Sequential Nonparametric Estimation via Hermite Series Estimators

Michael Jared Stephanou

A thesis submitted to the Faculty of Science, University of Cape Town, Cape Town in fulfillment of the requirements for the degree of Doctor of Philosophy in Statistics.

Supervisor: Doctor Melvin Varughese

January 2020

Declaration

I declare that all results presented in this thesis are original except where reference is made to the work of others. Some of the original results in this thesis have been published in the following works,

1. Michael Stephanou, Melvin Varughese, and Iain Macdonald. "Sequential quantiles via Hermite series density estimation." Electron. J. Stat., 11(1):570-607, 2017, arXiv preprint arXiv:1507.05073v2.

2. Michael Stephanou and Melvin Varughese. "On the properties of Hermite Series based Distribution Function Estimators." Metrika , 2020, doi: 10.1007/s00184-020-00785-z.

These papers are references in the bibliography ([62], [61]) and contribute to the subject matter of chapters 4, 5 (along with associated appendices A, B) respectively. This thesis is being submitted for the degree of Doctor of Philosophy in the University of the Cape Town, Cape Town. It has not been submitted before for any degree or examination in any other University.

Michael Stephanou

Signed by candidate

January 2020

Abstract

Sequential Nonparametric Estimation via Hermite Series Estimators

Michael Jared Stephanou

Algorithms for estimating the statistical properties of streams of data in real time, as well as for the efficient analysis of massive data sets, are becoming particularly pertinent given the increasing ubiquity of such data. In this thesis we introduce novel approaches to sequential (online) estimation in both stationary and non-stationary settings based on Hermite series density estimators.

In the univariate context we apply Hermite series based distribution function estimators to sequential cumulative distribution function estimation. These distribution function estimators are particularly useful because they allow the sequential estimation of the *full* cumulative distribution function. This is in contrast to the empirical distribution function estimator and smooth kernel distribution function estimator which only allow sequential cumulative probability estimation at predefined values on the support of the associated density function. We explore the asymptotic consistency and robustness properties of the Hermite series based cumulative distribution function estimator thereby redressing a gap in the literature. Given the sequential Hermite series based distribution function estimator, we obtain sequential quantile estimates numerically. Our algorithms go beyond existing sequential quantile estimation algorithms in that they allow arbitrary quantiles (as opposed to pre-specified quantiles) to be estimated at any point in time, in both the static and dynamic quantile estimation settings.

In the bivariate context we introduce a Hermite series based sequential estimator for the Spearman's rank correlation coefficient and provide algorithms applicable in both the stationary and non-stationary settings. To treat the the non-stationary setting, we introduce a novel, exponentially weighted estimator for the Spearman's rank correlation, which allows the local nonparametric correlation of a bivariate data stream to be tracked. To the best of our knowledge this is the first algorithm to be proposed for estimating a time-varying Spearman's rank correlation that does not rely on a moving window approach.

We explore the practical effectiveness of the Hermite series based estimators through real data and simulation studies, demonstrating competitive performance compared to leading existing algorithms. The potential applications of this work are manifold. Our sequential distribution function and quantile estimation algorithms can be applied to real time anomaly and outlier detection, real time provisioning for future demand as well as real time risk estimation for example. The Hermite series based Spearman's rank correlation estimator can be applied to fast and robust online calculation of correlation which may vary over time. Possible machine learning applications include fast feature selection and hierarchical clustering on massive data sets amongst others.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr Melvin Varughese for his untiring support and guidance over the years through all the demands of work and life. Your insights have been a wellspring of new and interesting avenues of exploration. I am grateful to Melvin and Iain Macdonald for the productive and enjoyable collaborations we have shared.

I would also like to express my deepest gratitude to my partner Laura, my parents Vias and Sue and my brothers Matthew and Jordan for their continual encouragement and support in every regard. Time with you makes all efforts worthwhile.

Finally, I would also like to thank Stuart Cullender, Norman Ives, Kavishin Pather, Bernhard Steinhardt, Matthew Stephanou, Nils Tania and Derrick Van Zyl for many helpful and engaging discussions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The statistical analysis of streaming data and one-pass analysis of massive data sets has become highly relevant in recent times. These data are being generated by a number of sources including the global financial markets, internet applications, sensors embedded in various devices (internet of things) and data-intensive scientific research endeavours such as the Large Hadron Collider and the Square Kilometre Array (see [14] for a survey of the field of Big Data). Ideally such algorithms should be able to process observations sequentially, without requiring the storage of all observations. In addition, the time taken to process each observation should not grow with the number of previous observations. The core aim of this thesis is to introduce novel estimators and algorithms to treat the sequential (online) estimation of fundamentally useful statistical properties of univariate data streams and massive data sets as well bivariate data streams and massive data sets.

In the univariate setting, certain statistical properties naturally lend themselves to efficient, sequential calculation such as the mean and variance [69, 50, 71, 12]. Depending on the application, these moments may not be sufficient however. This may be true for skewed data for example. Indeed, higher order moments are useful in a variety of applications and algorithms to calculate these moments in a sequential (online) manner have been in-

troduced (see [8, 52]). Similar developments include incremental formulae for cumulants up to fourth order [4, 19]. It may be necessary to obtain a more direct distributional summary of the data than moments can provide however. In addition, if the data stream being analysed is prone to outliers or gross errors then more robust statistics may be required. Quantiles are a natural choice in these settings. Examples of areas in which sequential quantile estimation is relevant include network traffic and latency analysis [11], real time fraud detection [10] and high frequency trading (see [47] for an introduction to sequential algorithms in high frequency trading). Other conceivable applications of sequential quantile estimation include real time detection of anomalies and flagging of noteworthy observations, real time outlier detection and removal, real time provisioning for future demand and load balancing as well as real time risk estimation. In many applications of interest one needs to determine whether a particular value or observation is greater than or less than a certain quantile. In such cases it is more direct to use the cumulative distribution function. Thus, closely linked to quantile estimation is distribution function estimation. In fact, the sequential estimation of the full cumulative distribution function is an important and useful development in itself. In this thesis we propose new distribution function and quantile estimators based on Hermite series estimators and study their properties. These results are novel and interesting in their own right. That said, the particular setting we consider is that of online estimation and thus existing nonparametric methods are weighed against our methods in this specific context. In the general context, there are of course a number of well established nonparametric distribution function and quantile estimators which we will briefly review below.

Consider a sequence of i.i.d. random variables, $\mathbf{x_i} \sim f(x)$, with cumulative distribution function $F(x)$ and associated continuous probability density function $f(x)$. The estimation of the distribution function is a fundamental problem in nonparametric statistics and has been well-studied. The ubiqui-

tous empirical distribution function (EDF) estimator is defined as

$$\hat{F}_1(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\mathbf{x}_i \leq x\}. \tag{1.1}$$

The Glivenko-Cantelli theorem demonstrates the uniform, almost sure convergence of this estimator [66]. The central limit theorem implies that pointwise, the EDF estimator is asymptotically normally distributed with standard $\sqrt{n}$ rate of convergence. $\hat{F}_1(x)$ is unbiased and has $\text{MSE}(\hat{F}_1(x)) = \frac{F(x)(1-F(x))}{n}$. The rate of convergence in mean square is thus $O(n^{-1})$. The EDF estimator has been demonstrated to be asymptotically deficient compared to the smooth kernel distribution function estimator, with a suitably chosen kernel type, in the MSE sense. This relative deficiency relates to the fact that more observations are required in order for the EDF to perform as well as the kernel distribution function estimator at a given x in the MSE sense [57]. The kernel distribution function estimator is defined as:

$$\hat{F}_2(x) = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{x} \frac{1}{h_n} K\left(\frac{\mathbf{x}_i - x'}{h_n}\right) dx', \tag{1.2}$$

where the kernel function $K(x)$ is a non-negative function that integrates to one and has mean zero and the bandwidth, $h_n > 0$, is a smoothing parameter. The almost sure uniform convergence of $\hat{F}_2(x)$ has also been proved [48]. In addition, $\hat{F}_2(x)$ has been shown to be asymptotically normally distributed [68].

More modern approaches to estimating distribution functions have been developed. An approach based on Bernstein polynomials is presented in [6] and further studied in [44] for distribution functions associated with probability densities supported on a closed interval. This estimator is less general than the previously described EDF and smooth kernel CDF estimator in that it cannot be applied in situations with unknown support. The Bernstein estimator of order $m > 0$ of $F(x)$ is defined for $x \in [0, 1]$ as:

$$\hat{F}_3(x) = \sum_{k=0}^{m} F_n(k/m) P_{k,m}(x),$$

where $P_{k,m}(x) = \binom{m}{k} x^k (1-x)^{m-k}$, $k = 0, 1, \cdots, m$, are binomial probabilities and F_n is the EDF estimator defined above for n observations. As noted in [44], this estimator has the advantageous property of generating estimates that are genuine distribution functions i.e. $\hat{F}_3(0) = 0$, $\hat{F}_3(1) = 1$ for any value of m and $\hat{F}_3(x)$ has a nonnegative first derivative on $[0, 1]$. In [44] it is shown that the Bernstein estimator of the distribution function is consistent in mean square and the rate provided. The mean integrated squared error rate is also provided. The relative MSE and MISE deficiency of the EDF with respect to the Bernstein estimator is also established. In addition, the pointwise asymptotic normality of the Bernstein estimator is demonstrated. A related approach to estimating distribution functions for probability densities supported on closed intervals is based on Bernstein-Durrmeyer operators (see [15]).

Closely related to the EDF, the sample quantile is a popular nonparametric estimator of the corresponding population quantile. Define the values $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)}$ to be a permutation of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ such that $\mathbf{x}_{(1)} \leq \mathbf{x}_{(2)} \leq \cdots \leq \mathbf{x}_{(n)}$. Here, $\mathbf{x}_{(i)}$, is known as the ith order statistic. The EDF can be written in terms of the order statistics as:

$$\hat{F}_4(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\mathbf{x}_{(i)} \leq x\}.$$

The inverse cumulative distribution function or quantile function is defined as:

$$q(p) = \inf\{x : F(x) \geq p\}. \tag{1.3}$$

The p-th quantile can be estimated from the order statistics. In particular if $p \in \left(\frac{i-1}{n}, \frac{i}{n}\right]$ then $\hat{q}(p) = \mathbf{x}_{(i)}$, the ith order statistic. The sample quantile

estimator is a function of at most two order statistics and thus may suffer a loss of efficiency for certain distributions. A natural way to improve efficiency is to form a weighted average of several order statistics under an appropriate weight function. Such estimators are called L-estimators. The most popular class of L-estimators uses a density function (kernel) as its weight function, these are known as kernel quantile estimators (see [60]). The kernel quantile estimator is defined as:

$$\hat{q}(p) = \sum_{i=1}^{n} \int_{\frac{i-1}{n}}^{\frac{i}{n}} \frac{1}{h_n} K\left(\frac{s-p}{h_n}\right) \mathbf{x}_{(i)} ds.$$

It has been shown that under suitable conditions on $F(x)$ and h_n the kernel quantile estimator is more efficient than the sample quantile estimator in the MSE sense [25]. Under comparable assumptions to those to prove joint asymptotic normality of a set of empirical quantiles, joint asymptotic normality of the kernel quantile estimator has also been proved [26]. The rate is also provided. In [60] the asymptotically optimal bandwidth for the kernel quantile estimator is derived and the corresponding MSE is provided. It is also shown that many different variants of the kernel quantile estimator are asymptotically equivalent in the MSE sense. In addition other L-estimators such as Harell-Davis, Kaigh-Lachenbruch and the Brewer estimators are shown to be asymptotically equivalent to the kernel quantile estimator above with a Gaussian kernel and certain smoothing parameters. A more modern approach to estimating quantiles based on the Bernstein-Durrmeyer operator is provided in [53]. Like other L-estimators, the BD estimator constitutes a weighted version of several order statistics. MSE and MISE consistency results are also provided.

In the context of sequential distribution function and quantile estimation, the aforementioned estimators have shortcomings. Both the EDF and kernel distribution function estimator only allow sequential estimation of the cumulative probability at a set of fixed x values (see chapters 4 and 5 of

[31] and chapter 7 of [20] for a discussion of recursive kernel estimators). See also, [41] for an approach using Bernstein polynomials. For quantile estimation, both the sample quantile estimator and L-estimators such as the kernel quantile estimator and the Bernstein-Durrmeyer estimator require the storage and updating of one or more order statistics (a sorted sequence of all observations seen so far). Updating the order statistics cannot in general be done in $O(1)$ time. Moreover, the addition of a new observation would in general change a number of order statistics. Finally, in the context of sequential quantile estimation on streaming data, non-stationarity cannot be naturally addressed since these estimators have no means of *forgetting* previous observations (other than windowing and resetting).

Approaches specifically for the sequential estimation of quantiles have been developed. Sequential quantile estimation algorithms can be differentiated on whether they seek to maintain an online estimate of a single quantile or multiple quantiles. They can be further differentiated on whether they are meant to estimate static quantiles of a stream of data or dynamic quantiles of a stream of data. In the case of static quantile estimation, online quantile estimates pertain to all the data observed so far and quantiles of the stream being analysed are assumed to be fixed. In the dynamic case, quantile estimates pertain to the current behaviour of the process and it is assumed that quantiles may vary over time. A number of algorithms have been proposed for sequential quantile estimation in these settings. In [40] the P^2 algorithm was proposed which utilises parabolic interpolation in order to estimate a particular quantile. In [55] and [56] this algorithm was extended to the simultaneous estimation of several quantiles. In [49] the P^2 algorithm was further extended to treat dynamic quantile estimation via exponentially weighted quantile estimators. Algorithms have also been proposed based on stochastic approximation [58], [64]. This approach was extended to dynamic quantile estimation in [13] by the introduction of Exponentially Weighted Stochastic Approximation (EWSA). Subsequent to (but independent of) the

developments in this thesis (published in [62]), other approaches to the online estimation of quantiles in static and dynamic settings have appeared in the literature ([73], [33], [32], [65]). All the aforementioned methods have a potentially limiting shortcoming however; online estimates can only be obtained for a pre-selected, fixed set of quantiles (e.g. $p = 0.5, 0.9, 0.99$ etc.).

In this thesis we propose novel techniques based on Hermite series *density* estimators (see [59, 67, 29, 30, 45] for the definition and properties of these density estimators) to maintain an online estimate of the full CDF and the full quantile function in both the static and dynamic settings and thus yield estimates of the cumulative probability at *arbitrary* x and estimates of *arbitrary* quantiles that can be updated in constant time ($O(1)$ time). This is the primary advantage of our suggested approach. Moreover, these estimators make it straightforward to calculate quantiles and cumulative probabilities in a decentralised (distributed) setting. Even if our proposed techniques only have comparable accuracy with existing techniques, they would still be valuable. We demonstrate using simulated and real data that our techniques may in fact be more accurate, particularly in the case of sequential quantile estimation. There are several motivating factors to using Hermite series estimators as opposed to other estimators as discussed in sections 2.2 and 2.3. In particular, we discuss advantages of Hermite series estimators over the Gram-Charlier Type A series and Edgeworth estimators in section 2.2 and advantages over various other orthogonal series estimators in section 2.3.

It is noteworthy that orthogonal polynomials have also been applied in treating other statistical problems in the streaming/distributed and massive data settings, namely computationally efficient parameter estimation and Bayesian inference in generalised linear models [38].

The particular contributions of this thesis with regards to sequential cumulative distribution function and quantile estimation are as follows:

- We define a sequential cumulative distribution function estimator based on Hermite series density estimators. In doing so, we furnish an analyt-

ical expression for the Hermite series based CDF estimator and provide a sequential update rule. In addition we propose a version of the estimator suitable for non-stationary data streams via the introduction of an exponential weighting scheme.

- We present novel algorithms for sequential quantile estimation based on the Hermite series CDF estimators applicable in both static and dynamic settings. In the dynamic setting, we leverage the Hermite cumulative distribution function estimator along with an exponential weighting scheme.

- We investigate the asymptotic properties of the Hermite series based CDF estimator. We begin by studying the special cases of distributions associated with probability density functions with bounded or non-negative support. For these cases, we prove mean squared error (MSE) and mean integrated squared error (MISE) convergence of the Hermite CDF estimator and provide the associated rates of convergence.

- We also provide mean absolute error (MAE) results for the Hermite series based quantile estimator for the aforementioned special cases.

- In the general setting corresponding to full real line support we derive new asymptotic convergence results in the mean squared error, mean integrated squared error and almost sure sense, along with rates, for the full real line Hermite series based CDF estimator. In doing so we derive novel asymptotic bias and variance results. While the general rate of MSE convergence for the full real line estimator is worse than the special case results we have derived, they are uniform. We also provide a concrete example of a distribution where the rate of MSE convergence approaches the optimal rate, $O(n^{-1})$, namely the normal distribution.

- We prove B-robustness (Bias-robustness) of the full real line Hermite

series distribution function estimator for finite N and contrast this with a closely related estimator based on the Gram-Charlier type A series.

- We investigate the practical effectiveness of the Hermite series based cumulative distribution function and quantile estimators through real data and simulation studies. In the case of the Hermite series based quantile estimator, we demonstrate competitiveness with a leading existing algorithm.

Thus far, we have discussed the univariate setting. In the context of the statistical analysis of bivariate streaming data and one-pass analysis of massive bivariate data sets, certain quantities again easily extend to online calculation, such as the Pearson product-moment correlation coefficient. By contrast, online algorithms for nonparametric measures of concordance such as the Spearman's rank correlation coefficient (Spearman's Rho) and Kendall's rank correlation coefficient (Kendall Tau) have only recently been proposed [72]. These nonparametric correlation measures are suitable for all monotonic relationships and not just linear relationships as in the case of the Pearson correlation coefficient [27]. In addition, these nonparametric correlation measures are more robust than the Pearson correlation estimator [17]. Applications of nonparametric correlation measures include eliciting relationships for financial instruments [2] amongst others.

We propose a novel approach to the sequential estimation of the most popular nonparametric correlation measure, Spearman's rank correlation, based on bivariate Hermite series density estimators and Hermite series based distribution function estimators. We make full use of the advantage presented by these estimators of maintaining a sequential estimate of the full CDF. The key idea is that we can utilise the Hermite series cumulative distribution function estimator and the bivariate Hermite series density estimator together with a large sample definition of the Spearman's rank correlation estimator to furnish online estimates of the Spearman's rank correlation.

These estimates can be updated in constant i.e. $O(1)$ time and require only a small and fixed amount of memory ($O(1)$ memory requirements with respect to number of observations).

This algorithm is useful in the stationary sequential estimation setting (i.i.d. observations from a bivariate distribution for example) as well as one-pass batch estimation in the setting of massive data sets. In the i.i.d. observation case, we are able to provide asymptotic guarantees on the rate of convergence in mean of this estimator to the large sample Spearman's rank estimate. The Hermite series based Spearman's rank correlation estimation algorithm can also be modified to estimate the Spearman's rank correlation for non-stationary bivariate data streams. To treat the case of sequential estimation in the non-stationary setting, we introduce a novel, exponentially weighted estimator for the Spearman's rank correlation, which allows the local nonparametric correlation of a bivariate data stream to be tracked. To the best of our knowledge this is the first algorithm to be proposed for estimating a time-varying Spearman's rank correlation that does not rely on a moving window approach.

The rest of the thesis is organised as follows, in chapter 2 we review some relevant background on univariate and bivariate Hermite series density estimators. In chapter 3 we review nonparametric correlation coefficients and Spearman's rank correlation in particular. In chapter 4 we present novel estimators and algorithms for sequential cumulative distribution function and quantile estimation, in both the static and dynamic settings. Also included are asymptotic convergence results in certain special cases, namely cumulative distribution functions associated with probability densities with non-negative and bounded support. This chapter also presents real data and simulation studies demonstrating the competitiveness of our Hermite series based sequential quantile estimators compared to a leading existing sequential quantile estimation algorithm. In chapter 5 we further study the Hermite series based cumulative distribution function estimator. In particu-

lar, we prove asymptotic convergence results for the general Hermite series based distribution function estimator associated with probability densities with full real line support. We also prove certain robustness properties of this estimator. This analysis redresses a gap in the literature on the properties of Hermite series based distribution function estimators and establishes results that we make use of in the next chapter. In chapter 6 we define a sequential Spearman's rank correlation estimator based on Hermite series cumulative distribution function and density estimators. This novel sequential Spearman's rank estimator is shown to be competitive with the only known existing algorithm through simulation studies. We also present a relevant real data example. We conclude in chapter 7. We collect some relevant technical details associated with the above chapters in the appendices. Specifically, the appendices, A, B and C are associated with the chapters 4, 5 and 6 respectively. Finally, we present illustrative R code demonstrating efficient implementations of our techniques presented for sequential estimation of cumulative distribution functions, quantiles and Spearman's rank correlation coefficients using Hermite series based estimators in appendix D.

Chapter 2

Hermite Series Estimators

2.1 Hermite Polynomials and Hermite Functions

In this section we introduce the Hermite polynomials and Hermite functions which will play a central role in the construction of orthogonal series estimators for probability density functions. The Hermite polynomials are a classical orthogonal polynomial sequence. Following standard notation [63] we define the Hermite polynomials:

$$H_k(x) = (-1)^k e^{x^2} \frac{d^k}{dx^k} e^{-x^2},$$

which are orthogonal under the weight function e^{-x^2} i.e.

$$\int_{-\infty}^{\infty} e^{-x^2} H_k(x) H_l(x) dx = \sqrt{\pi} 2^k k! \delta_{kl}. \tag{2.1}$$

The Hermite polynomials form an orthogonal basis for $L_2(\mathbb{R})$. From this point forward we will refer to $L_2(\mathbb{R})$ as L_2 where it is understood that the domain of integration is the full real line. The Hermite polynomials satisfy the following recurrence relation,

$$\begin{aligned} H_0(x) &= 1, \\ H_{k+1}(x) &= 2xH_k(x) - 2kH_{k-1}(x),\ k \geq 1, \end{aligned} \tag{2.2}$$

which can be used to calculate a sequence of Hermite polynomials, $H_k(x),\ k = 0, 1, \ldots, N$ at a particular value of x in a computationally efficient manner.

The following explicit expression may also be utilised:

$$H_k(x) = k! \sum_{m=0}^{\lfloor k/2 \rfloor} \frac{(-1)^m}{k!(k-2m)!} (2x)^{k-2m}. \tag{2.3}$$

This explicit expression proves useful in deriving an analytical cumulative distribution function estimator in chapter 4.

Finally, some useful inequalities for $H_k(x)$ are as follows [63] (used in [29] for example):

$$(2^k k! \sqrt{\pi})^{-1/2} \left| H_k(x) \right| e^{\frac{-x^2}{2}} \leq c_a (k+1)^{-1/4},\ |x| \leq a \tag{2.4}$$

for some positive constant c_a and non-negative a and

$$(2^k k! \sqrt{\pi})^{-1/2} \left| x^{-1/3} H_k(x) \right| e^{\frac{-x^2}{2}} \leq d_a (k+1)^{-1/4},\ |x| \geq a \tag{2.5}$$

for some positive constant d_a and positive a.

Normalised Hermite functions can be defined from the Hermite polynomials as,

$$h_k = \left(2^k k! \sqrt{\pi}\right)^{-\frac{1}{2}} e^{-\frac{x^2}{2}} H_k(x),\ k = 0, 1, \cdots, \tag{2.6}$$

satisfying $\int_{-\infty}^{\infty} h_k(x) h_j(x) dx = \delta_{ij}$, with δ_{ij} being the Kronecker delta function. The Hermite functions $h_k, k = 0, 1, \cdots$ form an orthonormal basis for

L_2. When stated in terms of the normalised Hermite functions, we have the following inequalities,

$$\max_{|x|\leq a} |h_k(x)| \leq c_a(k+1)^{-\frac{1}{4}}, \tag{2.7}$$

$$\max_{|x|\geq a} |h_k(x)|x^{\lambda} \leq d_a(k+1)^{s}, \tag{2.8}$$

where c_a and d_a are positive constants, $s = \max(\frac{\lambda}{2} - \frac{1}{12}, -\frac{1}{4})$ [63].

In the next section we establish the link between Hermite series expansions of L_2 functions and expansions of probability density functions in particular.

2.2 Hermite Series Expansion

If $f(x) \in L_2$ i.e. $\int_{-\infty}^{\infty} f^2(x)dx < \infty$ then $f(x)$ can be expressed as follows,

$$f(x) = \sum_{k=0}^{\infty} b_k h_k(x), \tag{2.9}$$

where the coefficients satisfy,

$$b_k = \int_{-\infty}^{\infty} f(x)h_k(x)dx. \tag{2.10}$$

The fact that $f(x)$ can be expanded in this manner is a consequence of the fact that the Hermite functions h_k, $k = 0, 1, \cdots$, are an orthonormal basis for L_2. This is the expansion used to define what is termed the Hermite series estimator in [59, 67, 29, 30, 45].

A different, but completely equivalent form of the expansion, termed the Gauss-Hermite expansion (following [9, 54]) is given by,

$$f(x) = \sum_{k=0}^{\infty} a_k H_k(x)Z(x), \tag{2.11}$$

where $Z(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is the standard normal probability density function and

$$a_k = \alpha_k \int_{-\infty}^{\infty} Z(x) f(x) H_k(x) dx, \tag{2.12}$$

with $\alpha_k = \frac{\sqrt{\pi}}{2^{k-1} k!}$.

The coefficients of the two forms of the Hermite series expansion are related by:

$$b_k = \sqrt{\frac{1}{\alpha_k}} a_k.$$

We refer to the two forms of Hermite series expansion interchangeably. We favour the Gauss-Hermite form in chapter 4, and the more commonly used form in chapters 5 and 6. The usefulness of the Gauss-Hermite form of the expansion is that it makes explicit the role of the normal distribution. Indeed it is explicit that one would expect near Gaussian distributions to be well represented with just a few coefficients. The expansion (2.9) is more common in the literature however and simplifies the presentation of certain calculations.

The $N + 1$ term truncated Hermite series expansion is defined as:

$$f_N(x) = \sum_{k=0}^{N} b_k h_k(x), \tag{2.13}$$

or equivalently,

$$f_N(x) = \sum_{k=0}^{N} a_k H_k(x) Z(x). \tag{2.14}$$

It is noteworthy that the coefficients a_k and b_k in the Hermite series expansions above are such that the L_2 distance between $f(x)$ and $f_N(x)$ is minimised i.e. no other choice of coefficients would lead to a better approximation of $f(x)$ in the L_2 distance sense. This follows from the fact that

$f(x) \in L_2$ and the fact that the first $N+1$ Hermite functions constitute an orthonormal basis for a $N+1$ dimensional subspace of L_2. See [18] for a succinct proof using these facts. At this point, $f(x)$ is an arbitrary function in L_2, it need not be a probability density function. For the purposes of density estimation however we will regard $f(x)$ as a probability density function (a non-negative function that integrates to one) that is also in L_2.

A number of other probability density expansions have been defined in terms of the Hermite polynomials, namely the Gram-Charlier type A series and the Edgeworth series. These expansions use a different form of the Hermite polynomials, namely the Chebyshev-Hermite polynomials. The Chebyshev-Hermite polynomials are defined as,

$$He_k(x) = (-1)^k e^{x^2/2} \frac{d^k}{dx^k} e^{-x^2/2},$$

which are orthogonal under the weight function $e^{-x^2/2}$ i.e.

$$\int_{-\infty}^{\infty} e^{-x^2/2} He_k(x) He_l(x) dx = \sqrt{2\pi} k! \delta_{kl}.$$

The infinite series Gram-Charlier type A and Edgeworth expansions, while formally identical to each other, differ in how truncation is performed. For a probability density $f(x)$ with mean, $\mu = 0$ and standard deviation, $\sigma = 1$, we have the following truncated expansion for the Gram-Charlier type A series (phrased in terms of cumulants and neglecting cumulants above the fourth cumulant),

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \left(1 + \frac{\kappa_3}{6} He_3(x) + \frac{\kappa_4}{24} He_4(x)\right), \tag{2.15}$$

and for the Edgeworth expansion,

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \left(1 + \frac{\kappa_3}{6} He_3(x) + \frac{\kappa_4}{24} He_4(x) + \frac{\kappa_3^2}{720} He_6(x)\right), \quad (2.16)$$

where κ_j, $j = 3, 4$ are the third and fourth cumulants of $f(x)$. The main advantage of the truncated Edgeworth expansion over the Gram-Charlier type A expansion is that it is an asymptotic expansion i.e. for the finite series, the remainder is of smaller order than the last term that has been incorporated.

The Hermite series expansions as defined above in (2.9), (2.11) have some important advantages over the Gram-Charlier type A and Edgeworth expansions. The first is that the Gram-Charlier type A series and Edgeworth expansions rarely converge as infinite series. Cramer has shown that the Gram-Charlier type A series and Edgeworth expansions only converge under very restrictive conditions on the tails of the distribution, for example $\int_{-\infty}^{\infty} f(x) e^{\frac{1}{2}x^2} dx < \infty$, where $f(x)$ is a probability density of bounded variation in every finite interval (see the discussion in [42]). Indeed, a ubiquitous distribution like the exponential distribution does not satisfy this condition. In addition, a convergent type A series requires that the distribution is uniquely determined by its moments. The log-normal distribution fails this assumption for example [37]. By contrast, the infinite Hermite series expansions (2.9), (2.11) above converge simply if $f(x) \in L_2$.

For the Gram-Charlier type A and Edgeworth series it has been noted that usually terms not higher than He_6 are incorporated. In fact, due to high sampling variance for cumulants beyond the fourth cumulant, usually only the first few terms are included, limiting the flexibility of describing more complex density functions. The Hermite series estimators based on the finite Hermite series expansion can be applied to very complex density functions however, and it remains meaningful to include a large number of terms (50 or more, see [54] for some examples) by comparison.

A potentially more serious constraint in applying the Gram-Charlier type A and Edgeworth expansions in practice is the fact that since the coefficients in these expansions involve cumulants, their sample estimates are sensitive to outliers and not robust (see [70] for example). Hermite series estimators are robust however. In fact, in [70] generalised Gram-Charlier type A and Edgeworth expansions are introduced that do exhibit robustness, of which the Hermite series expansion (2.9) is a special case (i.e. for scaling parameter $\alpha = \sqrt{2}$ in equation (16) and (19) in that article). The robustness property is attributed to the fact that the sample estimates of series coefficients are rendered less sensitive to outliers and thus the generalised expansions, including the Hermite series expansion, are more stable for heavy tailed distributions. See also [54] for similar analyses. In this thesis we prove in section 5.5 that a cumulative distribution function estimator based on the Gram-Charlier type A series is not robust in contrast to a CDF estimator based on the Hermite series expansion (2.13) above, which is robust. In the context of quantile estimation, the Cornish-Fisher expansion suffers the same dependence on non-robust cumulants.

As an aside, saddlepoint approximations would also be non-robust in general given their dependence on cumulants. In addition, in the case of the saddle-point expansion [28] it is well known that multimodal distributions prove a particular challenge. By contrast, the Gauss-Hermite based expansions can handle very complex multimodal distributions [54].

2.3 Hermite Series Density Estimators

2.3.1 Univariate Density Estimators

In the context of nonparametric density estimation, a natural estimator for the b_k coefficients is (following from equation (2.10) and the law of large numbers):

$$\hat{b}_k = \frac{1}{n}\sum_{i=1}^{n} h_k(\mathbf{x}_i), \quad \mathbf{x}_i \sim f(x). \tag{2.17}$$

Thus, the $N+1$ term Hermite series estimator for probability density functions has the following form:

$$\hat{f}_N(x) = \sum_{k=0}^{N} \hat{b}_k h_k(x). \tag{2.18}$$

Equivalently, for the Gauss-Hermite form of the Hermite series estimator, we have:

$$\hat{a}_k = \frac{\alpha_k}{n}\sum_{i=1}^{n} Z(\mathbf{x}_i) H_k(\mathbf{x}_i), \quad \mathbf{x}_i \sim f(x), \tag{2.19}$$

and

$$\hat{f}_N(x) = \sum_{k=0}^{N} \hat{a}_k H_k(x) Z(x). \tag{2.20}$$

One common measure of the quality of the estimate of an unknown probability density function is the mean integrated squared error (MISE). Let the true probability density function $f(x) \in L_2$. The mean integrated squared error associated with the truncated Hermite series estimate of $f(x)$ is:

$$
\begin{aligned}
E\int_{-\infty}^{\infty}(\hat{f}_N(x)-f(x))^2dx &= E\int_{-\infty}^{\infty}(\sum_{k=0}^{N}\hat{b}_k h_k(x)-f(x))^2 \\
&= E[\sum_{k,l=0}^{N}\hat{b}_k\hat{b}_l\int_{-\infty}^{\infty}h_k(x)h_l(x)dx \\
&- 2\sum_{k=0}^{N}\hat{b}_k\int_{-\infty}^{\infty}h_k(x)f(x)dx \\
&+ \int_{-\infty}^{\infty}[f(x)]^2dx] \\
&= E[\sum_{k=0}^{N}\hat{b}_k^2 - 2\sum_{k=0}^{N}\hat{b}_k b_k \\
&+ b_k^2], \qquad (2.21)
\end{aligned}
$$

thus,

$$
E\int_{-\infty}^{\infty}(\hat{f}_N(x)-f(x))^2dx = \sum_{k=0}^{N}E(\hat{b}_k-b_k)^2 + \sum_{k=N+1}^{\infty}b_k^2, \qquad (2.22)
$$

where we have made use of Parseval's identity, $\sum_{k=0}^{\infty}b_k^2 = \int_{-\infty}^{\infty}f^2(x)dx$ and the definition of b_k (2.10).

The first term is associated with the error due to using estimates of the coefficients, $\hat{b}_k$ instead of the true, unknown, coefficients b_k. This is the integrated variance term. The second term of the MISE is associated with the error due to truncation. This is the integrated squared bias term.

Similarly, we have for the Gauss-Hermite form of the Hermite series estimator, (2.20),

$$E\int_{-\infty}^{\infty}(\hat{f}_N(x)-f(x))^2dx = E\left[\sum_{k=0}^{N}\frac{2^{k-1}k!}{\sqrt{\pi}}(\hat{a}_k-a_k)^2\right]+\sum_{k=N+1}^{\infty}\frac{2^{k-1}k!}{\sqrt{\pi}}a_k^2. \tag{2.23}$$

The theoretical properties of the Hermite series density estimators have been well established [59, 67, 29, 30, 45]. These include MSE and MISE consistency under various conditions, uniform almost sure convergence and asymptotic normality.

A few further comments are in order. In practice, the Hermite series estimator provides a good fit to a wide variety of probability density functions [54]. For completeness however, the following shortcomings should be noted as they may be important depending on the application of the Hermite series estimate of the density. In principle, for the truncated series, the probability density function that results may be negative at certain values of x. Also, truncated Hermite series estimates should capture nearly Gaussian distributions well in a relatively small number of coefficients. However, for distributions that differ greatly from the Gaussian distribution, a large number of coefficients may be required for a satisfactory fit, even if convergence is guaranteed in principle.

It is noteworthy that several other orthogonal series estimators have been defined, those based on Legendre polynomials for example. The estimators based on other classical orthogonal polynomials are applicable only to probability densities with bounded or non-negative support however. The Hermite series estimators are the most natural in situations with unknown support since they are applicable to probability densities with support on the full real line. In this sense they are more general than the other orthogonal series estimators. More modern estimators such as wavelet density estimators are applicable to probability densities with full real line support (see [24] for example), however the coefficients cannot, in general, be evaluated in closed

form, unlike the Hermite series estimator's coefficients.

In the next section we define the bivariate extension of the Hermite series estimator discussed above.

2.3.2 Bivariate Density Estimators

The bivariate Hermite series probability density function estimator for a density, $f(x, y) \in L_2$, is given by:

$$\begin{aligned}\hat{f}_{N_1,N_2}(x,y) &= \sum_{k=0}^{N_1}\sum_{j=0}^{N_2} \hat{A}_{kj} h_k(x) h_j(y), \\ \hat{A}_{kj} &= \frac{1}{n}\sum_{i=1}^{n} h_k(\mathbf{x_i}) h_j(\mathbf{y_i}). \quad k = 0, \ldots, N_1,\ j = 0, \ldots, N_2,\end{aligned} \tag{2.24}$$

where $(\mathbf{x_i}, \mathbf{y_i}) \sim f(x, y)$. It is straightforward to show that the MISE associated with the bivariate Hermite series density estimator is:

$$E \int (\hat{f}_{N_1 N_2}(x,y) - f(x,y))^2 dxdy \tag{2.25}$$

$$= \sum_{k=0}^{N_1}\sum_{l=0}^{N_2} E(\hat{A}_{kl} - A_{kl})^2 + \sum_{k=N_1+1}^{\infty}\sum_{l=N_2+1}^{\infty} A_{kl}^2. \tag{2.26}$$

Theoretical properties of the univariate Hermite series density estimator easily generalise to the multivariate case as discussed in [59] and [67].

Chapter 3

Measures of Association

Consider two random variables x and y with a bivariate distribution $F(x, y)$ and probability density function, $f(x, y)$. There are various measures of association that have been defined. The ubiquitous Pearson product-moment correlation coefficient:

$$\rho(x, y) = \frac{cov(x, y)}{(var(x)var(y))^{1/2}}, \tag{3.1}$$

is perhaps the best known measure of association, with sample estimator:

$$\hat{\rho}(x, y) = \frac{\sum_{i=1}^{n}(\mathbf{x}_i - \bar{X})(\mathbf{y}_i - \bar{Y})}{(\sum_{i=1}^{n}(\mathbf{x}_i - \bar{X})^2 \sum_{i=1}^{n}(\mathbf{y}_i - \bar{Y})^2)^{1/2}}, \tag{3.2}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i$ are sample means. The Pearson product-moment correlation coefficient is a measure of the linear association between x and y. It is invariant under separate changes in scale and location in x and y. The value of the coefficient ranges between -1 and 1. If x and y are independent, $\rho(x, y) = 0$. The converse is not true in general however. A notable exception is where $f(x, y)$ is a bivariate normal distribution, in which case $\rho(x, y) = 0$ does imply independence. Thus, $\rho(x, y)$ is a particularly suitable measure of association for the bivariate normal distribution but potentially less so in the case of other distributions. One shortcoming of

the Pearson product-moment correlation coefficient is that it is not invariant under all order-preserving transformations of x and y. In addition, its usage is only justified for linear relationships between x and y. Finally, the sample estimator for $\hat{\rho}(x, y)$ is sensitive to outliers and thus not robust.

Nonparametric measures of concordance (association) that are appropriate when the relationship between x and y is not necessarily linear but rather monotonic are Spearman's rank correlation coefficient and Kendall's rank correlation coefficient. These nonparametric correlation coefficients are based on the ranks of the variables x and y as opposed to their actual values and as such are manifestly invariant to all order preserving transformations. In addition, the estimators for these correlation coefficients are not as sensitive to outliers and are therefore more robust measures of association [17]. Finally these nonparametric correlation coefficients satisfy a set of criteria for good relative measures of association between two random variables as defined in [27]. In this thesis we focus our attention on the most popular measure of nonparametric correlation, namely Spearman's rank correlation coefficient. Suppose we have a sample of n observations drawn from a continuous bivariate probability distribution, $F(x, y)$, with probability density function $f(x, y)$, i.e. $(\mathbf{x}_i,\mathbf{y}_i) \sim f(x, y)$. The Spearman's rank correlation coefficient is defined as the sample Pearson product-moment correlation of the ranks of the observations, (r_i,s_i), $i = 1 \dots n$,

$$R = \frac{\sum_{i=1}^{n}(r_i - \bar{R})(s_i - \bar{S})}{(\sum_{i=1}^{n}(r_i - \bar{R})^2 \sum_{i=1}^{n}(s_i - \bar{S})^2)^{1/2}} \tag{3.3}$$

where $\bar{R} = \frac{1}{n}\sum_{i=1}^{n} r_i$ and $\bar{S} = \frac{1}{n}\sum_{i=1}^{n} s_i$ are the sample means of the ranks. The coefficient R does not directly have a population analog. This is due to the fact that if we assume the marginal distributions of the random variables are continuous, the values which the random variables can take cannot be enumerated and ranked. However, we can define a constant which is a natural estimand as the sample size, $n \to \infty$.

