
           change-point?
63     CP.estim = make.k.times.WNArg(H.estim.usual, N, as.matrix
           (vec))
64
65     # way 1:
66     # estimate H on block before and after this change-point
67     H.estim.2blocks = estimate.H.piecewise(vec, CP.estim)
```

```
68    H.estim.2blocks = mean(H.estim.2blocks)
69    # assign new estimate to H? No. Because:
70    # from now on, cp-detection will not change,
71    # since H changes only scaling
72    # => new H maybe changes test decision, but not argmax
73
74    # way 2:
75    # remove jump and estimate H on whole time series
76    remove.jump <- function(vec){
77       N = length(vec)
78       # under H=H.estim.usual, where detects Wilcoxon test
             change-point?
79       H.estim.usual = estimate.H.try(vec)
80       CP.estim = make.k.times.WNArg(H.estim.usual, N, as.
             matrix(vec))
81
82       # if H.estim.usual returns NA, CP.estim = integer(0),
             i.e.\ length 0 vector
83       if(length(CP.estim)!=0){
84          # estimate jump height
85          h.estim = mean(vec[(CP.estim+1):N])-mean(vec[1:CP.
                estim])
86          # remove jump
87          vec.jumpfree = c(vec[1:CP.estim], vec[(CP.estim+1):
                N]-h.estim)
88       }else{
89          vec.jumpfree = NA
90       }
91       vec.jumpfree
92    }
93
94    # estimate H on whole time series without jump
95    vec.jumpfree = remove.jump(vec)
96    H.estim.jumpfree = estimate.H.try(vec.jumpfree)
97
98    # way 3:
99    # repeat way 2 iteratively
100   # drawback: many repetitions smoothe time series=> H
         underestimated
101   estimate.jumpfree.iterative  <- function(vect){
102      # set start values
103      H.iterativ = 2
104      H.Differenz = 2
105      counter = 0
106      while(abs(H.Differenz) > 0.01){
107         # remove jump from vec
```

```
108           vect.jumpfree = remove.jump(vect)
109           # estimate H on time series without jump
110           H.estim.jumpfree = estimate.H.try(vect.jumpfree)
111
112           # did H.estim.jumpfree change?
113           H.Differenz = H.iterativ-H.estim.jumpfree
114           # if yes, try another run, based on jumpfree series
115           vect = vect.jumpfree
116           H.iterativ = H.estim.jumpfree
117           counter = counter + 1
118        }
119        c(H.iterativ, counter)
120     }
121     # estimate.jumpfree.iterative(vec)
122
123     c(H.estim.jumpfree, H.estim.2blocks)
124  }
```

C.10 Estimating the first Hermite coefficient (section 7.3)

The program `simulateCoeff1` produces a table in which the estimations $\widetilde{a_1}$ for a_1, as defined in (7.3), are given – each the mean and the variance in 10,000 simulation runs for different sample sizes. For a better comparison, the mean is given relative to the true value a_1; if this value is not known, the program must be adapted.

```
1  simulateCoeff1 <- function(G, decreas=FALSE, trueCoef, zv.
       lrd=FALSE){
2     # G: function G that generates data
3     # decreas: Is G decreasing?
4     # trueCoef: true first Hermite coefficient of G
5     # zv.lrd: replace true xi by new xi iid or lrd
6
7     library(fArma)
8     H = 0.7
9     alpha = 2-2*H
10
11    simCoeff.norm.sort <- function(n, G, decreas, H){
12       # replace true xi by new xi\sim N(0,1) iid
13       xi = rnorm(n)
14       x = G(fgnSim(n, H))
15       mean(sort(xi)*sort(x, decreasing=decreas))
16    }
17
18    simCoeff.LRD.sort <- function(n, G, decreas, H){
19       # replace true xi by new xi\sim N(0,1) LRD
20       xi = fgnSim(n, H)
```

```
21        x = G(fgnSim(n, H))
22        mean(sort(xi)*sort(x, decreasing=decreas))
23     }
24
25     # k = number of repititions
26     k = 10000
27     Tnorm = rep(NA, k)
28
29     NN = c(10, 50, 100, 500, 1000, 2000, 5000, 10000)
30
31     Results = matrix(rep(NA,length(NN)*2), byrow = FALSE,
          ncol = 2)
32     colnames(Results) = c('mean/true', 'variance')
33     rownames(Results) =  NN
34
35     for(n in NN){
36        for(i in 1:k){
37           Tnorm[i] = if(zv.lrd) simCoeff.LRD.sort(n, G,
                decreas, H) else simCoeff.norm.sort(n, G,
                decreas, H)
38        }
39        Results[paste(n),1] = mean(Tnorm)
40        Results[paste(n),2] = var(Tnorm)
41     }
42
43     Results[,1] = Results[,1]/trueCoef
44
45     Results
46  }
```

Appendix D

Exact simulation results

D.1 $\bar{X} - \bar{Y}$ test (section 2.2 and 2.3)

D.1.1 One divided sample

We consider a series of N observations which is cut into two pieces,

$$X_1, X_2, \ldots, X_m \quad \text{and} \quad X_{m+1}, X_{m+2}, \ldots, X_{m+n}$$

with $N = m + n$. It is $m = [\lambda N]$ and $n = [(1 - \lambda)N]$ for a $\lambda \in (0, 1)$. We call the second sample the Y-sample ($Y_k := X_{m+k}$). We consider the scaled and normalized $\bar{X} - \bar{Y}$ test statistic

$$T := \sqrt{\frac{mn}{(m+n)^{2-D}}} \frac{\bar{X} - \bar{Y}}{\sigma_{\text{diff}}}.$$

Since T is a linear statistic of the X's which are standard normally distributed, T is always normally distributed with mean zero, so we only need to look at its variance. We have calculated the sample variance of T, based on a set of $10,000$ simulations

- for $H = 0.6$ (Table D.1),
- for $H = 0.7$ (Table D.2) and
- for $H = 0.9$ (Table D.3).

D.1.2 Two independent samples

Now instead of one sample of observations which is cut into two pieces, we consider two single stationary Gaussian processes $(X_i)_{i \geq 1}$ and $(Y_j)_{j \geq 1}$ which are independent of each other, and look at the performance of the scaled and normalized $\bar{X} - \bar{Y}$ test statistic

$$T := \sqrt{\frac{mn}{(m+n)^{2-D}}} \frac{\bar{X} - \bar{Y}}{\sigma_{\text{diff},2}}$$

which is asymptotically standard normally distributed, according to Theorem 2.4. Again, we have calculated the sample variance of T, based on $10,000$ simulation runs

$N \setminus \lambda$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	0.98	0.989	1.014	1.026	1.028	1.004	1.005	1.004	0.983
50	0.999	0.962	0.963	0.971	0.989	0.994	0.997	1.011	0.998
100	0.963	0.98	0.996	0.997	0.992	0.981	0.978	0.982	0.995
500	0.977	0.985	1.003	1.008	1.03	1.014	1.002	0.999	0.997
1000	1.004	1.016	1.013	1.004	0.996	1.002	1.009	0.991	0.986
2000	1.008	0.997	0.989	0.985	0.992	0.986	0.982	0.982	0.995

Table D.1: Sample variance of $10,000$ values of T for fGn with Hurst parameter $H = 0.6$ ($D = 0.8$). N is the overall sample size, the data was divided into two samples after the $[\lambda N]$-th observation.

$N \setminus \lambda$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	0.992	0.993	0.992	0.997	1.003	1.007	1.007	1.001	1.002
50	1.023	1.015	0.996	0.989	0.995	0.994	0.986	1.003	0.992
100	1.025	1.006	1	1.014	1.008	1.008	0.997	0.989	0.992
500	1.001	0.984	0.994	0.997	0.992	0.993	0.994	0.985	0.991
1000	0.984	1.004	1.014	1.015	1.017	1.004	1.005	1.008	1.008
2000	1.003	1.005	0.998	1.002	0.992	1	1.011	1.014	1.02

Table D.2: Sample variance of $10,000$ values of T for fGn with Hurst parameter $H = 0.7$ ($D = 0.6$). N is the overall sample size, the data was divided into two samples after the $[\lambda N]$-th observation.