We begin by noting that:

$$r_i = n\hat{F}_n^{(1)}(\mathbf{x}_{(i)}), \qquad s_i = n\hat{F}_n^{(2)}(\mathbf{y}_{(i)}), \tag{3.4}$$

where $\mathbf{x}_{(i)}, \mathbf{y}_{(i)}$ are the ith smallest observations of the respective separately ranked x and y samples, $\hat{F}_n^{(1)}(x) = \hat{F}_n(x, \infty)$ is the empirical distribution function estimate of the marginal cumulative distribution function $F^{(1)}(x) = F^{(1)}(x, \infty)$ and $\hat{F}_n^{(2)}(y) = \hat{F}_n(\infty, y)$ is the empirical distribution function estimate of the marginal cumulative distribution function $F^{(2)}(y) = F^{(2)}(\infty, y)$.

If we express (3.3) as a functional of the bivariate empirical distribution function estimate $\hat{F}_n(x, y)$ we obtain,

$$\begin{aligned} R &= T(\hat{F}_n) \\ &= \frac{\int_{-\infty}^{\infty} \left(\hat{F}_n^{(1)}(t) - \overline{\hat{F}_n^{(1)}}\right)\left(\hat{F}_n^{(2)}(u) - \overline{\hat{F}_n^{(2)}}\right) d\hat{F}_n(t,u)}{\left(\int_{-\infty}^{\infty} \left(\hat{F}_n^{(1)}(t) - \overline{\hat{F}_n^{(1)}}\right)^2 d\hat{F}_n^{(1)}(t)\right)^{1/2} \left(\int_{-\infty}^{\infty} \left(\hat{F}_n^{(2)}(u) - \overline{\hat{F}_n^{(2)}}\right)^2 d\hat{F}_n^{(2)}(u)\right)^{1/2}} \end{aligned} \tag{3.5}$$

where $\overline{\hat{F}_n^{(1)}} = \int_{-\infty}^{\infty} \hat{F}_n^{(1)}(v) d\hat{F}_n^{(1)}(v)$ and $\overline{\hat{F}_n^{(2)}} = \int_{-\infty}^{\infty} \hat{F}_n^{(2)}(x) d\hat{F}_n^{(2)}(x)$. If we replace $\hat{F}_n(x, y)$ with $F(x, y)$ we obtain,

$$T(F) = \rho(F^{(1)}(X), F^{(2)}(Y)), \tag{3.6}$$

where $\rho(x, y)$ is the Pearson product-moment correlation. The quantity $\rho(F^{(1)}(X), F^{(2)}(Y))$ is termed the grade correlation. Thus (3.3) is a Fisher consistent estimator of the grade correlation.

We can also define the grade correlation as the constant for which R is an unbiased estimator in large samples [27],

$$lim_{n\to\infty} E(R) = \rho(F^{(1)}(X), F^{(2)}(Y)). \tag{3.7}$$

The grade correlation can be simplified to (using the exposition in [27]),

$$\rho(F^{(1)}(X), F^{(2)}(Y)) = 12 \int \int (F^{(1)}(x) - 1/2)(F^{(2)}(y) - 1/2) f(x, y) dx dy, \tag{3.8}$$

an expression that will be useful in the developments in section 6.1 below where we define an estimator for the grade correlation based on Hermite series bivariate density estimators and Hermite series based cumulative distribution function estimators.

Chapter 4

Sequential Estimation of Quantiles using Hermite Series Estimators

In this chapter we introduce novel sequential CDF and quantile estimators, along with associated sequential algorithms, based on the Hermite series density estimator. The chapter is organised as follows, in section 4.1 we define a Hermite series based estimator (in Gauss-Hermite form) for the cumulative distribution function and discuss a numerical means of obtaining arbitrary quantiles. We then set about applying this to sequential quantile estimation in the settings of static and dynamic quantile estimation. In this chapter, we make our treatment of static and dynamic quantile estimation concrete by considering the cases of independent identically distributed (i.i.d.) data streams and non-identically distributed independent data streams respectively. Observations are continuous random variables that are revealed sequentially (i.e. one at a time). The basic algorithm for the static case is presented in section 4.2. We then proceed to treat the dynamic case by introducing an exponentially weighted moving average estimator for the Gauss-Hermite coefficients in section 4.3. In section 4.4 we investigate the quality

of the Gauss-Hermite CDF and quantile estimators theoretically in the special case of distributions associated with probability density functions with non-negative support (also applicable to bounded support). We compare the proposed techniques to a leading existing algorithm for both simulated data (in section 4.5) and real data (in section 4.6). The practical consideration of standardising the observations from the data stream being analysed is treated in appendix A.1. Useful MISE results for the exponentially weighted Gauss-Hermite expansion are derived in appendix A.2.

4.1 Estimating Quantiles using the Gauss-Hermite Expansion

4.1.1 Cumulative Distribution Function

In this section we derive an analytical expression for a Gauss-Hermite based cumulative distribution function estimator. To the best of our knowledge, this is the first such analytical derivation of a cumulative distribution function estimator based on the Gauss-Hermite expansion (i.e. based on Hermite series estimators). We utilise this expression to numerically obtain quantiles. We have discussed a number of well-established results on smooth CDF estimators based on other nonparametric techniques in the introduction.

Before we begin, we recall the definitions of the Gamma functions:

$\Gamma(a, x)$ is the upper incomplete Gamma function defined as:

$$\Gamma(a, x) = \int_x^\infty t^{a-1}\, e^{-t}\, \mathrm{d}t,$$

$\gamma(a, x)$ is the lower incomplete Gamma function defined as:

$$\gamma(a,x) = \int_0^x t^{a-1}\, e^{-t}\, \mathrm{d}t,$$

and $\Gamma(a)$ is the usual Gamma function defined as:

$$\Gamma(a) = \int_0^\infty t^{a-1}\, e^{-t}\, \mathrm{d}t.$$

Now, utilising (2.20) a natural estimator for the cumulative distribution function associated with $f(x)$ can be defined as:

$$\begin{aligned} \hat{F}_N(x) &= \int_{-\infty}^{x} \hat{f}_N(x')dx' \\ &= \sum_{k=0}^{N} \hat{a}_k \int_{-\infty}^{x} H_k(x')Z(x')dx'. \end{aligned} \tag{4.1}$$

We derive an analytical expression for (4.1) using (2.3) as follows:

For $x < 0$:

$$\begin{aligned} \hat{F}_N(x) &= \int_{-\infty}^{x} \hat{f}_N(x')dx' \\ &= \sum_{k=0}^{N} \hat{a}_k \int_{-\infty}^{x} H_k(x')Z(x')dx' \\ &= \sum_{k=0}^{N} \hat{a}_k k! \sum_{l=0}^{\lfloor k/2 \rfloor} \frac{(-1)^l 2^{k-2l}}{l!(k-2l)!\sqrt{2\pi}} \int_{-\infty}^{x} (x')^{k-2l} e^{-x'^2/2} dx' \\ &= \sum_{k=0}^{N} \hat{a}_k k! \sum_{l=0}^{\lfloor k/2 \rfloor} \frac{(-1)^l 2^{k-2l}}{l!(k-2l)!\sqrt{2\pi}} (-1)^{k-2l} 2^{\frac{k}{2}-l-\frac{1}{2}} \Gamma(-l+\frac{k}{2}+\frac{1}{2}, \frac{x^2}{2}) \end{aligned}$$

For $x \geq 0$:

$$\begin{aligned}
\hat{F}_N(x) &= \int_{-\infty}^{x} \hat{f}_N(x')dx' \\
&= \int_{-\infty}^{\infty} \hat{f}_N(x')dx' - \int_{x}^{\infty} \hat{f}_N(x')dx' \\
&= \sum_{k=0}^{N} \hat{a}_k \left[\int_{-\infty}^{\infty} H_k(x')Z(x')dx' - \int_{x}^{\infty} H_k(x')Z(x')dx' \right] \\
&= \sum_{k=0}^{N} \hat{a}_k k! \sum_{l=0}^{\lfloor k/2 \rfloor} \frac{(-1)^l 2^{k-2l}}{l!(k-2l)!\sqrt{2\pi}} \left[\int_{-\infty}^{\infty} (x')^{k-2l} e^{-x'^2/2} dx' \right. \\
&- \left. \int_{x}^{\infty} (x')^{k-2l} e^{-x'^2/2} dx' \right] \\
&= \sum_{k=0}^{N} \hat{a}_k k! \sum_{l=0}^{\lfloor k/2 \rfloor} \frac{(-1)^l 2^{k-2l}}{l!(k-2l)!\sqrt{2\pi}} 2^{\frac{k}{2}-l-\frac{1}{2}} \times \\
&\times \left[\left[(-1)^{k-2l} + 1 \right] \Gamma(-l + \frac{k}{2} + \frac{1}{2}) - \Gamma(-l + \frac{k}{2} + \frac{1}{2}, \frac{x^2}{2}) \right] \\
&= \sum_{k=0}^{N} \hat{a}_k k! \sum_{l=0}^{\lfloor k/2 \rfloor} \frac{(-1)^l 2^{k-2l}}{l!(k-2l)!\sqrt{2\pi}} 2^{\frac{k}{2}-l-\frac{1}{2}} \times \\
&\times \left[\left[(-1)^{k-2l} \right] \Gamma(-l + \frac{k}{2} + \frac{1}{2}) + \gamma(-l + \frac{k}{2} + \frac{1}{2}, \frac{x^2}{2}) \right]
\end{aligned}$$

Thus,

$$\hat{F}_N(x) = \begin{cases} \sum_{k=0}^{N} \hat{a}_k k! \sum_{l=0}^{\lfloor k/2 \rfloor} \frac{(-1)^l 2^{\frac{3k}{2}-3l-1} \left[\left[(-1)^{k-2l} \right] \Gamma(-l+\frac{k}{2}+\frac{1}{2}) + \gamma(-l+\frac{k}{2}+\frac{1}{2}, \frac{x^2}{2}) \right]}{l!(k-2l)!\sqrt{\pi}} \\ \quad \text{if } x \geq 0, \\ \sum_{k=0}^{N} \hat{a}_k k! \sum_{l=0}^{\lfloor k/2 \rfloor} \frac{(-1)^{-l+k} 2^{\frac{3k}{2}-3l-1} \Gamma(-l+\frac{k}{2}+\frac{1}{2}, \frac{x^2}{2})}{l!(k-2l)!\sqrt{\pi}} \\ \quad \text{if } x < 0. \end{cases} \tag{4.2}$$

The expression (4.2) allows us to directly estimate the cumulative dis-

tribution function without the need to numerically integrate the estimated probability density function (2.20). While the analytical expression (4.2) is useful in a number of settings, it may be more computationally efficient to utilise the following recurrence relation to calculate the cumulative probability estimate (derived using (2.2) and integration by parts):

$$\begin{aligned}
\int_{-\infty}^{x} H_0(x')Z(x')dx' &= \frac{1}{2}\text{erfc}\left(\frac{-1}{\sqrt{2}}x\right), \\
\int_{-\infty}^{x} H_1(x')Z(x')dx' &= -\sqrt{\frac{2}{\pi}}e^{-x^2/2}, \\
\int_{-\infty}^{x} H_{k+1}(x')Z(x')dx' &= -2H_k(x)Z(x) \\
&+ 2k\int_{-\infty}^{x} H_{k-1}(x')Z(x')dx',\ k = 1, \ldots, N,
\end{aligned} \tag{4.3}$$

where $\text{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-x'^2}dx'$ is the complementary error function. The CDF estimate is then obtained using (4.1). Note that using the recurrence relation (2.2) for the Hermite polynomials allows the coefficients, $\hat{a}_k$, $k = 0, \ldots, N$, to also be obtained in a computationally efficient manner.

An alternative estimator for the cumulative distribution function can be defined as follows:

$$\hat{F}_N(x) = \begin{cases} 1 - \sum_{k=0}^{N} \hat{a}_k k! \sum_{l=0}^{\lfloor k/2 \rfloor} \frac{(-1)^l 2^{\frac{3k}{2}-3l-1}\Gamma(-l+\frac{k}{2}+\frac{1}{2}, \frac{x^2}{2})}{l!(k-2l)!\sqrt{\pi}} & \text{if } x \geq 0 \\ \sum_{k=0}^{N} \hat{a}_k k! \sum_{l=0}^{\lfloor k/2 \rfloor} \frac{(-1)^{-l+k} 2^{\frac{3k}{2}-3l-1}\Gamma(-l+\frac{k}{2}+\frac{1}{2}, \frac{x^2}{2})}{l!(k-2l)!\sqrt{\pi}} & \text{if } x < 0 \end{cases} \tag{4.4}$$

This expression is derived in the same manner as (4.2) except that the integral $\int_{-\infty}^{\infty} \hat{f}_N(x')dx'$ is replaced with unity. We have found that empir-

ically this estimator yields more accurate results in the particular case of quantile estimation and thus we make use of (4.4) for this application. The estimator (4.4) is close but not precisely equal to the estimator (4.5) below (the estimators agree for $x \geq 0$):

$$\begin{aligned} \hat{F}_N(x) &= 1 - \int_x^\infty \hat{f}_N(x')dx' \\ &= 1 - \sum_{k=0}^{N} \hat{a}_k \int_x^\infty H_k(x')Z(x')dx', \end{aligned} \tag{4.5}$$

Both estimators appear advantageous in the quantile estimation setting (empirically) and it is noteworthy that the asymptotic consistency results we derive can be seen to apply to the above estimator with no additional effort.

It may be more computationally efficient to use (4.5) instead of (4.4) and make use of the recurrence relation,

$$\begin{aligned} \int_x^\infty H_0(x')Z(x')dx' &= \frac{1}{2}\text{erfc}\left(\frac{1}{\sqrt{2}}x\right), \\ \int_x^\infty H_1(x')Z(x')dx' &= \sqrt{\frac{2}{\pi}}e^{-x^2/2}, \\ \int_x^\infty H_{k+1}(x')Z(x')dx' &= 2H_k(x)Z(x) \\ &+ 2k\int_x^\infty H_{k-1}(x')Z(x')dx',\ k = 1, \ldots, N, \end{aligned} \tag{4.6}$$

along with the recurrence relation (2.2) to efficiently compute the coefficients, $\hat{a}_k$, $k = 0, \ldots, N$.

4.1.2 Inverse Cumulative Distribution Function

The inverse cumulative distribution function or quantile function is defined in equation (1.3). We can utilise (4.2) (or (4.4)) along with a numerical root-finding algorithm to determine the value of the p-th quantile, $x_p = q(p)$, $0 < p < 1$. Newton's method can be applied for example. In this case, the following equation is iteratively evaluated until sufficient accuracy is achieved:

$$\hat{x}_p^{(i+1)} = \hat{x}_p^{(i)} - \frac{\hat{F}_N(\hat{x}_p^{(i)}) - p}{\hat{f}_N(\hat{x}_p^{(i)})}, \tag{4.7}$$

where $\hat{f}_N(x)$ is given in (2.20) and $\hat{F}_N(x)$ is given in (4.2) (or (4.4)). Naturally, convergence behaviour will depend on the choice of initial value, $\hat{x}_p^{(0)}$, and the properties of $\hat{f}_N(x)$. In principle convergence may be slow, or the method may not converge at all. If techniques such as Newton's method and related methods prove unstable in a given setting, more robust numerical root-finding algorithms can be applied. The best choice of root-finding algorithm and optimal initial value selection are areas for future research.

It is important to note that in many cases of interest, the quantile itself is not required but rather it is necessary to determine whether an observation is above or below a particular quantile (consider outlier detection for example). In this case, no root finding is required. One simply plugs the observation into the cumulative distribution function and determines whether the cumulative probability is less than or greater than p.

4.2 Online Quantile Estimation: Static Quantiles

The problem we treat in this section involves estimating an unknown cumulative distribution function (and associated inverse cumulative distribution function) from a stream of independent and identically-distributed (i.i.d.) continuous random variable data. The Gauss-Hermite expansion furnishes an efficient means to achieve this. The primary reason for this is that the coefficients in the expansion can be updated with each new observation without recalculating the entire sum in (2.19). We can just incorporate a new term corresponding to the new observation and maintain a running average for each coefficient. Moreover, we can simply plug in these updated coefficients into the analytical expression for the cumulative distribution function we have derived (4.2) (or (4.4)). Any quantile can then be obtained by a simple numerical root finding procedure (section 4.1.2).

The basic algorithm can be summarised as follows:

1. Initialise $N+1$ coefficients as follows $\hat{a}_k^{(0)} = \alpha_k Z(\mathbf{x}_0) H_k(\mathbf{x}_0)$, where $\mathbf{x}_0$ is the first observation from the data stream and $k = 0 \dots N$.

2. For each new observation, $\mathbf{x}_i$, update $\hat{a}_0, \dots \hat{a}_N$ as follows:

$$\hat{a}_k^{(i)} = \frac{1}{i}\left[(i-1)\hat{a}_k^{(i-1)} + \alpha_k Z(\mathbf{x}_i) H_k(\mathbf{x}_i)\right], \; k = 0, \dots, N. \tag{4.8}$$

3. Plug the updated coefficients $\hat{a}_0, \dots \hat{a}_N$ into the expressions (2.20) and (4.2) (or (4.4)) to obtain updated estimates of the probability density function and cumulative distribution function respectively.

4. Utilise a numerical root finding algorithm, along with the updated cumulative distribution function (and potentially the updated probability density function) to determine any arbitrary quantile.

The above algorithm can also be applied to summarise the distribution of massive datasets in an efficient, one-pass manner which should be particularly useful when the size of the dataset is larger than the available memory.

The computational cost of *updating* each of the coefficients (2.19) is manifestly constant ($O(1)$) and does not depend on the number of previous observations. Also, since the cumulative distribution function only depends on the coefficients and has no explicit dependence on the observations, the time complexity of *updating* the cumulative distribution function following the arrival of a new observation is also $O(1)$. Similarly, the computational cost of the numerical root finding algorithm yielding any quantile of interest from the updated cumulative distribution does not depend on the number of previous observations. This is to be contrasted with deterministic approaches to obtaining quantiles such as efficient heap based median maintenance which has a time complexity $O(\log \text{i})$ at the i-th observation and has growing space requirements.

While the updating procedure for the coefficients is fully sequential, it is clear that since we use a fixed and constant N the resultant Gauss-Hermite estimate of the probability density function is biased and thus, the resultant CDF and quantile estimates will in general be biased too. Thus our quantile estimator is sequential but biased. This bias does not prevent the estimator from being useful however.

Data streams that have static quantiles are likely to be less prevalent than those that have dynamic quantiles (quantiles that vary over time). Indeed, many real-world data streams of interest exhibit non-stationarity. In section 4.3 we consider how the proposed algorithm can be modified to treat quantile estimation in the more realistic, dynamic setting.

4.2.1 Selection of N

It is natural to assume that the quality of the CDF and quantile estimates is related to the MISE of $\hat{f}_N(x)$. Indeed, we demonstrate in section 4.4 that the MISE of $\hat{f}_N(x)$ directly determines bounds on the MSE of the CDF and the MAE of the resultant quantile estimates for distributions with non-negative support under certain conditions. When viewed in the context of the MISE of $\hat{f}_N(x)$, the choice of N controls the trade-off between the integrated variance and integrated squared bias of the estimate of $f(x)$. For the Hermite series estimators, the integrated variance term vanishes as $n \to \infty$ for fixed N (under certain conditions on $f(x)$, see [29]). This leaves the contribution from the integrated bias term. The higher the value of N, the smaller the integrated bias. Thus in the setting of streaming data or one pass analysis of a massive data set, where we regard $n \to \infty$, N would naively be made as large as possible to minimise the bias and hence the MISE (as noted above the MISE controls the quality of the CDF and quantile estimates, see section 4.4.2). It is worth noting however that memory requirements and processing time increase with N since more coefficients have to be stored and updated. In addition, early quantile estimates could be poor for large N. If the intended application is sensitive to poor early quantile estimates, a small sample from the data stream or massive data set can be analysed in order to select N. While it is clear that in general the optimal N is different for the PDF, CDF and quantile estimates, our results suggest that it is a reasonable starting point to attempt to minimise the MISE of $\hat{f}_N(x)$. Principled techniques exist to select the (MISE) optimal N such as the Kronmal-Tarter optimal stopping rule algorithm [43], [51]. This algorithm must be applied with care though as it may perform poorly if $f(x)$ is multimodal or peaked as pointed out in [22]. The reader is referred to [22] and [36] for improvements to the algorithm. Data driven selection of N specific to the CDF and quantile estimates is an area of future research. In our simulation studies we demonstrate that above a minimum size of N, the effectiveness of the algorithm is in fact not critically

dependent on the choice of N. Good results can be obtained for a range of values of N. Through extensive empirical studies we have determined that a value of N = 6 yields good results for data with a unimodal distribution for example.

4.3 Online Quantile Estimation: Dynamic Quantiles

The problem we treat in this section involves obtaining a local estimate of an unknown cumulative distribution function (and associated inverse cumulative distribution function) from a stream of continuous random variable data with dynamic quantiles. In order to do this we replace the Gauss-Hermite coefficient estimator defined in (2.19) with an exponentially weighted moving average estimator for the coefficients. We will term the resulting expansion an exponentially weighted Gauss-Hermite expansion (EWGH expansion). The new estimator for the coefficients is given by:

$$\begin{aligned} \hat{a}_k^{(i)} &= \lambda \left[\alpha_k Z(\mathbf{x}_i) H_k(\mathbf{x}_i)\right] + (1-\lambda)\hat{a}_k^{(i-1)}, \\ \hat{a}_k^{(0)} &= \left[\alpha_k Z(\mathbf{x}_0) H_k(\mathbf{x}_0)\right], \; k = 0, \ldots, N, \end{aligned} \tag{4.9}$$

where $0 < \lambda \leq 1$ controls the weight of new observations (and controls how rapidly the weightings of older observations decrease). This weighting scheme allows the local behaviour of the data stream to be tracked. The algorithm for obtaining arbitrary quantiles presented in the previous section is essentially unchanged except that we replace (4.8) with (4.9). To re-iterate, the updated coefficients can then be plugged into into the expressions (2.20) and (4.2) (or (4.4)) to obtain updated estimates of the probability density function and cumulative distribution function respectively. A numerical root finding procedure can again be applied to obtain arbitrary quantiles.

4.3.1 Selection of the Parameters λ and N

For the choice of N, the same broad considerations apply as in section 4.2.1 i.e. the more complex the probability distribution of the data being analysed, the higher the appropriate N to ensure a sufficiently low bias. In our simulation studies we have observed that a value of $N = 6$ gives competitive performance for all choices in the set $\lambda = 0.01, 0.05, 0.1$, for unimodal distributions.

Given a choice of N, there are two factors to consider when selecting λ. The first is how quickly the quantiles of the non-stationary data stream are expected to vary. We expect that the optimal λ will be smaller for slowly varying quantiles and larger for rapidly varying quantiles. By selecting λ, one is essentially selecting an effective window size of previously observed data to include in the quantile estimation. This follows from the fact that more recent data is weighted more heavily than older data. Consider the fraction of the weight included in the most recent r terms.

$$\lambda \sum_{j=0}^{r-1} (1-\lambda)^j = 1 - (1-\lambda)^r.$$

This is to be contrasted with the weight of the first (oldest) term:

$$(1-\lambda)^r.$$

If we define the effective window size as that number of observations for which 99.9% of the weight is contained in the most recent r terms or equivalently that the remaining 0.1% of the weight is associated with the first (oldest) term, then we see that the effective window size is:

$$r = \frac{\log 0.001}{\log (1-\lambda)}.$$

We include below a tabulation of some commonly used values of λ in EWMA applications and their associated effective window size in number of observations.

λ	Effective Window Size
0.01	687
0.05	135
0.1	66
0.2	31

The ideal effective window size should be selected by judgement and domain specific knowledge of the data being analysed. For example, when analysing high frequency forex return data, one may expect that the most recent observations are the most pertinent and only the previous few minutes of observations would be relevant to estimating the current, local behaviour of the process. The second factor to consider is that λ cannot be too large. As we demonstrate in section 4.4, the MISE of $\hat{f}_N(x)$ determines the quality of both CDF and quantile estimates under certain conditions. Using the bound on the MISE that we derive in theorem 14 we see that we can achieve a small integrated variance term by ensuring:

$$\lambda N^{1/2}$$

is sufficiently small. As a starting point, we have found through extensive empirical studies that the common choices of $\lambda = 0.01, 0.05, 0.1$ yield good performance of the algorithm in a number of scenarios. Indeed, in our simulation studies we demonstrate that the EWGH algorithm provides good results over this range of values for λ.

4.4 Quality of CDF and Quantile estimates for Non-negative Random Variables

In this section we derive error bounds on the mean squared error of the Gauss-Hermite CDF estimator and the mean absolute error (MAE) of the quantile estimator for distributions with support on the positive half-real line, subject to some additional conditions. In particular, we demonstrate that these error bounds directly depend on the MISE of the associated probability density function estimates. This greatly simplifies obtaining the aforementioned error bounds and allows us to examine the asymptotic behaviour of the Gauss-Hermite based CDF estimator and the associated quantile estimator. While these results *do not* directly apply to the sequential quantile estimation setting (since N and λ are fixed in that case), they are novel and interesting in their own right. These asymptotic results provide general context and provide comfort that the behaviour of estimators constructed from the Hermite series probability density estimators have sensible asymptotic properties. To obtain asymptotic results we utilise existing MISE consistency results along with the associated rates for the standard Gauss-Hermite expansion [29] as well as novel MISE results for the exponentially weighted Gauss-Hermite expansion which we derive in appendix A.2. All the results derived below are easily extended to $f(x)$ with support on a bounded interval, $[a, b]$ with $\hat{F}_N(x) = \int_a^x \hat{f}_N(x')dx'$(omitted for brevity). Our results are obtained by utilising inequalities which apply specifically in the cases that the probability density function, $f(x)$, has bounded or non-negative support. These inequalities are derived by utilising the Cauchy-Schwarz inequality. This approach is not directly generalisable to the full real line support case (CDF estimator (4.1)). Indeed this particular fact motivated exploring the approach presented in chapter 5.

4.4.1 Quality of the Cumulative Distribution Function Estimate

In what follows we assume that $f(x)$ is supported on $[0, \infty)$.

For non-negative random variables we obtain the simpler estimator:

$$\begin{aligned} \hat{F}_N(x) &= \int_0^x \hat{f}_N(x')dx' \\ &= \sum_{k=0}^{N} \hat{a}_k k! \sum_{l=0}^{\lfloor k/2 \rfloor} \frac{(-1)^l 2^{\frac{3k}{2}-3l-1} \left[\gamma(-l+\frac{k}{2}+\frac{1}{2}, \frac{x^2}{2})\right]}{l!(k-2l)!\sqrt{\pi}}, \; x \geq 0 \end{aligned} \tag{4.10}$$

The expression (4.10) differs from (4.2) since the domain of integration is different. We consider the Mean Squared Error (MSE) criterion and an integrated weighted MSE criterion (inspired by the Cramer-von Mises criterion) as measures of the quality of the estimated cumulative distribution function. The integrated weighted MSE criterion is defined as follows:

$$\begin{aligned} \overline{\omega^2} &= E \int_0^\infty \left[\hat{F}_N(x) - F(x)\right]^2 f(x)dx \\ &= \int_0^\infty E\left[\hat{F}_N(x) - F(x)\right]^2 f(x)dx, \end{aligned} \tag{4.11}$$

where we have made use of Fubini's theorem which allows us to interchange the ordering of the integrals. Note that the PDF $f(x)$ is the weighting factor in (4.11). We begin by presenting two propositions that will be necessary in deriving consistency and rate results for the MSE and $\overline{\omega^2}$.

Proposition 1. *Suppose $f(x)$ is supported on $[0, \infty)$ and $f(x) \in L_2$ then we have:*

$$E\left|\hat{F}_N(x) - F(x)\right|^2 \leq x\, MISE(\hat{f}_N),$$

for fixed x.

Proof. For a non-negative random variable:

$$\left|\hat{F}_N(x) - F(x)\right| = \left|\int_0^x \left[\hat{f}_N(x') - f(x')\right] dx'\right|.$$

By the Cauchy-Schwarz inequality:

$$\left|\hat{F}_N(x) - F(x)\right|^2 \leq \left[\int_0^x \left|\hat{f}_N(x') - f(x')\right|^2 dx'\right] \left[\int_0^x dx'\right].$$

Thus:

$$\left|\hat{F}_N(x) - F(x)\right|^2 \leq x \left[\int_0^x \left|\hat{f}_N(x') - f(x')\right|^2 dx'\right].$$

Now, since $(\hat{f}_N(x') - f(x'))^2$ is non-negative we have,

$$\int_0^x (\hat{f}_N(x') - f(x'))^2 dx' \leq \int_{-\infty}^{\infty} (\hat{f}_N(x') - f(x'))^2 dx'.$$

Thus:

$$\begin{aligned} E\left|\hat{F}_N(x) - F(x)\right|^2 &\leq xE\left[\int_{-\infty}^{\infty} \left|\hat{f}_N(x') - f(x')\right|^2 dx'\right] \\ &= x\,\mathrm{MISE}(\hat{f}_N). \end{aligned} \tag{4.12}$$

This implies that if we have an upper bound for the MISE of $\hat{f}_N(x)$ we can bound the MSE of $\hat{F}_N(x)$.

□

Proposition 2. *Suppose $f(x)$ is supported on $[0, \infty)$ and $f(x)$ has a finite mean, $\mu < \infty$ then we have:*

$$\overline{\omega^2} \leq MISE(\hat{f}_N)\mu.$$

Proof. Utilising proposition 1 we have,

$$\begin{aligned} \overline{\omega^2} &\leq \text{MISE}(\hat{f}_N) \int_0^\infty x f(x) dx \\ &= \text{MISE}(\hat{f}_N)\mu, \end{aligned} \tag{4.13}$$

where μ is the mean of $f(x)$.

□

We now consider the consistency and associated rate of convergence for the CDF estimator (4.10) defined from the standard Gauss-Hermite coefficients (2.19).

Theorem 1. *Suppose $f(x)$ is supported on $[0, \infty)$ and $f(x) \in L_2$. In addition suppose that $\frac{N^{\frac{1}{2}}(n)}{n} \to 0$ as $N(n), n \to \infty$ and $E|X|^{\frac{2}{3}} < \infty$ then we have:*

$$E\left|\hat{F}_N(x) - F(x)\right|^2 \to 0.$$

Proof. The result follows from proposition 1 and the fact that $\text{MISE}(\hat{f}_N) \to 0$ under the conditions $\frac{N^{\frac{1}{2}}(n)}{n} \to 0$ as $N(n), n \to \infty$ and $E|X|^{\frac{2}{3}} < \infty$ [29]. □

Theorem 2. *Suppose $f(x)$ is supported on $[0, \infty)$, $f(x) \in L_2$, $r \geq 1$ derivatives of $f(x)$ exist and $(x - \frac{d}{dx})^r f(x) \in L_2$. Suppose in addition that $\frac{N^{\frac{1}{2}}(n)}{n} \to 0$ as $N(n), n \to \infty$ and $E|X|^{\frac{2}{3}} < \infty$ then if:*

$$N(n) \sim n^{2/(2r+1)}, \text{ we have}$$

$$E\left|\hat{F}_N(x) - F(x)\right|^2 = x\, O\left(n^{-2r/(2r+1)}\right).$$

Proof. In [29] it is established that provided $f(x) \in L_2$, $r \geq 1$ derivatives of $f(x)$ exist, $(x - \frac{d}{dx})^r f(x) \in L_2$, $\frac{N^{\frac{1}{2}}(n)}{n} \to 0$ as $N(n), n \to \infty$ and $E|X|^{\frac{2}{3}} < \infty$ then if:

$$N(n) \sim n^{2/(2r+1)}, \text{ we have}$$

$$\text{MISE}(\hat{f}_N) = O\left(n^{-2r/(2r+1)}\right).$$

Combining this with proposition 1 completes the proof.

□

Note that for $r = 1$ the rate is $O(n^{-2/3})$. It is important to note that this rate is suboptimal compared to the smooth kernel CDF estimate rate which is $O(n^{-1})$. For smooth probability density functions ($r \to \infty$) satisfying the appropriate conditions, the rate for the Gauss-Hermite CDF estimator approaches $O(n^{-1})$.

Theorem 3. *Suppose $f(x)$ is supported on $[0, \infty)$ and $f(x)$ has a finite mean, $\mu < \infty$. In addition suppose that $\frac{N^{\frac{1}{2}}(n)}{n} \to 0$ as $N(n), n \to \infty$ then we have:*

$$\overline{\omega^2} \to 0.$$

Proof. The result follows directly from proposition 2 and the fact that $\text{MISE}(\hat{f}_N) \to 0$ under the conditions $\frac{N^{\frac{1}{2}}(n)}{n} \to 0$ as $N(n), n \to \infty$ and $E|X|^{\frac{2}{3}} < \infty$ [29] (for positive random variables, we have $E(X) < \infty$ implying $E|X|^{\frac{2}{3}} < \infty$ by the Lyapunov inequality).

□

Theorem 4. *Suppose $f(x)$ is supported on $[0, \infty)$, $f(x)$ has a finite mean, $\mu < \infty$, $f(x) \in L_2$, $r \geq 1$ derivatives of $f(x)$ exist and $(x - \frac{d}{dx})^r f(x) \in L_2$. Suppose in addition that $\frac{N^{\frac{1}{2}}(n)}{n} \to 0$ as $N(n), n \to \infty$ then if:*

$$N(n) \sim n^{2/(2r+1)}, \text{ we have}$$

$$\overline{\omega^2} = O\left(n^{-2r/(2r+1)}\right).$$

Proof. This follows directly from proposition 2 and the MISE result referenced in the proof of theorem 2.

□

Remark 1. The condition, $f(x) \in L_2$, $r \geq 1$ derivatives of $f(x)$ exist and $(x - \frac{d}{dx})^r f(x) \in L_2$ for a particular r, plays a central role in deriving MSE and MISE convergence rates in this thesis. This condition is closely related to the space of rapidly decreasing functions (as alluded to in [67]), namely Schwartz space , $S(\mathbb{R})$. Functions in Schwartz space are those functions, $f(x)$, whose derivatives all exist everywhere on $\mathbb{R}$ and along with $f(x)$, decrease to zero faster than any inverse power of x as $|x| \to \infty$. Schwartz space can be expressed as a decreasing intersection of subspaces $S_p(\mathbb{R})$ for $p = 0, 1, 2, \ldots$ [7],

$$S(\mathbb{R}) = \cap_{p \geq 0} S_p(\mathbb{R}) \subset \cdots \subset S_2(\mathbb{R}) \subset S_1(\mathbb{R}) \subset L_2.$$

The subspaces $S_p(\mathbb{R})$ are comprised of all $2p$ differentiable functions, $f(x)$, for which $x^a f^{(b)}(x) \in L_2$ for every $a, b \in \{0, 1, 2, \ldots\}$ with $a + b \leq 2p$ ([7], Theorem 13). By virtue of this fact and the Minkowski inequality it is easy to see that all functions in $S_p(\mathbb{R})$ satisfy $(x - \frac{d}{dx})^r f(x) \in L_2$ where $r = 2p$. There is also an interesting connection to functions in a particular Sobolev space, namely, $W^{r,2}(\mathbb{R}) = H^r$. Functions in $W^{r,2}(\mathbb{R})$ are defined as the subset of functions $f(x) \in L_2$ whose weak derivatives up to order r are also in L_2. Probability densities, $f(x)$, with compact support that are in $W^{r,2}$ also satisfy the condition $(x - \frac{d}{dx})^r f(x) \in L_2$ (by the definition of $W^{r,2}(\mathbb{R})$ and the Minkowski inequality).

Remark 2. An important class of distributions for which we can gain further insight into the theoretical performance of the Gauss-Hermite CDF estimator (and quantile estimator in principle) is power-law distributions. These are heavy-tailed distributions and are highly relevant in a number of fields including finance [16]. We utilise the following definition of the power law distribution:

$$f(x) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\alpha},$$

where $\alpha > 1$ is a requirement for normalisability. In addition, we assume $x_{\min} > 0$. Thus $f(x)$ is supported on $[0, \infty)$, $f(x) \in L_2$ and all derivatives of $f(x)$ exist for $x \geq x_{\min}$. If we require in addition that $\alpha > 2$, then we have a finite mean $E(X) < \infty$. The condition in the theorems proven above that $(x - \frac{d}{dx})^r f(x) \in L_2$ can be related to α as follows:

Denote the operator $\frac{d}{dx}$ as D. Now:

$$\begin{aligned}(x - D)^r f(x) &= \frac{\alpha - 1}{\left(x_{\min}\right)^{1-\alpha}} \left(x^r - x^{r-1}D - x^{r-2}Dx + x^{r-2}D^2 + \cdots + (-1)^r D^r \right) x^{-\alpha} \\ &= O(x^{-(\alpha - r)}),\end{aligned}$$

since x and D are non-commutative operators. Thus $[(x - D)^r f(x)]^2 = O(x^{-2(\alpha - r)})$. This implies that for $[(x - D)^r f(x)]^2$ to be integrable, we must have $-2(\alpha - r) < -1$. Which implies $r < \alpha - \frac{1}{2}$ and thus $r = \lceil \alpha - \frac{1}{2} \rceil - 1$. Thus if we also have $\frac{N^{\frac{1}{2}}(n)}{n} \to 0$ as $N(n), n \to \infty$, theorem 1 and theorem 2 imply that the Gauss-Hermite CDF estimator is consistent for power-law distributions and the rate is $E\left|\hat{F}_N(x) - F(x)\right|^2 = x\, O\left(n^{-(2\lceil \alpha - \frac{1}{2} \rceil - 2)/(2\lceil \alpha - \frac{1}{2} \rceil - 1)}\right)$. Similarly theorem 3 and theorem 4 imply that $\overline{\omega^2} \to 0$ and the rate is $\overline{\omega^2} = O\left(n^{-(2\lceil \alpha - \frac{1}{2} \rceil - 2)/(2\lceil \alpha - \frac{1}{2} \rceil - 1)}\right)$. Thus for power-

law distributions that have a finite mean, the rates are $O\left(n^{-2/3}\right)$ or better. For power-law distributions that have a finite variance, $\alpha > 3$ and thus the rates are $O\left(n^{-4/5}\right)$ or better. As $\alpha \to \infty$ (along with the number of finite moments of the power-law distribution), the Gauss-Hermite rates approach $O\left(n^{-1}\right)$.

Remark 3. The results we have derived above apply to i.i.d. continuous (real-valued) random variable data. The MISE results in [46] for stationary mixing sequences along with proposition 1 and 2 that we have obtained suggest that similar results for the MSE and $\overline{\omega^2}$ of the Gauss-Hermite CDF estimator as derived above are easily obtainable for non-independently distributed data and stationary data in particular, under the appropriate conditions.

It is also feasible to apply the Gauss-Hermite CDF estimator to discrete random variables. Indeed, using this estimator may be regarded as a form of smoothing. Intuitively, the Gauss-Hermite estimator would work best for discrete distributions with countably infinite support or finite support with large cardinality. Investigating the effectiveness of the Gauss-Hermite approach in these applications is an area for future research.

To summarise, we have derived the asymptotic results above in the setting where N depends on n, i.e. $N = N(n)$. These results give comfort that the Gauss-Hermite based CDF estimator has sensible asymptotic behaviour. Our proposed online estimators have N fixed however. Thus results such as theorem 1 do not apply directly. Instead, as $n \to \infty$, these online estimators have MSE bounds determined by the integrated squared bias of the truncated Gauss-Hermite PDF estimators on which the plug-in CDF estimators are based. While this bias persists for fixed N even as $n \to \infty$, these estimators are nonetheless very useful in practice as we have demonstrated in our simulation and real data results. In addition, we can still gain insight into the behaviour of these estimators at fixed N. We now consider the behaviour of

the EWGH CDF estimator defined from utilising the exponentially weighted Gauss-Hermite coefficients (4.9) in the expression (4.10) for the CDF where λ is fixed, N is fixed and sufficiently large and $n \to \infty$. We begin with the case of i.i.d. data drawn from $f(x)$.

Theorem 5. *Suppose $f(x)$ is supported on $[0, \infty)$, $f(x)$ has a finite mean, $\mu < \infty$, $f(x) \in L_2$, $r \geq 1$ derivatives of $f(x)$ exist and $(x - \frac{d}{dx})^r f(x) \in L_2$. Then for N fixed and sufficiently large and $n \to \infty$.*

$$E \left| \hat{F}_N(x) - F(x) \right|^2 = x \left[O(N^{1/2}) \left[\frac{\lambda}{2 - \lambda} \right] + O(N^{-r}) \right],$$

and

$$\overline{\omega^2} = O(N^{1/2}) \left[\frac{\lambda}{2 - \lambda} \right] + O(N^{-r}).$$

Proof. The result follows from proposition 1 and 2 respectively along with theorem 14. □

We now treat the case of independent, non-identically distributed data. In particular, we consider the case of a change point where the distribution changes from $f_1(x)$ to $f_2(x)$. We consider this a fundamental example of non-identically distributed data.

Theorem 6. *Suppose $s+1$ observations are drawn from a probability distribution $f_1 \in L_2$ followed by a further t observations from a second distribution $f_2 \in L_2$ i.e. we assume an independent sequence of $s+1$ observations $\mathbf{x}_0, \ldots, \mathbf{x}_s \sim f_1(x)$ followed by an independent sequence of t observations, $\mathbf{x}_{s+1}, \ldots, \mathbf{x}_{t+s} \sim f_2(x)$. The total number of observations is thus $s+t+1$. If $r \geq 1$ derivatives of $f_2(x)$ exist and $(x - \frac{d}{dx})^r f_2(x) \in L_2$ and both distributions have a finite mean, then:*

$$E\left|\hat{F}_N(x) - F_2(x)\right|^2$$
$$= x\left[O(N^{1/2})\left[4(1-\lambda)^{2t} + \frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2(s+t)}\right] + (1-\lambda)^{2(s+t)}\right] + O(N^{-r})\right]$$

and

$$\overline{\omega^2} = O(N^{1/2})\left[4(1-\lambda)^{2t} + \frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2(s+t)}\right] + (1-\lambda)^{2(s+t)}\right] + O(N^{-r}),$$

where $F_2(x)$ is the CDF of $f_2(x)$ and $\overline{\omega^2} = \int_0^\infty E\left[\hat{F}_N(x) - F_2(x)\right]^2 f_2(x)dx$.

Proof. The result follows from proposition 1 and 2 respectively along with theorem 15. □

Note that asymptotically we can choose $\lambda(n) \to 0$, $n \to \infty$ such that $E\left|\hat{F}_N(x) - F(x)\right|^2 \to 0$ and $\overline{\omega^2} \to 0$ for both the i.i.d. case and the non-identically distributed, independent (change point) case, where $n = s+t+1$ is the total number of observations. We do not present the proof here for the sake of brevity.

4.4.2 Quality of Quantile Estimate

We now present a result pertaining to the mean absolute error (MAE) of Gauss-Hermite quantile estimates.

Theorem 7. *Suppose $f(x)$ is supported on $[0, \infty)$. In addition we suppose that the true quantile x_p lies between $[x_p^{min}, x_p^{max}]$ and that $f(x) \geq d$, $d > 0$ for $x \in [x_p^{min}, x_p^{max}]$. Finally we assume that we know x_p^{min}, x_p^{max}, d allowing us to refine the Gauss-Hermite quantile estimate as follows:*

$$\hat{x}_p = \begin{cases} \hat{F}_N^{-1}(p) & \text{if } \hat{F}_N^{-1}(p) \text{ exists in } [x_p^{min}, x_p^{max}] \\ & \text{and } \hat{f}_N(x) \geq d,\ x \in [x_p^{min}, x_p^{max}] \\ \text{undefined} & \text{otherwise.} \end{cases} \tag{4.14}$$

For well defined $\hat{x}_p$:

$$E\left|x_p - \hat{x}_p\right| \leq \frac{\sqrt{x_p}}{d}\sqrt{MISE(\hat{f}_N(x))}.$$

Proof. In the proof of proposition 1 we established:

$$\left|\hat{F}_N(x) - F(x)\right|^2 \leq x \int_{-\infty}^{\infty} (\hat{f}_N(x') - f(x'))^2 dx'.$$

Thus:

$$\left|\hat{F}_N(x_p) - F(x_p)\right| \leq \sqrt{x_p}\sqrt{\int_{-\infty}^{\infty} (\hat{f}_N(x') - f(x'))^2 dx'}, \tag{4.15}$$

$$\left|\hat{F}_N(x_p) - p\right| \leq \sqrt{x_p}\sqrt{\int_{-\infty}^{\infty} (\hat{f}_N(x') - f(x'))^2 dx'}, \tag{4.16}$$

where $x_p = q(p)$. Provided $\hat{x}_p = \hat{F}_N^{-1}(p)$ exists, this implies:

$$\left|\hat{F}_N(x_p) - \hat{F}_N(\hat{x}_p)\right| \leq \sqrt{x_p}\sqrt{\int_{-\infty}^{\infty} (\hat{f}_N(x') - f(x'))^2 dx'} \tag{4.17}$$

$$\left|\int_{\hat{x}_p}^{x_p} \hat{f}_N(x) dx\right| \leq \sqrt{x_p}\sqrt{\int_{-\infty}^{\infty} (\hat{f}_N(x') - f(x'))^2 dx'} \tag{4.18}$$

$$|x_p - \hat{x}_p|\left|\hat{f}_N(\hat{c})\right| \leq \sqrt{x_p}\sqrt{\int_{-\infty}^{\infty} (\hat{f}_N(x') - f(x'))^2 dx'}, \tag{4.19}$$

where $\hat{c}$ lies in the interval $(x_p, \hat{x}_p)$ if $\hat{x}_p > x_p$ or $(\hat{x}_p, x_p)$ if $\hat{x}_p < x_p$. We have

applied the mean value theorem and thus $\hat{f}_N(\hat{c})$ is equal to the *mean* value of $\hat{f}_N$ in the interval i.e. $\hat{f}_N(\hat{c}) = \frac{1}{x_p - \hat{x}_p} \int_{x_p}^{\hat{x}_p} \hat{f}_N(x)dx$.

Thus, provided $\hat{f}_N(\hat{c}) \neq 0$, we have:

$$|x_p - \hat{x}_p| \leq \frac{\sqrt{x_p}}{\left|\hat{f}_N(\hat{c})\right|} \sqrt{\mathrm{ISE}(\hat{f}_N)}, \tag{4.20}$$

To get a more concrete result we assume we utilise our refined quantile estimator. This estimator is quite natural as it restricts the quantile estimates to those formed from bona-fide probability densities i.e. those $\hat{f}_N(x)$ that are non-negative. Moreover, requiring $\hat{f}_N(x) > 0$ ensures that there is a unique solution for $\hat{x}_p = \hat{q}_N(p)$. Finally, it allows us to reject quantile estimates that are out of bounds. For the estimates that are not undefined we have, via the Cauchy-Schwarz inequality:

$$E\,|x_p - \hat{x}_p| \leq \frac{\sqrt{x_p}}{d} \sqrt{\mathrm{MISE}(\hat{f}_N)}. \tag{4.21}$$

□

This result again demonstrates the direct link to the MISE of the probability density function and allows one to, in principle, establish asymptotic properties under certain conditions (similar to above results for the CDF). As a note to the practitioner, this refined quantile estimator can be viewed as one where we reject estimates that are out of bounds and those that are not constructed from bona-fide probability density estimates. We expect this to be an infrequent scenario unless N is too small and the probability density estimator is heavily biased or too few observations have been incorporated into the estimate. We base this expectation on extensive empirical analysis.

4.4.3 Monotonicity of Cumulative Distribution Function Estimate

We can obtain insights into the monotonicity of CDF estimates obtained using the fixed and finite N Gauss-Hermite cumulative distribution function estimator by appealing to the properties of Hermite polynomials. In particular, the Hermite polynomials have regions of strict positivity beyond the largest positive root (this is easy to see in that $\lim_{x\to\infty} H_k(x) \to +\infty$). In [63] bounds on the largest positive root were obtained, namely,

$$h_{k1} \leq \sqrt{2k+1} - d(2k+1)^{-1/6} \tag{4.22}$$

where h_{k1} is the largest root and d is a constant which is approximately equal to 1.85575. Notice that the bound is strictly increasing in k.

Thus for $x > h_{N1}$, $H_k(x) > 0$, $k = 0, \ldots, N$. We tabulate some minimum values of x, denoted x_N^*, guaranteeing positivity of $H_k(x)$, $k = 0, \ldots, N$ for $x > x_N^*$ for various values of N in the table below, based on (4.22).

N	x_N^*
3	1.3040
10	3.4653
20	5.4038
50	9.1899

This has several useful implications which we discuss below.

Power Law Distributions

For a given fixed and finite N and a power law distribution with $x_{\min} > x_N^*$, the finite N Hermite series estimator $\hat{f}_N(x)$ is strictly positive for $x \geq x_{\min}$. This follows from the definition of the Gauss-Hermite density estimator, (2.19) and (2.20). Thus the cumulative distribution function estimator defined as $\hat{F}_N(x) = \int_{x_{\min}}^{x} \hat{f}_N(x')dx'$ is strictly monotonically increasing. For

power law distributions with $x_{\min}$ bigger than a certain constant we therefore have guarantees of better behaviour of this estimator than in the general case, at least in the monotonicity sense.

Guaranteeing Monotonicity

It is clear that for $f(x)$ with non-negative support or support on a (possibly unknown) bounded, non-negative interval, we can guarantee monotonicity of the CDF estimates by transforming the random variables as $\mathbf{x}_i + x_N^*$ prior to estimation of the Gauss-Hermite coefficients and using the estimator,

$$\hat{F}_N(x) = \int_{x_N^*}^{x+x_N^*} \hat{f}_N(x')dx', \tag{4.23}$$

for the CDF.

Uniqueness of Quantile Estimates

For CDF estimates that are strictly monotonically increasing, we are guaranteed the uniqueness of a quantile estimate derived from this CDF estimate.

4.5 Simulation Results

In this section we evaluate the behaviour of the Gauss-Hermite (GH) online quantile estimation algorithm presented in section 4.2 and the Exponentially Weighted Gauss Hermite (EWGH) algorithm presented in section 4.3 on simulated data. We also compare the performance of these algorithms to a leading existing algorithm for online quantile estimation, namely Exponentially Weighted Stochastic Approximation (EWSA), which has been shown to be competitive with a number of other algorithms for online quantile estimation [13]. The EWSA algorithm is an exponentially weighted version of the stochastic approximation algorithm of [64]. EWSA has two parameters, one that controls the size of the batches of data used to update the quantile

estimates (denoted M) and a weighting factor w that controls the weighting of updates to the density and quantile estimates in the stochastic approximation scheme (see [13] for a detailed description of the algorithm).

In our investigations both i.i.d and non-identically distributed simulated data are considered. In particular, i.i.d data from a chi-squared distribution with five degrees of freedom, χ_5, and an exponential distribution with mean and variance equal to one are considered. In the i.i.d. case, the data have static quantiles.

In the non-identically distributed setting we first consider simulated data drawn from distributions with non-stationary parameters. In particular we consider simulated data drawn from a normal distribution with variance one and a mean of $.006j$ at update j to simulate data with a linear trend in the mean. We then consider an exponential distribution with mean and variance equal to $1 + 0.006j$ at update j to simulate data with a non-stationary mean and standard deviation. This is an interesting test in that the dispersion of the data increases as the number of observations grows. Both of these models were studied in the simulations of [13]. In addition, we consider data drawn from a χ_5 distribution for the first half of the data set, switching to an exponential distribution with mean and variance equal to one for the second half of the data set. This is an example of a non-identically distributed data set with a change point. All these non-identically distributed simulated data have dynamic quantiles that change over time.

The choice of these test distributions can be motivated by the diversity of their properties and the frequent appearance of these distributions in statistical applications. For each distribution, three quantiles are estimated, namely the 0.5 (median), 0.9 and 0.99 quantiles following [13]. Note that in the case of the Gauss-Hermite based algorithms arbitrary quantiles can

be obtained at any point in time whereas algorithms such as the EWSA algorithm require the quantiles to be specified upfront. Online, arbitrary quantiles are not available for algorithms such as EWSA. The maximum number of observations, $m = 4000$ per run in the i.i.d. case and $m = 1000$ in the non-identically distributed case (corresponding to the maximum number of *updates* in [13]). There were 1000 runs in total for each distribution. We utilise the empirical root mean squared error (RMSE) to evaluate the performance of the online quantile estimation algorithms in estimating the quantile q after j observations (where we denote the quantile estimate $\hat{q}_j$). The RMSE at updating step j is defined by:

$$\text{RMSE}(\hat{q}_j) = \left[E(\hat{q}_j - q_j)^2\right]^{\frac{1}{2}}, \tag{4.24}$$

and is estimated by averaging the squared difference between q_j and $\hat{q}_j$ over the 1000 simulation runs and then taking the square root. As a measure of the error in the RMSE estimate at updating step j, we construct a 95th percentile bootstrap confidence interval:

$$(\text{RMSE}^*_{(0.025)}, \text{RMSE}^*_{(0.975)}),$$

where $\text{RMSE}^*_{(0.025)}$ and $\text{RMSE}^*_{(0.975)}$ denote the 0.025 and 0.975 quantiles of the bootstrap estimates of the RMSE respectively (1000 bootstrap estimates were utilised in constructing the intervals for each j).

In the non-identically distributed case we also introduce another measure to assess performance which will be utilised in our real data studies in section 4.6 as well. The measure is defined as follows: we count the number of times the $(j+1)$th observation is smaller than the online estimates of the 0.5 (median), 0.9 and 0.99 quantiles obtained up to observation j. These counts are then normalised by the total number of observations. Ideally, the out-of-sample observation should be smaller than the median quantile with

probability 0.5. Similarly, the out-of-sample observation should be smaller than the 0.9 quantile with probability 0.9 and it should be smaller than the 0.99 quantile with probability 0.99. We compare the observed frequencies to these probabilities and report 95th percentile bootstrap confidence intervals for the observed frequencies.

In [13] a range of values for the EWSA algorithm parameters M and w were investigated for performance. It was demonstrated that $w = 0.05$ gives the best trade-off between bias and long-run variability. In addition, the value of $M = 15$ was shown to be superior in long-run performance to other values of M tested. Since we evaluate the same i.i.d. models (except for the addition of the chi-squared i.i.d. model which is qualitatively similar) and the same non-identically distributed data stream models as [13] in our simulations (except for the addition of the change point case which switches from chi-squared to exponentially distributed), we regard these parameters as principled choices for good performance of the EWSA algorithm in these settings.

In our simulation studies for the GH and EWGH algorithms we demonstrate the effectiveness of the algorithms over a range of values for N and λ and identify particularly good choices. Concretely for the GH algorithm we consider $N = 4, 6, 8, 10, 12$ at $m = 100, 400, 4000$ observations to evaluate the bias-variance trade-off at different numbers of observations. We also present RMSE curves where we compare the results to the EWSA algorithm. For the EWGH algorithm we consider $\lambda = 0.01, 0.05, 0.1$ ($N = 6$) at $m = 100, 400, 1000$ observations and present RMSE curves comparing to the GH ($N = 6$) and EWSA algorithms.

Finally, in order to practically apply the Gauss-Hermite based algorithms more effectively, an online standardisation procedure was applied to the data

as outlined in appendix A.1. This is not a pre-processing step (this would defeat the purpose of an online algorithm) but rather part of the online algorithm. Also, this procedure may not always be necessary, depending on the application. It is also noteworthy that we utilise the alternative CDF estimator (4.4) in estimating quantiles for the GH and EWGH algorithms as we have found this estimator to yield better results empirically.

4.5.1 IID Data

For the i.i.d. simulated data, we evaluate the performance of the GH algorithm for various choices of N and we use the EWGH algorithm ($\lambda = 0.05$) and the EWSA algorithm ($M = 15$, $w = 0.05$) for comparison. The GH algorithm performs well for most choices of N, illustrating that the effectiveness of the algorithm is not critically dependent on this choice. That said, the value of $N = 6$ appears to be the best choice in most cases when viewed from a RMSE perspective. In addition, $N = 6$ is a good choice from a computational speed and efficiency viewpoint since there are fewer coefficients to store and update than for higher choices of N. When compared to the EWSA algorithm, the GH algorithm performs better in most cases. The EWGH algorithm (with $\lambda = 0.05$) has a larger error than the GH algorithm in all cases. This is not entirely surprising however. The EWGH algorithm trades extra variance in the estimates of the coefficients for the ability to track dynamic quantiles. The individual tests are discussed below.

The Chi-Squared Distribution

The chi-squared distribution appears frequently in statistics and is a useful test distribution in that it has support on the half real line $[0, \infty)$ and not the full real line. This distribution is used to simulate skewed data. This is a challenging test case for the Gauss-Hermite based algorithms since the chi-squared distribution is a considerable departure from the normal dis-

tribution which undergirds the Gauss-Hermite expansion. We consider the chi-squared distribution with five degrees of freedom in particular. See figure 4.1 for plots of the GH RMSE for $N = 4, 6, 8, 10, 12$ at $m = 100, 400, 4000$ observations. The results for the EWGH and EWSA algorithms are also included for comparison. These figures illustrate that good results are achieved for all values of N considered for the GH algorithm. $N = 6$ appears to provide the best results. It is interesting to note that upon first inspection, it is counter-intuitive that the RMSE increases at higher values of N even when the number of observations is large. We suspect that this is due to the additional bias introduced by the online standardisation procedure as well as by using the CDF estimator (4.4) instead of (4.2). See figure 4.3 for a comparison of the GH algorithm ($N = 6$) and the EWSA algorithm. Note that the EWSA results were excluded from figure 4.1(i) and figure 4.3(c) since they were disproportionately large and would obscure the GH results when presented on a common scale. This may indicate instability in the EWSA algorithm for estimating tail quantiles such as $p = 0.99$.

The Exponential Distribution

The exponential distribution is another commonly occurring distribution. The distribution also has support on the half real line $[0, \infty)$ and is used to simulate skewed data. The exponential distribution is an even more challenging test case for the Gauss-Hermite based algorithms than the chi-squared distribution. This is due to the fact that the exponential distribution's mode occurs at the start of its domain which is to be contrasted with the mode of the normal distribution which is equal to its median. We consider the exponential distribution with mean and variance equal to one. See figure 4.2 for plots of the GH RMSE for $N = 4, 6, 8, 10, 12$ at $m = 100, 400, 4000$ observations. The results for the EWGH and EWSA algorithms are also included for comparison. These figures again illustrate that good results are achieved for all values of N considered and that the value of $N = 6$ appears to provide

the best results. See figure 4.4 for a comparison of the GH algorithm ($N = 6$) and the EWSA algorithm. Note that the EWSA results were excluded from figure 4.2(i) and figure 4.4(c) since they were again disproportionately large and would obscure the GH results when presented on a common scale.

4.5.2 Non-identically Distributed Data

For the non-identically distributed simulated data, we evaluate the performance of the EWGH algorithm for various choices of λ and we use the GH algorithm ($N = 6$) and the EWSA algorithm ($M = 15$, $w = 0.05$) for comparison. Motivated by the analysis of the GH algorithm in the i.i.d. setting we set $N = 6$ for all tests of the EWGH algorithm. The EWGH algorithm compares favourably with the EWSA algorithm for all values of λ, illustrating that the effectiveness of the algorithm is not critically dependent on this choice in the models we studied. The GH and EWGH algorithms outperform the EWSA algorithm in almost all cases. The dynamic quantile tracking ability of the EWGH algorithm is apparent in that it achieves better results than the GH algorithm when using an appropriate value of λ. The individual tests are discussed below.

Normal Distribution with Drift

The normal distribution is ubiquitous. We consider simulated data drawn from a normal distribution with variance one and a mean of $.006j$ at update j to simulate data with a linear trend in the mean. See figure 4.5 for plots of the EWGH RMSE for $\lambda = 0.01, 0.05, 0.1$ at $m = 100, 400, 1000$ observations. The results for the GH and EWSA algorithms are also included for comparison. These figures illustrate that competitive results are achieved for all values of λ considered for the EWGH algorithm. The value of $\lambda = 0.01$ appears to provide the best results. See figure 4.8 for the RMSE curve for the EWGH algorithm with $N = 6$, $\lambda = 0.01$ compared to the GH and EWSA algorithms.

The results for the observed out-of-sample frequencies of observations smaller than or equal to current quantile estimates for the normal distribution with drift are presented in table A.1 in appendix A.3. Competitive results are again achieved for all values of λ compared to the GH and EWSA algorithms. Both $\lambda = 0.01$ and $\lambda = 0.05$ appear to be good choices overall for the EWGH algorithm.

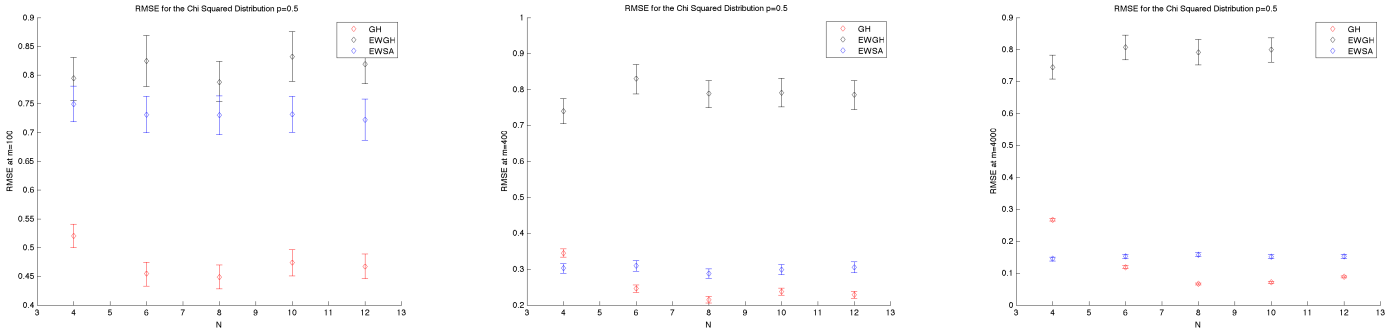
Exponential Distribution with Drift

In this section we consider an exponential distribution with mean and variance equal to $1 + 0.006j$ at update j to simulate data with a non-stationary mean and standard deviation. The dispersion of the data increases as the number of observations grows. See figure 4.6 for plots of the EWGH RMSE for $\lambda = 0.01, 0.05, 0.1$ at $m = 100, 400, 1000$ observations. The results for the GH and EWSA algorithms are also included for comparison. These figures illustrate that competitive results are again achieved for all values of λ considered for the EWGH algorithm. The value of $\lambda = 0.01$ appears to provide the best results. See figure 4.9 for the RMSE curve for the EWGH algorithm with $N = 6$, $\lambda = 0.01$ compared to the GH and EWSA algorithms. The results for the observed out-of-sample frequencies of observations smaller than or equal to current quantile estimates for the exponential distribution with drift are presented in table A.2 in appendix A.3. As previously, competitive results are achieved for all values of λ compared to the GH and EWSA algorithms. Both $\lambda = 0.01$ and $\lambda = 0.05$ appear to be good choices overall for the EWGH algorithm.

Chi-Squared to Exponential distribution with Change Point

In order to simulate data with a change point, we consider data drawn from a chi-squared distribution with five degrees of freedom for the first half of the data switching to an exponential distribution with mean and variance equal to one for the second half of the data. See figure 4.7 for plots of the EWGH

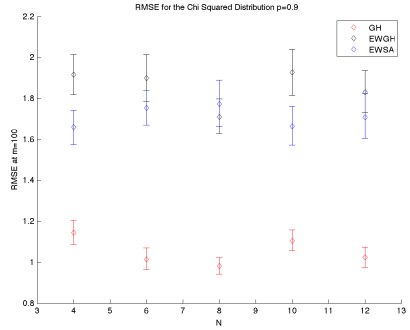
RMSE for $\lambda = 0.01, 0.05, 0.1$ at $m = 100, 400, 1000$ observations. The results for the GH and EWSA algorithms are also included for comparison. Competitive results are again achieved for essentially all values of λ considered for the EWGH algorithm. Here we see that the best value of λ depends on the number of observations. This makes sense in that the change point occurs at half the total number of observations and thus the fewer the number of total observations, m, the more quickly the quantile estimation method needs to adapt to preserve accuracy. The values of $\lambda = 0.05$ or $\lambda = 0.1$ appear to provide the best results for $m = 100$, $\lambda = 0.01$ or $\lambda = 0.05$ provide best results for $m = 400$ and $\lambda = 0.01$ provides best results for $m = 1000$. See figure 4.10 for the RMSE curve for the EWGH algorithm with $N = 6$, $\lambda = 0.01$ compared to the GH and EWSA algorithms. The results for the observed out-of-sample frequencies of observations smaller than or equal to current quantile estimates for the simulated data with a change point are presented in table A.3 in appendix A.3. It is apparent that the EWGH algorithm is competitive with the GH and EWSA algorithms, particularly with appropriate choices of λ. These results echo the RMSE results with respect to change point adaptation, $\lambda = 0.1$ is best for $m = 100$ and $\lambda = 0.05$ is best for $m = 400, 1000$.



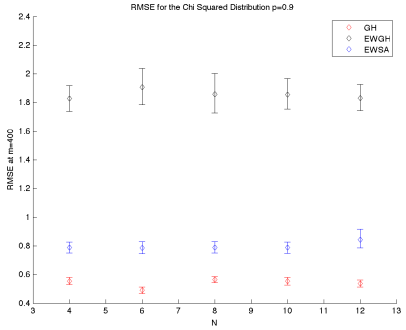
(a) m=100, p=0.5

(b) m=400, p=0.5

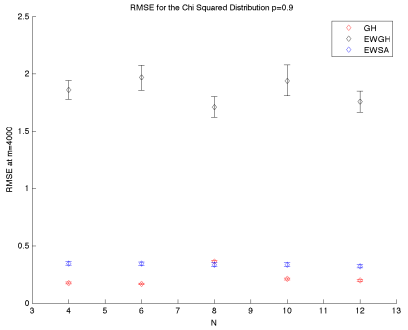
(c) m=4000, p=0.5



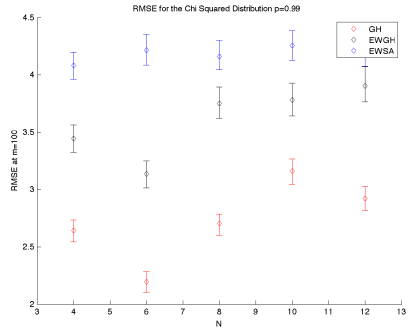
(d) m=100, p=0.9



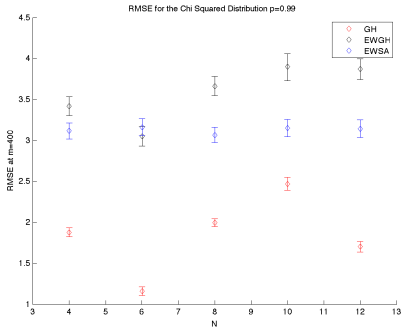
(e) m=400, p=0.9



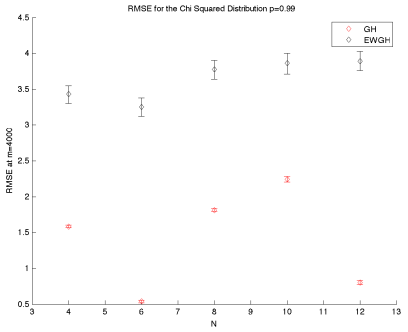
(f) m=4000, p=0.9



(g) m=100, p=0.99

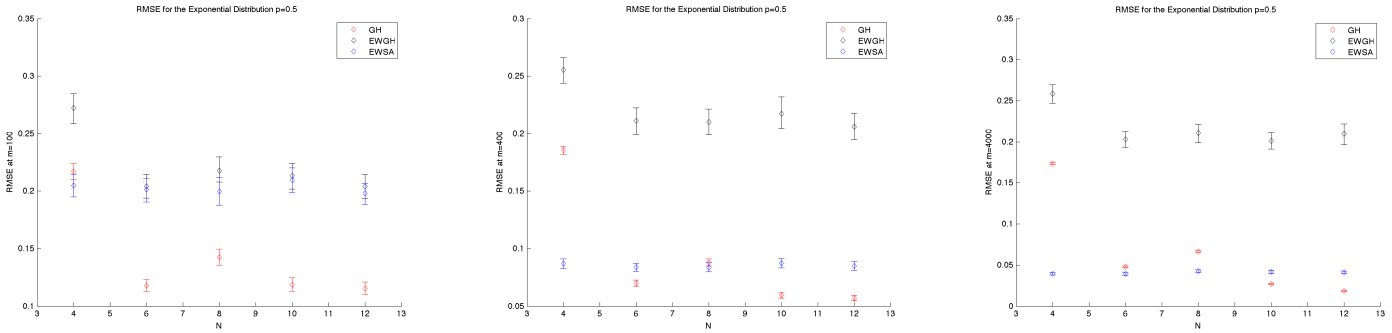


(h) m=400, p=0.99



(i) m=4000, p=0.99

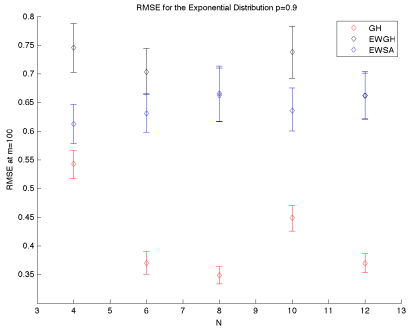
Figure 4.1: Chi-Squared Distribution: GH RMSE for $N = 4, 6, 8, 10, 12$ at $m = 100, 400, 4000$ observations for the $p = 0.5, 0.9, 0.99$ quantiles (including 95% percentile bootstrap confidence intervals). The exact quantiles are 4.3515, 9.2364 and 15.0863 for comparison.



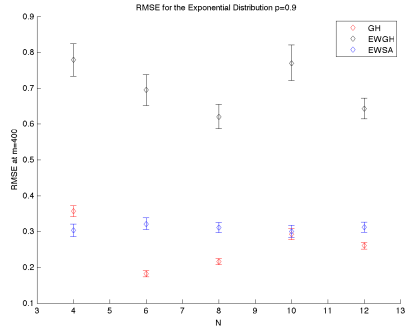
(a) m=100, p=0.5

(b) m=400, p=0.5

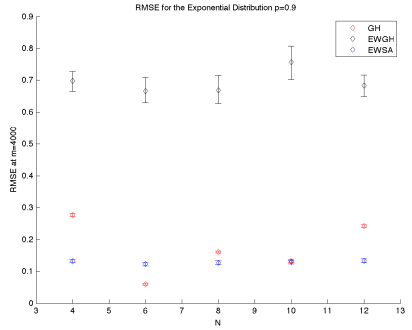
(c) m=4000, p=0.5



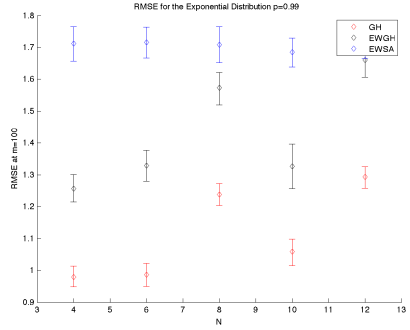
(d) m=100,p=0.9



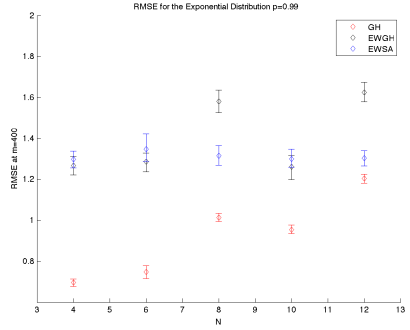
(e) m=400, p=0.9



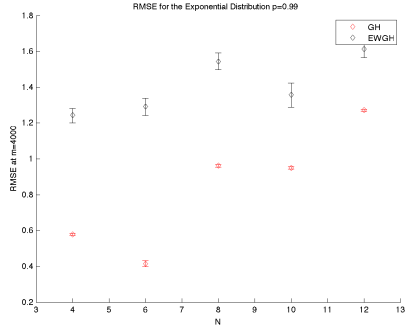
(f) m=4000, p=0.9



(g) m=100, p=0.99

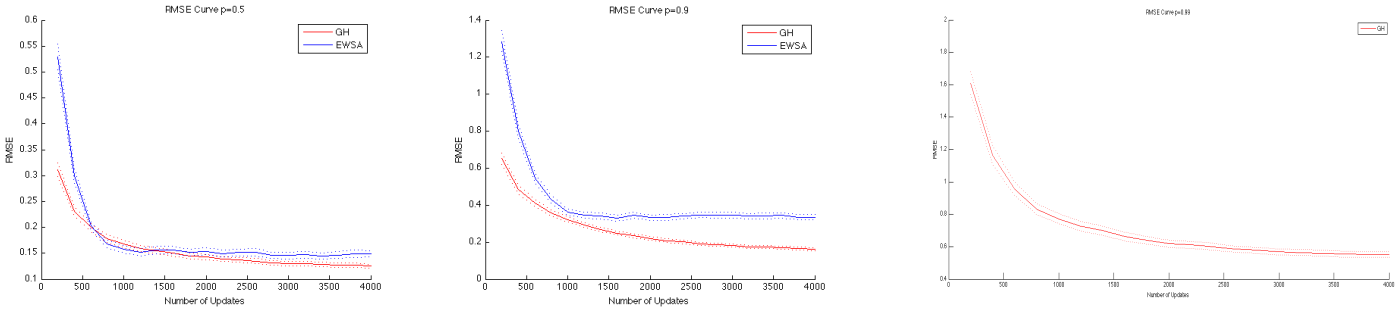


(h) m=400, p=0.99



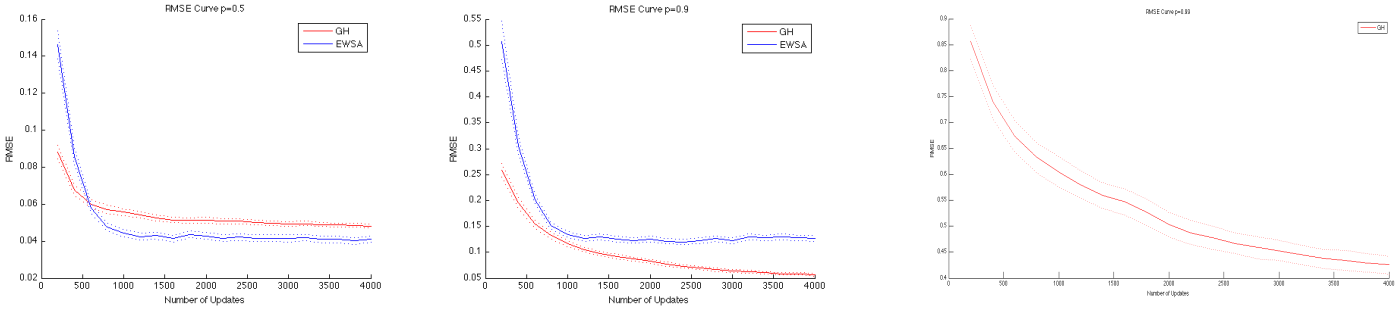
(i) m=4000, p=0.99

Figure 4.2: Exponential Distribution: GH RMSE for $N = 4, 6, 8, 10, 12$ at $m = 100, 400, 4000$ observations for the $p = 0.5, 0.9, 0.99$ quantiles (including 95% percentile bootstrap confidence intervals). The exact quantiles are 0.6931, 2.3026 and 4.6052 for comparison.



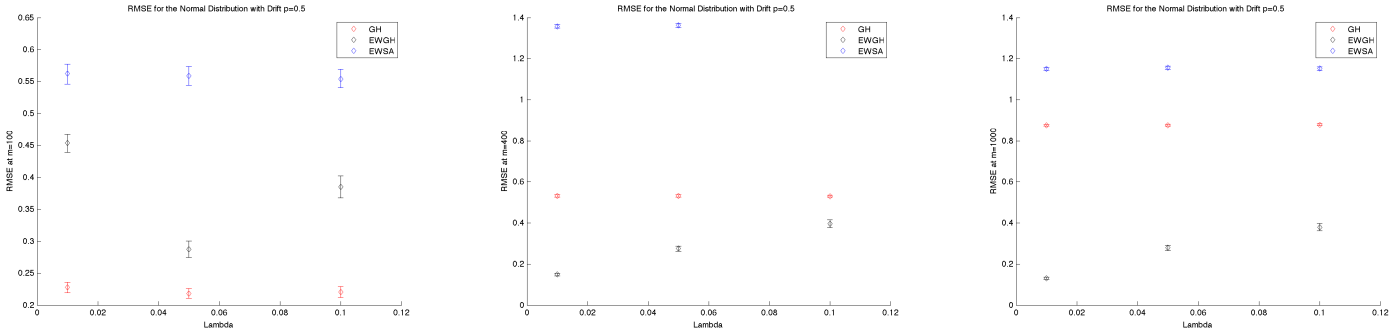
(a) p=0.5 (b) p=0.9 (c) p=0.99

Figure 4.3: RMSE curves associated with the chi-squared distribution with five degrees of freedom for the 0.5, 0.9 and 0.99 quantiles (including 95% percentile bootstrap confidence intervals). The exact quantiles are 4.3515, 9.2364 and 15.0863 respectively. The Gauss-Hermite algorithm utilises $N = 6$.



(a) p=0.5 (b) p=0.9 (c) p=0.99

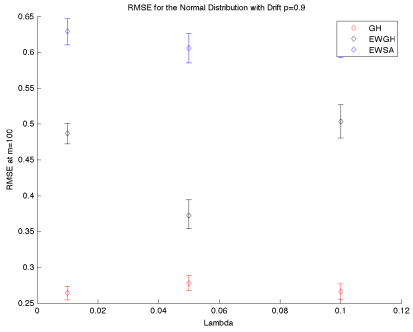
Figure 4.4: RMSE curves associated with the exponential distribution for the 0.5, 0.9 and 0.99 quantiles (including 95% percentile bootstrap confidence intervals). The exact quantiles are 0.6931, 2.3026 and 4.6052 respectively. The Gauss-Hermite algorithm utilises $N = 6$.