$N \setminus \lambda$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	0.988	0.995	1.001	1.014	1.007	1.012	1.003	0.99	1.005
50	1	0.995	0.997	0.992	0.99	0.999	1.001	1.007	1.012
100	1.002	0.996	0.996	0.995	0.991	0.997	1.007	1.008	1.012
500	1	0.996	0.985	0.986	0.981	0.99	0.991	0.993	0.997
1000	0.999	0.99	1.001	1.007	1.011	1.02	1.011	0.998	1
2000	0.99	1.003	1.011	1.004	1.007	0.998	1.003	1.011	0.999

Table D.3: Sample variance of $10,000$ values of T for fGn with Hurst parameter $H = 0.9$ ($D = 0.2$). N is the overall sample size, the data was divided into two samples after the $[\lambda N]$-th observation.

- for $H = 0.6$ (Table D.4),
- for $H = 0.7$ (Table D.5) and
- for $H = 0.9$ (Table D.6).

$N \setminus \lambda$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	0.967	0.979	0.991	0.993	0.995	0.981	0.98	0.985	0.982
50	0.986	0.966	0.976	0.979	0.982	0.989	0.993	1.001	0.995
100	0.957	0.978	0.978	0.985	0.978	0.975	0.973	0.987	1.003
500	0.99	0.995	1.003	1.009	1.013	1.013	1.005	1.01	1.01
1000	1.003	1.013	1.007	1.007	1	1.007	1.014	1.009	1.008
2000	1.01	1.006	1.012	1.018	1.02	1.021	1.004	0.997	1.014

Table D.4: Sample variance of $10,000$ values of T for two independent samples of fGn with Hurst parameter $H = 0.6$ ($D = 0.8$). N is the overall sample size, the first sample has length $[\lambda N]$.

$N \setminus \lambda$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	0.985	0.992	0.995	1.003	1.008	1.007	1.014	1.019	1.028
50	1.009	1.003	0.992	0.981	0.976	0.975	0.978	0.986	0.977
100	1.012	0.999	0.994	1.002	1.003	1.007	1.002	1	0.99
500	1.009	1.014	1.025	1.025	1.021	1.019	1.01	0.993	0.989
1000	1.009	1.027	1.027	1.021	1.019	1.013	1.013	1.016	1.016
2000	1.027	1.028	1.015	1.017	1.014	1.014	1.023	1.023	1.019

Table D.5: Sample variance of $10,000$ values of T for two independent samples of fGn with Hurst parameter $H = 0.7$ ($D = 0.6$). N is the overall sample size, the first sample has length $[\lambda N]$.

$N \setminus \lambda$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	0.994	1.002	1.005	1.008	1.006	1.006	1.004	1	1.004
50	1.003	1.003	1.003	1.002	1.001	1	0.999	0.998	0.996
100	1.007	1.005	1.004	1.002	1	1.002	1.002	0.998	0.993
500	1.007	1.004	1.001	1.003	1.004	1.007	1.006	1.007	1.007
1000	0.992	0.993	0.996	0.997	0.996	0.996	0.993	0.99	0.991
2000	0.984	0.985	0.981	0.974	0.969	0.964	0.962	0.96	0.954

Table D.6: Sample variance of $10,000$ values of T for two independent samples of fGn with Hurst parameter $H = 0.9$ ($D = 0.2$). N is the overall sample size, the first sample has length $[\lambda N]$.

D.2 Estimation of the variance of $\bar{X}$ (section 2.4)

Here the simulations for the estimated variance of $\bar{X}$ relative to the true variance, i.e. for the quotient $\widehat{\mathrm{Var}}[X^{(r)}]/\mathrm{Var}[\bar{X}_N]$, are presented,

- for $H = 0.6$ (Table D.7),
- for $H = 0.7$ (Table D.8) and
- for $H = 0.9$ (Table D.9),

each for different sample sizes N and different block sizes (r is either relative, fixed or $r = N^\beta$). The parameter which yields the best result is highlighted in each table.

$N \setminus r$	$1/\sqrt{N} \cdot N$	$1/50 \cdot N$	$1/10 \cdot N$	$1/5 \cdot N$
50	0.79	0.955	0.842	0.722
100	0.836	0.955	0.836	0.718
500	0.927	0.955	0.84	0.718
1000	0.923	0.954	0.836	0.725
2000	0.944	0.954	0.838	0.719

$N \setminus r$	10	50	100
50	0.722	–	–
100	0.836	0.422	–
500	0.955	0.84	0.718
1000	0.973	0.906	0.836
2000	0.984	0.946	0.903

$N \setminus \beta$	0.1	0.3	0.5	0.7	0.9
50	1.306	0.942	0.79	0.597	–
100	1.41	1.176	0.836	0.663	–
500	1.631	1.028	0.927	0.76	–
1000	1.73	1.083	0.948	0.812	–
2000	1.05	1.053	0.963	0.823	0.421

Table D.7: Relative results $\widehat{\mathrm{Var}}[X^{(r)}]/\mathrm{Var}[\bar{X}_N]$ of variance estimation of $\bar{X}$, each value averaged over $10,000$ simulations, sample size N, block size r, $H = 0.6$ ($D = 0.8$).

$N \setminus r$	$1/\sqrt{N} \cdot N$	$1/50 \cdot N$	$1/10 \cdot N$	$1/5 \cdot N$
50	0.695	0.904	0.75	0.619
100	0.746	0.904	0.746	0.615
500	0.852	0.903	0.746	0.611
1000	0.867	0.906	0.748	0.619
2000	0.893	0.905	0.747	0.619

$N \setminus r$	10	50	100
50	0.619	–	–
100	0.746	0.343	–
500	0.903	0.746	0.611
1000	0.938	0.836	0.748
2000	0.959	0.891	0.833

$N \setminus \beta$	0.1	0.3	0.5	0.7	0.9
50	1.143	0.848	0.695	0.489	–
100	1.235	1.037	0.746	0.566	–
500	1.417	0.97	0.852	0.658	–
1000	1.489	1.024	0.887	0.715	–
2000	1.025	1.011	0.907	0.729	0.337

Table D.8: Relative results $\widehat{\mathrm{Var}}[X^{(r)}]/\mathrm{Var}[\bar{X}_N]$ of variance estimation of $\bar{X}$, each value averaged over $10,000$ simulations, sample size N, block size r, $H = 0.7$ ($D = 0.6$).

$N \setminus r$	$1/\sqrt{N} \cdot N$	$1/50 \cdot N$	$1/10 \cdot N$	$1/5 \cdot N$
50	0.323	0.544	0.37	0.276
100	0.369	0.543	0.369	0.275
500	0.461	0.542	0.367	0.274
1000	0.496	0.543	0.369	0.275
2000	0.531	0.544	0.37	0.276

$N \setminus r$	10	50	100
50	0.276	–	–
100	0.369	0.128	–
500	0.542	0.367	0.274
1000	0.602	0.451	0.369
2000	0.654	0.523	0.452

$N \setminus \beta$	0.1	0.3	0.5	0.7	0.9
50	0.588	0.434	0.323	0.198	–
100	0.66	0.533	0.369	0.241	–
500	0.805	0.594	0.461	0.3	–
1000	0.86	0.645	0.502	0.34	–
2000	0.759	0.672	0.535	0.357	0.13

Table D.9: Relative results $\widehat{\mathrm{Var}}[X^{(r)}]/\mathrm{Var}[\bar{X}_N]$ of variance estimation of $\bar{X}$, each value averaged over $10,000$ simulations, sample size N, block size r, $H = 0.9$ ($D = 0.2$).