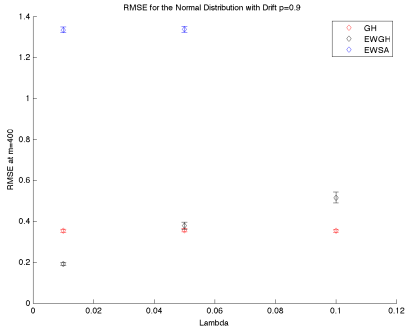
(a) m=100, p=0.5

(b) m=400, p=0.5

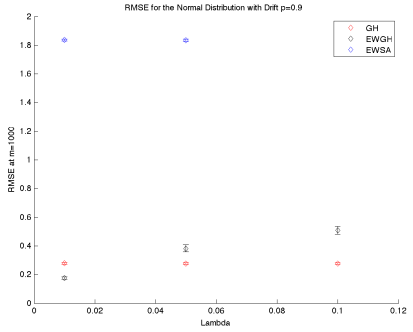
(c) m=1000, p=0.5



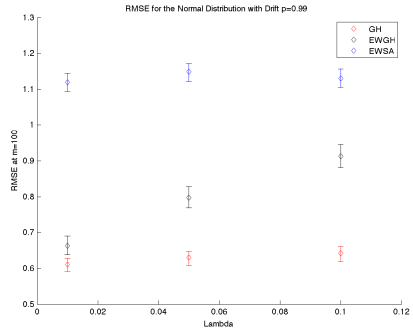
(d) m=100, p=0.9



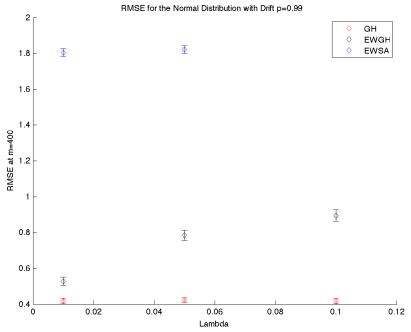
(e) m=400, p=0.9



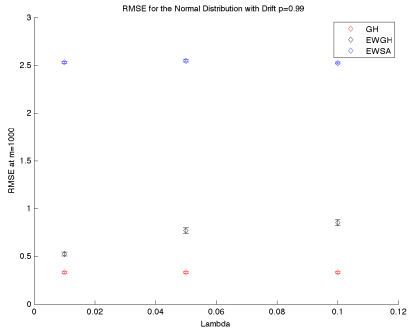
(f) m=1000, p=0.9



(g) m=100, p=0.99

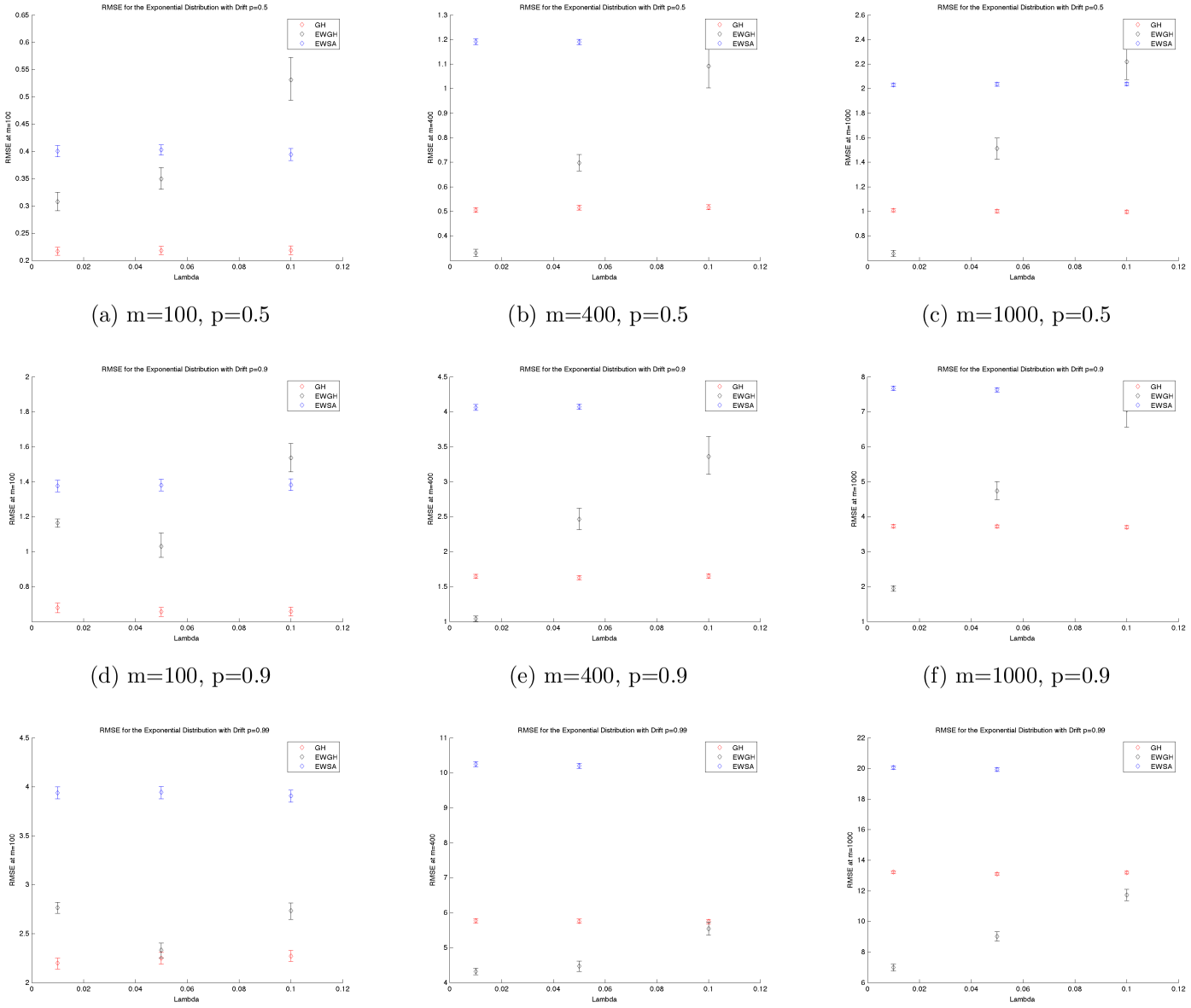


(h) m=400, p=0.99



(i) m=1000, p=0.99

Figure 4.5: Normal Distribution with Drift: EWGH RMSE for $\lambda = 0.01, 0.05, 0.1$ at $m = 100, 400, 1000$ observations for the $p = 0.5, 0.9, 0.99$ quantiles (including 95% percentile bootstrap confidence intervals). The GH and EWGH algorithms utilise $N = 6$.

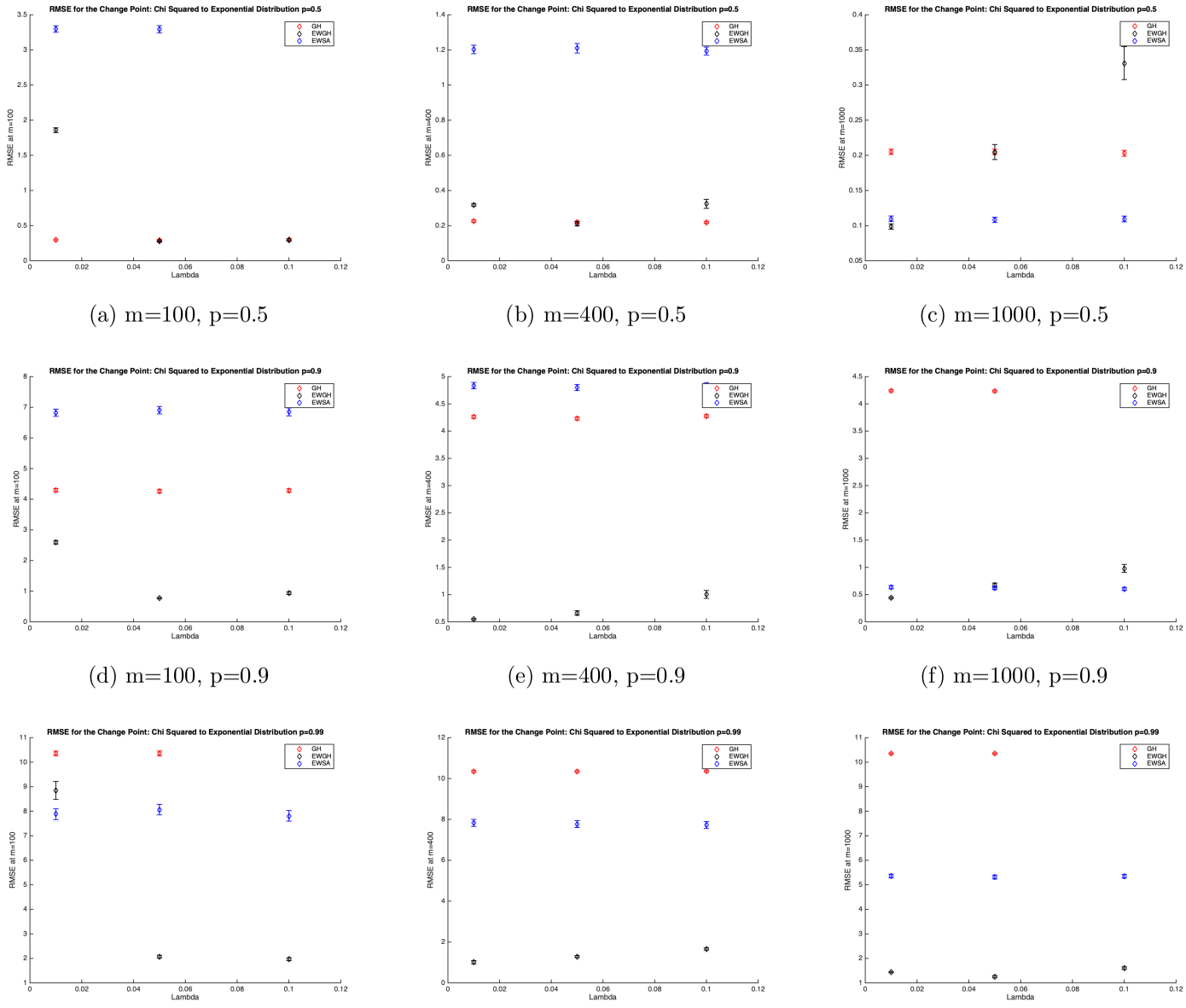


(a) m=100, p=0.5 (b) m=400, p=0.5 (c) m=1000, p=0.5

(d) m=100, p=0.9 (e) m=400, p=0.9 (f) m=1000, p=0.9

(g) m=100, p=0.99 (h) m=400, p=0.99 (i) m=1000, p=0.99

Figure 4.6: Exponential Distribution with Drift: EWGH RMSE for $\lambda = 0.01, 0.05, 0.1$ at $m = 100, 400, 1000$ observations for the $p = 0.5, 0.9, 0.99$ quantiles (including 95% percentile bootstrap confidence intervals). The GH and EWGH algorithms utilise $N = 6$.

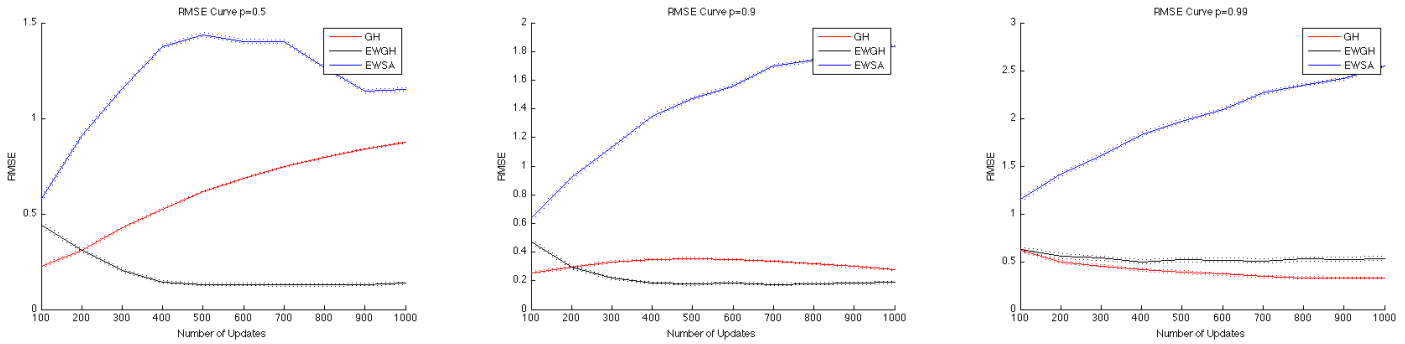


(a) m=100, p=0.5 (b) m=400, p=0.5 (c) m=1000, p=0.5

(d) m=100, p=0.9 (e) m=400, p=0.9 (f) m=1000, p=0.9

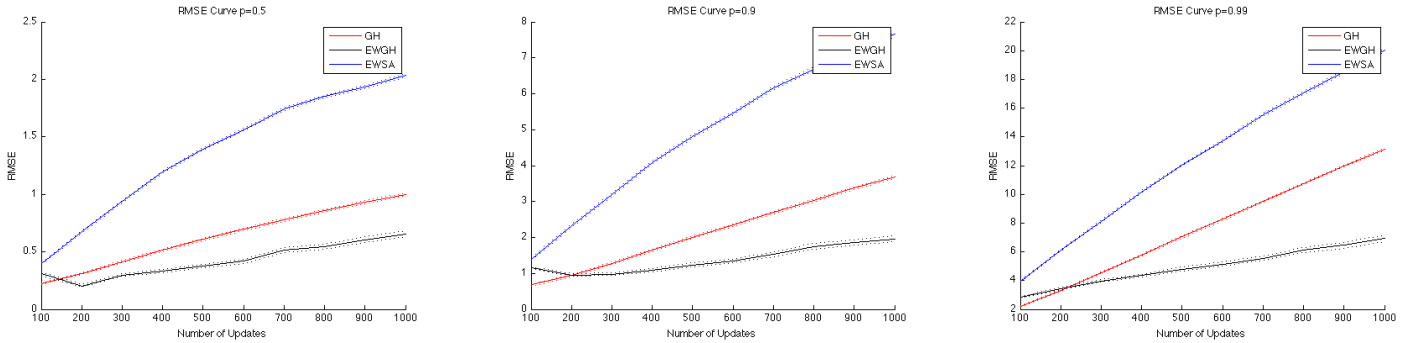
(g) m=100, p=0.99 (h) m=400, p=0.99 (i) m=1000, p=0.99

Figure 4.7: Change Point, Chi-Squared to Exponential: EWGH RMSE for $\lambda = 0.01, 0.05, 0.1$ at $m = 100, 400, 1000$ observations for the $p = 0.5, 0.9, 0.99$ quantiles (including 95% percentile bootstrap confidence intervals). The GH and EWGH algorithms utilise $N = 6$.



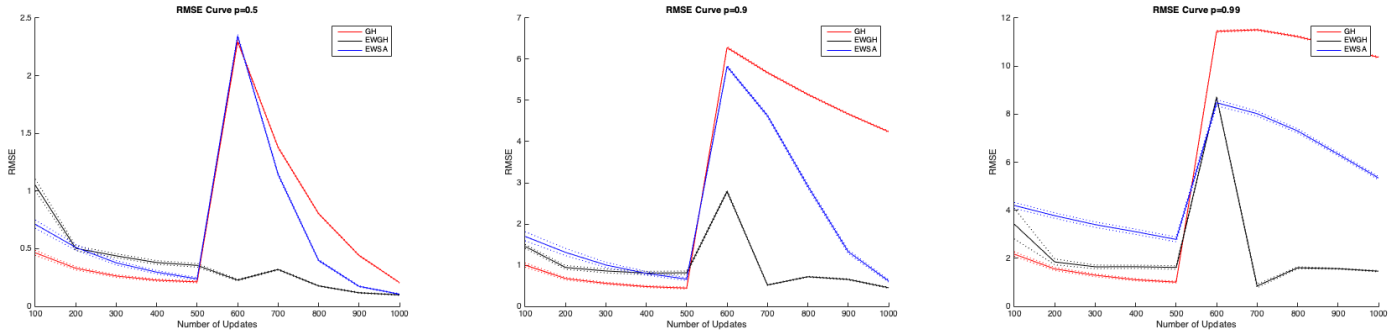
(a) p=0.5 (b) p=0.9 (c) p=0.99

Figure 4.8: RMSE curves associated with the normal distribution with mean $= 0.006j$ and standard deviation $= 1$ for the 0.5, 0.9 and 0.99 quantiles. The final exact quantiles at $j = m = 1000$ are 6, 7.2816 and 8.3263 respectively. The GH and EWGH algorithms utilise $N = 6$. The EWGH algorithm utilises $\lambda = 0.01$.



(a) p=0.5 (b) p=0.9 (c) p=0.99

Figure 4.9: RMSE curves associated with the exponential distribution with mean and standard deviation $= 1 + 0.006j$ for the 0.5, 0.9 and 0.99 quantiles. The final exact quantiles at $j = m = 1000$ are 4.8562, 16.1319 and 32.2638 respectively. The GH and EWGH algorithms utilise $N = 6$. The EWGH algorithm utilises $\lambda = 0.01$.



(a) p=0.5 (b) p=0.9 (c) p=0.99

Figure 4.10: RMSE curves associated with simulated data with a change point, switching from chi-squared with five degrees of freedom to exponential distribution with mean and standard deviation equal to one for the 0.5, 0.9 and 0.99 quantiles. The final exact quantiles at $j = m = 1000$ are the quantiles of the standard exponential distribution, namely 0.6931, 2.3026 and 4.6052 respectively. The GH and EWGH algorithms utilise $N = 6$. The EWGH algorithm utilises $\lambda = 0.01$.

4.6 Real Data Results

In this section we test the GH and EWGH algorithms on a real data set to evaluate their effectiveness in a non-idealised setting. In particular we consider one month (January 2013) of high frequency forex return data, namely EURUSD spot mid-price returns intervalled at 15 seconds. We begin with the mid-price series:

$$p_{t_{(1)}}, p_{t_{(2)}}, \dots p_{t_{(m)}}, \quad t_{(i+1)} - t_{(i)} \approx 15 \text{ seconds},$$

which we transform online to obtain arithmetic returns:

$$r_{(1)}, r_{(2)}, \dots r_{(m-1)}, \quad r_{(j)} = p_{t_{(j+1)}} - p_{t_{(j)}}.$$

The summary statistics of the arithmetic returns (in pips, 1 pip = 0.0001) are as follows (data set size = 71797):

Statistic	Value (Pips)
Mean	1.0449e-05
Standard Deviation	1.0852
Skewness	-0.0222
Kurtosis	16.5736

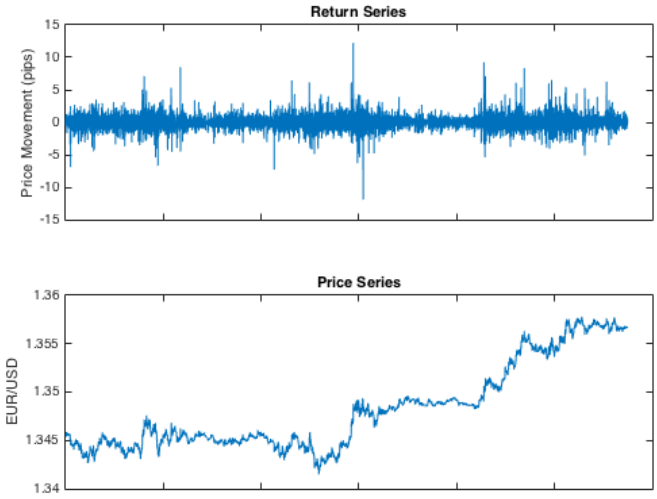


Figure 4.11: An extract of the forex returns and price series for 2013-01-28 to 2013-01-31. Noteworthy features include the prevalence of return outliers and distinctive changes in return variance. Also apparent are periods of pronounced, but temporary trending behaviour corresponding to changes in the mean of the return distribution.

Accurately tracking the quantiles of this return series is a non-trivial check of our methods - the distribution of these returns is non-stationary and the frequency of outliers is high. The frequency of outliers is related to the fact that the distribution of the returns is heavy-tailed as evidenced by the high kurtosis. This will probe the dynamic quantile estimation performance in the setting of general non-stationarity as well as evaluating the robustness of the algorithms. Applications of online quantile estimation for financial price series include identifying high frequency trading opportunities, real-time risk estimation (such as the calculation of real time Value at Risk, VaR) and outlier detection.

Our test utilises the performance measure we introduced in the non-identically distributed simulation studies conducted previously. To reiter-

ate we count the number of times the $(i+1)$th observation is smaller than the online estimates of the 0.5 (median), 0.9 and 0.99 quantiles obtained up to observation i. These counts are then normalised by the total number of observations. We compare the observed frequencies to the probabilities $0.5, 0.9, 0.99$ respectively and report 95% percentile bootstrap confidence intervals for the observed frequencies. The bootstrap confidence intervals are created by calculating the observed frequencies for each day in the period in question (January 2013), resampling the resultant daily frequencies with replacement (1000 resamples) and providing the 0.025 and 0.975 quantiles of the resampled distribution. We also include the results for the EWSA algorithm for comparison. The parameters used for the algorithms are as follows: $N = 6$ for the GH and EWGH algorithms. This choice of N was motivated by the simulation results and computational efficiency considerations. For the EWGH algorithm, $\lambda = 0.05$. This choice was motivated by an analysis of an initial sample (not contained within the January 2013 data set) and by the fact that we expect forex return quantiles to vary rapidly around news events and thus a choice of λ bigger than 0.01 (the best choice in the simulation studies) appeared appropriate. The parameters for the EWSA algorithm are the same as in section 4.5.

Algorithm	p=0.5	p=0.9	p=0.99
GH	0.469 (0.463,0.475)	0.860 (0.849,0.873)	0.974 (0.971,0.978)
EWGH	0.500 (0.499,0.502)	0.894 (0.892,0.895)	0.992 (0.991,0.992)
EWSA	0.501 (0.500,0.502)	0.897 (0.895,0.899)	0.976 (0.973,0.979)

The EWGH algorithm performs well overall and is the only algorithm that does not underestimate the p=0.99 quantile. This provides evidence that not only does the EWGH algorithm perform well in a non-stationary environment, but also that it handles heavy-tailed distributions well. Note that we have performed the same test on other months to check the consistency of the results and the same behaviour emerges (omitted for brevity).

We re-iterate that for the GH and EWGH algorithms, we can obtain online estimates of any quantile of interest.

Chapter 5

Properties of Hermite Series based Distribution Function Estimators

In this chapter we study the properties of the general Hermite series based distribution function estimator defined on the full real line. The chapter is organised as follows, in section 5.1 we restate the Hermite series based distribution function estimator in terms of the standard Hermite series coefficients as opposed to the Gauss-Hermite form. This is motivated by the fact that using the standard form of the Hermite series estimator makes the derivations that follow simpler. In section 5.2 we establish MSE consistency for the Hermite series distribution function estimator defined on the full real line (with rate) and demonstrate a specific case where the rate approaches the optimal $O(n^{-1})$, namely the normal distribution. In section 5.3 we establish MISE consistency (with rate) of the Hermite series distribution function estimator defined on the full real line. In section 5.4 we demonstrate the almost sure convergence of the Hermite series distribution estimator for densities with support on the full line, including the case where the number of terms in the series is random. In section 5.5 we prove that the Hermite series distri-

bution function estimate is B-robust for finite N. A real data example and simulation study are presented in sections 5.6 and 5.7 respectively. Finally we collect technical details of relevant known results and novel proofs in the appendices B.1 and B.2.

5.1 Hermite Distribution Function Estimators

In chapter 4 we applied distribution function estimators based on Hermite series density estimators. In that chapter we utilised the Gauss-Hermite form of the density estimator to derive distribution function estimators. Restated in terms of the standard Hermite series coefficients ((2.17) associated with (2.18)), the following distribution function estimator was obtained for densities with support on the positive half real line,

$$\begin{aligned}\hat{F}_N(x) &= \int_0^x \hat{f}_N(t)dt \\ &= \sum_{k=0}^{N} \hat{a}_k \sqrt{k!} \sum_{l=0}^{\lfloor k/2 \rfloor} \frac{(-1)^l 2^{k-3l-\frac{1}{2}} \left[\gamma(-l+\frac{k}{2}+\frac{1}{2}, \frac{x^2}{2})\right]}{l!(k-2l)!(\pi)^{\frac{1}{4}}}, \; x \geq 0, \end{aligned} \tag{5.1}$$

where the integration has been performed analytically. The MSE and MISE consistency have been demonstrated and rates provided for the estimator above in chapter 4. These results were obtained by utilising inequalities which apply specifically in the cases that the probability density function, $f(x)$, has bounded or non-negative support. Namely:

$$E|\hat{F}_N(x) - F(x)|^2 \leq (x-a)E\int (\hat{f}_N(t) - f(t))^2 dt, \tag{5.2}$$

for the CDF estimator, $\hat{F}_N(x) = \int_a^x \hat{f}_N(t)dt$ applicable where $f(x)$ has support on $[a, b]$ (bounded support case),

$$E|\hat{F}_N(x) - F(x)|^2 \leq xE \int (\hat{f}_N(t) - f(t))^2 dt, \tag{5.3}$$

for the CDF estimator, $\hat{F}_N(x) = \int_0^x \hat{f}_N(t)dt$ applicable where $f(x)$ has support on $[0, \infty)$ (non-negative support case).

The above inequalities are derived by utilising the Cauchy-Schwarz inequality. Attempting to apply the same approach in the full real line support case fails however. We utilise a different approach to obtain MSE and MISE consistency results for the full real line support case in this chapter.

Now, for the full real line:

$$\hat{F}_N(x) = \int_{-\infty}^x \hat{f}_N(t)dt. \tag{5.4}$$

The integration can again be performed analytically and the following expression results (where the expression obtained in chapter 4 has been restated in terms of standard Hermite series coefficients),

$$\hat{F}_N(x) = \begin{cases} \sum_{k=0}^N \hat{a}_k \sqrt{k!} \sum_{l=0}^{\lfloor k/2 \rfloor} \frac{(-1)^l 2^{k-3l-\frac{1}{2}} \left[\left[(-1)^{k-2l} \right] \Gamma(-l+\frac{k}{2}+\frac{1}{2}) + \gamma(-l+\frac{k}{2}+\frac{1}{2}, \frac{x^2}{2}) \right]}{l!(k-2l)!(\pi)^{\frac{1}{4}}} \\ \quad \text{if } x \geq 0, \\ \sum_{k=0}^N \hat{a}_k \sqrt{k!} \sum_{l=0}^{\lfloor k/2 \rfloor} \frac{(-1)^{-l+k} 2^{k-3l-\frac{1}{2}} \Gamma(-l+\frac{k}{2}+\frac{1}{2}, \frac{x^2}{2})}{l!(k-2l)!(\pi)^{\frac{1}{4}}} \\ \quad \text{if } x < 0, \end{cases} \tag{5.5}$$

where $\Gamma(a, x) = \int_x^\infty t^{a-1} e^{-t} dt$ is the upper incomplete Gamma function, $\gamma(a, x) = \int_0^x t^{a-1} e^{-t} dt$ is the lower incomplete Gamma function and $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$ is the usual Gamma function. Note that the aforementioned suitability of using the estimator (5.5) in sequential (online) estimation follows from the fact that, in the fixed N case, the coefficients, $\hat{a}_k$, $k = 0, \ldots N$,

can be updated in a sequential manner. Concretely, the following update rule can be applied (see the previous chapter for more details, extensions and applications):

$$\hat{a}_k^{(i)} = \frac{1}{i}\left[(i-1)\hat{a}_k^{(i-1)} + h_k(x_i)\right], \; k = 0, \ldots, N. \tag{5.6}$$

5.2 Mean Squared Error

In this section we present bias, variance and MSE consistency results of the full real-line estimator (5.5).

We begin by deriving a bound on the bias under certain regularity conditions on the probability density function. Under these conditions, the bound is a decreasing function of N for sufficiently large N and moreover the estimator is asymptotically unbiased.

Proposition 3. *Suppose $f(x) \in L_2$ is r times continuously differentiable and $(x - \frac{d}{dx})^r f(x) \in L_2$ where $r > 2$. As $N \to \infty$, the squared bias of the estimator (5.5) is:*

$$\left|E[\hat{F}_N(x)] - F(x)\right|^2 = O\left(N^{-r+2}\right),$$

uniformly in x and thus the estimator (5.5) is asymptotically unbiased:

$$\left|E[\hat{F}_N(x)] - F(x)\right|^2 \to 0.$$

Proof. See B.2.1. □

Next, we derive a bound on the variance of the estimator (5.5) under certain absolute moment conditions. This bound is an increasing function of N and decreasing function of n. A suitable choice of $N(n)$ ensures that the variance approaches zero as $n \to \infty$.

Proposition 4. *Suppose* $E|X|^{2/3} < \infty$. *As* $N = N(n), n \to \infty$, *the variance of the estimator (5.5) is:*

$$E\left|\hat{F}_N(x) - E[\hat{F}_N(x)]\right|^2 = O\left(\frac{N^{\frac{5}{2}}}{n}\right),$$

uniformly in x. *If in addition we have* $\frac{N^{\frac{5}{2}}(n)}{n} \to 0$ *then:*

$$E\left|\hat{F}_N(x) - E[\hat{F}_N(x)]\right|^2 \to 0.$$

Proof. See B.2.2. □

Finally, we present the main result of this section pertaining to the MSE of the estimator (5.5).

Theorem 8. *Suppose* $f(x) \in L_2$ *is* r *times continuously differentiable and* $(x - \frac{d}{dx})^r f(x) \in L_2$ *where* $r > 2$. *In addition, suppose* $E|X|^{2/3} < \infty$ *and that* $N(n) \sim n^{2/(2r+1)}$ *as* $n \to \infty$, *then we have:*

$$E\left|\hat{F}_N(x) - F(x)\right|^2 = O\left(n^{-2(r-2)/(2r+1)}\right) \tag{5.7}$$

$$= o(1) \text{ uniformly in } x. \tag{5.8}$$

Proof. The theorem above follows from propositions 3 and 4 and the fact that the mean squared error is the sum of the squared bias and the variance of the estimator $\hat{F}_N(x)$. Also note that the choice $N(n) \sim n^{2/(2r+1)}$ is motivated by the fact that the squared bias term is asymptotically monotonically decreasing in N and the variance term is asymptotically monotonically increasing in N, thus $N(n)$ is selected to asymptotically minimise the MSE by making the rates of the squared bias and the variance of the estimator $\hat{F}_N(x)$ equal. □

Note that the rate provided and the mean squared consistency is uniformly in x. It is noteworthy that the rate is worse than the Hermite CDF

estimator defined on the positive half real line in chapter 4 which had a mean squared convergence rate of $xO(n^{-2r/(2r+1)})$. The rate obtained in theorem 8 is uniform however. Furthermore, the rate is suboptimal compared to the $O(n^{-1})$ of smooth kernel distribution estimators (making a relative deficiency analysis redundant). The rate does however approach optimal for $r >> 1$ where $r > 2$ derivatives of $f(x)$ exist and $(x - \frac{d}{dx})^r f(x) \in L_2$. How realistic are these conditions? One ubiquitous distribution for which this condition is satisfied is the normal distribution.

5.2.1 An Example

Normal Distribution

We have for the standard normal distribution:

$$\int_{-\infty}^{\infty} \left[(x - \frac{d}{dx})^r \frac{e^{-x^2/2}}{\sqrt{2\pi}} \right]^2 dx \tag{5.9}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[e^{-x^2/2} e^{x^2/2} (x - \frac{d}{dx})^r e^{-x^2/2} \right]^2 dx \tag{5.10}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[e^{-x^2/2} H_r(x) \right]^2 dx \tag{5.11}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x^2} H_r^2(x) dx \tag{5.12}$$

$$= \frac{1}{2\sqrt{\pi}} 2^r r! \tag{5.13}$$

$$< \infty \tag{5.14}$$

where we have used the fact that $e^{x^2/2}(x - \frac{d}{dx})^r e^{-x^2/2} = H_r(x)$ i.e. the r-th Hermite polynomial [5], and the orthogonality relation for Hermite polynomials. Now since r derivatives of $f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$ exist and $(x - \frac{d}{dx})^r f(x) \in L_2$ for all finite $r \geq 1$, we have that the rate derived above approaches $O(n^{-1})$. Since the leading term in the Hermite series PDF estimator is proportional

to the standard normal distribution, this behaviour makes sense.

Remark 4. In practice, sequential Hermite CDF estimators, which are applicable to streaming data and the one-pass analysis of massive data sets, have a fixed and finite N. In these settings, N is fixed and independent of the number of observations, n. Moreover $n \to \infty$. Thus while theorem 8 above - with $N(n)$ as a function of n - gives us comfort that the asymptotic properties of the Hermite series CDF estimators are sensible, it is not directly applicable to the fixed N scenario presented in chapter 4. We can however still obtain some insight into the performance of the Hermite series based CDF estimators in these practical scenarios. The equation (B.16) implies that for N fixed and sufficiently large and $r >> 1$, the Hermite CDF estimator is approximately unbiased since the bias is $O(N^{-r/2+1})$. With regards to the variance of the Hermite CDF estimator, for N fixed and sufficiently large (B.20) states that the variance is $O\left(\frac{N^{\frac{5}{2}}}{n}\right)$ and thus approaches zero when $n \to \infty$. Therefore, for N fixed and sufficiently large, if we have $n \to \infty$ and $r >> 1$, the MSE is approximately zero. The same holds true for the MISE treated below.

5.3 Mean Integrated Squared Error

In this section we consider the mean integrated squared error, MISE:

$$\text{MISE}(\hat{F}_N) = \int_{-\infty}^{\infty} E\left|\hat{F}_N(x) - F(x)\right|^2 f(x)dx. \tag{5.15}$$

This definition of the MISE (up to an additional non-negative weight function, $w(x)$) is commonly applied in bandwidth selection for smooth kernel distribution function estimators (e.g.[3] amongst others). This definition of the MISE also corresponds to the integrated weighted MSE criterion we utilised in chapter 4, differing only in the domain of integration. In the theorem below, we prove convergence of the Hermite series distribution function

estimator in MISE and provide the associated rate under the same regularity and absolute moment conditions previously specified in theorem 8.

Theorem 9. *Suppose $f(x) \in L_2$ is r times continuously differentiable and $(x - \frac{d}{dx})^r f(x) \in L_2$ where $r > 2$. In addition, suppose $E|X|^{2/3} < \infty$ and that $N(n) \sim n^{2/(2r+1)}$ as $n \to \infty$, then we have:*

$$\int_{-\infty}^{\infty} E\left|\hat{F}_N(x) - F(x)\right|^2 f(x)dx = O\left(n^{-2(r-2)/(2r+1)}\right) \tag{5.16}$$

$$= o(1). \tag{5.17}$$

Proof. The result follows directly from Theorem 8 which implies that the MSE convergence is uniform in x. □

5.4 Almost Sure Convergence

In this section we study the almost sure convergence of the Hermite series distribution function estimator, firstly for deterministic $N(n)$ and then for random $N(n)$.

Theorem 10. *Suppose $f(x) \in L_2$ is r times continuously differentiable and $(x - \frac{d}{dx})^r f(x) \in L_2$ where $r > 2$. In addition suppose $E|X|^s < \infty$, $s > 8(r+1)/3(2r+1)$ and that $N(n) \sim n^{2/(2r+1)}$ as $n \to \infty$, then we have:*

$$\left|\hat{F}_N(x) - F(x)\right| = O(n^{-(r-2)/(2r+1)} \log n)\, a.s. \tag{5.18}$$

$$\to 0\, a.s.\ \textit{uniformly in } x. \tag{5.19}$$

Proof. See B.2.3. □

As noted in [29], $s > 8(r+1)/3(2r+1)$ is satisfied independently of r for $s > \frac{16}{9}$.

In the next theorem, we prove almost sure convergence for Hermite series distribution function estimators with random length i.e. $N(n)$ is a random variable with values in $\mathbb{N}$. Here $N(n)$ could be a measurable function of $\mathbf{x_i}, i = 1, \ldots, n$ and thus theorem 11 below could apply when data-driven estimators for $N(n)$ are used (under certain conditions).

Theorem 11. *Suppose $f(x) \in L_2$ is r times continuously differentiable and $(x - \frac{d}{dx})^r f(x) \in L_2$ where $r > 2$. In addition suppose that $N(n) \to \infty$ a.s. as $n \to \infty$, $\sum_{n=1}^{\infty} P\left(\frac{N(n)}{n^\gamma} > \epsilon\right) < \infty$, for all $\epsilon > 0$ with $0 < \gamma < 6/17$, then we have:*

$$\left|\hat{F}_{N(n)}(x) - F(x)\right| \to 0 \quad a.s. \text{ uniformly in } x. \tag{5.20}$$

Proof. See B.2.4. □

5.5 Robustness

An important property of an estimator is its performance in the presence of outlying observations corresponding to contamination of the data, known as its robustness. More precisely, robustness of an estimator at distribution F, refers to its performance in the presence of contamination by observations from a distribution different from F.

The influence function [34, 35] measures the standardised effect of an infinitesimal contamination at a particular point, x', on the estimator. For a function estimator, $T(x, F)$, evaluated at x and distribution F, the influence function is defined as:

$$IF(x, x'; T, F) = lim_{\epsilon \to 0} \frac{T(x, (1-\epsilon)F + \epsilon\delta_{x'}) - T(x, F)}{\epsilon}, \tag{5.21}$$

where the point mass $\delta_{x'}$ is the distribution such that $P(t = x') = 1$. The most important robustness measure derived from the influence function is the

supremum of its absolute value, $sup_{x'}|IF(x, x'; T, F)|$, known as the gross-error sensitivity [35]. If the gross-error sensitivity is finite, this implies that the maximal (approximate) influence that a small, fixed proportion of identical outliers at x' can have on the value of the estimator is bounded. The gross-error sensitivity can be regarded as an upper bound on the standardised asymptotic bias of the estimator introduced by contamination by outliers. Thus if the gross-error sensitivity is finite, the estimator is said to be Bias-robust or B-robust. The empirical influence function is defined as:

$$IF(x, x'; T, \hat{F}_n) = lim_{\epsilon \to 0} \frac{T(x, (1-\epsilon)\hat{F}_n + \epsilon\delta_{x'}) - T(x, \hat{F}_n)}{\epsilon}, \tag{5.22}$$

where $\hat{F}_n$ is the empirical distribution function. For linear functionals,

$$\lim_{n \to \infty} IF(x, x'; T, \hat{F}_n) \to IF(x, x'; T, F) \, a.s.$$

(essentially due to the strong law of large numbers). This is not true in general however. In this section we present B-robustness results for the finite N Hermite series based cumulative distribution function estimator along with the smooth kernel distribution function estimator and an estimator based on integrating the Gram-Charlier type A series density estimator.

Proposition 5. *The fixed N Hermite series cumulative distribution function estimator (5.5) is Bias-robust.*

Proof. See B.2.5. □

Proposition 6. *The smooth kernel distribution function estimator (1.2) is Bias-robust.*

Proof. See B.2.6. □

Proposition 7. *The cumulative distribution function estimator (B.42) based on the integrating the Gram-Charlier type A series density estimator is not Bias-robust.*

Proof. See B.2.7. □

It is noteworthy that we have only proved that the Hermite series based cumulative distribution function estimator (5.5) is B-robust for the finite N case and not in the case where $N \to \infty$ which is required for the estimator to be consistent. On the other hand, we have proved that the smooth kernel distribution function estimator is B-robust for any h including $h \to 0$ (therefore applicable to the consistent estimator). Thus the B-robustness result we have obtained for the smooth kernel distribution function estimator is more general than that obtained for the Hermite series based estimator. However, a very closely related estimator based on integrating the Gram-Charlier type A series density estimator with a finite number of terms is not B-robust indicating that this property is not necessarily trivial.

5.6 Real Data Example

In this section we present an illustrative example of applying the Hermite series distribution function estimator to a real data set, namely forex return data. In particular, we consider one minute EURUSD spot mid-price returns (successive price movements) for the period of 2018-04-02 to 2018-06-30. Using the mid-price series:

$$p_{t_{(1)}}, p_{t_{(2)}}, \ldots p_{t_{(n)}}, \quad t_{(i+1)} - t_{(i)} = 1 \text{ minute},$$

we obtain the arithmetic returns via differencing:

$$r_{(1)}, r_{(2)}, \ldots r_{(n-1)}, \quad r_{(j)} = p_{t_{(j+1)}} - p_{t_{(j)}}.$$

We convert the arithmetic returns to pips, where 1 pip = 0.0001, for clarity. A one day sample of the mid-price series and associated arithmetic returns are plotted in figures 5.1 and 5.2 respectively.

Figure 5.1: EURUSD mid-price series for 2018-04-02.

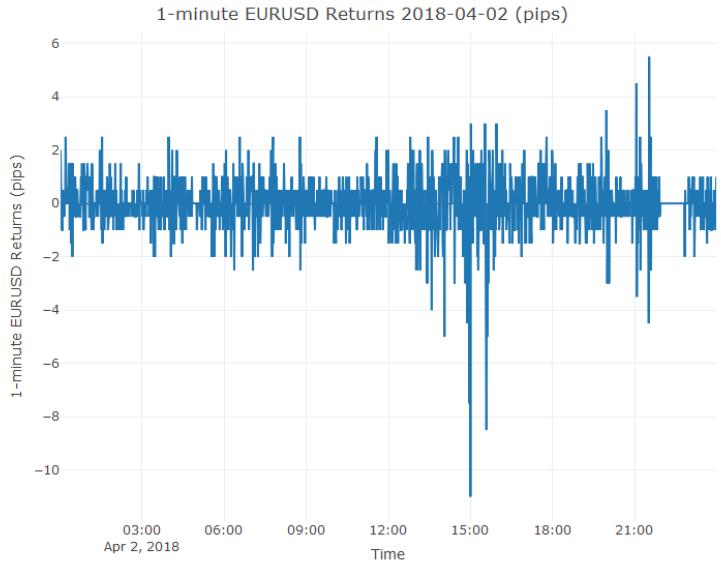


Figure 5.2: 1-minute EURUSD arithmetic returns (in pips) for 2018-04-02.

The arithmetic returns for the period 2018-04-02 to 2018-06-30 have the

following summary statistics in pips:

Summary Statistic	Value
Count	65944
Mean	-0.01
Standard Deviation	1.64
Skewness	-0.09
Kurtosis	28.03

Table 5.1: Summary statistics of 1-minute EURUSD arithmetic returns.

It is noteworthy that the kurtosis of the return data is high indicating heavy tails for the distribution of returns and frequent outlying values. The histogram of returns is presented in figure 5.3.

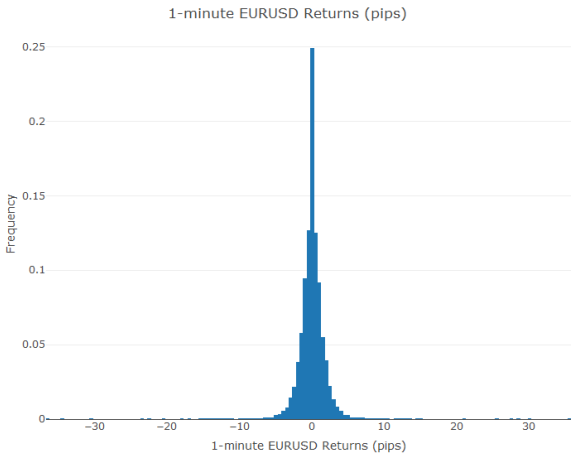


Figure 5.3: Histogram of 1-minute EURUSD arithmetic returns over the period 2018-04-02 to 2018-06-30.

In figure 5.4, we present the result of applying the Hermite series distribution function estimator to the EURUSD returns data set, using different values of N. We also include the empirical distribution function for comparison.

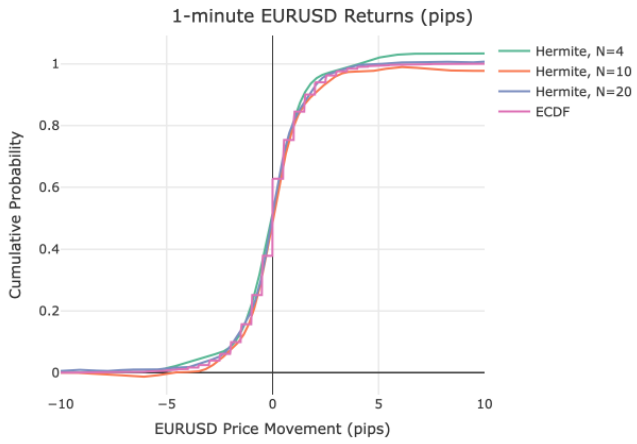


Figure 5.4: Hermite series cumulative distribution function estimator applied to EURUSD 1 minute return data for N=4, 10, 20 compared to empirical distribution function.

The associated mean and maximum absolute differences of the Hermite series distribution function estimates compared to the empirical distribution function estimates (at each observation in the data) are presented in the table below.

Method	Mean Absolute Difference	Max Absolute Difference
Hermite, N=4	0.044	0.139
Hermite, N=10	0.058	0.136
Hermite, N=20	0.045	0.130

It is evident from figure 5.4 that all values of N produce a reasonably good fit to the data and for sufficiently large values of N there is a very good fit to the data. This effectiveness is particularly noteworthy given the heavy-tailed nature of the data. However, certain limitations are also evident from the figure. The Hermite series distribution function estimator may produce distribution function estimates which are not monotonically increasing and may have cumulative probability values that are outside the range of $[0, 1]$.

Indeed, the limiting value of the Hermite CDF estimator as $x \to \infty$ is not guaranteed to be 1 and can be smaller than 1 as we see for $N = 10$, or greater than 1, as seen for $N = 4$. This behaviour is related to the fact that the Hermite series distribution function estimator is obtained by integrating the Hermite series density estimator which does not necessarily produce bona fide density function estimates (i.e. density function estimates that are non-negative and integrate to one). Cumulative probability estimates below 0 and greater than 1 can be interpreted as being close to the true cumulative probability in the range $[0, 1]$ asymptotically in various senses (MSE etc). It is worth noting that unless a particular application specifically requires a bona fide cumulative distribution function estimate, this does not necessarily limit the usefulness of the Hermite CDF estimator in practice. If a bona fide estimate is required, several options to ameliorate invalid estimates are feasible. One option is to discard estimates at particular values of x which lie outside $[0, 1]$, another is to replace estimates below zero with 0 and estimates greater than one with 1. The performance implications of these methods are an area for future research.

5.7 Simulation Study

In this section we compare the performance of the Hermite series based distribution function estimator to the empirical distribution function estimator (1.1) and the smooth kernel distribution function estimator (1.2) with a Gaussian kernel. The aim of this study is to assess the performance of the various estimators in a finite sample setting. While we have demonstrated in section 5.2 that the Hermite series distribution function estimator is, in general, *asymptotically* deficient with respect to the EDF and Kernel estimators in the MSE sense, the relative finite sample performance is worth investigating. In particular, we seek to gain insight into whether the asymptotic deficiency of the Hermite series estimator manifests in samples of practical

size. In the present simulation study, we use the MISE as a measure of performance and we compare the estimators for three distributions: the Laplace distribution (location=0, scale =1), the log-normal distribution (meanlog=0, sdlog=1) and finally the exponential distribution (rate=1). These distributions are ubiquitous and have a diverse range of properties. The study is conducted as follows:

For the empirical distribution function estimator, $\hat{F}_n$, the following is repeated for each test distribution (log-normal, Laplace and exponential):

1. Draw $n = 20, 50, 100, 500, 1000$ i.i.d. observations, x_i, $i = 1, \ldots, n$ from the test distribution.

2. Plug the observations x_i into the EDF estimator and evaluate $\int_{-\infty}^{\infty}(\hat{F}_n(x) - F(x))^2 f(x)dx$ numerically, where $F(x)$ is the exact CDF of the test distribution and $f(x)$ is the exact PDF of the same distribution.

3. Repeat $m = 10000$ times and estimate $E\int_{-\infty}^{\infty}((\hat{F}_n(x) - F(x))^2)f(x)dx$ using $\widehat{\text{MISE}}(\hat{F}_n(x)) = \frac{1}{m}\sum_{j=1}^{m}\int_{-\infty}^{\infty}(\hat{F}_n^{(j)}(x) - F(x))^2 f(x)dx$, where j indexes a particular set of n observations used to construct the estimators.

Similarly for the kernel distribution function estimator, $\hat{K}_h$ and Hermite series distribution function estimator, $\hat{F}_N(x)$. For the kernel distribution function estimator we repeat for each value of $h = i/1000$ for $i = 20 \ldots 250$ and identify the value of h that minimises the estimated MISE. For the Hermite series distribution function estimator we repeat for each each value of $N = 2, 3, \ldots, 20$ and identify the value of N that minimises the estimated MISE. The results are summarised in table 5.2.

Note that the above procedure estimates the performance of the estimators with optimal choices for the N and h parameters. We have not utilised data-driven estimators for N, h as we wish to study the performance of the distribution function estimators in isolation. We observe that for samples of size up to 1000 observations, the Hermite series estimator with optimal N is

generally superior to the EDF estimator but is inferior to the kernel distribution function estimator with optimal h in all cases studied. One advantage of the Hermite series estimator is that it yields a much more compact representation of the CDF, requiring only $N + 1$ coefficients versus requiring the entire data set for the EDF and Kernel distribution function estimators in order to evaluate the cumulative probability at arbitrary x. In order to assess the sensitivity of the Hermite series estimator MISE performance to choice of N, we plot the MISE (and associated standard errors) for each choice of N across the previously studied sample sizes and distributions in figure 5.5. It is clearly apparent that the performance of the Hermite series cumulative distribution function estimator is not critically dependent on the choice of N and is robust to the choice of this parameter, a useful property in practice.

	MISE (x10^-3)			Standard Error (x10^-3)			Optimal Parameter	
	Hermite	Kernel	Empirical	Hermite	Kernel	Empirical	Hermite (N)	Kernel (h)
Log-Normal								
20	6.47	6.37	8.31	0.07	0.07	0.07	6	0.244
50	2.85	2.68	3.31	0.03	0.03	0.03	6	0.202
100	1.46	1.39	1.67	0.01	0.01	0.01	14	0.166
500	0.34	0.30	0.33	0.00	0.00	0.00	18	0.110
1000	0.19	0.15	0.17	0.00	0.00	0.00	18	0.077
Laplace								
20	7.05	6.23	8.24	0.07	0.07	0.07	8	0.240
50	2.91	2.58	3.31	0.03	0.03	0.03	8	0.243
100	1.50	1.35	1.66	0.02	0.01	0.02	12	0.202
500	0.31	0.30	0.33	0.00	0.00	0.00	20	0.108
1000	0.16	0.15	0.16	0.00	0.00	0.00	20	0.080
Exponential								
20	6.25	6.15	8.31	0.06	0.06	0.07	6	0.244
50	2.89	2.67	3.36	0.03	0.03	0.03	14	0.186
100	1.60	1.42	1.67	0.02	0.01	0.01	18	0.148
500	0.50	0.31	0.33	0.00	0.00	0.00	19	0.069
1000	0.35	0.16	0.17	0.00	0.00	0.00	19	0.040

Table 5.2: MISE performance comparison of Hermite series CDF estimator, smooth Kernel CDF estimator with Gaussian kernel and EDF.

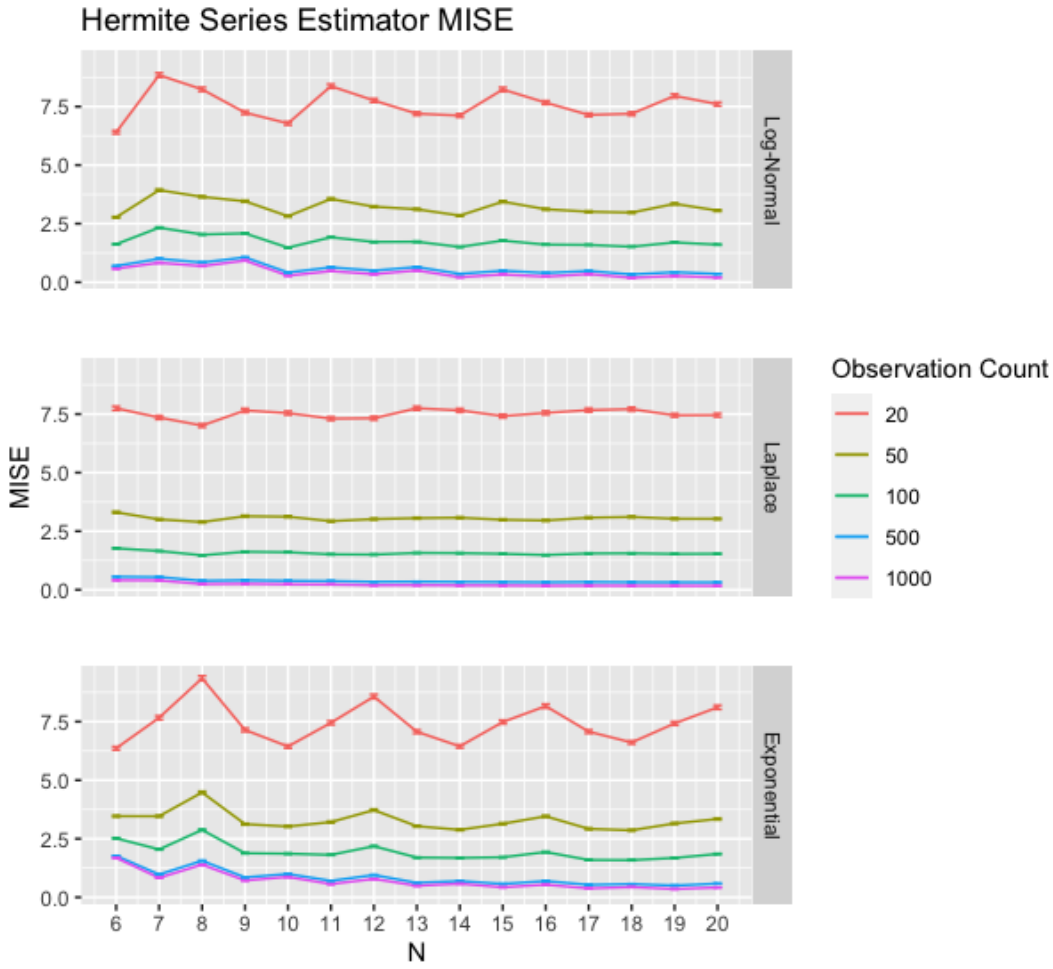


Figure 5.5: Hermite series cumulative distribution function estimator MISE performance at $n = 20, 50, 100, 500, 1000$ observations for N=6,...,20 across log-normal, Laplace and exponential distributions.

5.8 Numerical Check of MISE Theoretical Results

In this section we present a numerical check of Theorem 9. In order to perform the check, we draw observations from a beta distribution with probabil-

ity density function, $f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$, $\alpha, \beta = 6$ where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function which we have defined in terms of the gamma function. The beta distribution probability density function is supported on $x \in [0, 1]$. We have chosen this particular distribution since the relevant quantities in the assumptions of the theorem are readily apparent.

Considering each of the assumptions of Theorem 9 in turn, it is easy to see that the first assumption, $f(x) \in L_2$ is satisfied. It is also straightforward to see that $f(x)$ is $r = 4$ times continuously differentiable, *on the full real line*, and $(x - \frac{d}{dx})^r f(x) \in L_2$ with $r = 4$. Thus the requirement that $r > 2$ is satisfied. Note also that since the mean of the distribution is finite, $E|X|^{2/3} < \infty$ is also satisfied. Finally we set $N(n) = n^{2/9}$ to ensure $N(n) \sim n^{2/(2r+1)}$ as $n \to \infty$. In our check we consider $N(n) = 18, 20, \ldots, 40$ which yields the following sample sizes, n (using $n = N(n)^{9/2}$):

N(n)	n
18	445,375
20	715,542
22	1,098,758
24	1,625,364
26	2,330,130
28	3,252,454
30	4,436,553
32	5,931,642
34	7,792,111
36	10,077,696
38	12,853,642
40	16,190,862

The following steps are repeated for each value of n in the table above:

1. Draw n i.i.d. observations, x_i, $i = 1, \ldots, n$ from the beta distribution with $\alpha, \beta = 6$.

2. Plug the observations x_i into the Hermite series distribution function estimator and evaluate $\int_{-\infty}^{\infty} (\hat{F}_{N(n)}(x) - F(x))^2 f(x) dx$ numerically, where $F(x)$

is the exact CDF of the Beta(6,6) distribution and $f(x)$ is the exact PDF of the same distribution.

3. Repeat $m = 100$ times and estimate $E\int_{-\infty}^{\infty}((\hat{F}_{N(n)}(x)-F(x))^2)f(x)dx$ using $\widehat{\text{MISE}}(\hat{F}_{N(n)}(x)) = \frac{1}{m}\sum_{j=1}^{m}\int_{-\infty}^{\infty}(\hat{F}^{(j)}_{N(n)}(x) - F(x))^2 f(x)dx$, where j indexes a particular set of n observations used to construct the estimators.

According to Theorem 9, the MISE rate should be bounded as $O\left(n^{-2(r-2)/(2r+1)}\right) = O(n^{-4/9})$ for sufficiently large n (and $N(n)$). This implies that $\log(\text{MISE}) \leq -4/9\log(n)+\log(c)$ for some positive constant c when n (and $N(n)$) are sufficiently large. We plot log(MISE) versus $\log(n)$ to assess this in figure 5.6 along with a continuous piecewise linear approximation as a visual aid.

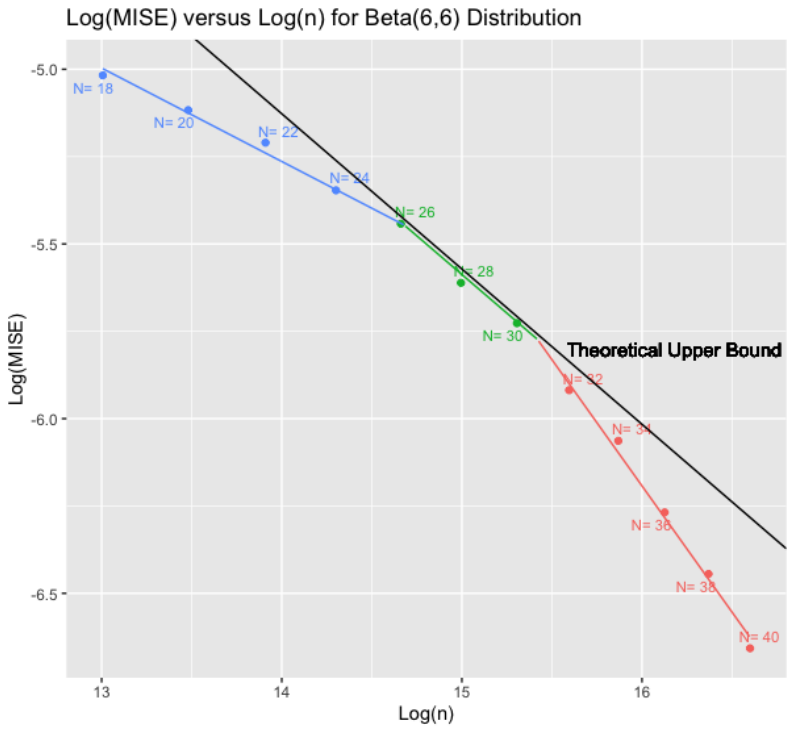


Figure 5.6: Log(MISE) versus Log(n) for a beta distribution with $\alpha, \beta = 6$ along with standard error bars, a continuous piecewise linear approximation and the least upper bound line with the expected theoretical gradient.

Note that a result consistent with Theorem 9 would imply that there should exist a line with gradient of $-4/9$ that bounds all log(MISE) values for sufficiently large values of n and $N(n)$. Comfortingly, our numerical results appear consistent with the bounds that arise from Theorem 9. As can be seen in figure 5.6, the log(MISE) results for $N = 18, \ldots, 24$ (in blue) appear to be in the pre-asymptotic regime and are decreasing more slowly than a line with gradient $-4/9$. For $N \geq 26$, the results suggest that n and $N(n)$ are sufficiently large to be in the asymptotic regime. The log(MISE) results for $N = 26, 28, 30$ (in green) decrease at a rate similar to a line with a gradient $-4/9$ and the log(MISE) results for $N \geq 32$ (in red) decrease more quickly than a line with gradient $-4/9$ which is still consistent with the bound. Obtaining numerical verification in general is challenging. The reasons for this are primarily that the actual MISE need not be a polynomial function of n (or even smoothly depend on n) and the bound only applies for sufficiently large $n, N(n)$ values which are difficult to determine in practice.

Chapter 6

Sequential Estimation of Nonparametric Correlation using Hermite Series Estimators

In this chapter, we introduce sequential estimators for nonparametric correlation based on the Hermite series density estimators we have reviewed in chapter 2 along with the Hermite series based CDF estimators we have studied in chapters 4 and 5. The chapter is organised as follows, in section 6.1 we link the Hermite series based density and CDF estimators to Spearman's rank correlation coefficient estimation. In section 6.2.1 we present an algorithm for calculating the Spearman's rank correlation coefficient applicable to the stationary sequential setting. In section 6.2.2 we present an algorithm suitable for non-stationary data streams, based on an exponentially weighted Spearman's rank correlation estimator. We provide the rate of convergence in mean of the Hermite series based Spearman's rank correlation estimator applicable to the stationary setting for i.i.d data streams in section 6.3. In section 6.4 we investigate the variance (and hence standard error) properties

of the exponentially weighted Spearman's rank estimator. In section 6.5 we present simulation studies which demonstrate the effectiveness of our algorithms in practice both in the stationary and non-stationary settings. In the stationary setting, we demonstrate that our algorithm is competitive with an existing algorithm. In addition, we present an application of the non-stationary Spearman's rank algorithm to real data in the form of streaming forex data in section 6.6. The proof of mean absolute error (MAE) convergence of the Hermite series based estimator is presented in appendix C.1. In appendix C.2 we present a proof of the variance of the exponentially weighted Hermite series based Spearman's estimator. Finally we define an exponentially weighted version of the standard Pearson's product-moment correlation estimator in appendix C.4 which is useful in comparatively assessing the robustness of the Hermite series based estimator in section 6.5.

6.1 Hermite Series Estimators for Spearman's Rank Correlation Estimation

Hermite series estimators for univariate probability density functions are given by (2.18). Associated univariate distribution function estimators were studied in chapters 4 and 5. Bivariate Hermite series probability density function estimators are given by (2.24).

In the context of online estimation of the Spearman's rank correlation coefficient we make use of the expression (3.8), the Hermite series bivariate density function estimator (2.24) and the univariate Hermite series based distribution function estimator (5.4) to define:

$$\hat{R}_{N_1,N_2,N_3,N_4} = 12 \int \int (\hat{F}^{(1)}_{N_1}(x) - 1/2)(\hat{F}^{(2)}_{N_2}(y) - 1/2)\hat{f}_{N_3N_4}(x,y)dxdy. \quad (6.1)$$

As discussed in chapter 3, the grade correlation is that constant which (3.3)

unbiasedly estimates in large samples. Given that the online streaming and massive data set scenarios are large sample situations, the above estimator is natural. Note that in principle N_1, N_2, N_3 and N_4 can all take on distinct values, however for simplicity of the explication of the algorithms below (along with certain computational advantages), we set $N_1 = N_2 = N_3 = N_4 = N$.

$$\hat{R}_N = 12 \int \int (\hat{F}_N^{(1)}(x) - 1/2)(\hat{F}_N^{(2)}(y) - 1/2)\hat{f}_{NN}(x,y)dxdy. \tag{6.2}$$

For computational efficiency it is advantageous to phrase the estimator above in terms of linear algebra operations:

$$\begin{aligned} \hat{R}_N = & 12\hat{\mathbf{a}}_{(\mathbf{1})}^{(n)}\mathbf{W}^T\hat{\mathbf{A}}^{(n)}\mathbf{W}\hat{\mathbf{a}}_{(\mathbf{2})}^{(n)} \\ & - 6\hat{\mathbf{a}}_{(\mathbf{1})}^{(n)}\mathbf{W}^T\hat{\mathbf{A}}^{(n)}\mathbf{z} \\ & - 6\mathbf{z}\hat{\mathbf{A}}^{(n)}\mathbf{W}\hat{\mathbf{a}}_{(\mathbf{2})}^{(n)} \\ & + 3\mathbf{z}\hat{\mathbf{A}}^{(n)}\mathbf{z}, \end{aligned} \tag{6.3}$$

where $\hat{\mathbf{a}}_{(\mathbf{1})}^{(n)}$ are the coefficients $(\hat{\mathbf{a}}_{(\mathbf{1})}^{(n)})_k$ associated with the Hermite series cumulative distribution function estimator $\hat{F}_N^{(1)}(x)$, $\hat{\mathbf{a}}_{(\mathbf{2})}^{(n)}$ are the coefficients $(\hat{\mathbf{a}}_{(\mathbf{2})}^{(n)})_k$ associated with the Hermite series cumulative distribution function estimator $\hat{F}_N^{(2)}(y)$ and $\hat{\mathbf{A}}^{(n)}$ are the coefficients $(\hat{\mathbf{A}}^{(n)})_{kl}$ associated with the Hermite series bivariate density function estimator $\hat{f}_{NN}(x,y)$. The matrix $(\mathbf{W})_{kl} = \int_{-\infty}^{\infty} h_k(u) \int_{-\infty}^{u} h_l(v)dvdu$ and vector $(\mathbf{z})_k = \int_{-\infty}^{\infty} h_k(u)du$. For all matrices and vectors thus defined, $k, l = 0, \ldots, N$. Note that these integrals can be evaluated numerically once and the values stored for rapid calculation.

In the next section we utilise this estimator in order to define algorithms for sequential (online) Spearman's rank correlation estimation in both stationary and non-stationary data settings.

6.2 Sequential Spearman's Rank Correlation Estimation

6.2.1 Sequential Analysis of Stationary Data

In this section, we propose an algorithm for sequential (online) estimation of the Spearman's rank correlation coefficient in the setting of stationary data, namely stationary streaming data and massive data sets.

Algorithm 1

1. For each observation from the data stream $(\mathbf{x}_i, \mathbf{y}_i)$, where $i = 1, \ldots, n$, apply the update rules:

$$\begin{aligned} \hat{\mathbf{a}}_{(\mathbf{1})}^{(1)} &= \mathbf{h}(\mathbf{x}_1) \\ \hat{\mathbf{a}}_{(\mathbf{1})}^{(i)} &= \frac{1}{i}\left[(i-1)\hat{\mathbf{a}}_{(\mathbf{1})}^{(i-1)} + \mathbf{h}(\mathbf{x}_i)\right],\ i = 2, \ldots, n \end{aligned} \tag{6.4}$$

$$\begin{aligned} \hat{\mathbf{a}}_{(\mathbf{2})}^{(1)} &= \mathbf{h}(\mathbf{y}_1) \\ \hat{\mathbf{a}}_{(\mathbf{2})}^{(i)} &= \frac{1}{i}\left[(i-1)\hat{\mathbf{a}}_{(\mathbf{2})}^{(i-1)} + \mathbf{h}(\mathbf{y}_i)\right],\ i = 2, \ldots, n \end{aligned} \tag{6.5}$$

and

$$\begin{aligned} \hat{\mathbf{A}}^{(1)} &= \mathbf{h}(\mathbf{x}_1) \otimes \mathbf{h}(\mathbf{y}_1) \\ \hat{\mathbf{A}}^{(i)} &= \frac{1}{i}\left[(i-1)\hat{\mathbf{A}}^{(i-1)} + \mathbf{h}(\mathbf{x}_i) \otimes \mathbf{h}(\mathbf{y}_i)\right],\ i = 2, \ldots, n. \end{aligned} \tag{6.6}$$

2. Plug the coefficients $\hat{\mathbf{a}}_{(\mathbf{1})}^{(i)}$, $\hat{\mathbf{a}}_{(\mathbf{2})}^{(i)}$ and $\hat{\mathbf{A}}^{(i)}$ into the expression (6.3) to obtain an updated online estimate of the Spearman's rank correlation

coefficient, $\hat{R}_N$.

Here $\mathbf{h}(\mathbf{x}_i)$ and $\mathbf{h}(\mathbf{y}_i)$ are the vectors $h_k(\mathbf{x}_i)$, $k = 0, \ldots, N$ and $h_l(\mathbf{y}_i), l = 0, \ldots, N$ respectively. For computational efficiency it is advantageous to calculate $\mathbf{h}(\mathbf{x}_i)$ and $\mathbf{h}(\mathbf{y}_i)$ making use of the recurrence relation for the Hermite polynomials $H_{k+1}(x) = 2xH_k(x) - 2kH_{k-1}(x)$.

In section 6.3 we prove asymptotic convergence results of the estimator (6.3) in the i.i.d. case which is a stationary scenario.

6.2.2 Sequential Analysis of Non-Stationary Data

In this section we describe an algorithm for tracking the Spearman's rank correlation coefficient for a dynamically varying (non-stationary) data stream. To the best of our knowledge the algorithm we present below is the only online algorithm for the Spearman's rank correlation coefficient applicable to non-stationary data streams that does not rely on maintaining a moving/sliding window of previous observations. Our approach is based on using an exponentially weighted moving average version of the Hermite series coefficients. The parameter λ in the algorithm below controls the weighting of new observations (and controls how rapidly the weights of older observations decrease). This weighting scheme allows the local nonparametric correlation of a bivariate data stream to be tracked.

Algorithm 2

1. For each observation from the data stream $(\mathbf{x}_i, \mathbf{y}_i)$, where $i = 1, \ldots, n$, apply the update rules:

$$\begin{aligned} \hat{\mathbf{a}}_{(\mathbf{1})}^{(1)} &= \mathbf{h}(\mathbf{x}_1) \\ \hat{\mathbf{a}}_{(\mathbf{1})}^{(i)} &= (1-\lambda)\hat{\mathbf{a}}_{(\mathbf{1})}^{(i-1)} + \lambda \mathbf{h}(\mathbf{x}_i),\ i = 2, \ldots, n \end{aligned} \tag{6.7}$$

$$\hat{\mathbf{a}}_{(\mathbf{2})}^{(1)} = \mathbf{h}(\mathbf{y}_1)$$
$$\hat{\mathbf{a}}_{(\mathbf{2})}^{(i)} = (1-\lambda)\hat{\mathbf{a}}_{(\mathbf{2})}^{(i-1)} + \lambda \mathbf{h}(\mathbf{y}_i),\ i = 2, \ldots, n \tag{6.8}$$

and

$$\hat{\mathbf{A}}^{(1)} = \mathbf{h}(\mathbf{x}_1) \otimes \mathbf{h}(\mathbf{y}_1)$$
$$\hat{\mathbf{A}}^{(i)} = (1-\lambda)\hat{\mathbf{A}}^{(i-1)} + \lambda \mathbf{h}(\mathbf{x}_i) \otimes \mathbf{h}(\mathbf{y}_i),\ i = 2, \ldots, n. \tag{6.9}$$

2. Plug the coefficients $\hat{\mathbf{a}}_{(\mathbf{1})}^{(i)}$, $\hat{\mathbf{a}}_{(\mathbf{2})}^{(i)}$ and $\hat{\mathbf{A}}^{(i)}$ into the expression (6.3) to obtain an updated online estimate of the Spearman's rank correlation coefficient, $\hat{R}_N$.

In section 6.4 we elucidate the relationship between λ and the variance of the exponentially weighted version of the Hermite series based Spearman's rank correlation estimator in the i.i.d. scenario.

6.3 Mean Absolute Error of Hermite based Estimator

In this section, we present a theorem concerning the asymptotic convergence of the estimator (6.1) to the grade correlation in an i.i.d. setting as the number of observations $n \to \infty$. In particular, we prove convergence in mean and provide the rate. While these results do not directly apply to the fixed and finite N_1, N_2, N_3, N_4 scenario, they do give assurances that the asymptotic properties of the estimator (6.1) are sensible.

Theorem 12. *For a sample of n i.i.d. bivariate observations $(\mathbf{x_i}, \mathbf{y_i}) \sim f(x, y)$, suppose the following assumptions are satisfied:*

1. $f^{(1)}(x) \in L_2$, $r_1 \geq 8$ *derivatives of* $f^{(1)}(x)$ *exist,* $(x - \frac{d}{dx})^{r_1} f^{(1)}(x) \in L_2$ *and* $E|X|^{4/3} < \infty$, *where* $f^{(1)}(x)$ *is the probability density function associated with the marginal cumulative distribution* $F^{(1)}(x)$.

2. $f^{(2)}(y) \in L_2$, $r_2 \geq 8$ *derivatives of* $f^{(2)}(y)$ *exist,* $(y - \frac{d}{dy})^{r_2} f^{(2)}(y) \in L_2$ *and* $E|Y|^{4/3} < \infty$, *where* $f^{(2)}(y)$ *is the probability density function associated with the marginal cumulative distribution* $F^{(2)}(y)$.

3. $f(x,y) \in L_2$ *and* $(x - \partial_x)^{r_3}(y - \partial_y)^{r_3} f(x,y) \in L_2$, $r_3 \geq 8$.

4. $N_1(n) = O(n^{2/(2r_1+1)})$, $N_2(n) = O(n^{2/(2r_2+1)})$ *and* $N_3(n) = N_4(n) = O(n^{1/(2r_3+1)})$ *as* $n \to \infty$.

For $n \to \infty$*:*

$$
\begin{aligned}
&E|\hat{R}_{N_1,N_2,N_3,N_4} - \rho(F^{(1)}(X), F^{(2)}(Y))| \\
&= O(n^{34/12(2r_1+1)+34/12(2r_2+1)-(r_3-2)/(2r_3+1)}) \\
&+ O(n^{34/12(2r_1+1)-(r_2-2)/(2r_2+1)}) \\
&+ O(n^{-(r_1-2)/(2r_1+1)}) \\
&= o(1) \text{ uniformly in } x,y.
\end{aligned}
$$

See appendix C.1 for the proof of theorem 12.

6.4 Variance of Exponentially Weighted Hermite based Estimator

The estimator introduced in section 6.2.2 based on exponentially weighted versions of the Hermite series coefficients should be applicable in the general non-stationary scenario. In the general case, we cannot necessarily obtain direct theoretical insights into the bias and variance of the Hermite series based Spearman's rank correlation estimator (e.g. useful bounds etc.).

We can however get insights into the variance of the estimator introduced in 6.2.2 in the i.i.d. scenario. In particular we can establish the relationship between the variance of $\hat{R}_N$ and λ for fixed N as $n \to \infty$.

Theorem 13. *For fixed N, fixed $0 < \lambda < 1$ and $n \to \infty$ we have:*

$$\begin{aligned}
Var(\hat{R}_N) = & \left[\frac{\lambda}{2-\lambda}\right] \times \\
& \times [\sum_{r=0}^{N} (g_r^{(1)})^2 \, Var(h_r(x)) + \sum_{s=0}^{N} (g_s^{(2)})^2 \, Var(h_s(y)) \\
& + \sum_{u,v=0}^{N} (g_{uv}^{(3)})^2 \, Var(h_u(x) h_v(y)) \\
& + \sum_{r,s=0}^{N} g_r^{(1)} g_s^{(2)} \, Cov(h_r(x), h_s(y)) \\
& + \sum_{r,u,v=0}^{N} g_r^{(1)} g_{uv}^{(3)} \, Cov(h_r(x), h_u(x) h_v(y)) \\
& + \sum_{s,u,v=0}^{N} g_s^{(2)} g_{uv}^{(3)} \, Cov(h_s(y), h_u(x) h_v(y))] \\
& + o(\lambda),
\end{aligned}$$

where

$$g_r^{(1)} = \frac{\partial \hat{R}_N}{\partial (\hat{a}_{(1)})_r}|_{a_{(1)},a_{(2)},A} = 12 \sum_{l,m,o} W_{lr} A_{lm} W_{mo} (a_{(2)})_o - 6 \sum_{l,m} W_{lr} A_{lm} z_m$$

$$g_s^{(2)} = \frac{\partial \hat{R}_N}{\partial (\hat{a}_{(2)})_s}|_{a_{(1)},a_{(2)},A} = 12 \sum_{k,l,m} (a_{(1)})_k W_{lk} A_{lm} W_{ms} - 6 \sum_{k,m} z_k A_{km} W_{ms}$$

$$g_{uv}^{(3)} = \frac{\partial \hat{R}_N}{\partial (\hat{A}_{uv})}|_{a_{(1)},a_{(2)},A} = 12 \sum_{k,o} (a_{(1)})_k W_{uk} W_{vo} (a_{(2)})_o - 6 \sum_{k} (a_{(1)})_k W_{uk} z_v$$
$$- 6 \sum_{o} z_u W_{vo} (a_{(2)})_o + 3 z_u z_v$$

Thus $Var(\hat{R}_N) = O(\lambda)$ *and the standard error of* $\hat{R}_N$ *is* $O(\lambda^{1/2})$.

See appendix C.2 for a proof of theorem 13.

As is fruitful in settings like bandwidth selection for kernel density estimation, we can obtain specific results for the normal distribution and use these as a rough guide to the behaviour of similar distributions (e.g. Silverman's rule of thumb). In particular, we have for $n \to \infty$:

$$\mathrm{Var}(\hat{R}_N) = \left[\frac{\lambda}{2 - \lambda} \right] g_{N,\rho} + o(\lambda).$$

We tabulate numerically obtained values of $g_{N,\rho}$ for the multivariate normal distribution for various values of N and correlation parameter of the distribution, ρ, in table 6.1 below (standard errors obtained via bootstrap) .

N=4			N=6			N=8			N=10		
ρ	$g_{N,\rho}$	SE	ρ	$g_{N,\rho}$	SE	ρ	$g_{N,\rho}$	SE	ρ	$g_{N,\rho}$	SE
-0.75	0.701	0.007	-0.75	0.474	0.005	-0.75	0.407	0.004	-0.75	0.357	0.004
-0.50	0.891	0.009	-0.50	0.845	0.008	-0.50	0.757	0.008	-0.50	0.747	0.007
-0.25	1.106	0.011	-0.25	1.104	0.011	-0.25	1.027	0.010	-0.25	1.003	0.010
0.00	1.182	0.012	0.00	1.190	0.012	0.00	1.129	0.011	0.00	1.100	0.011
0.25	1.112	0.012	0.25	1.086	0.011	0.25	1.028	0.010	0.25	1.010	0.010
0.50	0.886	0.009	0.50	0.848	0.009	0.50	0.747	0.007	0.50	0.734	0.007
0.75	0.718	0.007	0.75	0.475	0.005	0.75	0.395	0.004	0.75	0.364	0.004

Table 6.1: Values for $g_{N,\rho}$ for multivariate normal distribution.

These results should be of use beyond the stationary scenario. Specifically, a practically relevant example of non-stationarity is that of a data stream which switches between different stationary regimes. We can directly interpret the results we have derived in this context. The larger the value of λ, the more responsive the estimator will be to the new conditions after a regime switch (namely, the distribution of the bivariate data stream and its concordance properties). However, this comes at a cost of increased variance in the estimator $\hat{R}_N$ in each stationary regime. The above gives a rough guide as to how large one can set λ to increase responsiveness whilst staying within a chosen variance "budget". The variance budget determines the preciseness of the correlation estimate in each stationary regime. Roughly speaking, the variance for sufficiently small lambda is approximately $\left[\frac{\lambda}{2-\lambda}\right] g_{N,\rho}$ for distributions similar to the bivariate normal distribution with correlation parameter ρ in the i.i.d. scenario. Given the table above, it is clear that the maximum variance across all values of N is approximately $\left[\frac{\lambda}{2-\lambda}\right]$ (obtained at $\rho = 0$). Thus, λ should be selected according to,

$$\lambda \lesssim \frac{2\sigma_{tol}^2}{1+\sigma_{tol}^2}, \tag{6.10}$$

in order to stay within the variance budget defined by σ_{tol}^2. In fact, given

that one would reasonably set σ_{tol} to a small value e.g. $\sigma_{tol} \leq 0.2$, we can apply:

$$\lambda \lesssim 2\sigma_{tol}^2, \tag{6.11}$$

furnishing a simple rough guide for selecting λ. These results may even be potentially applied in a weakly serially correlated scenario. This is an area for further research.

6.5 Simulation Studies

6.5.1 Stationary Data

In this section, we evaluate the effectiveness of the proposed sequential Hermite series based Spearman's rank correlation estimation algorithm and compare to an existing algorithm [72] which we will refer to as the count matrix based approach. The count matrix algorithm provides a neat and straightforward way to approximate the Spearman's rank correlation coefficient for streaming data in a sequential manner. In essence, observations are placed in two-dimensional buckets, similar to constructing a two dimensional histogram with fixed cut-points. These cut-points are chosen to be quantiles of the univariate standard normal distribution. Upon the arrival of a new observation, the appropriate bucket's count is incremented. The associated matrix of observation counts forms an input to a tied-observation form of the Spearman's rank correlation coefficient estimator, phrased in terms of linear algebra operations for computational efficiency. It is worth briefly noting here that the count matrix algorithm can only be applied in non-stationary scenarios using a moving window approach, where memory requirements grow linearly with window size. By contrast, the Hermite series based algorithm can be easily extended to the non-stationary scenario (i.e. algorithm 2 in section 6.2.2) and does not rely on a moving window approach. This implies

that even if the accuracy of the algorithms was similar, the Hermite series based approach is still a valuable development.

The simulation study is as follows, we draw i.i.d. samples from the ubiquitous bivariate normal distribution for several sample sizes, namely $n = 10^4, 5 \times 10^4, 10^5$ at various values of the correlation parameter and compare the accuracy of the sequential Hermite series based Spearman's rank correlation estimation algorithm to the count matrix based algorithm. We utilise the following parameters for the mean vector and covariance matrix respectively, $\mu = (0, 0)$, $\Sigma = (\sigma_1, \rho\sigma_1\sigma_2; \rho\sigma_1\sigma_2, \sigma_2)$ where $\sigma_1 = 1, \sigma_2 = 1$ and $\rho = -0.75, -0.5, -0.25, 0.25, 0.5, 0.75$.

For each sample size $n = 10^4, 5 \times 10^4, 10^5$ and each value of the correlation parameter, $\rho = -0.75, -0.5, -0.25, 0.25, 0.5, 0.75$, the following steps are repeated $m = 10^3$ times for the Hermite series based Spearman's rank correlation estimation algorithm,

1. Draw n i.i.d. observations, $(x_i, y_i), \quad i = 1, \ldots, n$ from the bivariate normal distribution with mean vector μ and correlation parameter ρ.

2. Iterate through the sample, updating the Hermite series based Spearman's rank correlation estimate, $\hat{R}_N$, in accordance with algorithm 1 in section 6.2.1. The Spearman's rank correlation is then recorded at the last observation in the sample, $i = n$, for the Hermite series based algorithm.

3. Calculate the exact Spearman's correlation coefficient, R, for the sample.

4. Calculate the absolute error between the Hermite series based Spearman's estimate and the exact Spearman's correlation coefficient.