D.3 Change-point test comparison in (section 3.6)

We consider the model $X_i = \mu_i + G(\xi_i)$, $i = 1, \ldots, n$, where $(\xi_i)_{i\geq 1}$ is a mean-zero Gaussian process with $\mathrm{Var}[\xi_i] = 1$ and auto-covariance function (1.1) and a transformation $G : \mathbb{R} \to \mathbb{R}$, $G \in \mathcal{G}^1$ or $G \in \mathcal{G}^2$. We wish to test the hypothesis

$$H : \mu_1 = \ldots = \mu_n$$

against the alternative

$$A : \mu_1 = \ldots = \mu_k \neq \mu_{k+1} = \ldots = \mu_n \text{ for some } k \in \{1, \ldots, n-1\}.$$

with the "Wilcoxon-type" test which rejects H for large values of

$$W_n = \frac{1}{n\, d_n} \max_{1\leq k\leq n-1} \left| \sum_{i=1}^{k} \sum_{j=k+1}^{n} \left(I_{\{X_i \leq X_j\}} - \frac{1}{2} \right) \right|.$$

and with a "difference-of-means" test which rejects H for large values of

$$D_n = \frac{1}{n\, d_n} \max_{1\leq k\leq n-1} \left| \sum_{i=1}^{k} \sum_{j=k+1}^{n} (X_i - X_j) \right|.$$

We consider the performance of both tests under null hypothesis, i.e. no change in the mean, and under certain alternatives, i.e. different level shifts.

D.3.1 Normally distributed data

We consider

$$G(t) = t,$$

so that $(X_i)_{i\geq 1}$ is fGn.

- Table D.10 shows the level of both tests under Gaussian data.
- Table D.11 shows the power of both tests under Gaussian data.

D.3.2 Symmetric, normal-tailed data

We consider

$$\begin{aligned} G(t) &= \frac{-b \operatorname{sgn}\left(\Phi(t) - \frac{1}{2}\right) \log\left(1 - 2\left|\Phi(t) - \frac{1}{2}\right|\right)}{\sqrt{2b^2}} \\ &= \begin{cases} \frac{1}{\sqrt{2}} \log(2\Phi(t)) & \text{if } t \leq 0 \\ -\frac{1}{\sqrt{2}} \log\left(2(1 - \Phi(t))\right) & \text{else} \end{cases} \end{aligned}$$

so that $(X_i)_{i\geq 1}$ follows a standardised Laplace (or double exponential) distribution.

- Table D.12 shows the level of both tests under standardised Laplace(0,4) distributed data.
- Table D.13 shows the power of both tests under standardised Laplace(0,4) distributed data.

n / H	0.6	0.7	0.9
10	0.024	0.031	0.038
50	0.039	0.042	0.047
100	0.042	0.046	0.044
500	0.047	0.052	0.049
1000	0.047	0.052	0.053

n / H	0.6	0.7	0.9
10	0.013	0.026	0.326
50	0.042	0.050	0.167
100	0.044	0.051	0.140
500	0.051	0.052	0.100
1000	0.051	0.054	0.095

Table D.10: Level of "difference-of-means" test (left) and level of "Wilcoxon-type" test (right) for fGn time series with LRD parameter H; $10,000$ simulation runs. Both tests have asymptotically level 5%.

h / λ	0.05	0.1	0.3	0.5
0.5	0.060	0.090	0.388	0.524
1	0.090	0.254	0.952	0.985
2	0.261	0.965	1.000	1.000

h / λ	0.05	0.1	0.3	0.5
0.5	0.058	0.086	0.386	0.525
1	0.077	0.184	0.948	0.985
2	0.119	0.763	1.000	1.000

Table D.11: Power of "difference-of-means" test (left) and power of "Wilcoxon-type" test (right) for $n = 500$ observations of fGn with LRD parameter $H = 0.7$, different break points $[\lambda n]$ and different level shifts h. Both tests have asymptotically level 5%. The calculations are based on 10,000 simulation runs.

n / H	0.6	0.7	0.9
10	0.036	0.046	0.057
50	0.043	0.05	0.057
100	0.046	0.051	0.056
500	0.051	0.051	0.054
1000	0.049	0.053	0.054

n / H	0.6	0.7	0.9
10	0.013	0.026	0.326
50	0.042	0.050	0.167
100	0.044	0.051	0.140
500	0.051	0.052	0.100
1000	0.051	0.054	0.095

Table D.12: Level of "difference-of-means" test (left) and level of "Wilcoxon-type" test (right) for standardised Laplace(0,4)-transformed fGn with LRD parameter H, $10,000$ repetitions. Both tests have asypmtotically level 5%.

h / λ	0.05	0.1	0.3	0.5
0.5	0.060	0.090	0.401	0.540
1	0.090	0.262	0.960	0.988
2	0.269	0.971	1.000	1.000

h / λ	0.05	0.1	0.3	0.5
0.5	0.061	0.101	0.544	0.704
1	0.084	0.238	0.987	0.998
2	0.120	0.809	1.000	1.000

Table D.13: Power of "difference-of-means" test (left) and power of "Wilcoxon-type" test (right) for $n = 500$ observations of standardised Laplace(0,4)-transformed fGn with LRD parameter $H = 0.7$, $10,000$ repetitions, different break points $[\lambda n]$ and different break height h. Both tests have asypmtotically level 5%.

D.3.3 Heavy-tailed data

We consider the transformation

$$G(t) = \left(\frac{\beta k^2}{(\beta-1)^2(\beta-2)}\right)^{-1/2}\left(k(\Phi(t))^{-1/\beta} - \frac{\beta k}{\beta-1}\right)$$

which yields Pareto(β, k) distributed observations which exhibit heavy tails. For $\beta = 3$, $k = 1$, i.e. for

$$G(t) = \frac{1}{\sqrt{3/4}}\left((\Phi(t))^{-1/3} - \frac{3}{2}\right),$$

the data have finite expectation and finite variance.

- Table D.14 shows the level of both tests under standardised Pareto(3,1) distributed data. The tests are based on asymptotic critical values.
- Table D.15 shows the power of both tests under standardised Pareto(3,1) distributed data. The tests are based on asymptotic critical values.
- Since the "difference-of-means" test does not reach its asymptotic level, the above power comparison is not meaningful. Table D.16 presents the power of the test under standardised Pareto(3,1) distributed data when the test is based on finite sample quantiles as critical values.

n / H	0.6	0.7	0.9
10	0.104	0.109	0.117
50	0.138	0.127	0.126
100	0.145	0.125	0.126
500	0.140	0.103	0.119
1000	0.131	0.101	0.123
2000	0.120	0.086	0.115
10000	0.106	0.069	0.101

n / H	0.6	0.7	0.9
10	0.013	0.026	0.326
50	0.042	0.050	0.167
100	0.044	0.051	0.140
500	0.051	0.052	0.100
1000	0.051	0.054	0.095

Table D.14: Level of "difference-of-means" test (left) and level of "Wilcoxon-type" test (right) for standardised Pareto(3,1)-transformed fGn with LRD parameter H; $10,000$ simulation runs. Both tests have asymptotically level 5%.

h / λ	0.05	0.1	0.3	0.5
0.5	0.116	0.177	0.756	0.864
1	0.177	0.693	0.998	1.000
2	0.815	0.998	1.000	1.000

h / λ	0.05	0.1	0.3	0.5
0.5	0.088	0.294	0.983	0.998
1	0.115	0.655	1.000	1.000
2	0.138	0.944	1.000	1.000

Table D.15: Power of "difference-of-means" test (left) and power of "Wilcoxon-type" test (right) for $n = 500$ observations of standardised Pareto(3,1)-transformed fGn with LRD parameter $H = 0.7$, different break points $[\lambda n]$ and different level shifts h. Both tests have asymptotically level 5%. The calculations are based on 10,000 simulation runs.

h / λ	0.05	0.1	0.3	0.5
0.5	0.053	0.078	0.566	0.733
1	0.078	0.379	0.994	0.999
2	0.423	0.994	1.000	1.000

Table D.16: Power of the "difference-of-means" test, based on the finite sample quantiles, for $n = 500$ observations of Pareto(3,1)-transformed fGn with LRD parameter $H = 0.7$, different break points $[\lambda n]$ and different level shifts h. The calculations are based on 10,000 simulation runs.

A second set of simulations was made with Pareto(β, k) distributed observations for $\beta = 3$, $k = 1$; these data have finite expectation, but an infinite variance, so that we consider the centralized transformation

$$G(t) = \frac{1}{\sqrt{\Phi(t)}} - 2.$$

- Table D.17 shows the level of both tests under standardised Pareto(2,1) distributed data. The tests are based on asymptotic critical values.
- Table D.18 shows the power of both tests under standardised Pareto(2,1) distributed data. The tests are based on asymptotic critical values.
- Since the "difference-of-means" test again does not reach its asymptotic level, we considered its performance when it is based on finite sample quantiles as critical values in order to compare it to the "Wilcoxon-type" test (which again already reaches its asymptotic level when it is based on the asymptotical critical values). The resulting power of the test is shown in Table D.16.

n / H	0.6	0.7	0.9
10	0.104	0.104	0.107
50	0.159	0.138	0.120
100	0.181	0.151	0.122
500	0.223	0.148	0.124
1000	0.232	0.151	0.130

n / H	0.6	0.7	0.9
10	0.013	0.026	0.326
50	0.042	0.050	0.167
100	0.044	0.051	0.140
500	0.051	0.052	0.100
1000	0.051	0.054	0.095

Table D.17: Level of "difference-of-means" test (left) and level of "Wilcoxon-type" test (right) for Pareto(2,1)-transformed fGn with LRD parameter H; $10,000$ simuation runs.

h / λ	0.05	0.1	0.3	0.5
0.5	0.148	0.156	0.272	0.350
1	0.156	0.200	0.741	0.853
2	0.199	0.651	0.996	0.999

h / λ	0.05	0.1	0.3	0.5
0.5	0.075	0.180	0.878	0.960
1	0.097	0.401	0.997	1.000
2	0.122	0.744	1.000	1.000

Table D.18: Power of "difference-of-means" test (left) and power of "Wilcoxon-type" test (right) for $n = 500$ observations of Pareto(2,1)-transformed fGn with LRD parameter $H = 0.7$, different break points $[\lambda n]$ and different level shifts h. The calculations are based on 10,000 simulation runs.

H / n	10	50	100	500	1,000	∞
0.6	2.30	2.66	2.68	2.58	2.50	1.55
0.7	1.87	2.06	1.96	1.77	1.71	1.23
0.9	1.05	1.06	0.99	1.01	1.00	0.62

Table D.19: 5%-quantiles of the finite sample distribution of the "difference-of-means" test under the null hypothesis for Pareto(2,1)-transformed fGn with different LRD parameter H and different sample sizes n. The calculations are based on 10,000 simulation runs.

n / H	0.6	0.7	0.9
50	0.020	0.046	0.114
100	0.028	0.045	0.065
500	0.041	0.044	0.039
1000	0.042	0.044	0.043

n / H	0.6	0.7	0.9
50	0.024	0.087	0.447
100	0.038	0.076	0.282
500	0.049	0.052	0.114
1000	0.047	0.049	0.093

Table D.20: Level of "difference-of-means" test (left) and level of "Wilcoxon-type" test (right) for fGn time series with LRD parameter H, estimated by the Whittle estimator; $10,000$ simulation runs.

h / λ	0.05	0.1	0.3	0.5
0.5	0.039	0.050	0.250	0.389
1	0.035	0.068	0.775	0.917
2	0.018	0.078	0.998	1.000

h / λ	0.05	0.1	0.3	0.5
0.5	0.045	0.054	0.242	0.381
1	0.035	0.052	0.666	0.873
2	0.007	0.007	0.663	0.936

Table D.21: Power of "difference-of-means" test (left) and power of "Wilcoxon-type" test (right) for $n = 500$ observations of fGn with LRD parameter $H = 0.7$, estimated by the Whittle estimator; different break points $[\lambda n]$ and different level shifts h. The calculations are based on 10,000 simulation runs.

D.4 Change-point tests with estimated Hurst parameter (section 7.1)

We redo the simulation for the change-point tests from Section 3.6, but this time the Hurst parameter H is estimated from the data befor the test is applied.

D.4.1 Normally distributed data

We consider $G(t) = t$, so that $(X_i)_{i\geq 1}$ is fGn.

- Table D.20 shows the level of both tests under Gaussian data and estimated H.
- Table D.21 shows the power of both tests under Gaussian data and estimated H.

D.4.2 Heavy-tailed data

We consider the transformation

$$G(t) = \left(\frac{\beta k^2}{(\beta-1)^2(\beta-2)}\right)^{-1/2}\left(k(\Phi(t))^{-1/\beta} - \frac{\beta k}{\beta-1}\right)$$

which yields Pareto(β, k) distributed observations which exhibit heavy tails. For $\beta = 3$, $k = 1$ the data have finite expectation and finite variance.

n / H	0.6	0.7	0.9
50	0.175	0.241	0.273
100	0.212	0.286	0.269
500	0.315	0.392	0.282
1000	0.359	0.441	0.309

n / H	0.6	0.7	0.9
50	0.098	0.258	0.643
100	0.126	0.299	0.550
500	0.207	0.366	0.469
1000	0.243	0.413	0.468

Table D.22: Level of "difference-of-means" test (left) and level of "Wilcoxon-type" test (right) for standardised Pareto(3,1)-transformed fGn with LRD parameter H, estimated by the Whittle estimator; 10,000 simulation runs.

h / λ	0.05	0.1	0.3	0.5
0.5	0.378	0.472	0.882	0.939
1	0.397	0.686	0.994	0.998
2	0.437	0.794	0.999	1.000

h / λ	0.05	0.1	0.3	0.5
0.5	0.415	0.698	0.991	0.997
1	0.344	0.679	0.997	0.999
2	0.124	0.215	0.755	0.874

Table D.23: Power of "difference-of-means" test (left) and power of "Wilcoxon-type" test (right) for $n = 500$ observations of standardised Pareto(3,1)-transformed fGn with LRD parameter $H = 0.7$, estimated by the Whittle estimator; different break points $[\lambda n]$ and different level shifts h. The calculations are based on 10,000 simulation runs.

- Table D.22 shows the level of both tests under standardised Pareto(3,1) distributed data and estimated H.
- Table D.23 shows the power of both tests under standardised Pareto(3,1) distributed data and estimated H.

D.5 Estimating Hermite coefficients (section 7.3)

We consider the model $X_i = G(\xi_i)$, $i = 1, \ldots, n$, where $(\xi_i)_{i \geq 1}$ is a stationary mean-zero Gaussian process with $\mathrm{Var}[\xi_i] = 1$ and auto-covariance function (1.1) and a transformation $G \in \mathcal{G}^2$ which has Hermite rank $m = 1$. We have observed n data $X_1, \ldots, X_n$ and estimate the first Hermite coefficient

$$a_1 = E[\xi\, G(\xi)],$$

where $\xi \sim \xi_1$ by

$$\widetilde{a_1} = \frac{1}{n} \sum_{i=1}^{n} \xi'_{(i)} X_{(i)},$$

where $\xi'_1, \ldots, \xi'_n$ are n i.i.d. standard normal distributed random variables and $X_{(i)}$ is the i-th variable in the sorted sample $X_{(1)} \leq \ldots \leq X_{(n)}$ ($\xi'_{(i)}$ analogously).