The mean absolute error (MAE) between the Hermite series based estimate and the exact Spearman's rank correlation coefficient for a partic-

ular value of ρ and sample size n is then estimated through $\widehat{\text{MAE}}(\hat{R}_N) = \frac{1}{m}\sum_{j=1}^{m}|\hat{R}_N^{(j)} - R^{(j)}|$, where j indexes a particular set of n observations.

The exact same steps as above are repeated for the count matrix algorithm, where we iterate through each sample, updating the count matrix based Spearman's rank correlation estimate in accordance with algorithm 2 in [72]. In terms of the choice of parameters, for the Hermite series based algorithm, we searched a coarse grid of N values, up to a maximum of $N = 20$ to identify the value of N that minimised the average estimated MAE across all values of ρ for a given sample size. We evaluated two choices of cut-points for the count matrix based algorithm, namely $c = 20$ and $c = 30$. The choices are motivated as follows, $c = 20$ yields a similar number of values to maintain in memory for the count matrix based algorithm as the Hermite series based algorithm with $N = 20$ i.e. this constitutes a like-for-like comparison. In [72] a rule of thumb of $c = 30$ cut-points is suggested for the count matrix based algorithm to give good performance. Larger numbers of cut-points are expected to give even better performance. We limit the maximum cut-points to 30 however since this already implies storing 960 values for the count matrix versus the Hermite series algorithm where the maximum number of values to store is 483 for $N = 20$. It is also worth noting that the execution times of the Hermite series based and count matrix based algorithms are roughly comparable (in our implementation). The standard error of the MAE is also evaluated for both the Hermite series based algorithm and the count matrix based algorithm.

The optimal value of N for the Hermite series based algorithm was found to be $N = 20$ for all sample sizes. Tabulated results for this section are presented in appendix C.3. The comparative MAE results are presented in table C.1 for the Hermite series based algorithm with $N = 20$ and count matrix algorithm with $c = 20$. The results for the Hermite series based algorithm with $N = 20$ and count matrix algorithm with $c = 30$ are presented

in table C.2. These results are summarised in table C.3 and C.4 as the average MAE (and standard deviation of MAE) across all the values of ρ considered.

These results suggest that the Hermite series based algorithm performs better than the count matrix algorithm as the sample size increases which is particularly relevant in the streaming data and massive data set scenarios. We have repeated the same simulation analysis for other values of μ in addition to $\mu = (0, 0)$, namely $\mu = (1, 1), (2, 2), (3, 3)$. We have observed that while both the Hermite series based and count matrix based Spearman's rank correlation estimation algorithms perform well for the mean vector at the origin, performance deteriorates for both algorithms with mean-vectors μ away from the origin. Interestingly, the accuracy of the Hermite series based algorithm deteriorates less and performs better than the count matrix algorithm in all these scenarios. However, the degradation in performance suggests that both algorithms would benefit from standardising the observations. Thus we implemented an online standardisation procedure and applied it identically for both algorithms (see appendix A.1 for a description of calculating the mean and standard deviation in a stable, online manner). Comfortingly, the results for different values of μ and σ_1, σ_2 are very similar to table C.1 and C.2 above with the online standardisation procedure in place.

Next we simulate from a non-normal bivariate distribution by first drawing i.i.d. observations from a standard bivariate normal distribution, $(x_i, y_i), \quad i = 1, \ldots, n$ and then transforming these observations with a strictly monotonically increasing function, $f(x)$, as $(f(x_i), f(y_i)), \quad i = 1, \ldots, n$. Since $f(x)$ preserves orderings of x_i and y_i individually, the Spearman's rho will be unchanged. We choose $f(x) = \exp(x)$. The marginal distributions of x and y are then log-normal distributions. The results are summarised in table C.5 and C.6.

We see that the performance of the Hermite series based algorithm is better than the count matrix algorithm with $c = 20$ (a like-for-like comparison) and somewhat worse compared to the count matrix algorithm with $c = 30$.

6.5.2 Non-Stationary Data

In this section, we evaluate our proposed online Spearman's algorithm applicable to non-stationary streams. The count matrix algorithm described previously can only be applied in the non-stationary scenario by applying a moving window approach. The larger the moving window, the greater the memory requirements. By contrast, the exponentially weighted Hermite series based Spearman's rank estimation algorithm has fixed memory requirements irrespective of the "effective" window size implied by the weighting parameter λ. This is a distinct advantage of the Hermite series based approach and the count matrix approach is thus not directly comparable in this setting. As such we do not include comparisons to the count matrix approach in this section.

The non-stationary models we evaluate are the following,

1. We draw $n = 10^4$ observations from a bivariate normal distribution with mean vector $\mu = (0, 0)$, covariance matrix $\Sigma = (1, \rho; \rho, 1)$, where $\rho^{(i)} = -1 + 2\frac{i-1}{n-1}$, $i = 1, \ldots, n$. Thus the correlation begins at -1 and ends at 1. This models a bivariate stream that begins perfectly anti-correlated and ends perfectly correlated. In addition, we replace 0.5% i.e. 50 observations uniformly at random with gross errors modelled by a bivariate normal with mean vector $\mu = (0, 0)$ and covariance matrix $\Sigma = (10^4, 0; 0, 10^4)$. This allows us to also explore the robustness of the Hermite series based Spearman's rank correlation estimation algorithm.

2. We draw $n = 10^4$ observations from a bivariate normal distribution with mean vector $\mu = (0, 0)$, covariance matrix $\Sigma = (1, \rho; \rho, 1)$, where $\rho^{(i)} = \sin(2\pi\frac{i-1}{n-1})$, $i = 1, \ldots, n$. This models a bivariate stream that oscillates between correlated and anti-correlated regimes. This could represent the price return innovations of two financial time series which switch between momentum and mean-reversion regimes for example. In addition, we replace 0.5% i.e. 50 observations uniformly at random

with gross errors modelled by a bivariate normal with mean vector $\mu = (0, 0)$, $\Sigma = (10^4, 0; 0, 10^4)$.

In figure 6.1 we plot the evolution of the correlation parameter ρ for the two models described above.

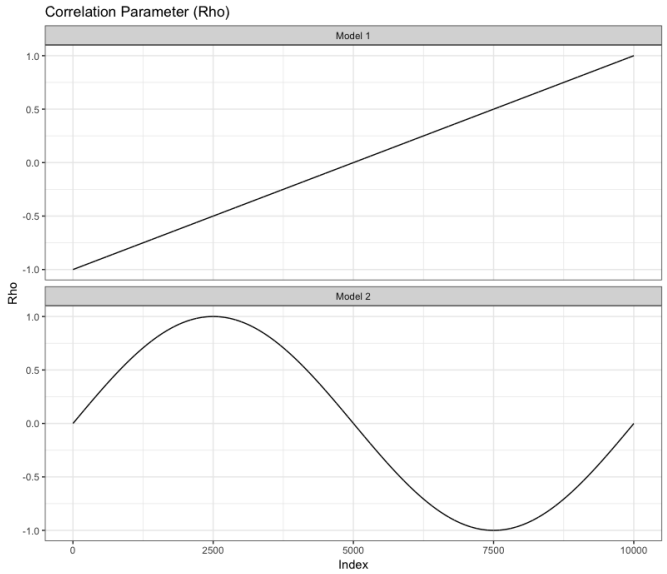


Figure 6.1: Evolution of the correlation parameter rho for the non-stationary models.

To assess the robustness of the exponentially weighted Hermite series based Spearman's rank correlation estimator, we also transform the Spearman's rank to Pearson's product-moment correlation (using the relationship $\hat{\rho}_S = 2\sin\left(\frac{\pi}{6}\hat{R}_N\right)$, applicable to bivariate normal distributions) and compare to an online, exponentially weighted version of the standard Pearson's correlation estimator, $\hat{\rho}(x, y)_\lambda$, which we define in appendix C.4. While the Spearman's rank based estimator for the Pearson's product-moment correlation is not as efficient as the standard Pearson's correlation estimator, it allows us to assess robustness and accuracy in a comparable manner.

For each of the two models, the following steps are repeated:

1. Draw $n = 10^4$ observations from the model.

2. Iterate through the sample, updating the exponentially weighted Hermite series based Spearman's rank correlation estimate in accordance with algorithm 2 in section 6.2.2 at each observation. In addition, the Pearson's correlation is estimated from the Hermite series based Spearman's rank estimate as $\hat{\rho}_S = 2\sin\left(\frac{\pi}{6}\hat{R}_N\right)$.

3. Calculate the exact Spearman's correlation coefficient at each observation through $R = \frac{6}{\pi}\arcsin(\frac{\rho^{(i)}}{2})$, where $\rho^{(i)}$ is the exact Pearson's correlation at each observation generated by the model under consideration.

4. We repeat the aforementioned steps $m = 10^3$ times and estimate two MAE curves. The first MAE curve tracks the MAE at the ith observation between the Hermite based Spearman's correlation estimate and the exact Spearman's rank coefficient. The second MAE curve tracks the MAE at the ith observation between the Pearson's correlation estimate obtained by transforming the Hermite based Spearman's correlation estimate and the exact Pearson's correlation. The latter MAE curve is compared to the MAE performance of an exponentially weighted version of the standard Pearson's correlation estimator.

5. Repeat this procedure at a grid of N and λ values to assess the performance of the Hermite series based estimator for different values of these parameters.

The Spearman's rank correlation results for model 1 are presented in figure 6.2. The Pearson's product-moment results for model 1 are presented in figure 6.3.

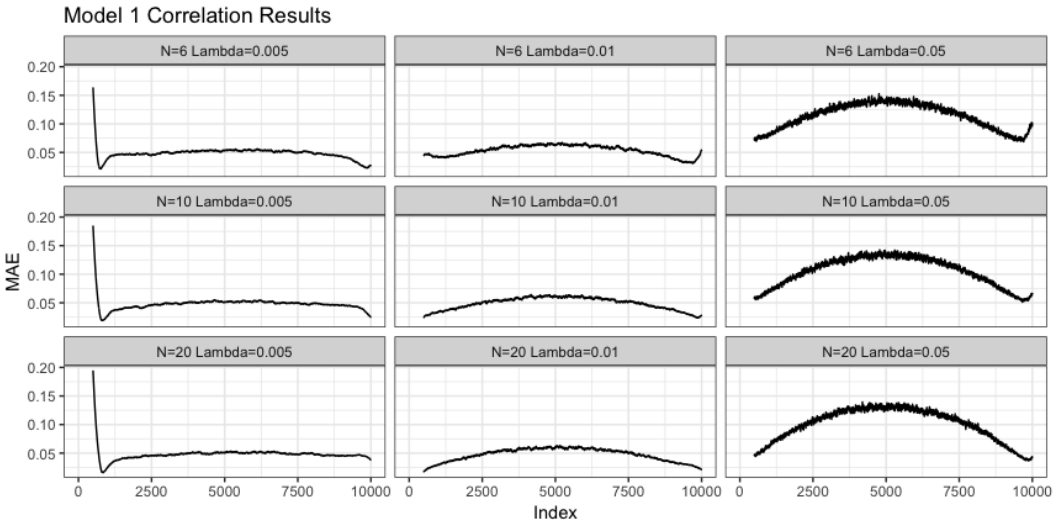


Figure 6.2: Mean absolute error results for the Hermite series based Spearman's rank correlation estimator for Model 1 on a grid of N and λ values.

The best results are achieved for $\lambda = 0.005$ and the MAE is relatively insensitive to the choice of N.

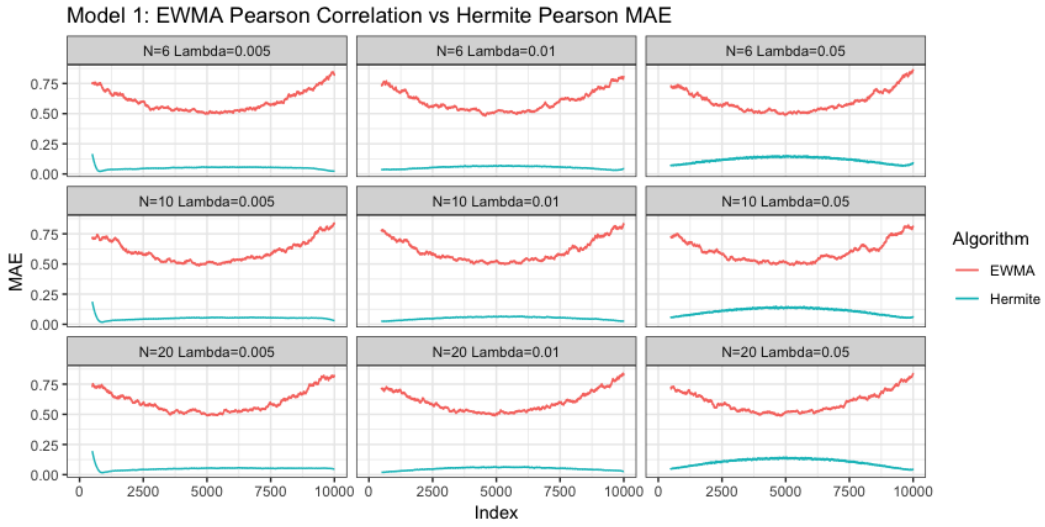


Figure 6.3: Mean absolute error results for the Hermite series based Spearman's rank correlation estimator transformed to give Pearson's product-moment correlation compared to an exponentially weighted version of the standard Pearson's product-moment correlation estimator (λ is the same for both algorithms).

It is clear that the Pearson's correlation derived from the Hermite series based Spearman's rank correlation estimator is more accurate and robust than an exponentially weighted version of the standard Pearson's product-moment correlation estimator.

The Spearman's rank correlation results for model 2 are presented in figure 6.4. The Pearson's product-moment results for model 2 are presented in figure 6.5.

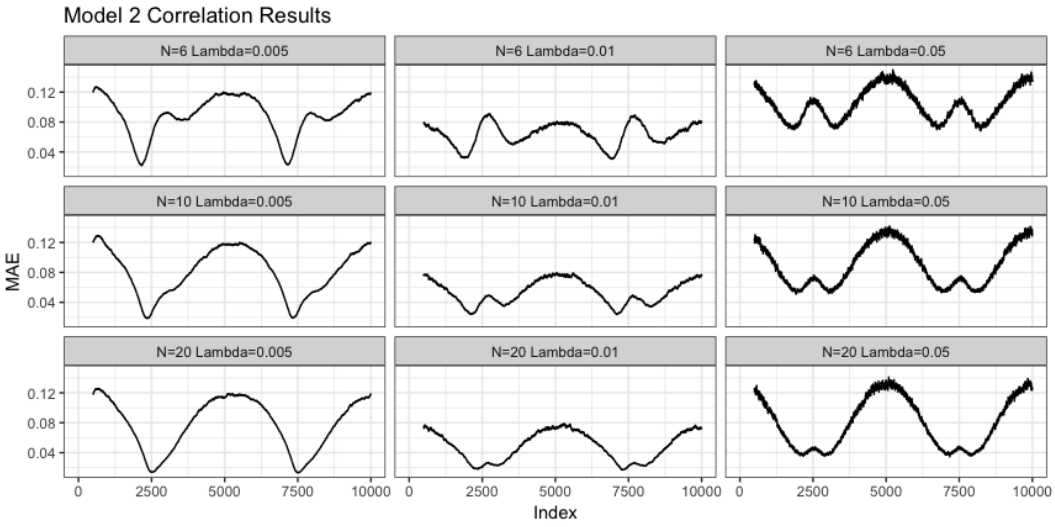


Figure 6.4: Mean absolute error results for the Hermite series based Spearman's rank correlation estimator for Model 2 on a grid of N and λ values.

The best results are achieved for $\lambda = 0.01$ with higher values of N performing slightly better than lower values ($N = 20$ yields the best performance).

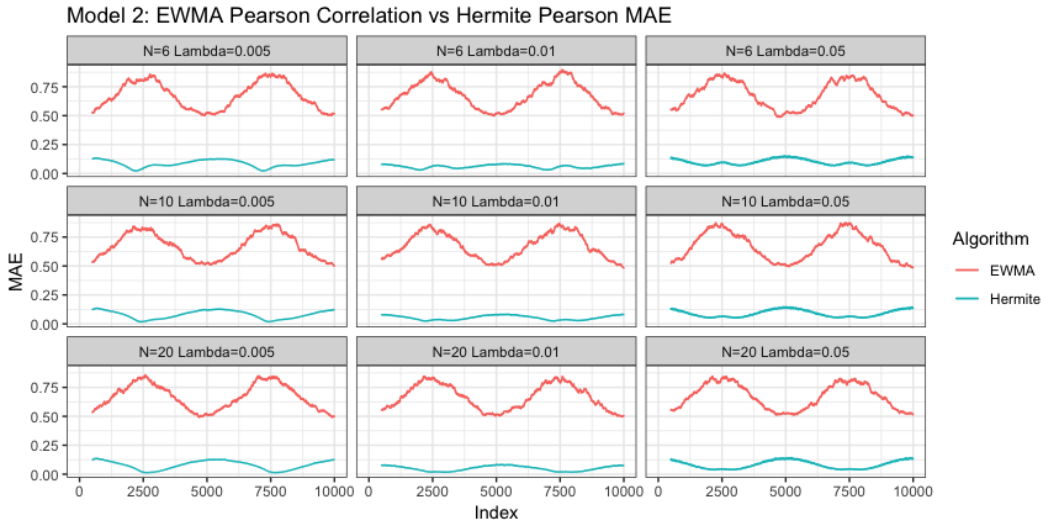


Figure 6.5: Mean absolute error results for the Hermite series based Spearman's rank correlation estimator transformed to give Pearson's product-moment correlation compared to an exponentially weighted version of the standard Pearson's product-moment correlation estimator (λ is the same for both algorithms).

Again the Pearson's correlation derived from the Hermite series based Spearman's rank correlation estimator is more accurate and robust than an exponentially weighted version of the standard Pearson's product-moment correlation estimator.

6.6 Real Data Example

In this section, we assess the exponentially weighted Hermite series based Spearman's rank correlation estimator applied to tick-by-tick forex data (sourced from [39]). The association between two currency pairs (EURUSD and GBPUSD in this instance) is expected to vary over time and is thus a good example of a non-stationary setting.

6.6.1 Data Description

The forex data is sourced from [39] and is comprised of tick-by-tick, top-of-book bid and offer quotes (aggregated across several bank liquidity providers) for EURUSD and GBPUSD for April 2019. The mid-price series for EURUSD, $p_i^{(1)}$, and GBPUSD, $p_j^{(2)}$, are calculated from the bid and offer quotes as follows:

$$p_i^{(1)} = (b_i^{(1)} + a_i^{(1)})/2, \quad i = 1, \ldots, n_1,$$

$$p_j^{(2)} = (b_j^{(2)} + a_j^{(2)})/2, \quad j = 1, \ldots, n_2,$$

where $a^{(1)}, a^{(2)}$ are EURUSD and GBPUSD offer quotes respectively and $b^{(1)}, b^{(2)}$ are EURUSD and GBPUSD bid quotes respectively. These mid-prices are then sampled on a minutely basis using the previous-tick methodology (i.e. the most recent tick in each currency pair is recorded as the price at a given minute), see [1] for example for a description of this methodology in the context of other approaches to data synchronisation. The mid-price minutely sampled series are thus:

$$p^{(1)}_{t_{(1)}}, p^{(1)}_{t_{(2)}}, \ldots p^{(1)}_{t_{(n)}}, \quad t_{(i+1)} - t_{(i)} = 1 \text{ minute},$$

$$p^{(2)}_{t_{(1)}}, p^{(2)}_{t_{(2)}}, \ldots p^{(2)}_{t_{(n)}}, \quad t_{(i+1)} - t_{(i)} = 1 \text{ minute},$$

and we obtain the basis point log returns returns via:

$$r^{(1)}_{(1)}, r^{(1)}_{(2)}, \ldots r^{(1)}_{(n-1)}, \quad r^{(1)}_{(j)} = 10^4 \log \frac{p^{(1)}_{t_{(j+1)}}}{p^{(1)}_{t_{(j)}}}.$$

$$r^{(2)}_{(1)}, r^{(2)}_{(2)}, \ldots r^{(2)}_{(n-1)}, \quad r^{(2)}_{(j)} = 10^4 \log \frac{p^{(2)}_{t_{(j+1)}}}{p^{(2)}_{t_{(j)}}}.$$

The descriptive statistics of the returns are presented in table 6.2.

Summary Statistic	EURUSD Returns	GBPUSD Returns
Count	30572.00	30572.00
Mean	0.00	0.01
Standard Deviation	0.77	1.21
Skewness	-0.66	0.08
Kurtosis	21.06	26.53
Min	-17.50	-26.67
Max	7.49	21.97

Table 6.2: Descriptive statistics of forex basis point log returns.

A 2-d histogram of the returns is presented in figure 6.6.

Figure 6.6: 2-d histogram of EURUSD and GBPUSD log returns.

A density plot of the returns is presented in figure 6.7.

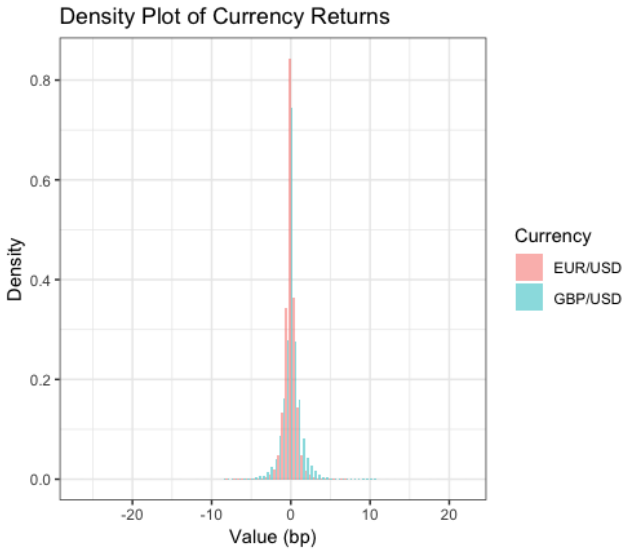


Figure 6.7: Density plot of EURUSD and GBPUSD log returns.

The data presents a mild violation of the assumption of a continuous bivariate distribution in that some values of basis point log returns are repeated (due to the fact that the minimum increment of quotes i.e. tick size is 1e-5). Indeed 17.3% of EURUSD log returns and 11% of GBPUSD log returns are repeated.

6.6.2 Results

The initial analysis we conduct is as follows, we iterate through the bivariate forex returns data set described above and update the exponentially weighted Hermite series based Spearman's rank correlation estimator at each minutely return update $(r_{(j)}^{(1)}, r_{(j)}^{(2)}), \quad j = 1, \ldots, n$. These estimates are then compared to the exact Spearman's rank correlation estimates based on a moving window of a certain size. In choosing the appropriate value for moving window size, w, we utilise the commonly used relation between simple moving averages and exponentially weighted moving averages which asserts $\lambda = 2/(w+1)$. This relation is based on equating the average "age"/lag of an observation in

the moving window, $(w+1)/2$, and the average "age"/lag of an observation in the exponentially weighted setting, $1/\lambda$. For analysing the forex returns data at an average "age"/lag of approximately two hours, we set $\lambda = 0.01$, corresponding to a moving window size $w = 200$. Based on the simulation studies conducted previously, we have seen that the Hermite series based algorithm is reasonably insensitive to the choice of N, we thus chose $N = 10$ for illustrative purposes. The results are presented in figure 6.8.

Figure 6.8: Hermite series based exponentially weighted Spearman's rank estimator ($\lambda = 0.01$) versus moving window exact Spearman's correlation ($w = 200$).

It is clear that the exponentially weighted Hermite series based Spearman's rank correlation estimator and the moving window exact Spearman's rank correlation estimator track the association between the currency pairs similarly. Indeed the mean absolute difference between the estimates across all observations is 0.048.

In terms of forecasting ability, we evaluate the two hour forward, realised Spearman's rank correlation coefficient formed from observations $j +$

$1, \ldots, j+120$ where $j = 1000, \ldots, (n-120)$ compared to the moving window and exponentially weighted Hermite series based Spearman's correlation estimates at observation j where $j = 1000, \ldots, (n-120)$. To select the best λ, N values for the Hermite series based estimator and the best w value for the moving window estimator we analysed forex data for the previous month of March 2019. In particular, we performed a grid-search for λ and N, covering $\lambda = 0.001, 0.002, 0.005, 0.01, 0.05$ and $N = 6, 8, 10, 20$ for the exponentially weighted Hermite series Spearman's rank correlation estimator. The best performing parameters were selected according to MAE of the predicted Spearman's rank correlation versus realised forward Spearman's rank correlation for the next two hours. These values were $\lambda = 0.002, N = 10$. Note that we again applied the online standardisation procedure that we found advantageous in our simulations studies in section 6.5.1. For the moving window approach, we evaluated $w = 50, 100, \ldots, 1000$. The best value for the moving window size was $w = 850$. Even at $w = 200$ the moving window approach was more computationally expensive and slower than the Hermite series based approach. We thus terminated the parameter search at $w = 1000$. We also evaluated the constant prediction based on the batch estimate formed from calculating the Spearman's rank correlation coefficient for the whole March 2019 data set.

The out of sample results for the MAE of the estimates compared to the forward realised Spearman's rank correlation coefficient in April 2019 are:

Algorithm	MAE
Hermite EW ($\lambda = 0.002$)	**0.1103**
Moving Window ($w = 850$)	0.1120
Constant Estimate	0.1211

Table 6.3: Out of sample MAE results for forecasting forward, realised two hour Spearman's rank correlation.

The Hermite series based exponentially weighted Spearman's estimator

is slightly more accurate than the moving window estimate and significantly more accurate than the batch estimate formed from March 2019.

The main advantages of the exponentially weighted estimator are as follows:

- The Hermite based Spearman's rank correlation estimator has $O(1)$ time and memory complexity with respect to λ. By contrast, the moving window approach grows in time complexity and memory requirements with moving window size. For large moving window sizes, the Hermite series based estimator will be significantly faster. As an example, with our implementation in R, the moving window estimator with $w = 1000$ is roughly 5-10 times slower than the Hermite series based algorithm.

- Observations older than a certain threshold do not sharply drop off, but are rather incorporated with decreasing weighting.

Chapter 7

Conclusion

In this thesis we have tackled the challenge of the sequential (online) estimation of univariate and bivariate statistics of streaming data (both stationary and non-stationary) and massive data sets. The estimators and algorithms we have introduced to treat stationary scenarios are also applicable to decentralised settings where estimates are formed from several distributed subsets of the data and subsequently combined into a centralised estimate.

In the univariate context, we have applied sequential cumulative distribution function estimators based on Hermite series density estimators which allow arbitrary quantiles to be obtained numerically. We have also redressed a gap in the literature on the theoretical properties of Hermite series based distribution function estimators and associated quantile estimators. For probability densities with support on the positive half-real line, we have proved asymptotic MSE consistency (pointwise consistency) as well as asymptotic MISE consistency for these Hermite CDF estimators. We have also provided the associated rates. In addition, we have demonstrated that the MAE of quantile estimates obtained from these estimators directly depends on the MISE of the Hermite series density estimator under certain conditions.

The particular usefulness of Hermite distribution function estimators is for distributions associated with probability densities with support on the

full real line however. We have also proved consistency of the Hermite distribution function estimators in this setting, i.e. for estimators defined for densities with support on the full real line, in the MSE and MISE sense. In doing so, we have furnished results on the bias and variance of these estimators. Additionally, we have provided a specific example where the rate of convergence in MSE approaches the optimal rate, $O(n^{-1})$, namely the normal distribution. In addition, we have proved almost sure convergence of the full real line estimators, including the case of Hermite series distribution function estimators of random length. We have also explored the robustness of the finite N Hermite series based cumulative distribution function estimator and have proved that the estimator is Bias-robust, implying that the approximate maximum bias due to data contamination is bounded. Finally, we have explored the finite sample properties of these Hermite series based estimators.

Our investigations reveal inferiority of the Hermite series based distribution function estimator compared to the smooth kernel distribution function estimator with regards to asymptotic rate of convergence and finite sample performance. This suggests that in all scenarios other than massive data sets and streaming data, the kernel distribution function estimator should be favoured over the Hermite series based estimator. However, an important additional interpretation of our results is that in streaming/massive data set scenarios, the Hermite series based estimator still has sufficiently good performance and approaches the optimal rate of MSE convergence in some practically relevant cases.

For sequential quantile estimation, we have introduced algorithms - based on the Hermite series CDF estimators - for online quantile estimation in the settings of static quantile estimation and dynamic quantile estimation. These algorithms have $O(1)$ time complexity for *updating* the distribution and quantile function estimates. In the static quantile estimation setting we have exploited the fact that Hermite series coefficients can be updated in

a sequential manner. To treat dynamic quantile estimation, we have introduced a novel expansion with an exponentially weighted estimator for the Hermite series coefficients. This expansion allows the local behaviour of a non-stationary stream of data to be tracked.

To make our analysis concrete, we have considered i.i.d data streams and independent, non-identically distributed data streams in our simulation studies and our theoretical analysis. The simulation studies revealed the Hermite series based algorithms to be competitive with a leading existing algorithm for online quantile estimation. In addition, a test on real forex data confirmed the effectiveness of the exponentially weighted algorithm in a more general and realistic setting and provided evidence that our techniques are effective for heavy-tailed distributions. The particular usefulness of our algorithms is that they allow *arbitrary* quantiles to be estimated in an online manner. They do not require a particular set of quantiles to be specified upfront, which is a limitation of existing algorithms. In obtaining these novel algorithms, we have thus provided a solution to the problem of online distribution function and online quantile function estimation for both stationary and non-stationary data streams. Online estimates of these functions are in fact useful in a broader context in that any function of these quantities can be calculated in an online manner. These online estimates could be used in online machine learning applications for example.

In the bivariate setting, we have introduced novel, sequential (online) algorithms to estimate the most popular measure of nonparametric correlation, the Spearman's rank correlation coefficient, in both stationary (static) and non-stationary (dynamic) settings. These algorithms are based on bivariate Hermite series density estimators and Hermite series based cumulative distribution function estimators. The stationary setting corresponds to bivariate data streams where the Spearman's rank correlation coefficient is constant. This setting also includes the one-pass estimation of the Spearman's rank correlation for massive data sets. In addition, this algorithm could be applied

in decentralised settings where estimates of the Spearman's rank correlation, generated on separate portions of a larger dataset, can be combined to form an overall estimate. We have proved that in the i.i.d. setting, the Hermite series based Spearman's rank correlation estimator converges to the grade correlation which is the constant which is unbiasedly estimated by the standard Spearman's rank correlation estimator in large samples.

The non-stationary (dynamic) setting corresponds to situations where the Spearman's rank correlation of a bivariate data stream is expected to vary over time. In order to treat these scenarios we have introduced an exponentially weighted version of the Hermite series based Spearman's rank correlation estimator. To the best of our knowledge this is the first algorithm allowing the online tracking of a time-varying Spearman's rank correlation of a bivariate data stream that does not rely on moving/sliding windows. We have derived variance (and standard error) results for the exponentially weighted estimator in the i.i.d. scenario.

The effectiveness of the Hermite series based Spearman's rank estimator has been demonstrated in the stationary setting by means of a simulation study which revealed the Hermite series based estimator to be competitive with an existing approach. We have also demonstrated the effectiveness of the exponentially weighted Hermite estimator through simulation and real data studies. We expect our algorithms to be useful in a variety of settings, particularly those where robustness to outliers and errors is required in addition to fast online calculation of correlation which may vary over time (financial applications involving high frequency data streams for example).

Potential applications in the machine learning domain include:

- Feature selection on massive datasets. In particular, a univariate filter approach based on Spearman's rank correlation can be applied. The Spearman's rank correlation is suitable for all monotonic relationships and is more robust than the Pearson's correlation and could be a more effective measure on which to rank and filter out features. The ability of

Spearman's rank correlation to capture more general relationships than linear relationships between a given feature and the response variable is shared by a popular alternative, namely information gain (mutual information). Unlike the information gain (mutual information), it is not necessary to discretise the features however and it could be significantly faster to apply the Hermite series based Spearman's rank correlation measure. The effectiveness and speed of feature selection using the Hermite series based Spearman's rank correlation estimator compared to feature selection based on the Pearson's correlation or information gain is an area for future research.

- Hierarchical clustering on massive data sets or streaming data. A dissimilarity matrix based on Spearman's correlation can be calculated in real time (for streaming data), or in one pass (for massive data sets) using the Hermite series based Spearman's rank correlation estimator. This dissimilarity matrix can then be used along with standard agglomerative or divisive hierarchical clustering algorithms. For a reasonably small number of variables this should facilitate real-time hierarchical clustering. This is again a potentially fruitful area for future research.

Appendix A

Sequential Quantile Estimation

A.1 Standardising the observations

It is reasonable to assume that in practice, the quality of the fit yielded by the truncated Gauss-Hermite expansion should be better if applied to standardised random variables (with mean equal to zero and standard deviation equal to one). For a random variable x with mean μ and standard deviation σ we have the standardised random variable:

$$\tilde{x} = \frac{x - \mu}{\sigma},$$

with probability density $\tilde{f}(\tilde{x}) = \sigma f(\sigma\tilde{x}+\mu)$. The associated truncated Gauss-Hermite expansion is:

$$\tilde{f}(\tilde{x}) = \sum_{k=0}^{N} a_k H_k(\tilde{x}) Z(\tilde{x}),$$

where

$$a_k = \alpha_k \int_{-\infty}^{\infty} Z(\tilde{x}) \tilde{f}(\tilde{x}) H_k(\tilde{x}) d\tilde{x}.$$

Ideally, we would utilise:

$$\hat{a}_k(\mu, \sigma) = \alpha_k \frac{1}{n} \sum_{i=1}^{n} Z(\frac{x_i - \mu}{\sigma}) H_k(\frac{x_i - \mu}{\sigma}),$$

to estimate the coefficients. In practice however, we do not know the true values of μ and σ and thus we must use estimates of these values, $\hat{\mu}$ and $\hat{\sigma}$. In general, the effect of using these estimates is to *bias* the estimate of $\hat{a}_k$. This can be seen by considering the Taylor expansion of $\hat{a}_k(\hat{\mu}, \hat{\sigma})$ which implies $E(\hat{a}_k(\hat{\mu}, \hat{\sigma})) \neq E(\hat{a}_k(\mu, \sigma))$.

Despite this bias, standardising using the estimated mean and standard deviation improves the quality of the fit in many cases. In the sections below we provide online algorithms to estimate the mean and standard deviation in both the static and dynamic setting. These estimates can then be plugged into the standard Gauss-Hermite coefficients and the exponentially weighted Gauss-Hermite coefficients respectively.

A.1.1 Static Estimation

The usual estimators of the mean and standard deviation can be calculated in an online way. The mean ($\hat{\mu}_k$) can be updated with each new observation as:

$$\hat{\mu}_1 = x_1,$$

$$\hat{\mu}_k = \frac{1}{k} \left((k-1)\hat{\mu}_{k-1} + x_k \right), \quad k \geq 2.$$

The standard deviation ($\hat{\sigma}_k$) can also be estimated in an online manner. To avoid numerical precision problems an algorithm such as that originating from Welford [69] can be applied.

$$M_1 = x_1,$$

$$S_1 = 0,$$

$$M_k = M_{k-1} + \frac{(x_k - M_{k-1})}{k},$$

$$S_k = S_{k-1} + (x_k - M_{k-1})\,(x_k - M_k)\,,$$

$$\hat{\sigma}_k = \sqrt{\frac{S_k}{k-1}}, \quad k \geq 2.$$

A.1.2 Dynamic Estimation

For the purposes of obtaining a local estimate of the mean and standard deviation we can utilise EWMA estimators:

$$\hat{\mu}_1 = x_1,$$
$$\hat{\mu}_k = (1-\lambda)\hat{\mu}_{k-1} + \lambda x_k, \quad k \geq 2,$$

$$\hat{V}_1 = 1,$$
$$\hat{V}_k = (1-\lambda)\hat{V}_{k-1} + \lambda(x_k - \hat{\mu}_k)^2, \quad k \geq 2,$$
$$\hat{\sigma}_k = \sqrt{\hat{V}_k}.$$

$$0 < \lambda \leq 1.$$

A.2 MISE of the Exponentially Weighted Gauss-Hermite Expansion

A.2.1 MISE of the Exponentially Weighted Gauss-Hermite Expansion for IID Data

In this section we derive results pertaining to i.i.d. data for the EWGH density estimator defined from the exponentially weighted Gauss-Hermite coefficients (4.9). We begin by deriving a bound on the MSE of the coefficients (4.9) which is then used to obtain a bound on the MISE of the EWGH density estimator.

Proposition 8. *Let the true probability density function $f(x) \in L_2$ and $E|X|^{\frac{2}{3}} < \infty$. The MSE of the coefficients (4.9) has the following upper bound for a sample of $n+1$ observations in the i.i.d. case:*

$$MSE(\hat{a}_k) \leq \left[\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2n}\right]+(1-\lambda)^{2n}\right]\frac{\sqrt{\pi}c(k+1)^{-1/2}}{2^{k-1}k!},$$

where c is a constant.

Proof. The truncated EWGH expansion of $f(x)$ is given by:

$$\hat{f}_N(x) = \sum_{k=0}^{N} \hat{a}_k H_k(x) Z(x),$$

where the coefficient estimates are given by (4.9). In what follows we assume an i.i.d sequence of $n+1$ observations has been drawn from $f(x)$ i.e. $\mathbf{x}_i \sim f(x)$.

The estimator (4.9) can be equivalently written as:

$$\hat{a}_k = \lambda \sum_{j=0}^{n-1} (1-\lambda)^j \left[\alpha_k Z(\mathbf{x}_{n-j}) H_k(\mathbf{x}_{n-j})\right] + (1-\lambda)^n \left[\alpha_k Z(\mathbf{x}_0) H_k(\mathbf{x}_0)\right].$$

The expected value of this estimator, $E(\hat{a}_k) = a_k$ and thus the estimator is unbiased.

The variance, $\mathrm{Var}(\hat{a}_k) = E\left[(\hat{a}_k - a_k)^2\right]$ of the estimator is:

$$\begin{aligned}
\mathrm{Var}(\hat{a}_k) &= \lambda^2 \sum_{j=0}^{n-1} (1-\lambda)^{2j} \mathrm{Var}\left(\left[\alpha_k Z(\mathbf{x}_{n-j}) H_k(\mathbf{x}_{n-j})\right]\right) \\
&+ (1-\lambda)^{2n} \mathrm{Var}\left(\left[\alpha_k Z(\mathbf{x}_0) H_k(\mathbf{x}_0)\right]\right) \\
&= \left[\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2n}\right] + (1-\lambda)^{2n}\right] \sigma_X^2(k), \qquad \text{(A.1)}
\end{aligned}$$

where $\sigma_X^2(k) = \mathrm{Var}\left[\alpha_k Z(x) H_k(x)\right] = E\left[\frac{\pi}{\left[2^{k-1}k!\right]^2}\left[Z(x)\right]^2\left[H_k(x)\right]^2\right] - a_k^2$. Note that we have exploited the independence of the observations $\mathbf{x}_i$.