	Gauß		Laplace(0,4)		Pareto(3,1)		Pareto(2,1)	
n	rel	var	rel	var	rel	var	rel	var
10	0.720	0.091	0.702	0.112	0.671	0.192	0.671	3.287
50	0.911	0.021	0.904	0.031	0.885	0.109	0.844	1.085
100	0.946	0.011	0.936	0.017	0.922	0.064	0.913	1.204
500	0.982	0.002	0.981	0.004	0.971	0.021	0.955	0.270
1000	0.989	0.001	0.988	0.002	0.983	0.014	0.967	0.159
2000	0.993	0.001	0.993	0.001	0.987	0.008	0.978	0.116
5000	0.996	0.000	0.996	0.000	0.993	0.004	0.988	0.071
10000	0.998	0.000	0.998	0.000	0.995	0.003	0.994	0.039

Table D.24: Estimated Hermite coefficients $\widetilde{a_1}$, the mean relative to the true value a_1 (rel) and the variance (var), based on $10,000$ repetitions, for different G

For $10,000$ realizations of a fGn series of length n (with Hurst parameter $H = 0.7$ and varying sample size n) and for different transformations G (such that normal, Laplace and two different Pareto distributed observations arise) I have calculated $\widetilde{a_1}$. Based on these each $10,000$ values, I have calculated the sample mean and the sample variance.

- Table D.24 shows the mean of $\widetilde{a_1}$ relative to the true coefficient a_1 as well as the variance of $\widetilde{a_1}$.

D.6 Wilcoxon's two-sample statistic (section 5.4)

For realizations $\xi_1, \ldots, \xi_n$ of fGn for different Hurst parameters H, for different cutting points $\lambda \in [0,1]$ and different sample sizes n, I have calculated the two-sample Wilcoxon test statistic

$$U_{W,[\lambda n],n} = \frac{1}{\sigma_W} \left(n^{-2+\frac{D}{2}} \sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda N]+1}^{n} \left(I_{\{\xi_i \leq \xi_j\}} - \frac{1}{2} \right) \right)$$

with

$$\sigma_W^2 := \frac{1}{4\pi} \left(\lambda^2 - \lambda + (1-\lambda)\lambda^{2-D} + \lambda(1-\lambda)^{2-D} \right).$$

I have repeated this $10,000$ times for each choice of parameters n, λ, H. The results are given

- for $H = 0.6$ in Table D.25,
- for $H = 0.7$ in Table D.26 and
- for $H = 0.9$ in Table D.27.

$n \setminus \lambda$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	1.196	1.195	1.2	1.213	1.212	1.188	1.189	1.191	1.199
50	1.065	1.025	1.019	1.023	1.045	1.051	1.061	1.075	1.076
100	1.009	1.017	1.036	1.035	1.027	1.022	1.018	1.023	1.035
500	1.003	1.008	1.025	1.031	1.053	1.035	1.023	1.021	1.025
1000	1.024	1.038	1.031	1.022	1.016	1.023	1.029	1.016	1.007
2000	1.028	1.011	1.005	1.001	1.009	1.005	1.001	1.001	1.011

Table D.25: Sample variance of $10,000$ values of $U_{W,[\lambda n],n}$ for fGn with Hurst parameter $H = 0.6$ ($D = 0.8$). n is the overall sample size, the data was divided into two samples after the $[\lambda n]$-th observation.

$N \setminus \lambda$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	1.331	1.302	1.285	1.285	1.284	1.289	1.292	1.299	1.339
50	1.123	1.108	1.079	1.068	1.076	1.073	1.070	1.091	1.087
100	1.087	1.054	1.048	1.066	1.059	1.062	1.051	1.043	1.059
500	1.021	1.005	1.014	1.017	1.013	1.015	1.013	1.005	1.018
1000	1.000	1.019	1.029	1.027	1.030	1.016	1.018	1.022	1.022
2000	1.015	1.014	1.008	1.014	1.003	1.011	1.022	1.025	1.032

Table D.26: Sample variance of $10,000$ values of $U_{W,[\lambda n],n}$ for fGn with Hurst parameter $H = 0.7$ ($D = 0.6$). n is the overall sample size, the data was divided into two samples after the $[\lambda n]$-th observation.

$N \setminus \lambda$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	2.539	2.430	2.394	2.398	2.380	2.396	2.413	2.45	2.584
50	1.710	1.680	1.680	1.666	1.663	1.673	1.680	1.694	1.731
100	1.564	1.541	1.532	1.529	1.523	1.538	1.551	1.565	1.585
500	1.339	1.327	1.316	1.316	1.308	1.319	1.319	1.325	1.340
1000	1.282	1.270	1.278	1.285	1.288	1.299	1.288	1.273	1.278
2000	1.229	1.238	1.243	1.231	1.236	1.226	1.231	1.240	1.237

Table D.27: Sample variance of $10,000$ values of $U_{W,[\lambda n],n}$ for fGn with Hurst parameter $H = 0.9$ ($D = 0.2$). n is the overall sample size, the data was divided into two samples after the $[\lambda n]$-th observation.

List of Figures

1.1 Annual minima of the water level in the Nile river 7
1.2 Auto-correlation function of Nile river data 7
1.3 Fractional Brownian motion, looks from different distances 13
1.4 Fractional Brownian motion, different Hurst parameters 13
1.5 Fractional Gaussian noise, different Hurst parameters 14
1.6 Relation between LRD behaviour of the ξ_i and $G(\xi_i)$ 14

2.1 Density of $\bar{X} - \bar{Y}$-statistic 31
2.2 Unnormalized variance of $\bar{X} - \bar{Y}$ 31
2.3 $\bar{X} - \bar{Y}$ for one-sample and two-sample case 32
2.4 Estimation of Var$[\bar{X}]$ 37
2.5 Auto-covariance estimator $\widehat{\gamma}_h$, average and confidence belts 38
2.6 Auto-covariance estimator $\widehat{\gamma}_h$, different realizations 38
2.7 Estimations of Var $[\bar{X}_m - \bar{Y}_n]$ 45
2.8 Estimations of Var $[\bar{X}_m - \bar{Y}_n]$, with trimmed auto-covariances 45
2.9 Covariance estimation influences estimation of Var $[\bar{X}_m - \bar{Y}_n]$ 46

3.1 Estimated p.d.f. and c.d.f. of $\sup_{0\leq\lambda\leq 1} |Z(\lambda) - \lambda Z(1)|$ 61
3.2 fGn with breaks 67
3.3 Level of "diff.-of-means" and "Wilcoxon-type" test, Gaussian data . . . 68
3.4 Power of "diff.-of-means" and "Wilcoxon-type" test, Gaussian data . . . 68
3.5 Level of "diff.-of-means" and "Wilcoxon-type" test, Laplace data 69
3.6 Power of "diff.-of-means" and "Wilcoxon-type" test, Laplace data 69
3.7 Level of "diff.-of-means" and "Wilcoxon-type" test, Pareto(3,1) data . . 70
3.8 Power of "diff.-of-means" and "Wilcoxon-type" test, Pareto(3,1) data . . 70
3.9 Power of "diff.-of-means" and "Wilcoxon-type" test, Pareto(3,1) data, finite sample quantiles 75
3.10 Level of "diff.-of-means" and "Wilcoxon-type" test, Pareto(2,1) data . . 75
3.11 Power of "diff.-of-means" and "Wilcoxon-type" test, Pareto(2,1) data . . 76
3.12 Power of "diff.-of-means" and "Wilcoxon-type" test, Pareto(2,1) data, finite sample quantiles 76

4.1 Power of "diff.-of-means" and "Wilcoxon-type" test, Gaussian data, sample size $n = 2,000$ . . . 103
4.2 Change in power of "diff.-of-means" and "Wilcoxon-type" test, when sample size increases . . . 103

5.1 Density of scaled and normalized Wilcoxon statistic $U_{W,[\lambda n],n}$ . . . 124

6.1 Applying the Phragmén-Lindelöf principle . . . 151

7.1 Boxplots of Hurst parameter estimates . . . 163
7.2 MSE of Hurst parameter estimates . . . 163
7.3 Boxplots of rescaling $n^{2-D/2}/n^{2-\hat{D}/2}$ under null hypothesis . . . 164
7.4 Boxplots of rescaling $n^{2-D/2}/n^{2-\hat{D}/2}$ under alternative . . . 164
7.5 Level of "diff.-of-means" and "Wilcoxon-type" test, Gaussian data, estimated H . . . 165
7.6 Power of "diff.-of-means" and "Wilcoxon-type" test, Gaussian data, estimated H . . . 165
7.7 Level of "diff.-of-means" and "Wilcoxon-type" test, Pareto(3,1) data, estimated H . . . 166
7.8 Power of "diff.-of-means" and "Wilcoxon-type" test, Pareto(3,1) data, estimated H . . . 166
7.9 Estimators for H under jump, Whittle Estimator . . . 175
7.10 Estimators for H under jump, Box-Periodogram Estimator . . . 176
7.11 Difference of e.d.f. and c.d.f. . . . 181
7.12 Estimated Hermite coefficients $\widetilde{a_1}$ . . . 182

B.1 Monotonicity in higher dimensions . . . 207
B.2 $I_{\{x \leq y\}}$ has infinite variation over any rectangle crossing the diagonal. . . . 211

List of Tables

3.1 α-quantiles of $\sup_{0\leq\lambda\leq 1}|Z(\lambda)-\lambda Z(1)|$ 62
3.2 Finite sample quantiles of "diff.-of-means" test, Pareto(3,1) data 77

4.1 Power of "diff.-of-means" test, relative to "Wilcoxon-type" test 102
4.2 Power of two-sample Gauß test and Wilcoxon test, Gaussian data 106
4.3 Power of two-sample Gauß test, relative to Wilcoxon test 106
4.4 Power of two-sample Gauß test and Wilcoxon test, Pareto data 107

7.1 Estimators for H under jump 173
7.2 Estimators for H under jump, Box-Periodogram Estimator 177

D.1 Sample variance of Gauß test statistic, one divided sample, $H = 0.6$ . . 236
D.2 Sample variance of Gauß test statistic, one divided sample, $H = 0.7$ . . 236
D.3 Sample variance of Gauß test statistic, one divided sample, $H = 0.9$ . . 236
D.4 Sample variance of Gauß test statistic, two indep. samples, $H = 0.6$ . . 238
D.5 Sample variance of Gauß test statistic, two indep. samples, $H = 0.7$ . . 238
D.6 Sample variance of Gauß test statistic, two indep. samples, $H = 0.6$ . . 238
D.7 Estimation of $\mathrm{Var}[\bar{X}]$, $H = 0.6$ 239
D.8 Estimation of $\mathrm{Var}[\bar{X}]$, $H = 0.7$ 240
D.9 Estimation of $\mathrm{Var}[\bar{X}]$, $H = 0.9$ 241
D.10 Level of "diff.-of-means" and "Wilcoxon-type" test, Gaussian data . . . 243
D.11 Power of "diff.-of-means" and "Wilcoxon-type" test, Gaussian data . . . 243
D.12 Level of "diff.-of-means" and "Wilcoxon-type" test, Laplace data 244
D.13 Power of "diff.-of-means" and "Wilcoxon-type" test, Laplace data 244
D.14 Level of "diff.-of-means" and "Wilcoxon-type" test, Pareto(3,1) data . . 245
D.15 Power of "diff.-of-means" and "Wilcoxon-type" test, Pareto(3,1) data . . 245
D.16 Power of "diff.-of-means" and "Wilcoxon-type" test, Pareto(3,1) data, finite sample quantiles 245
D.17 Level of "diff.-of-means" and "Wilcoxon-type" test, Pareto(2,1) data . . 247
D.18 Power of "diff.-of-means" and "Wilcoxon-type" test, Pareto(2,1) data . . 247
D.19 Finite sample quantiles of "difference-of-means" test, Pareto(2,1) data . 247
D.20 Level of "diff.-of-means" and "Wilcoxon-type" test, Gaussian data, estimated H 248

D.21 Power of "diff.-of-means" and "Wilcoxon-type" test, Gaussian data, estimated H 248
D.22 Level of "diff.-of-means" and "Wilcoxon-type" test, Pareto(3,1) data, estimated H 249
D.23 Power of "diff.-of-means" and "Wilcoxon-type" test, Pareto(3,1) data, estimated H 249
D.24 Estimated Hermite coefficients $\widetilde{a_1}$ 250
D.25 Sample variance of Wilcoxon test statistic, one divided sample, $H = 0.6$ 252
D.26 Sample variance of Wilcoxon test statistic, one divided sample, $H = 0.7$ 252
D.27 Sample variance of Wilcoxon test statistic, one divided sample, $H = 0.9$ 252

Bibliography

C. R. Adams, J. A. Clarkson (1934): Properties of functions $f(x, y)$ of bounded variation, *Trans. Amer. Math. Soc.*, 1934, 36(4), 711–730

J. Antoch, M. Hušková, A. Janic, T. Ledwina (2008): Data driven rank tests for the change point problem, *Metrika*, 2008, 68, 1–15

A. Aue, S. Hörmann, L. Horváth, M. Reimherr (2009): Break detection in the covariance structure of multivariate time series, *The Annals of Statistics*, 2009, 37, 4046–4087.

J. T. Barkoulas, C. F. Baum, N. Travlos (2000): Long memory in the Greek stock market, *Applied Financial Economics*, 2000, 10(2), 177–184

R. T. Baillie (1996): Long memory processes and fractional integration in econometrics, *Journal of Econometrics*, 1996, 73(1), 5–59

M. Basseville, I. V. Nikiforov (1993): *Detection of Abrupt Changes: Theory and Application*, Prentice-Hall Inc., Englewood Cliffs 1993, ISBN 0-13-126780-9, `http://http://www.irisa.fr/sisthem/kniga/`

H. Bateman, A. Erdélyi (1953): *Bateman Manuscript Project: Higher transcendental functions, 2*, McGraw-Hill Book Company Inc., New York 1953

J. Beran (1994): *Statistics for Long-Memory Processes*, Chapman & Hall/CRC, Boca Raton (Florida) 1994, ISBN 0-412-04901-5

J. Beran (2010): Long-range dependence, *Wiley Interdisciplinary Reviews: Computational Statistics*, 2010, 2, 26–35

I. Berkes, L. Horváth, P. Kokoszka, Q.-M. Shao (2006): On discriminating between long-range dependence and changes in the mean, *The Annals of Statistics*, 2006, 34, 1140–1165.