Now, we can obtain an upper bound on $\sigma_X^2(k)$ using the properties of Hermite polynomials (following from (2.4) and (2.5), see [29]):

$$\sigma_X^2(k) \leq \frac{\sqrt{\pi} c (k+1)^{-1/2}}{2^{k-1}k!}.$$

This yields the following bound on the MSE of $\hat{a}_k$ (since $\hat{a}_k$ is unbiased, the bound on the MSE of $\hat{a}_k$ is equal to the bound on the variance):

$$\begin{aligned}
\mathrm{MSE}(\hat{a}_k) &= E\left[(\hat{a}_k - a_k)^2\right] \\
&\leq \left[\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2n}\right] + (1-\lambda)^{2n}\right] \frac{\sqrt{\pi} c (k+1)^{-1/2}}{2^{k-1}k!} \qquad \text{(A.2)}
\end{aligned}$$

□

Theorem 14. *Let the true probability density function $f(x) \in L_2$ and $E|X|^{\frac{2}{3}} < \infty$. If $r \geq 1$ derivatives of $f(x)$ exist and $(x - \frac{d}{dx})^r f(x) \in L_2$ then the MISE for the EWGH expansion in the i.i.d. case behaves as follows for N sufficiently large and $n \to \infty$:*

$$MISE\,(\hat{f}_N) = O(N^{1/2}) \left[\frac{\lambda}{2-\lambda}\right] + O(N^{-r}).$$

Proof. If $r \geq 1$ derivatives of $f(x)$ exist and $(x - \frac{d}{dx})^r f(x) \in L_2$ then $\sum_{k=N+1}^{\infty} \frac{2^{k-1}k!}{\sqrt{\pi}} a_k^2 = O(N^{-r})$ (see [67],[29]). Combining this fact and the bound on MSE($\hat{a}_k$) (proposition 8) with the expression for the MISE (2.23) we obtain the following:

$$\begin{aligned} \text{MISE}\,(\hat{f}_N) &= \sum_{k=0}^{N} \frac{2^{k-1}k!}{\sqrt{\pi}} E\left[(\hat{a}_k - a_k)^2\right] + \sum_{k=N+1}^{\infty} \frac{2^{k-1}k!}{\sqrt{\pi}} a_k^2 \\ &\leq \left[\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2n}\right] + (1-\lambda)^{2n}\right] c \sum_{k=0}^{N} (k+1)^{-1/2} \\ &+ O(N^{-r}) \end{aligned} \tag{A.3}$$

For N sufficiently large and $n \to \infty$ we have:

$$\text{MISE}\,(\hat{f}_N) = O(N^{1/2}) \left[\frac{\lambda}{2-\lambda}\right] + O(N^{-r}).$$

□

A.2.2 MISE of the Exponentially Weighted Gauss-Hermite Expansion for Non-identically Distributed Data

In this section we extend the results derived for i.i.d data to non-identically distributed, independent data. We consider the case where we observe $s+1$ observations from a probability distribution $f_1 \in L_2$ followed by a further t observations from a second distribution $f_2 \in L_2$ i.e. we assume an independent sequence of $s+1$ observations $\mathbf{x}_0, \ldots, \mathbf{x}_s \sim f_1(x)$ followed by an independent sequence of t observations, $\mathbf{x}_{s+1}, \ldots, \mathbf{x}_{t+s} \sim f_2(x)$. We consider this a fundamental example of non-identically distributed data. We evaluate and bound the MISE of $\hat{f}_N(x)$ compared to $f_2(x)$ at observation $t+s$ by first deriving a bound on the MSE of the exponentially weighted Gauss-Hermite coefficients. We denote the true Gauss-Hermite coefficients of $f_1(x)$ as $a_k^{(1)}$ and the true Gauss-Hermite coefficients of $f_2(x)$ as $a_k^{(2)}$.

Proposition 9. *The MSE of the coefficients (4.9) compared to $a_k^{(2)}$ has the following upper bound after $s+1$ independent observations from $f_1 \in L_2$ followed by t independent observations from $f_2 \in L_2$ provided $E|X|^{\frac{2}{3}} < \infty$ for both distributions:*

$$\begin{aligned}
&\left[E\left(\hat{a}_k - a_k^{(2)}\right)^2\right] \\
&\leq \frac{c\sqrt{\pi}(k+1)^{-1/2}}{2^{k-1}k!}\left[4(1-\lambda)^{2t} + \frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2(s+t)}\right] + (1-\lambda)^{2(s+t)}\right].
\end{aligned}$$

Proof. The expected value of the exponentially weighted estimator for the Gauss-Hermite coefficients is:

$$
\begin{aligned}
E\left(\hat{a}_k\right) &= \lambda \sum_{j=0}^{s+t-1} (1-\lambda)^j E\left([\alpha_k Z(\mathbf{x}_{s+t-j}) H_k(\mathbf{x}_{s+t-j})]\right) \\
&+ (1-\lambda)^{s+t} E\left([\alpha_k Z(\mathbf{x}_0) H_k(\mathbf{x}_0)]\right) \\
&= \lambda a_k^{(1)} \sum_{j=t}^{s+t-1} (1-\lambda)^j + \lambda a_k^{(2)} \sum_{j=0}^{t-1} (1-\lambda)^j + (1-\lambda)^{s+t} a_k^{(1)} \\
&= a_k^{(2)} + (1-\lambda)^t [a_k^{(1)} - a_k^{(2)}].
\end{aligned} \tag{A.4}
$$

Thus the squared bias of the estimator compared to the true Gauss-Hermite coefficient $a_k^{(2)}$ is:

$$
\left[E\left(\hat{a}_k\right) - a_k^{(2)}\right]^2 = (1-\lambda)^{2t} [a_k^{(1)} - a_k^{(2)}]^2.
$$

Similarly, the variance $\mathrm{Var}(\hat{a}_k) = E\left[(\hat{a}_k - E\left(\hat{a}_k\right))^2\right]$ of the estimator is:

$$
\begin{aligned}
\mathrm{Var}(\hat{a}_k) &= \lambda^2 [\sigma_X^{(1)}(k)]^2 \sum_{j=t}^{s+t-1} (1-\lambda)^{2j} + \lambda^2 [\sigma_X^{(2)}(k)]^2 \sum_{j=0}^{t-1} (1-\lambda)^{2j} \\
&+ (1-\lambda)^{2(s+t)} [\sigma_X^{(1)}(k)]^2 \\
&= [\sigma_X^{(1)}(k)]^2 \left[\frac{\lambda}{2-\lambda}(1-\lambda)^{2t} + \left[1 - \frac{\lambda}{2-\lambda}\right](1-\lambda)^{2(s+t)}\right] \\
&+ [\sigma_X^{(2)}(k)]^2 \frac{\lambda}{2-\lambda}\left[1 - (1-\lambda)^{2t}\right],
\end{aligned} \tag{A.5}
$$

where $[\sigma_X^{(1)}(k)]^2 = \mathrm{Var}\left[\alpha_k Z(x) H_k(x)\right], x \sim f_1(x)$ and
$[\sigma_X^{(2)}(k)]^2 = \mathrm{Var}\left[\alpha_k Z(x) H_k(x)\right], x \sim f_2(x)$.

Thus the mean squared error of $\hat{a}_k$ compared to $a_k^{(2)}$ is:

$$\begin{aligned}\left[E\left(\hat{a}_k - a_k^{(2)}\right)^2\right] &= \left[E\left(\hat{a}_k\right) - a_k^{(2)}\right]^2 + \mathrm{Var}(\hat{a}_k) \\ &= (1-\lambda)^{2t}[a_k^{(1)} - a_k^{(2)}]^2 \\ &+ [\sigma_X^{(1)}(k)]^2\left[\frac{\lambda}{2-\lambda}(1-\lambda)^{2t} + \left[1 - \frac{\lambda}{2-\lambda}\right](1-\lambda)^{2(s+t)}\right] \\ &+ [\sigma_X^{(2)}(k)]^2\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2t}\right]. \end{aligned} \tag{A.6}$$

The MSE therefore depends on the difference between the true Gauss-Hermite coefficients of $f_1(x)$ and $f_2(x)$ and the number of observations since the switch between the distributions (along with the value of λ). We can bound the MSE above using the properties of Hermite polynomials:

$$\begin{aligned}\left[E\left(\hat{a}_k - a_k^{(2)}\right)^2\right] \leq \frac{c\sqrt{\pi}(k+1)^{-1/2}}{2^{k-1}k!}\times \\ \times\left[4(1-\lambda)^{2t} + \frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2(s+t)}\right] + (1-\lambda)^{2(s+t)}\right].\end{aligned} \tag{A.7}$$

□

Theorem 15. *Given $s+1$ independent observations from $f_1 \in L_2$ followed by t independent observations from $f_2 \in L_2$, if $r \geq 1$ derivatives of $f_2(x)$ exist and $(x - \frac{d}{dx})^r f_2(x) \in L_2$ and $E|X|^{\frac{2}{3}} < \infty$ for both distributions then the MISE for the EWGH expansion behaves as follows for N sufficiently large:*

$$\begin{aligned}MISE(\hat{f}_N) &= O(N^{1/2})\left[4(1-\lambda)^{2t} + \frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2(s+t)}\right] + (1-\lambda)^{2(s+t)}\right] \\ &+ O(N^{-r}).\end{aligned}$$

Proof. If r derivatives of $f(x)$ exist and $(x - \frac{d}{dx})^r f(x) \in L_2$ then $\sum_{k=N+1}^{\infty} \frac{2^{k-1}k!}{\sqrt{\pi}} a_k^2 = O(N^{-r})$ (see [67],[29]). Combining this fact with the bound on the MSE in proposition 9 and the expression for the MISE (2.23), we obtain:

$$\begin{aligned}\mathrm{MISE}\,(\hat{f}_N) &= O(N^{1/2}) \left[4(1-\lambda)^{2t} + \frac{\lambda}{2-\lambda} \left[1 - (1-\lambda)^{2(s+t)} \right] + (1-\lambda)^{2(s+t)} \right] \\ &\quad + O(N^{-r}).\end{aligned}$$

□

A.3 Additional Simulation Results for Non-identically Distributed Data

In this section we collect tables comparing performance of the GH, EWGH and EWSA algorithms for non-identically distributed data. These results are associated with section 4.5.2.

	p	Lambda	Frequency GH	Frequency EWGH	Frequency EWSA
Distribution: Normal (Drift)					
m=100					
	0.50	0.01	0.471 (0.439,0.502)	0.394 (0.363,0.426)	0.344 (0.315,0.374)
		0.05	0.471 (0.439,0.502)	0.505 (0.475,0.539)	0.344 (0.315,0.374)
		0.10	0.471 (0.439,0.502)	0.511 (0.481,0.542)	0.344 (0.315,0.374)
	0.90	0.01	0.884 (0.864,0.902)	0.817 (0.792,0.84)	0.768 (0.741,0.794)
		0.05	0.884 (0.864,0.902)	0.911 (0.892,0.928)	0.768 (0.741,0.794)
		0.10	0.884 (0.864,0.902)	0.905 (0.886,0.923)	0.768 (0.741,0.794)
	0.99	0.01	0.986 (0.979,0.993)	0.985 (0.977,0.992)	0.882 (0.862,0.902)
		0.05	0.986 (0.979,0.993)	0.997 (0.993,1)	0.882 (0.862,0.902)
		0.10	0.986 (0.979,0.993)	1 (1,1)	0.882 (0.862,0.902)
m=400					
	0.50	0.01	0.287 (0.258,0.316)	0.457 (0.426,0.488)	0.077 (0.062,0.094)
		0.05	0.287 (0.258,0.316)	0.483 (0.454,0.515)	0.077 (0.062,0.094)
		0.10	0.287 (0.258,0.316)	0.503 (0.474,0.533)	0.077 (0.062,0.094)
	0.90	0.01	0.831 (0.807,0.853)	0.902 (0.885,0.92)	0.472 (0.44,0.502)
		0.05	0.831 (0.807,0.853)	0.886 (0.867,0.906)	0.472 (0.44,0.502)
		0.10	0.831 (0.807,0.853)	0.923 (0.907,0.939)	0.472 (0.44,0.502)
	0.99	0.01	0.984 (0.976,0.991)	0.992 (0.986,0.997)	0.709 (0.679,0.736)
		0.05	0.984 (0.976,0.991)	0.999 (0.997,1)	0.709 (0.679,0.736)
		0.10	0.984 (0.976,0.991)	1 (1,1)	0.709 (0.679,0.736)
m=1000					
	0.50	0.01	0.18 (0.158,0.204)	0.505 (0.475,0.535)	0.113 (0.094,0.131)
		0.05	0.18 (0.158,0.204)	0.489 (0.461,0.518)	0.113 (0.094,0.131)
		0.10	0.18 (0.158,0.204)	0.493 (0.459,0.525)	0.113 (0.094,0.131)
	0.90	0.01	0.865 (0.843,0.886)	0.904 (0.885,0.921)	0.305 (0.276,0.332)
		0.05	0.865 (0.843,0.886)	0.914 (0.895,0.93)	0.305 (0.276,0.332)
		0.10	0.865 (0.843,0.886)	0.907 (0.889,0.925)	0.305 (0.276,0.332)
	0.99	0.01	0.985 (0.977,0.992)	0.989 (0.982,0.995)	0.42 (0.388,0.449)
		0.05	0.985 (0.977,0.992)	0.994 (0.989,0.998)	0.42 (0.388,0.449)
		0.10	0.985 (0.977,0.992)	1 (1,1)	0.42 (0.388,0.449)

Table A.1: Observed out-of-sample frequencies of observations smaller than or equal to current quantile estimates for normal distribution with drift in mean. Bootstrap 95% confidence intervals provided in brackets.

	p	Lambda	Frequency GH	Frequency EWGH	Frequency EWSA
Distribution: Exponential (Drift)					
m=100					
	0.50	0.01	0.451 (0.42,0.479)	0.514 (0.483,0.543)	0.37 (0.339,0.399)
		0.05	0.451 (0.42,0.479)	0.461 (0.431,0.492)	0.37 (0.339,0.399)
		0.10	0.451 (0.42,0.479)	0.485 (0.454,0.516)	0.37 (0.339,0.399)
	0.90	0.01	0.872 (0.85,0.893)	0.822 (0.797,0.845)	0.779 (0.753,0.805)
		0.05	0.872 (0.85,0.893)	0.913 (0.895,0.929)	0.779 (0.753,0.805)
		0.10	0.872 (0.85,0.893)	0.932 (0.915,0.946)	0.779 (0.753,0.805)
	0.99	0.01	0.961 (0.949,0.973)	0.947 (0.933,0.96)	0.86 (0.837,0.882)
		0.05	0.961 (0.949,0.973)	0.99 (0.983,0.996)	0.86 (0.837,0.882)
		0.10	0.961 (0.949,0.973)	0.999 (0.997,1)	0.86 (0.837,0.882)
m=400					
	0.50	0.01	0.431 (0.401,0.462)	0.487 (0.456,0.518)	0.296 (0.271,0.324)
		0.05	0.431 (0.401,0.462)	0.462 (0.433,0.493)	0.296 (0.271,0.324)
		0.10	0.431 (0.401,0.462)	0.491 (0.46,0.522)	0.296 (0.271,0.324)
	0.90	0.01	0.851 (0.829,0.871)	0.886 (0.866,0.906)	0.68 (0.652,0.711)
		0.05	0.851 (0.829,0.871)	0.913 (0.894,0.931)	0.68 (0.652,0.711)
		0.10	0.851 (0.829,0.871)	0.934 (0.919,0.948)	0.68 (0.652,0.711)
	0.99	0.01	0.938 (0.921,0.953)	0.961 (0.949,0.972)	0.769 (0.743,0.795)
		0.05	0.938 (0.921,0.953)	0.995 (0.99,0.999)	0.769 (0.743,0.795)
		0.10	0.938 (0.921,0.953)	1 (1,1)	0.769 (0.743,0.795)
m=1000					
	0.50	0.01	0.4 (0.369,0.428)	0.466 (0.437,0.497)	0.313 (0.283,0.342)
		0.05	0.4 (0.369,0.428)	0.496 (0.465,0.526)	0.313 (0.283,0.342)
		0.10	0.4 (0.369,0.428)	0.492 (0.461,0.521)	0.313 (0.283,0.342)
	0.90	0.01	0.842 (0.819,0.864)	0.902 (0.883,0.922)	0.724 (0.694,0.751)
		0.05	0.842 (0.819,0.864)	0.91 (0.892,0.927)	0.724 (0.694,0.751)
		0.10	0.842 (0.819,0.864)	0.937 (0.922,0.952)	0.724 (0.694,0.751)
	0.99	0.01	0.938 (0.923,0.952)	0.974 (0.963,0.984)	0.832 (0.808,0.854)
		0.05	0.938 (0.923,0.952)	0.994 (0.988,0.998)	0.832 (0.808,0.854)
		0.10	0.938 (0.923,0.952)	0.999 (0.997,1)	0.832 (0.808,0.854)

Table A.2: Observed out-of-sample frequencies of observations smaller than or equal to current quantile estimates for exponential distribution with drift in mean and variance. Bootstrap 95% confidence intervals provided in brackets.

	p	Lambda	Frequency GH	Frequency EWGH	Frequency EWSA
Distribution: Change Point					
m=100					
	0.50	0.01	0.574 (0.545,0.605)	0.908 (0.891,0.925)	0.973 (0.962,0.983)
		0.05	0.574 (0.545,0.605)	0.348 (0.319,0.377)	0.973 (0.962,0.983)
		0.10	0.574 (0.545,0.605)	0.453 (0.422,0.485)	0.973 (0.962,0.983)
	0.90	0.01	0.999 (0.997,1)	0.988 (0.981,0.994)	1 (1,1)
		0.05	0.999 (0.997,1)	0.815 (0.789,0.839)	1 (1,1)
		0.10	0.999 (0.997,1)	0.893 (0.872,0.912)	1 (1,1)
	0.99	0.01	1 (1,1)	1 (1,1)	1 (1,1)
		0.05	1 (1,1)	0.973 (0.963,0.982)	1 (1,1)
		0.10	1 (1,1)	0.996 (0.991,0.999)	1 (1,1)
m=400					
	0.50	0.01	0.598 (0.566,0.628)	0.305 (0.277,0.337)	0.846 (0.824,0.868)
		0.05	0.598 (0.566,0.628)	0.44 (0.407,0.472)	0.846 (0.824,0.868)
		0.10	0.598 (0.566,0.628)	0.486 (0.453,0.517)	0.846 (0.824,0.868)
	0.90	0.01	0.997 (0.993,1)	0.847 (0.824,0.869)	0.997 (0.993,1)
		0.05	0.997 (0.993,1)	0.904 (0.885,0.923)	0.997 (0.993,1)
		0.10	0.997 (0.993,1)	0.931 (0.915,0.946)	0.997 (0.993,1)
	0.99	0.01	1 (1,1)	0.983 (0.975,0.99)	1 (1,1)
		0.05	1 (1,1)	0.996 (0.992,1)	1 (1,1)
		0.10	1 (1,1)	1 (1,1)	1 (1,1)
m=1000					
	0.50	0.01	0.566 (0.534,0.596)	0.457 (0.425,0.489)	0.525 (0.495,0.557)
		0.05	0.566 (0.534,0.596)	0.446 (0.416,0.479)	0.525 (0.495,0.557)
		0.10	0.566 (0.534,0.596)	0.465 (0.435,0.496)	0.525 (0.495,0.557)
	0.90	0.01	1 (1,1)	0.84 (0.818,0.862)	0.902 (0.883,0.92)
		0.05	1 (1,1)	0.903 (0.885,0.921)	0.902 (0.883,0.92)
		0.10	1 (1,1)	0.936 (0.921,0.951)	0.902 (0.883,0.92)
	0.99	0.01	1 (1,1)	0.968 (0.957,0.978)	1 (1,1)
		0.05	1 (1,1)	0.994 (0.989,0.998)	1 (1,1)
		0.10	1 (1,1)	0.999 (0.997,1)	1 (1,1)

Table A.3: Observed out-of-sample frequencies of observations smaller than or equal to current quantile estimates for simulated data with change point. Bootstrap 95% confidence intervals provided in brackets.

Appendix B

Properties of Hermite Series based Distribution Function Estimators

B.1 Lemmas

The first lemma is due to [30] (Lemma 1) which is restated without proof as Lemma 1 below:

Lemma 1 ([30])**.**

$$\lim_{N\to\infty} \sum_{k=0}^{N} a_k h_k(x) = f(x) \tag{B.1}$$

at every differentiability point of f*. If* $f \in L_p$*,* $p > 1$*, the convergence holds for almost all* $x \in \mathbb{R}$*.*

The second Lemma is due to [29] (equation (12) in that paper) which is restated without proof as Lemma 2 below:

Lemma 2 ([29]). *For the Hermite series estimators* (2.18)*, if* $E|X|^{2/3} < \infty$*:*

$$\sum_{k=0}^{N} E\left(\hat{a}_k - a_k\right)^2 = O\left(\frac{N^{1/2}}{n}\right). \tag{B.2}$$

The third Lemma is due to [29] (following from equation (15) in Theorem 4 of [29]), which we restate without proof as Lemma 3 below:

Lemma 3 ([29]). *For the Hermite series estimators* (2.18)*, if* $E|X|^s < \infty$*,* $s > 8(r+1)/3(2r+1)$ *then:*

$$\sum_{k=0}^{N} \left(\hat{a}_k - a_k\right)^2 = O(n^{-2r/(2r+1)} \log n) a.s. \tag{B.3}$$

Finally, we present an important novel result with proof in Lemma 4 below. We make use of Lemma 4 several times in this thesis.

Lemma 4.

$$\int_{-\infty}^{x} |h_k(t)|dt \leq 2c_1(k+1)^{-\frac{1}{4}} + 12d_1(k+1)^{\frac{1}{2}}, \tag{B.4}$$

where c_1 *and* d_1 *are positive constants.*

Proof.

$$\begin{aligned}
\int_{-\infty}^{x} |h_k(t)|dt &\leq \int_{-\infty}^{\infty} |h_k(t)|dt && \text{(B.5)}\\
&= \int_{-\infty}^{-1} |h_k(t)|dt + \int_{-1}^{1} |h_k(t)|dt + \int_{1}^{\infty} |h_k(t)|dt && \text{(B.6)}\\
&= \int_{-1}^{1} |h_k(t)|dt + 2\int_{1}^{\infty} |h_k(t)|dt && \text{(B.7)}\\
&\leq 2c_1(k+1)^{-\frac{1}{4}} + \frac{2d_1}{b}(k+1)^{\frac{5}{12}+\frac{b}{2}},\ b > 0. && \text{(B.8)}
\end{aligned}$$

This follows from the inequalities implied by Theorem 8.91.3 of [63], namely, $\max_{|x|\leq a} h_k(x) \leq c_a(k+1)^{-\frac{1}{4}}$ and $\max_{|x|\geq a} |h_k(x)|x^{\lambda} \leq d_a(k+1)^s$ where c_a

and d_a are positive constants depending only on a, $s = \max(\frac{\lambda}{2} - \frac{1}{12}, -\frac{1}{4})$ and we have set $\lambda = 1 + b$, $b > 0$. In addition, we have made use of $\int_1^\infty x^{-1-b} = \frac{1}{b}$, $b > 0$. For concreteness we have set $b = \frac{1}{6}$. □

B.2 Proofs of Propositions and Theorems

B.2.1 Proof of Proposition 3

Proof.

$$\left|E[\hat{F}_N(x)] - F(x)\right| = \left|E\left[\int_{-\infty}^{x} \hat{f}_N(t)dt\right] - \int_{-\infty}^{x} f(t)dt\right| \tag{B.9}$$

$$\leq \int_{-\infty}^{x} \sum_{k=N+1}^{\infty} |a_k||h_k(t)|dt. \tag{B.10}$$

This follows from (2.18), (5.4), the fact that $E(\hat{a}_k) = a_k$ and Lemma 1. By the monotone convergence theorem we have,

$$\int_{-\infty}^{x} \sum_{k=N+1}^{\infty} |a_k||h_k(t)|dt = \sum_{k=N+1}^{\infty} |a_k| \int_{-\infty}^{x} |h_k(t)|dt. \tag{B.11}$$

Utilising Lemma 4 we have:

$$\sum_{k=N+1}^{\infty} |a_k| \int_{-\infty}^{x} |h_k(t)|dt \tag{B.12}$$

$$\leq \; 2c_1 \sum_{k=N+1}^{\infty} |a_k|(k+1)^{-\frac{1}{4}} + 12d_1 \sum_{k=N+1}^{\infty} |a_k|(k+1)^{\frac{1}{2}} \tag{B.13}$$

$$\leq \; 2c_1 \sum_{k=N+1}^{\infty} |b_{k+r}|(k+1)^{-\frac{1}{4}-\frac{r}{2}} + 12d_1 \sum_{k=N+1}^{\infty} |b_{k+r}|(k+1)^{\frac{1}{2}-\frac{r}{2}} \tag{B.14}$$

$$\leq \; 2c_1 ||(x - \frac{d}{dx})^r f(x)|| \sqrt{\sum_{k=N+1}^{\infty} (k+1)^{-\frac{1}{2}-r}}$$

$$+ \; 12d_1 ||(x - \frac{d}{dx})^r f(x)|| \sqrt{\sum_{k=N+1}^{\infty} (k+1)^{1-r}}, \tag{B.15}$$

where we have also used the fact that by assumption, $(x - \frac{d}{dx})^r f(x) \in L_2$ and [67] has shown $a_k^2 \leq \frac{b_{k+r}^2}{(k+1)^r}$, where b_k is the k-th coefficient of the expansion of $(x - \frac{d}{dx})^r f(x)$. In addition, $||(x - \frac{d}{dx})^r f(x)||^2 = \sum_{k=0}^{\infty} b_k^2$ by Parseval's theorem and we have made use of the Cauchy-Schwarz inequality in the last line. Using the well known properties of the Hurwitz Zeta function, $\zeta(s, a) = \sum_{k=0}^{\infty}(k+a)^{-s}$, (see [23], 25.11.43), we have:

$$\sum_{k=N+1}^{\infty} |a_k| \int_{-\infty}^{x} |h_k(t)|dt = O(N^{-r/2+1}), \tag{B.16}$$

completing the proof. □

B.2.2 Proof of Proposition 4

Proof. It is easy to see that

$$\left|\hat{F}_N(x) - E[\hat{F}_N(x)]\right| = \left|\sum_{k=0}^{N}(\hat{a}_k - a_k)\int_{-\infty}^{x} h_k(t)dt\right| \tag{B.17}$$

$$\leq \sqrt{\sum_{k=0}^{N}(\hat{a}_k - a_k)^2}\sqrt{\sum_{k=0}^{N}\left|\int_{-\infty}^{x} h_k(t)dt\right|^2}. \tag{B.18}$$

Now by virtue of Lemma 4 we have:

$$\left[\hat{F}_N(x) - E[\hat{F}_N(x)]\right]^2 = \sum_{k=0}^{N}(\hat{a}_k - a_k)^2 O\left(N^2\right). \tag{B.19}$$

Making use of Lemma 2 we have,

$$E\left[\hat{F}_N(x) - E[\hat{F}_N(x)]\right]^2 = O\left(\frac{N^{\frac{5}{2}}}{n}\right). \tag{B.20}$$

□

B.2.3 Proof of Theorem 10

Proof. We begin by restating the definition of the rate of almost sure convergence provided in [29]: for a sequence of random variables Y_n, we say that $Y = O(a_n)$ almost surely if $\frac{\beta n Y_n}{a_n} \to 0$ almost surely as $n \to \infty$, for all (non-negative) sequences $\{\beta_n\}$ convergent to zero. Now,

$$\left|\hat{F}_N(x) - F(x)\right| \leq \left|E[\hat{F}_N(x)] - F(x)\right| + \left|\hat{F}_N(x) - E[\hat{F}_N(x)]\right|. \tag{B.21}$$

By proposition 3,

$$\left|E[\hat{F}_N(x)] - F(x)\right| = O(N^{-r/2+1}) \tag{B.22}$$

$$= \quad O(n^{-(r-2)/(2r+1)}). \tag{B.23}$$

In addition, via (B.19) we have:

$$\left|\hat{F}_N(x) - E[\hat{F}_N(x)]\right| = \sqrt{\sum_{k=0}^{N} (\hat{a}_k - a_k)^2 O\left(N\right)}.$$

We make use of lemma 3 to obtain,

$$\left|\hat{F}_N(x) - E[\hat{F}_N(x)]\right| = O(n^{-(r-2)/(2r+1)} \log n)\, a.s., \tag{B.24}$$

and finally:

$$\left|\hat{F}_N(x) - F(x)\right| = O(n^{-(r-2)/(2r+1)} \log n)\, a.s. \tag{B.25}$$

□

B.2.4 Proof of Theorem 11

Proof. It suffices to prove $\sum_{n=1}^{\infty} P\left(|\hat{F}_{N(n)}(x) - F(x)| > \epsilon\right) < \infty$ for all $\epsilon > 0$ (Borel-Cantelli). We have via the law of total probability,

$$\begin{aligned}
&\sum_{n=1}^{\infty} P\left(|\hat{F}_{N(n)}(x) - F(x)| > \epsilon\right) \\
&\quad = \sum_{n=1}^{\infty} P\left(|\hat{F}_{N(n)}(x) - F(x)| > \epsilon \middle| N(n) > cn^{\gamma}\right) P\left(N(n) > cn^{\gamma}\right) \\
&\quad + \sum_{n=1}^{\infty} P\left(|\hat{F}_{N(n)}(x) - F(x)| > \epsilon \middle| N(n) \leq cn^{\gamma}\right) P\left(N(n) \leq cn^{\gamma}\right),
\end{aligned} \tag{B.26}$$

where c is a constant. By the assumption that $\sum_{n=1}^{\infty} P\left(\frac{N(n)}{n^{\gamma}} > \epsilon\right) < \infty$ for all $\epsilon > 0$, it is clear that the first term is finite. It remains to show that $\sum_{n=1}^{\infty} P\left(|\hat{F}_{N(n)}(x) - F(x)| > \epsilon \middle| N(n) \leq cn^{\gamma}\right) < \infty$ for all $\epsilon > 0$. By the conditional Markov inequality we have:

$$P\left(|\hat{F}_{N(n)}(x) - F(x)| > \epsilon | N(n) = q(n)\right) \leq \epsilon^{-p} E \left| \int_{-\infty}^{x} \sum_{k=0}^{q(n)} (\hat{a}_k - a_k) h_k(t) dt - \int_{-\infty}^{x} \sum_{k=q(n)+1}^{\infty} a_k h_k(t) dt \right|^p \quad \text{(B.27)}$$

for all $\epsilon > 0$. Using the fact that $|f+g|^p \leq 2^{p-1}(|f|^p + |g|^p)$ along with the Hölder inequality, lemma 4 and proposition 3 we have,

$$\begin{aligned}
& P\left(|\hat{F}_{N(n)}(x) - F(x)| > \epsilon | N(n) = q(n)\right) \\
\leq\ & \epsilon^{-p} 2^{p-1} \left(\sum_{k=0}^{q(n)} E|\hat{a}_k - a_k|^p \right) \left(\sum_{k=0}^{q(n)} \left(\int_{-\infty}^{x} |h_k(t)| dt \right)^{p/(p-1)} \right)^{p-1} \\
+\ & \epsilon^{-p} 2^{p-1} \left| \int_{-\infty}^{x} \sum_{k=q(n)+1}^{\infty} a_k h_k(t) dt \right|^p \\
\leq\ & \epsilon^{-p} 2^{p-1} b_1 \left(\sum_{k=0}^{q(n)} E|\hat{a}_k - a_k|^p \right) \left(\sum_{k=0}^{q(n)} (k+1)^{p/2(p-1)} \right)^{p-1} \\
+\ & \epsilon^{-p} 2^{p-1} b_2 q(n)^{-rp/2+p}, \qquad \text{(B.28)}
\end{aligned}$$

for all $\epsilon > 0$, where b_1, b_2 are positive constants. Now, the results of [21] imply $E|\hat{a_k} - a_k|^p = n^{-p} E|\sum_{i=1}^{n} (h_k(\mathbf{x_i}) - a_k)|^p \leq F_p n^{-p/2-1} \sum_{i=1}^{n} E|h_k(\mathbf{x_i}) - a_k|^p$, where F_p is a constant depending only on p. Also noting that $max_x |h_k(x)| \leq C(k+1)^{-1/12}$ where C is a positive constant (implied by Theorem 8.91.3 of [63]), we have:

$$
\begin{aligned}
& P\left(|\hat{F}_{N(n)}(x) - F(x)| > \epsilon | N(n) = q(n)\right) \\
\leq \;& \epsilon^{-p} 2^{p-1} b_3 n^{-p/2} q(n)^{-p/12+1} q(n)^{3p/2-1} + \epsilon^{-p} 2^{p-1} b_2 q(n)^{-rp/2+p}
\end{aligned} \tag{B.29}
$$

for all $\epsilon > 0$, where b_3 depends only on p. It is easy to see that for $r > 2$ and $q(n) = O(n^{\gamma})$, $0 < \gamma < 6/17$, we can choose p such that
$\sum_{n=1}^{\infty} P\left(|\hat{F}_{N(n)}(x) - F(x)| > \epsilon | N(n) = q(n)\right) < \infty$ for all $\epsilon > 0$. □

B.2.5 Proof of Proposition 5

Proof. The fixed N hermite series estimator (5.4) (equal to (5.5)) can be represented as:

$$
T(x, \hat{F}_n) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{x} d_N(t, y) dy \right] d\hat{F}_n(t), \tag{B.30}
$$

where $d_N(t, y) = \sum_{k=0}^{N} h_k(t) h_k(y)$. The influence function and empirical influence function are:

$$
IF(x, x'; T, F) = \int_{-\infty}^{x} d_N(x', y) dy - \int_{-\infty}^{\infty} \int_{-\infty}^{x} d_N(t, y) dy dF(t), \tag{B.31}
$$

$$
IF(x, x'; T, \hat{F}_n) = \int_{-\infty}^{x} d_N(x', y) dy - \int_{-\infty}^{\infty} \int_{-\infty}^{x} d_N(t, y) dy d\hat{F}_n(t), \tag{B.32}
$$

Now, for fixed N,

$$
\begin{aligned}
|\int_{-\infty}^{x} d_N(t,y)dy| &\leq \sum_{k=0}^{N} |h_k(t)| \int_{\infty}^{x} |h_k(y)| \\
&\leq u_1 \sum_{k=0}^{N} (k+1)^{-1/12-1/4} + v_1 \sum_{k=0}^{N} (k+1)^{1/2-1/12} \\
&< \infty, \qquad \text{(B.33)}
\end{aligned}
$$

where u_1 and v_1 are constants. The result (B.33) follows from Lemma 4 and the fact that $max_x|h_k(t)| \leq C(k+1)^{-1/12}$. Thus, the gross-error sensitivities, $sup_{x'}|IF(x,x';T,F)| < \infty$ and $sup_{x'}|IF(x,x';T,\hat{F}_n)| < \infty$ and the fixed N Hermite series cumulative distribution function estimator is Bias-robust. □

B.2.6 Proof of Proposition 6

Proof. The Kernel distribution function estimator is defined as:

$$
\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{x} \frac{1}{h} K\left(\frac{\mathbf{x}_i - y}{h}\right) dy. \qquad \text{(B.34)}
$$

This has the representation:

$$
T(x,\hat{F}_n) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{x} \frac{1}{h} K\left(\frac{t-y}{h}\right) dy\right] d\hat{F}_n(t), \qquad \text{(B.35)}
$$

where $\hat{F}_n$ is the empirical distribution function. The influence function and empirical influence function are easily seen to be:

$$
IF(x,x';T,F) = \left[\int_{-\infty}^{x} \frac{1}{h} K\left(\frac{x'-y}{h}\right) dy\right] - \int_{-\infty}^{\infty} \left[\int_{-\infty}^{x} \frac{1}{h} K\left(\frac{t-y}{h}\right) dy\right] dF(t), \qquad \text{(B.36)}
$$

$$IF(x, x'; T, \hat{F}_n) = \left[\int_{-\infty}^{x} \frac{1}{h} K\left(\frac{x'-y}{h}\right) dy\right] - \int_{-\infty}^{\infty} \left[\int_{-\infty}^{x} \frac{1}{h} K\left(\frac{t-y}{h}\right) dy\right] d\hat{F}_n(t). \tag{B.37}$$

Since $\int_{-\infty}^{\infty} K(u)du = 1$, it is clear that $sup_{x'}|IF(x, x'; T, F)| \leq 2 < \infty$, $sup_{x'}|IF(x, x'; T, \hat{F}_n)| \leq 2 < \infty$ and thus the smooth kernel distribution function estimator is Bias-robust. □

B.2.7 Proof of Proposition 7

Proof. Suppose a density function, $f(x)$, can be expanded formally as:

$$f(x) = \sum_{k=0}^{\infty} c_k He_k(x)\phi(x), \tag{B.38}$$

$$c_k = \frac{1}{k!} \int_{-\infty}^{\infty} f(x) He_k(x) dx, \tag{B.39}$$

where $\phi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$ and $He_k(x)$ are the Chebyshev-Hermite polynomials (following the notation of [63]). The truncated expansion has the form:

$$f(x) = \sum_{k=0}^{N} c_k He_k(x)\phi(x), \tag{B.40}$$

usually truncated to obtain:

$$f(x) = \phi(x)(1+\frac{1}{2}(\mu_2-1)He_2(x)+\frac{1}{6}\mu_3 He_3(x)+\frac{1}{24}(\mu_4-6\mu_2+3)He_4), \tag{B.41}$$

where μ_2, μ_3, μ_4 are non-central moments. This is the Gram-Charlier series of Type A [42]. A natural cumulative distribution function estimator based on the Gram-Charlier series is:

$$\hat{F}_N(x) = \sum_{k=0}^{N} \hat{c}_k \int_{-\infty}^{x} He_k(y)\phi(y)dy, \tag{B.42}$$

this has a representation

$$T(x, \hat{F}_n) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{x} \sum_{k=0}^{N} \frac{1}{k!} He_k(t) He_k(y)\phi(y)dy \right] d\hat{F}_n(t). \tag{B.43}$$

Now:

$$IF(x, x'; T, F)$$
$$= \sum_{k=0}^{N} \frac{1}{k!} He_k(x') \int_{-\infty}^{x} He_k(y)\phi(y)dy - \int_{-\infty}^{\infty} \int_{-\infty}^{x} \frac{1}{k!} He_k(t) He_k(y)\phi(y)dydF(t), \tag{B.44}$$

$$IF(x, x'; T, \hat{F}_n)$$
$$= \sum_{k=0}^{N} \frac{1}{k!} He_k(x') \int_{-\infty}^{x} He_k(y)\phi(y)dy - \int_{-\infty}^{\infty} \int_{-\infty}^{x} \frac{1}{k!} He_k(t) He_k(y)\phi(y)dyd\hat{F}_n. \tag{B.45}$$

Since $He_k(x')$ is not bounded, whereas the second term of $IF(x, x'; T, F)$ and $IF(x, x'; T, \hat{F}_n)$ is bounded, the gross-error sensitivities, $sup_{x'}|IF(x, x'; T, F)|$ and $sup_{x'}|IF(x, x'; T, \hat{F}_n)|$ are not bounded and thus the CDF estimator (B.42) is not B-robust. □

Appendix C

Sequential Spearman's Correlation Estimation

C.1 Proof of Theorem 12

In the proof of theorem 12, we make use of the following two lemmas proved below.

Lemma 5. *For the Hermite series based distribution function estimators of the marginal cumulative distribution functions $F^{(1)}(x)$ and $F^{(2)}(y)$, we have,*

$$\max_x(|\hat{F}_N^{(1)}(x)|) = O(N^{17/12}) \tag{C.1}$$

and

$$\max_y(|\hat{F}_N^{(2)}(y)|) = O(N^{17/12}). \tag{C.2}$$

Proof. Using (5.4) and lemma 4,

$$
\begin{aligned}
&|\hat{F}_N^{(1)}(x)| \\
&\leq \sum_{k=0}^{N} |\hat{a}_k| \int_{-\infty}^{x} |h_k(t)| dt \\
&\leq u_1 \sum_{k=0}^{N} (k+1)^{-\frac{1}{3}} + v_1 \sum_{k=0}^{N} (k+1)^{\frac{5}{12}} \\
&= O(N^{17/12}), \qquad \text{(C.3)}
\end{aligned}
$$

where u_1 and v_1 are positive constants. Note that we have used the fact that $max_x|h_k(x)| \leq C(k+1)^{-1/12}$ where C is a positive constant (implied by Theorem 8.91.3 of [63]) which implies $max_x|\hat{a}_k| \leq C(k+1)^{-1/12}$. Similarly $\max_y(|\hat{F}_N^{(2)}(y)|) = O(N^{17/12})$.

□

Lemma 6. *Suppose $f(x,y) \in L_2$ and $(x-\partial_x)^{r_3}(y-\partial_y)^{r_3} f(x,y) \in L_2$, $r_3 > 2$. In addition, suppose $E|XY|^{2/3} < \infty$ and that $N_3(n) = N_4(n) = O(n^{1/(2r_3+1)})$ as $n \to \infty$, then we have,*

$$
\begin{aligned}
E \int |\hat{f}_{N_3N_4}(x,y) - f(x,y)| dxdy &= O(n^{-(r_3-2)/(2r_3+1)}) \\
&= o(1) \text{ uniformly in } x \text{ and } y, \qquad \text{(C.4)}
\end{aligned}
$$

where $\hat{f}_{N_3N_4}(x,y)$ (2.24) is the bivariate Hermite series estimator.