E. Beutner, W. B. Wu, H. Zähle (2012): Asymptotics for statistical functionals of long-memory sequences, *Stochastic Processes and their Applications*, 122(3), 910–929

E. Beutner, H. Zähle (2011): Deriving the asymptotic distribution of U- and V-statistics of dependent data using weighted empirical processes, *Bernoulli*, to appear

J. Bhan, S. Kim, J. Kim, Y. Kwon, S. Yang, K. Lee (2006): Long-range correlations in Korean literary corpora, *Chaos, Solitons & Fractals*, 2006, 29(1), 69–81

N. H. Bingham, C. M. Goldie, J. L. Teugels (1989): *Regular Variation*, Cambridge University Press, Cambridge 1989, ISBN 0-521-37943-1

F. J. Breidt, N. Crato, P. de Lima (1998): The detection and estimation of long memory in stochastic volatility, *Journal of Econometrics*, 1998, 83(1-2), 325–348

P. Breuer, P. Major (1983): Central limit theorems for non-linear functionals of Gaussian fields, *Journal of Multivariate Analysis*, 1983, 13(3), 425–441

B. E. Brodsky, B. S. Darkhovsky (1993): *Nonparametric methods in change-point problems*, Kluwer Academic Publishers, Dordrecht 1993, ISBN 0-7923-2122-7

S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C.-K. Peng, M. Simons, H. E. Stanley (1995): Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis, *Phys. Rev. E*, 1995, 51(5), 5084–5091

R. Caballero, S. Jewson, A. Brix (2002): Long memory in surface air temperature: detection, modeling, and application to weather derivative valuation, *Climate Research*, 2002, 21, 127–140

O. Cappé, E. Moulines, J.-C. Pesquet, A. Petropulu, X. Yang (2002): Long-range dependence and heavy-tail modeling for teletraffic data, *IEEE Signal Processing Magazine*, 2002, 19(3), 14–27

Y. Chen, M. Ding, J. A. S. Kelso (1997): Long Memory Processes ($1/f^\alpha$ Type) in Human Coordination, *Physical Review Letters*, 79(22), 4501–4504

Y.-W. Cheung (1993): Long memory in foreign exchange rates, *Journal of Business and Economic Statistics*, 1993, 11, 93–101

Y.-W. Cheung, K. S. Lai (1995): A search for long memory in international stock market returns, *Journal of International Money and Finance*, 1995, 14(4), 597–615

J. A. Clarkson, C. R. Adams (1933): On definitions of bounded variation for functions of two variables, *Trans. Amer. Math. Soc.*, 1933, 35(4), 824–854

M. Csörgő, L. Horváth (1988): Invariance Principle for Changepoint Problems, *Journal of Multivariate Analysis*, 1988, 27, 151–168

M. Csörgő, L. Horváth (1997): *Limit Theorems in Change-Point Analysis*, John Wiley & Sons, Chichester 1997, ISBN 0-471-95522-1

H. Dehling, R. Fried (2010): Asymptotic Distribution of Two-Sample Empirical U-Quantiles with Applications to Robust Tests for Structural Change, *SFB 823 Discussion Paper*, 2010, 43

H. Dehling, A. Rooch, M. S. Taqqu (2012): Nonparametric Change-Point Tests for Long-Range Dependent Data. *Scand. J. Stat.*, 2012, to appear

H. Dehling, A. Rooch, M. S. Taqqu (2013): Power of Change-Point Tests for Long-Range Dependent Data. *preprint*, 2013

H. Dehling, M. S. Taqqu (1989): The Empirical process of some long-range dependent sequences with an application to U-statistics, *Ann. Stat.*, 1989, 17(4), 1767–1783

H. Dehling, M. S. Taqqu (1991): Bivariate symmetric statistics of long-range dependent observations, *Journal of Statistical Planning and Inference*, 1991, 28, 153–165

R. L. Dobrushin, P. Major (1979): Non-Central Limit Theorems for Non-Linear Functionals of Gaussian Fields, *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 1979, 50(1), 27–52

A. Erramilli, O. Narayan, W. Willinger (1996): Experimental queueing analysis with long-range dependent packet traffic, *IEEE/ACM Transactions on Networking*, 1996, 4(2), 209–223

K. Fritzsche, H. Grauert (2002): *From Holomorphic Functions to Complex Manifolds*, Springer, New York 2002, ISBN 978-0-387-95395-3

T. Geisel, J. Nierwetberg, A. Zacherl (1985): Accelerated Diffusion in Josephson Junctions and Related Chaotic Systems, *Phys. Rev. Lett.*, 1985, 54(7), 616–619

L. Giraitis, R. Leipus, D. Surgailis (1996): The change-point problem for dependent observations, *J. Statist. Plann. Inference*, 53, 297–310

L. Giraitis, M. S. Taqqu (1999): Whittle Estimator for Finite-Variance Non-Gaussian Time Series with Long Memory, *The Annals of Statistics*, 1999, 27(1), 178–203

K. Gröchenig (2001): *Foundations of Time-Frequency Analysis*, Birkhäuser, Boston 2001, ISBN 978-0-817-64022-4

H. Heuser (2003): *Lehrbuch der Analysis, Teil 1*, 15th edition, B.G. Teubner GmbH, Stuttgart 2003, ISBN 3-519-62233-5

G. H. Hardy (1905): On double Fourier series, and especially those with represent the double zeta-function with real and incommensurable parameters, *Quart. J. of Math.*, 1905, 37, 53–79

U. Hassler, M. Olivares (2007): Long memory and structural change: New evidence from German stock market returns, *unpublished*

U. Hassler, J. Scheithauer (2009): Detecting Changes from Short to Long Memory. *Statistical Papers*, 2011, 52(4), 847–870

L. Horváth, P. Kokoszka (1997): The effect of long-range dependence on change-point estimators, *J. Statist. Plann. Inference*, 64, 57–81

T. Hsing, W. B. Wu (2004): On weighted U-statistics for stationary processes, *Ann. Probab.*, 2004, 32(2), 1600–1631

C.-C. Hsu (2005): Long memory or structural changes: An empirical examination on inflation rates, *Economics Letters*, 2005, 88(2), 289–294

H. Hurst (1951): Long Term Storage Capacity of Reservoirs, *Transactions of the American Society of Civil Engineers*, 1951, 116, 770–799

H. Hurst (1955): Methods of using long-term storage in reservoirs, *Proceedings of the Institution of Civil Engineers*, 1955, part I, 519–577

K. Itō (1951): Multiple Wiener Integral, *J. Math. Soc. Japan*, 1951, 3, 157–169

R. Kannan, C. K. Krueger (1996): *Advanced Analysis: On the Real Line*, Springer, New York 1996, ISBN 978-0-387-94642-9

T. Karagiannis, M. Faloutsos, R. H. Riedi (2002): Long-Range Dependence: Now you see it, now you don't!, *Global Telecommunications Conference*, 2002, 3, 2165–2169

T. Karagiannis, M. Molle, M. Faloutsos (2004): Long-Range Dependence, Ten Years of Internet Traffic Modeling, *IEEE Internet Computing*, 2004, 8(5), 57–64

P. Kokoszka, R. Leipus (1998): Change-point in the mean of dependent observations, *Stat. & Prob. Letters*, 40, 385–393

W. Krämer, P. Sibbertsen (2002): Testing for Structural Changes in the Persistence of Long Memory. *International Journal of Business and Economics*, 1, 235–242

W. Krämer, P. Sibbertsen, C. Kleiber (2002): Long memory versus structural change in financial time series, *Allgemeines Statistisches Archiv*, 2002, 86, 83–96

H. R. Künsch (1987): Statistical aspects of self-similar processes, in: Yu. A. Prohorov, V.V. Sazonov (eds.), *Proceedings of the First World Congres of the Bernoulli Society*, 1, 67–74, VNU Science Press, Utrecht

H.-H. Kuo (2006): *Introduction to Stochastic Integration*, Springer, New York 2006, ISBN 978-0-387-28720-1

H. Kuswanto (2009): A New Simple Test Against Spurious Long Memory Using Temporal Aggregation Heri Kuswanto, *Discussion Paper, Wirtschaftswissenschaftliche Fakultät der Leibniz Universität Hannover*, No. 425

J. W. Lamperti (1962): Semi-stable stochastic processes, *Trans. Am. Math. Soc.*, 1962, 104, 62–78

M. J. Lebo, R. W. Walker, H. D. Clarke (2000): You must remember this: dealing with long memory in political analyses, *Electoral Studies*, 2000, 19(1), 31–48

E. Lehmann (1975): *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, San Francisco

J. B. Levy, M. S. Taqqu (2000): Renewal Reward Processes with Heavy-Tailed Inter-Renewal Times and Heavy-Tailed Rewards, *Bernoulli*, 2000, 6(1), 23–44

Q. Li, D. L. Mills (1998): On the Long-range Dependence of Packet Round-trip Delays in Internet, *In Proceedings of IEEE ICC98*, 1998, 1185–1191

S. Ling (2007): Testing for change points in time series models and limiting theorems for NED sequences, *The Annals of Statistics*, 2007, 35, 1213–1227

F. Lillo, S. Mike, J. D. Farmer (2005): Theory for long memory in supply and demand, *Physical Review E*, 2005, 71(6), 066122-1–066122-11

P. Major (1981a): *Multiple Wiener-Itō Integrals*, Lecture Notes in Mathematics 849, Springer, Berlin Heidelberg 1981

P. Major (1981b): Limit theorems for non-linear functionals of Gaussian sequences, *Z.Wahrscheinlichkeitstheorie und verw. Gebiete*, 1981, 57, 129–158

B. B. Mandelbrot, J. W. van Ness (1968): Fractional Brownian Motions, Fractional Noises and Applications, *SIAM Review*, 1968, 10(4), 422–437

B. B. Mandelbrot, J. R. Wallis (1969a): Computer Experiments With Fractional Gaussian Noises, *Water Ressources Research*, 1969, 5(1), 228–241

B. B. Mandelbrot, J. R. Wallis (1969b): Some long-run properties of geophysical records, *Water Ressources Research*, 1969, 5(2), 321–340

F. Móricz (1976): Moment inequalities and the strong laws of large numbers, *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 1976, 35(4), 299–314

M. A. Montemurro, P. A. Pury (2002): Long-range fractal correlations in literary corpora, *Fractals*, 2002, 10(4), 451–461

B. Øksendal (1998): *Stochastic Differential Equations*, 5th edition, Springer, Berlin Heidelberg 1998, ISBN 3-540-63720-6

A. Ott, J.P. Bouchaud, D. Langevin, W. Urbach (1990): Anomalous diffusion in "living polymers": A genuine Levy flight?, *Phys. Rev. Lett.*, 1990, 65(17), 2201–2204

A. B. Owen (2004): Multidimensional variation for quasi-Monte Carlo, *Technical Report, Department of Statistics, Stanford University*, 2004, no. 2004-02

C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons and H. E. Stanley (1992): Long-range correlations in nucleotide sequences, *Nature*, 1992, 359, 168–170

C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger (1994): Mosaic organization of DNA nucleotides, *Phys. Rev. E*, 1994, 49(2), 1685–1689

V. Pipiras, M. S. Taqqu (2011): *Long Range Dependence and Self-Similarity*, preprint

Y. A. Rozanov (1967): *Stationary Random Processes*, Holden-Day, San Francisco 1967

G. Samorodnitsky (2007): *Long Range Dependence*, Now Publishers, Hanover 2007, ISBN 978-1-60198-090-8

G. Samorodnitsky, M. S. Taqqu (1994): *Non-Gaussian Stable Processes: Stochastic Models with Infinite Variance*, Chapman and Hall, London, 1994

X. Shao (2011): A simple test of changes in mean in the possible presence of long-range dependence, *Journal of Time Series Analysis*, 2011, 32(6), 598–606

G. R. Shorack, J. A. Wellner (1986): *Empirical Processes with Applications to Statistics*, John Wiley & Sons, New York 1986, ISBN 0-471-86725-X

P. Sibbertsen (2004): Long Memory versus Structural Breaks: An Overview, *Statistical Papers*, 2004, 45, 465–515

P. Sibbertsen, J. Willert (2010): Testing for a Break in Persistence under Long-Range Dependencies and Mean Shifts, *Statistical Papers*, 2010

B. Simon (1974): *The $P(\Phi)_2$ Euclidean (Quantum) field theory*, Princeton University Press, Princeton 1974, ISBN 0-691-08144-1

R. L. Smith (1993): Long-Range Dependence and Global Warming, in: R. A. Madden, R. W. Katz: *Applications of Statistics to Modeling the Earth's Climate System*, 1994, 89–92

D. Surgailis (2003): CLTs for Polynomials of Linear Sequences: Diagram Formula with Illustrations, in: P. Doukhan, G. Oppenheim, M. S. Taqqu (eds.): *Theory and Applications of Long-Range Dependence*, Birkhäuser, Boston Basel Berlin 2003, ISBN 0-817-64168-8

M. S. Taqqu (1975): Weak Convergence to Fractional Brownian Motion and to the Rosenblatt Process, *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 1975, 31, 287–302

M. S. Taqqu (1979): Convergence of Integrated Processes of Arbitrary Hermite Rank, *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 1979, 50, 53–83

M. S. Taqqu (2003): *Fractional Brownian Motion and Long-Range Dependence*, in: P. Doukhan, G. Oppenheim, M. S. Taqqu (eds.): Theory and Applications of Long-Range Dependence, Birkhäuser, Boston Basel Berlin 2003, ISBN 0-817-64168-8

M. S. Taqqu (2012): *personal communication*

M. S. Taqqu, V. Teverovsky (1998): On Estimating the Intensity of Long-Range Dependence in Finite and Infinite Variance Time Series, in: R. Adler, R. Feldmann, M. S. Taqqu (eds.): *A Practical Guide To Heavy Tails: Statistical Techniques and Applications*, Birkhäuser, Boston 1998, 177–217

M. S. Taqqu, V. Teverovsky, W. Willinger (1995): Estimators for long-range dependence: an emprical study, *Fractals*, 1995, 3, 785–798

M. S. Taqqu, W. Willinger, R. Sherman (1997): Proof of a fundamental result in self-similar traffic modelling, *ACM/SIGCOMM Computer Communications Review*, 1997, 27, 5–23

V. Teverovsky, M. Taqqu (1997): Testing for long-range dependence in the presence of shifting means or a slowly declining trend, using a variance-type estimator, *J. Time Ser. Anal.*, 1997, 18, 279–304

E. C. Titchmarsh (1964): *The Theory of Functions*, 2nd edition, Oxford University Press 1964, ISBN 0-198-53349-7

C. Varotsos, D. Kirk-Davidoff (2006): Long-memory processes in ozone and temperature variation at the region 60° S–60° N, *Atmospheric Chemistry and Physics*, 2006, 6(12), 4093–4100

M. K. Vemuri (2008): Hermite expansions and Hardy's theorem, *arXiv:0801.2234v1 [math.AP]*, 2008

W. H. Young (1913): On Multiple Fourier Series, *Proc. London Math. Soc. (2)*, 1913, 11, 133–184

W. H. Young (1916): On multiple integration by parts and the second theorem of the mean, *Proc. London Math. Soc. (2)*, 1916, 16, 273–293

L. Wang (2003): Limit theorems in change-point problems with multivariate long-range dependent observations, *Statistics & Decisions*, 21(3), 283–300

L. Wang (2008a): Change-in-mean problem for long memory time series models with applications, *J. Statist. Computat. Simul.*, 78(7), 653–668

L. Wang (2008b): Change-point detection with rank statistics in long-memory time-series models, *Aust. N. Z. J. Stat.*, 50(3), 241–256

E. W. Weisstein (2010): Hermite Polynomial, *MathWorld–A Wolfram Web Resource*, `http://mathworld.wolfram.com/HermitePolynomial.html`, 2010-09-01

D. Wied, W. Krämer, H. Dehling (2011): Testing for a change in correlation at an unknown point in time using an extended functional delta method, *Econometric Theory*, to appear

N. Wiener (1938): The homogeneous chaos, *Amer. J. Math.*, 1938, 60, 897–936

W. Willinger, M. S. Taqqu, R. Sherman, D. V. Wilson (1997): Self-similarity through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level, *IEEE/ACM Transaction on Networking*, 1997, 5(1), 71–86

W. Willinger, M. S. Taqqu, V. Teverovsky (1999): Stock market prices and long-range dependence, *Finance and Stochastics*, 1999, 3(1), 1–13