Proof. From the inequalities implied by Theorem 8.91.3 of [63], namely,

$$
\max_{|x| \leq a} h_k(x) \leq c_a (k+1)^{-\frac{1}{4}} \qquad \text{(C.5)}
$$

and

$$\max_{|x|\geq a} |h_k(x)|x^{\lambda} \leq d_a(k+1)^s, \tag{C.6}$$

where c_a and d_a are constants, $s = \max(\frac{\lambda}{2} - \frac{1}{12}, -\frac{1}{4})$, we have:

$$\begin{aligned}
&E(\hat{A}_{kl} - A_{kl})^2 \\
&= \mathrm{Var}(\hat{A}_{kl}) \\
&= \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} h_k(x_i)h_l(y_i)) \\
&= \frac{1}{n}\mathrm{Var}(h_k(x)h_l(y)) \\
&\leq \frac{1}{n}E\left[(h_k(x))^2(h_l(y))^2\right] \\
&\leq \frac{c}{n}E(|XY|^{2/3})(k+1)^{-1/2}(l+1)^{-1/2},
\end{aligned} \tag{C.7}$$

where we have set $\lambda = -1/3$ in the inequality (C.6) and c is a constant. Note that this is satisfied when $E|X|^{4/3} < \infty$ and $E|Y|^{4/3} < \infty$ by the Cauchy-Schwarz inequality.

Now using the definition of (2.24) along with the monotone convergence theorem and the Lyapunov inequality we have,

$$E\int|\hat{f}_{N_3N_4}(x,y)-f(x,y)|dxdy \tag{C.8}$$

$$=\sum_{k=0}^{N_3}\sum_{l=0}^{N_4}E|A_{kl}-\hat{A}_{kl}|\int_{-\infty}^{\infty}|h_k(x)|dx\int_{-\infty}^{\infty}|h_l(y)|dy \tag{C.9}$$

$$+\sum_{k=N_3+1}^{\infty}\sum_{l=N_4+1}^{\infty}|A_{kl}|\int_{-\infty}^{\infty}|h_k(x)|dx\int_{-\infty}^{\infty}|h_l(y)|dy \tag{C.10}$$

$$\leq\sum_{k=0}^{N_3}\sum_{l=0}^{N_4}\sqrt{E|A_{kl}-\hat{A}_{kl}|^2}\int_{-\infty}^{\infty}|h_k(x)|dx\int_{-\infty}^{\infty}|h_l(y)|dy \tag{C.11}$$

$$+\sum_{k=N_3+1}^{\infty}\sum_{l=N_4+1}^{\infty}|A_{kl}|\int_{-\infty}^{\infty}|h_k(x)|dx\int_{-\infty}^{\infty}|h_l(y)|dy. \tag{C.12}$$

In a similar way to [67], if $(x-\partial_x)^{r_3}(y-\partial_y)^{r_3}f(x,y)\in L_2$, then,

$$|A_{kl}|\leq|B_{kl}|2^{-r_3}(k+1)^{-r_3/2}(l+1)^{-r_3/2}, \tag{C.13}$$

where B_{kl} are the bivariate Hermite expansion coefficients of $(x-\partial_x)^{r_3}(y-\partial_y)^{r_3}f(x,y)$.

Utilising the results for the variance of $\hat{A}_{kl}$ above along with lemma 4, (C.13) and the Cauchy-Schwarz inequality we have,

$$E\int|\hat{f}_{N_3N_4}(x,y)-f(x,y)|dxdy \tag{C.14}$$

$$=O\left(\frac{N_3^{5/4}N_4^{5/4}}{\sqrt{n}}\right)+O\left(N_3^{-r_3/2+1}N_4^{-r_3/2+1}\right). \tag{C.15}$$

Now if we set $N_3(n)=N_4(n)=O(n^{1/(2r_3+1)})$, we obtain:

$$E\int|\hat{f}_{N_3N_4}(x,y)-f(x,y)|dxdy=O(n^{-(r_3-2)/(2r_3+1)}). \tag{C.16}$$

□

We now present the proof of **Theorem** 12 below.

Proof. From (3.8) and (6.1) we have,

$$
\begin{aligned}
&|\hat{R}_{N_1,N_2,N_3,N_4} - \rho(F^{(1)}(X), F^{(2)}(Y))| \\
&= |12 \int (\hat{F}^{(1)}_{N_1}(x) - 1/2)(\hat{F}^{(2)}_{N_2}(y) - 1/2)\hat{f}_{N_3N_4}(x,y)dxdy \\
&- 12 \int (F^{(1)}(x) - 1/2)(F^{(2)}(y) - 1/2)f(x,y)dxdy| \\
&\leq 12| \int \hat{F}^{(1)}_{N_1}(x)\hat{F}^{(2)}_{N_2}(y)(\hat{f}_{N_3N_4}(x,y) - f(x,y))dxdy| \\
&+ 12| \int \hat{F}^{(1)}_{N_1}(x)(\hat{F}^{(2)}_{N_2}(y) - F^{(2)}(y))f(x,y))dxdy| \\
&+ 12| \int F^{(2)}(y)(\hat{F}^{(1)}_{N_1}(x) - F^{(1)}(x))f(x,y))dxdy| \\
&+ 6| \int \hat{F}^{(1)}_{N_1}(x)(\hat{f}_{N_3N_4}(x,y) - f(x,y))dxdy| \\
&+ 6| \int (\hat{F}^{(1)}_{N_1}(x) - F^{(1)}(x))f(x)dx| \\
&+ 6| \int \hat{F}^{(2)}_{N_2}(y)(\hat{f}_{N_3N_4}(x,y) - f(x,y))dxdy| \\
&+ 6| \int (\hat{F}^{(2)}_{N_2}(y) - F^{(2)}(y))f(y)dy| \\
&+ 3| \int (\hat{f}_{N_3N_4}(x,y) - f(x,y))dxdy|.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&|\hat{R}_{N_1,N_2,N_3,N_4} - \rho(F^{(1)}(X), F^{(2)}(Y))| \\
&\leq 12 \max_x(|\hat{F}^{(1)}_{N_1}(x)|) \max_y(|\hat{F}^{(2)}_{N_2}(y)|) \int |(\hat{f}_{N_3N_4}(x,y) - f(x,y))| dxdy \\
&+ 12 \max_x(|\hat{F}^{(1)}_{N_1}(x)|) \sqrt{\int (\hat{F}^{(2)}_{N_2}(y) - F^{(2)}(y))^2 f(y) dy} \\
&+ 12 \sqrt{\int (\hat{F}^{(1)}_{N_1}(x) - F^{(1)}(x))^2 f(x) dx} \\
&+ 6 \max_x(|\hat{F}^{(1)}_{N_1}(x)|) \int |(\hat{f}_{N_3N_4}(x,y) - f(x,y))| dxdy \\
&+ 6 \sqrt{\int (\hat{F}^{(1)}_{N_1}(x) - F^{(1)}(x))^2 f(x) dx} \\
&+ 6 \max_y(|\hat{F}^{(2)}_{N_2}(y)|) \int |(\hat{f}_{N_3N_4}(x,y) - f(x,y))| dxdy \\
&+ 6 \sqrt{\int (\hat{F}^{(2)}_{N_2}(y) - F^{(2)}(y))^2 f(y) dy} \\
&+ 3 \int (|\hat{f}_{N_3N_4}(x,y) - f(x,y))| dxdy,
\end{aligned}
$$

where we have applied the Cauchy-Schwarz inequality. Using lemma 5 and applying Jensen's inequality we obtain the following,

$$
\begin{aligned}
&E|\hat{R}_{N_1,N_2,N_3,N_4} - \rho(F^{(1)}(X), F^{(2)}(Y))| \\
&\leq O\left((N_1N_2)^{17/12}\right) E\int |\hat{f}_{N_3N_4}(x,y) - f(x,y)|dxdy \\
&+ O\left(N_1^{17/12}\right)\sqrt{E\int (\hat{F}^{(2)}_{N_2}(y) - F^{(2)}(y))^2 f(y)dy} \\
&+ 12\sqrt{E\int (\hat{F}^{(1)}_{N_1}(x) - F^{(1)}(x))^2 f(x)dx} \\
&+ O\left(N_1^{17/12}\right) E\int |\hat{f}_{N_3N_4}(x,y) - f(x,y)|dxdy \\
&+ 6\sqrt{E\int (\hat{F}^{(1)}_{N_1}(x) - F^{(1)}(x))^2 f(x)dx} \\
&+ O\left(N_2^{17/12}\right) E\int |\hat{f}_{N_3N_4}(x,y) - f(x,y)|dxdy \\
&+ 6\sqrt{E\int (\hat{F}^{(2)}_{N_2}(y) - F^{(2)}(y))^2 f(y)dy} \\
&+ 3E\int |\hat{f}_{N_3N_4}(x,y) - f(x,y)|dxdy.
\end{aligned}
$$

Using Lemma 6, we have,

$$
E\int |\hat{f}_{N_3N_4}(x,y) - f(x,y)|dxdy = O(n^{-(r_3-2)/(2r_3+1)}). \tag{C.17}
$$

In addition, given the assumptions 1 and 2 (also note that $E|X|^{4/3} < \infty$ and $E|Y|^{4/3} < \infty$ implies that $E|X|^{2/3} < \infty$ and $E|Y|^{2/3} < \infty$ by the Lyapunov inequality), theorem 9 implies,

$$
E\int (\hat{F}^{(1)}_{N_1}(x) - F^{(1)}(x))^2 f(x)dx = O(n^{-2(r_1-2)/(2r_1+1)}) \tag{C.18}
$$

and

$$E\int(\hat{F}_{N_2}^{(2)}(y)-F^{(2)}(y))^2 f(y)dy = O(n^{-2(r_2-2)/(2r_2+1)}). \tag{C.19}$$

Thus for $n \to \infty$:

$$\begin{aligned}
&E|\hat{R}_{N_1,N_2,N_3,N_4} - \rho(F^{(1)}(X), F^{(2)}(Y))| \\
&= O(n^{34/12(2r_1+1)+34/12(2r_2+1)-(r_3-2)/(2r_3+1)}) \\
&+ O(n^{34/12(2r_1+1)-(r_2-2)/(2r_2+1)}) \\
&+ O(n^{-(r_1-2)/(2r_1+1)}).
\end{aligned}$$

This implies that if we have $r_1, r_2, r_3 \geq 8$:

$$\begin{aligned}
&E|\hat{R}_{N_1,N_2,N_3,N_4} - \rho(F^{(1)}(X), F^{(2)}(Y))| \\
&= o(1).
\end{aligned}$$

□

C.2 Proof of Theorem 13

The proof of **Theorem** 13 is presented below.

Proof. We begin by noting that:

$$\begin{aligned}
\mathrm{Var}((\hat{a}_{(1)})_r) &= \lambda^2 \sum_{j=0}^{n-1} (1-\lambda)^{2j} \mathrm{Var}(h_r(x_{n-j})) + (1-\lambda)^{2n} \mathrm{Var}(h_r(x_0)) \\
&= \left[\frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2n}) + (1-\lambda)^{2n} \right] \mathrm{Var}(h_r(x)).
\end{aligned}$$

Thus for $n \to \infty$, we have $\mathrm{Var}((\hat{a}_{(1)})_r) \to \frac{\lambda}{2-\lambda}\mathrm{Var}(h_r(x))$.

Similarly, for $n \to \infty$:

$$\mathrm{Var}((\hat{a}_{(2)})_s) \to \frac{\lambda}{2-\lambda}\mathrm{Var}(h_s(y))$$

$$\mathrm{Var}((\hat{A})_{uv}) \to \frac{\lambda}{2-\lambda}\mathrm{Var}(h_u(x)h_v(y))$$

$$\mathrm{cov}((\hat{a}_{(1)})_r, (\hat{a}_{(2)})_s) \to \frac{\lambda}{2-\lambda}cov(h_r(x), h_s(y))$$

$$\mathrm{cov}((\hat{a}_{(1)})_r, (\hat{A}_{uv})) \to \frac{\lambda}{2-\lambda}cov(h_r(x), h_u(x)h_v(y))$$

$$\mathrm{cov}((\hat{a}_{(2)})_s, (\hat{A}_{uv})) \to \frac{\lambda}{2-\lambda}cov(h_s(y), h_u(x)h_v(y))$$

Thus, all the variances and covariances above are $O(\lambda)$. An exact Taylor series expansion of $\hat{R}_N$ implies $E(\hat{R}_N) = \hat{R}_N|_{a_{(1)},a_{(2)},A} + O(\lambda)$ and facilitates a convenient reorganisation of the variance expression, $E\left(\hat{R}_N - E(\hat{R}_N)\right)^2$.

A straightforward but lengthy calculation yields the following for $n \to \infty$,

$$\begin{aligned}
\mathrm{Var}(\hat{R}_N) = &\left[\frac{\lambda}{2-\lambda}\right] \times \\
&\times [\sum_{r=0}^{N} (g_r^{(1)})^2 \mathrm{Var}(h_r(x)) + \sum_{s=0}^{N} (g_s^{(2)})^2 \mathrm{Var}(h_s(y)) \\
&+ \sum_{u,v=0}^{N} (g_{uv}^{(3)})^2 \mathrm{Var}(h_u(x) h_v(y)) \\
&+ \sum_{r,s=0}^{N} g_r^{(1)} g_s^{(2)} \mathrm{Cov}(h_r(x), h_s(y)) \\
&+ \sum_{r,u,v=0}^{N} g_r^{(1)} g_{uv}^{(3)} \mathrm{Cov}(h_r(x), h_u(x) h_v(y)) \\
&+ \sum_{s,u,v=0}^{N} g_s^{(2)} g_{uv}^{(3)} \mathrm{Cov}(h_s(y), h_u(x) h_v(y))] \\
&+ o(\lambda),
\end{aligned}$$

where

$$\begin{aligned}
g_r^{(1)} &= \frac{\partial \hat{R}_N}{\partial (\hat{a}_{(1)})_r}|_{a_{(1)}, a_{(2)}, A} = 12 \sum_{l,m,o} W_{lr} A_{lm} W_{mo} (a_{(2)})_o - 6 \sum_{l,m} W_{lr} A_{lm} z_m \\
g_s^{(2)} &= \frac{\partial \hat{R}_N}{\partial (\hat{a}_{(2)})_s}|_{a_{(1)}, a_{(2)}, A} = 12 \sum_{k,l,m} (a_{(1)})_k W_{lk} A_{lm} W_{ms} - 6 \sum_{k,m} z_k A_{km} W_{ms} \\
g_{uv}^{(3)} &= \frac{\partial \hat{R}_N}{\partial (\hat{A}_{uv})}|_{a_{(1)}, a_{(2)}, A} = 12 \sum_{k,o} (a_{(1)})_k W_{uk} W_{vo} (a_{(2)})_o - 6 \sum_{k} (a_{(1)})_k W_{uk} z_v \\
&\quad - 6 \sum_{o} z_u W_{vo} (a_{(2)})_o + 3 z_u z_v.
\end{aligned}$$

Note that this essentially corresponds to the delta method as described in [42]. □

C.3 Simulation results for Stationary Data

In this section we collect simulation results associated with section 6.5.1.

	MAE (x10^-2)		Standard Error MAE (x10^-2)		Parameter	
Rho	Hermite	Matrix	Hermite	Matrix	Hermite (N)	Matrix (c)
n=10,000						
-0.75	**0.185**	0.275	0.004	0.002	20	20
-0.50	**0.181**	0.234	0.004	0.002	20	20
-0.25	0.168	**0.135**	0.004	0.002	20	20
0.00	0.183	**0.057**	0.004	0.001	20	20
0.25	0.182	**0.141**	0.004	0.002	20	20
0.50	**0.182**	0.235	0.004	0.002	20	20
0.75	**0.178**	0.275	0.004	0.002	20	20
n=50,000						
-0.75	**0.08**	0.277	0.002	0.001	20	20
-0.50	**0.078**	0.236	0.002	0.001	20	20
-0.25	**0.078**	0.134	0.002	0.001	20	20
0.00	0.078	**0.025**	0.002	0.001	20	20
0.25	**0.076**	0.133	0.002	0.001	20	20
0.50	**0.079**	0.235	0.002	0.001	20	20
0.75	**0.082**	0.276	0.002	0.001	20	20
n=100,000						
-0.75	**0.054**	0.276	0.001	0.000	20	20
-0.50	**0.054**	0.236	0.001	0.001	20	20
-0.25	**0.057**	0.133	0.001	0.001	20	20
0.00	0.056	**0.017**	0.001	0.000	20	20
0.25	**0.056**	0.132	0.001	0.001	20	20
0.50	**0.053**	0.235	0.001	0.001	20	20
0.75	**0.055**	0.275	0.001	0.000	20	20

Table C.1: MAE results for Hermite series based Spearman's rank correlation estimator ($N = 20$) versus count matrix based Spearman's rank correlation estimator ($c = 20$) at $\mu = (0, 0)$ for $\sigma_1 = 1, \sigma_2 = 1$ and $\rho = -0.75, -0.5, -0.25, 0.25, 0.5, 0.75$. Lowest MAE values for a given n, ρ are presented in bold.

	MAE (x10^-2)		Standard Error MAE (x10^-2)		Parameter	
Rho	Hermite	Matrix	Hermite	Matrix	Hermite (N)	Matrix (c)
n=10,000						
-0.75	0.185	**0.125**	0.004	0.001	20	30
-0.50	0.181	**0.11**	0.004	0.001	20	30
-0.25	0.168	**0.068**	0.004	0.001	20	30
0.00	0.183	**0.039**	0.004	0.001	20	30
0.25	0.182	**0.064**	0.004	0.001	20	30
0.50	0.182	**0.111**	0.004	0.001	20	30
0.75	0.178	**0.123**	0.004	0.001	20	30
n=50,000						
-0.75	**0.08**	0.125	0.002	0.000	20	30
-0.50	**0.078**	0.11	0.002	0.001	20	30
-0.25	0.078	**0.062**	0.002	0.001	20	30
0.00	0.078	**0.017**	0.002	0.000	20	30
0.25	0.076	**0.063**	0.002	0.001	20	30
0.50	**0.079**	0.11	0.002	0.001	20	30
0.75	**0.082**	0.125	0.002	0.000	20	30
n=100,000						
-0.75	**0.054**	0.126	0.001	0.000	20	30
-0.50	**0.054**	0.109	0.001	0.000	20	30
-0.25	**0.057**	0.062	0.001	0.000	20	30
0.00	0.056	**0.012**	0.001	0.000	20	30
0.25	**0.056**	0.063	0.001	0.000	20	30
0.50	**0.053**	0.11	0.001	0.000	20	30
0.75	**0.055**	0.126	0.001	0.000	20	30

Table C.2: MAE results for Hermite series based Spearman's rank correlation estimator ($N = 20$) versus count matrix based Spearman's rank correlation estimator ($c = 30$) at $\mu = (0, 0)$ for $\sigma_1 = 1, \sigma_2 = 1$ and $\rho = -0.75, -0.5, -0.25, 0.25, 0.5, 0.75$. Lowest MAE values for a given n, ρ are presented in bold.

	Average MAE (x10^-2)		Standard Deviation MAE (x10^-2)	
	Hermite (N=20)	Matrix (c=20)	Hermite (N=20)	Matrix (c=20)
n=10,000				
0.18	0.193		0.006	0.083
n=50,000				
0.079	0.188		0.002	0.094
n=100,000				
0.055	0.187		0.001	0.096

Table C.3: Summarised MAE results for Hermite series based Spearman's rank correlation estimator ($N = 20$) versus count matrix based Spearman's rank correlation estimator ($c = 20$) at $\mu = (0, 0)$ for $\sigma_1 = 1, \sigma_2 = 1$ across all values of ρ. Lowest average MAE values for a given n are presented in bold.

	Average MAE (x10^-2)		Standard Deviation MAE (x10^-2)	
	Hermite (N=20)	Matrix (c=30)	Hermite (N=20)	Matrix (c=30)
n=10,000				
0.18		**0.091**	0.006	0.034
n=50,000				
0.079	0.087		0.002	0.041
n=100,000				
0.055	0.087		0.001	0.042

Table C.4: Summarized MAE results for Hermite series based Spearman's rank correlation estimator ($N = 20$) versus count matrix based Spearman's rank correlation estimator ($c = 30$) at $\mu = (0, 0)$ for $\sigma_1 = 1, \sigma_2 = 1$ across all values of ρ. Lowest average MAE values for a given n are presented in bold.

	Average MAE (x10^-2)		Standard Deviation MAE (x10^-2)	
	Hermite (N=20)	Matrix (c=20)	Hermite (N=20)	Matrix (c=20)
n=10,000				
0.897	1.072		0.527	0.592
n=50,000				
0.812	0.939		0.634	0.547
n=100,000				
0.81	0.909		0.675	0.540

Table C.5: Summarized MAE results for Hermite series based Spearman's rank correlation estimator ($N = 20$) versus count matrix based Spearman's rank correlation estimator ($c = 20$) for bivariate normal variables transformed as $(f(x_i), f(y_i))$ with $f(x) = \exp(x)$, across all values of ρ. Lowest average MAE values for a given n are presented in bold.

	Average MAE (x10^-2)		Standard Deviation MAE (x10^-2)	
	Hermite (N=20)	Matrix (c=30)	Hermite (N=20)	Matrix (c=30)
n=10,000				
0.897	**0.69**		0.527	0.361
n=50,000				
0.812	**0.549**		0.634	0.300
n=100,000				
0.81	**0.517**		0.675	0.289

Table C.6: Summarized MAE results for Hermite series based Spearman's rank correlation estimator ($N = 20$) versus count matrix based Spearman's rank correlation estimator ($c = 30$) for bivariate normal variables transformed as $(f(x_i), f(y_i))$ with $f(x) = \exp(x)$, across all values of ρ. Lowest average MAE values for a given n are presented in bold.

C.4 Exponentially weighted Pearson's correlation Estimator

The exponentially weighted version of the standard Pearson's product-moment correlation estimator is defined as follows:

$$\begin{aligned}\bar{X}_\lambda^{(1)} &= \mathbf{x}_1 \\ \bar{X}_\lambda^{(i)} &= \lambda \mathbf{x}_i + (1-\lambda)\bar{X}_\lambda^{(i-1)}, \quad i = 2, \ldots, n\end{aligned}$$

$$\begin{aligned}\bar{Y}_\lambda^{(1)} &= \mathbf{y}_1 \\ \bar{Y}_\lambda^{(i)} &= \lambda \mathbf{y}_i + (1-\lambda)\bar{Y}_\lambda^{(i-1)}, \quad i = 2, \ldots, n\end{aligned}$$

$$\begin{aligned}V(x)_\lambda^{(1)} &= 1 \\ V(x)_\lambda^{(i)} &= \lambda(\mathbf{x}_i - \bar{X}_\lambda^{(i)})^2 + (1-\lambda)V(x)_\lambda^{(i-1)}, \quad i = 2, \ldots, n\end{aligned}$$

$$\begin{aligned}V(y)_\lambda^{(1)} &= 1 \\ V(y)_\lambda^{(i)} &= \lambda(\mathbf{y}_i - \bar{Y}_\lambda^{(i)})^2 + (1-\lambda)V(y)_\lambda^{(i-1)}, \quad i = 2, \ldots, n\end{aligned}$$

$$\begin{aligned}C(x,y)_\lambda^{(1)} &= 1 \\ C(x,y)_\lambda^{(i)} &= \lambda(\mathbf{x}_i - \bar{X}_\lambda^{(i)})(\mathbf{y}_i - \bar{Y}_\lambda^{(i)}) + (1-\lambda)C(x,y)_\lambda^{(i-1)}, \quad i = 2, \ldots, n\end{aligned}$$

$$\hat{\rho}(x,y)^{i}_{\lambda} = \frac{C(x,y)^{(i)}_{\lambda}}{\sqrt{V(x)^{(i)}_{\lambda} V(y)^{(i)}_{\lambda}}} \tag{C.20}$$

Appendix D

R Code

The code below is written in R and illustrates how to practically implement the algorithms described in this thesis.

D.1 Helper Functions

The function below calculates the Hermite function values for a particular vector of x values and $k = 0, \ldots, N$ using the recurrence relation (2.2).

```
HermiteFunction <- function(N, x) {
    b <- matrix(nrow = N + 1, ncol = length(x), 0)
    normalization <- matrix(nrow = N + 1, ncol = length(x), 0)
    b[1, 1:length(x)] <- rep(1, length(x))
    for (i in c(1:(N + 1))) {
        normalization[i, 1:length(x)] <- 1/sqrt(2^(i - 1) *
            ↪ factorial(i - 1) * sqrt(pi)) *
            exp(-x^2/2)
    }
    if (N == 0) {
        return(b * normalization)
```

```
    }
    b[2, 1:length(x)] <- 2 * x
    if (N == 1) {
        return(b * normalization)
    }
    for (i in c(3:(N + 1))) {
        b[i, 1:length(x)] <- (2 * x * b[i - 1, 1:length(x)] - 2
            ↪  * (i - 2) * b[i -
            2, 1:length(x)])
    }
    return(b * normalization)
}
```

The next function calculates the integral of the Hermite functions for a particular vector of x values, $\int_{-\infty}^{x} h_k(t)dt$, $k = 0, \ldots, N$, using a the normalised form of the recurrence relation (4.3). This is used to calculate a Hermite series based estimate of the CDF at the associated vector of x values using the estimator (5.4). Note that we utilise the erfc function from the R package *pracma*.

```
HermiteIntegralVal <- function(N,x,h_vec=NULL) {
  if (is.null(h_vec)){
    h_vec <- HermiteFunction(N,x)
  }
  b<-matrix(nrow=N+1,ncol=length(x),0)
  b[1,1:length(x)]<- pi^(1/4)/sqrt(2) *erfc(-1*(1/sqrt(2))*x)
  if (N==0){
    return(b)
  }
  b[2,1:length(x)]<- -sqrt(2)/(pi^(1/4)) *exp(-(1/2)*x^2)
  if (N==1){
```

```
    return(b)
  }
  for (i in c(3:(N+1))) {
    b[i,1:length(x)]<- -sqrt(2/(i-1))*h_vec[i-1,] + sqrt((i-2)/
        ↪ (i-1)) *b[i-2,]
  }
  return(b)
}
```

The next function calculates the integral of the Hermite functions for a particular vector of x values, $\int_x^\infty h_k(t)dt$, $k = 0, \ldots, N$, using a normalised form of the recurrence relation (4.6). This is used to calculate a Hermite series based estimate of the CDF at the associated vector of x values using the estimator (4.5) restated in terms of the standard Hermite series coefficients. We have found this alternative estimator to be more accurate for quantile estimation empirically.

```
HermiteIntegralValQuantileEst <- function(N,x,h_vec=NULL) {
  if (is.null(h_vec)){
    h_vec <- HermiteFunction(N,x)
  }
  b<-matrix(nrow=N+1,ncol=length(x),0)
  b[1,1:length(x)]<- pi^(1/4)/sqrt(2) *erfc((1/sqrt(2))*x)
  if (N==0){
    return(b)
  }
  b[2,1:length(x)]<- sqrt(2)/(pi^(1/4)) *exp(-(1/2)*x^2)
  if (N==1){
    return(b)
  }
  for (i in c(3:(N+1))) {
```

```
    b[i,1:length(x)]<- sqrt(2/(i-1))*h_vec[i-1,] + sqrt((i-2)/(
        ↪ i-1)) *b[i-2,]
  }
  return(b)
}
```

Finally, the function below calculates the integral of the Hermite functions, $\int_{-\infty}^{\infty} h_k(t)dt$, $k = 0, \ldots, N$, using the normalised form of the recurrence relation (4.3).

```
HermiteIntegralValInf <- function(N) {
  b<-matrix(nrow=N+1,ncol=1,0)
  b[1,1]<- sqrt(2) * pi^(1/4)
  if (N==0){
    return(b)
  }
  b[2,1]<- 0
  if (N==1){
    return(b)
  }
  for (i in c(3:(N+1))) {
    b[i,1]<- sqrt((i-2)/(i-1)) *b[i-2,]
  }
  return(b)
}
```

D.2 Sequential Hermite Estimator Classes

D.2.1 CDF Estimation

For sequential and one-pass estimation of the CDF using Hermite series estimators as introduced in Chapter 4, we have:

```
HermiteCDFEstimator <- function(N=6, normalise=FALSE,
    ↪ exponential_weighting_lambda=NA)
{
  this <- list(
    numberOfCoeffs = N,
    coeffVec=rep(0,N+1),
    nObs=0,
    normaliseObs = normalise,
    runningMean=0,
    runningVariation=0,
    exponential_weighting=exponential_weighting_lambda
  )
  class(this) <- append(class(this),"HermiteCDFEstimator")
  return(this)
}

updateSequential <- function(this, x, h_k)
{
  UseMethod("updateSequential",this)
}

updateSequential.HermiteCDFEstimator <- function(this, x, h_k=
    ↪ NULL)
{
```

```
this$nObs <- this$nObs +1
if (this$normaliseObs==TRUE){
  if (is.na(this$exponential_weighting)){
    prev_runningMean <- this$runningMean
    this$runningMean <- (this$runningMean*(this$nObs-1) + x)/
       ↪ this$nObs
    if (this$nObs < 2){
      return(this)
    }
    this$runningVariation <- this$runningVariation + (x-prev_
       ↪ runningMean)*(x-this$runningMean)
    running_std <- sqrt(this$runningVariation/(this$nObs-1))
  } else {
    if (this$nObs < 2){
      this$runningMean <- x
      this$runningVariation <- 1
      return(this)
    }
    this$runningMean <- (1-this$exponential_weighting)* this$
       ↪ runningMean + this$exponential_weighting*x
    this$runningVariation <- (1-this$exponential_weighting)*
       ↪ this$runningVariation + this$exponential_weighting*
       ↪ (x-this$runningMean)^2
    running_std <- sqrt(this$runningVariation)
  }
  x <- (x-this$runningMean)/running_std
}
if (is.null(h_k)){
  h_k <- HermiteFunction(this$numberOfCoeffs,x)
}
```

```
  if (is.na(this$exponential_weighting)){
    this$coeffVec <- (this$coeffVec*(this$nObs-1) + h_k)/this$
        ↪ nObs
  } else {
    this$coeffVec <- this$coeffVec*(1-this$exponential_
        ↪ weighting)+ h_k*this$exponential_weighting
  }
  return(this)
}

updateBatch<- function(this, x)
{
  UseMethod("updateBatch",this)
}

updateBatch.HermiteCDFEstimator <- function(this, x)
{
  this$nObs <- this$nObs + length(x)
  if (this$normaliseObs==TRUE){
    this$runningMean <- (this$runningMean * (this$nObs-length(x
        ↪ )) + length(x)*mean(x))/this$nObs
    this$runningVariation <- (this$runningVariation * (this$
        ↪ nObs-length(x)) + length(x)*var(x)*(length(x)-1))/
        ↪ this$nObs
    x <- (x - this$runningMean)/sqrt(this$runningVariation/(
        ↪ this$nObs-1))
  }
  h_k <- HermiteFunction(this$numberOfCoeffs,x)
  this$coeffVec <- (this$coeffVec*(this$nObs - length(x)) +
      ↪ rowSums(h_k))/this$nObs
```

```
  return(this)
}

getCumProbEst <- function(this, x)
{
  UseMethod("getCumProbEst",this)
}

getCumProbEst.HermiteCDFEstimator <- function(this, x)
{
  if (this$normaliseObs==TRUE){
    if (this$nObs < 2){
      return(NaN)
    }
    if (is.na(this$exponential_weighting)){
      running_std <- sqrt(this$runningVariation/(this$nObs-1))
    } else {
      running_std <- sqrt(this$runningVariation)
    }
    x <- (x-this$runningMean)/running_std
  }
  integrals_hermite <- HermiteIntegralVal(this$numberOfCoeffs,x
     ↪ )
  cdf_val<-as.vector(as.vector(this$coeffVec)%*%integrals_
     ↪ hermite)
  return(cdf_val)
}

getCumProbEstQuantileAdapted <- function(this, x)
{
```

```
  UseMethod("getCumProbEstQuantileAdapted",this)
}

getCumProbEstQuantileAdapted.HermiteCDFEstimator <- function(
    ↪ this, x)
{
  if (this$normaliseObs==TRUE){
    if (this$nObs < 2){
      return(NaN)
    }
    if (is.na(this$exponential_weighting)){
      running_std <- sqrt(this$runningVariation/(this$nObs-1))
    } else {
      running_std <- sqrt(this$runningVariation)
    }
    x <- (x-this$runningMean)/running_std
  }
  integrals_hermite <- HermiteIntegralValQuantileEst(this$
      ↪ numberOfCoeffs,x)
  cdf_val<- 1-as.vector(as.vector(this$coeffVec)%*%integrals_
      ↪ hermite)
  return(cdf_val)
}
```

D.2.2 Quantile Estimation

For sequential and one-pass estimation of quantiles using Hermite series estimators as described in chapter 4, we have:

```
HermiteQuantileEstimator <- function(N=6,exponential_weighting_
    ↪ lambda=NA)
{
  this <- list(
    numberOfCoeffs = N,
    exponential_weighting=exponential_weighting_lambda,
    cdfEstimator=HermiteCDFEstimator(N = N,normalise = T,
        ↪ exponential_weighting_lambda = exponential_weighting_
        ↪ lambda)
  )
  class(this) <- append(class(this),"HermiteQuantileEstimator")
  return(this)
}

updateSequential <- function(this, x)
{
  UseMethod("updateSequential",this)
}

updateSequential.HermiteQuantileEstimator <- function(this, x)
{
  this$cdfEstimator <- updateSequential(this$cdfEstimator,x)
  return(this)
}

updateBatch<- function(this, x)
{
  UseMethod("updateBatch",this)
}
```

```
updateBatch.HermiteQuantileEstimator <- function(this, x)
{
  this$cdfEstimator <- updateBatch(this$cdfEstimator,x)
  return(this)
}

getCumProbEstQuantile<- function(this, x)
{
  UseMethod("getCumProbEstQuantile",this)
}

getCumProbEstQuantile.HermiteQuantileEstimator <- function(this
    ↪ , x)
{
  integrals_hermite <- HermiteIntegralValQuantileEst(this$
      ↪ cdfEstimator$numberOfCoeffs,x)
  cdf_val<- 1-as.vector(as.vector(this$cdfEstimator$coeffVec)%*
      ↪ %integrals_hermite)
  return(cdf_val)
}

getQuantileEst<- function(this, p)
{
  UseMethod("getQuantileEst",this)
}

getQuantileEst.HermiteQuantileEstimator <- function(this, p)
{
  est <- uniroot(f = function(x){getCumProbEstQuantile(this,x)-
      ↪ p},interval = c(-100,100))$root
```

```
  if (is.na(this$exponential_weighting)){
    est <- est*sqrt(this$cdfEstimator$runningVariation/(this$
        ↪ cdfEstimator$nObs-1)) + this$cdfEstimator$runningMean
  } else {
    est <- est * sqrt(this$cdfEstimator$runningVariation) +
        ↪ this$cdfEstimator$runningMean
  }
  return(est)
}
```

D.2.3 Spearman's Rank Correlation Estimation

For sequential and one-pass estimation of Spearman's rank correlation coefficients using Hermite series estimators as described in chapter 6, we have:

```
HermiteSpearmansEstimator <- function(N1=6,N2=6,NCDF=6,
    ↪ normalise=FALSE, exponential_weighting_lambda=NA)
{
  this <- list(
    numberOfCoeffsX = N1,
    numberOfCoeffsY = N2,
    numberOfCoeffsCDF = NCDF,
    exponential_weighting = exponential_weighting_lambda,
    coeffMat=matrix(rep(0,(N1+1)*(N2+1)),nrow = (N1+1), ncol=(
        ↪ N2+1)),
    nObs=0,
    cdf_estimator_x=HermiteCDFEstimator(N=NCDF, exponential_
        ↪ weighting_lambda = exponential_weighting_lambda),
    cdf_estimator_y=HermiteCDFEstimator(N=NCDF,exponential_
        ↪ weighting_lambda = exponential_weighting_lambda),
```

```
    W=c(),
    z=c(),
    W_transpose=c(),
    z_transpose=c()
  )
  class(this) <- append(class(this),"HermiteSpearmansEstimator"
     ↪ )
  this <- initialise_estimator(this)
  return(this)
}

integrand <- function(x,r,s){
  hermite_vec <- HermiteFunction(max(s,r),x)
  hermite_integral_vec <- HermiteIntegralVal(r,x,hermite_vec)
  result <- hermite_integral_vec[r+1,]*hermite_vec[s+1,]
  return(result)
}

W_matrix <- function(max_r,max_s){
  result <- matrix(rep(0,(max_r+1)*(max_s+1)), nrow=(max_r+1),
     ↪ ncol=(max_s+1), byrow = TRUE)
  for (r in c(0:max_r)) {
    for (s in c(0:max_s)) {
      result[r+1,s+1] <- integrate(function(t){integrand(t,r,s)
         ↪ },lower=-Inf,upper=Inf)$value
    }
  }
  return(result)
}
```

```
z_vector <- function(max_r){
  result <- HermiteIntegralValInf(max_r)
  return(result)
}

initialise_estimator <- function(this)
{
  UseMethod("initialise_estimator",this)
}

initialise_estimator.HermiteSpearmansEstimator <- function(this
    ↪ )
{
  n_max <- max(c(this$numberOfCoeffsX,this$numberOfCoeffsY,this
      ↪ $numberOfCoeffsCDF))
  this$W <- W_matrix(n_max,n_max)
  this$W_transpose <- t(this$W)
  this$z <- z_vector(n_max)
  this$z_transpose <- t(this$z)
  return(this)
}

updateSequential <- function(this, x,y)
{
  UseMethod("updateSequential",this)
}

updateSequential.HermiteSpearmansEstimator <- function(this, x,
    ↪  y)
{
```

```
  this$nObs <- this$nObs +1
  h_x <- HermiteFunction(this$numberOfCoeffsX,x)
  h_y <- HermiteFunction(this$numberOfCoeffsY,y)
  if (this$numberOfCoeffsX == this$numberOfCoeffsCDF & this$
      ↪ numberOfCoeffsY == this$numberOfCoeffsCDF){
    this$cdf_estimator_x<-updateSequential(this$cdf_estimator_x
        ↪ ,x,h_x)
    this$cdf_estimator_y <-updateSequential(this$cdf_estimator_
        ↪ y,y,h_y)
  } else {
    this$cdf_estimator_x<-updateSequential(this$cdf_estimator_x
        ↪ ,x)
    this$cdf_estimator_y <-updateSequential(this$cdf_estimator_
        ↪ y,y)
  }
  if (is.na(this$exponential_weighting)){
    this$coeffMat <- (this$coeffMat*(this$nObs-1) + h_x %*% t(h
        ↪ _y))/(this$nObs)
  } else {
    this$coeffMat<- this$coeffMat*(1-this$exponential_weighting
        ↪ ) + h_x %*% t(h_y)*this$exponential_weighting
  }
  return(this)
}

getSpearmansCorrelation <- function(this)
{
  UseMethod("getSpearmansCorrelation",this)
}
```

```
getSpearmansCorrelation.HermiteSpearmansEstimator <- function(
    ↪ this)
{
  term1 <- 12*t(this$cdf_estimator_x$coeffVec)%*%this$W_
     ↪ transpose[1:length(this$cdf_estimator_x$coeffVec),1:
     ↪ nrow(this$coeffMat)]%*%this$coeffMat%*%this$W[1:ncol(
     ↪ this$coeffMat),1:length(this$cdf_estimator_y$coeffVec)]
     ↪ %*%this$cdf_estimator_y$coeffVec
  term2 <- -6 * t(this$cdf_estimator_x$coeffVec)%*%this$W_
     ↪ transpose[1:length(this$cdf_estimator_x$coeffVec),1:
     ↪ nrow(this$coeffMat)]%*%this$coeffMat%*%this$z[1:ncol(
     ↪ this$coeffMat)]
  term3 <- -6 *this$z[1:nrow(this$coeffMat)]%*% this$coeffMat%*
     ↪ %this$W[1:ncol(this$coeffMat),1:length(this$cdf_
     ↪ estimator_y$coeffVec)]%*%this$cdf_estimator_y$coeffVec
  term4 <- 3 * this$z[1:nrow(this$coeffMat)]%*% this$coeffMat%*
     ↪ %this$z[1:ncol(this$coeffMat)]
  return(term1[1]+term2[1]+term3[1]+term4[1])
}
```

Bibliography

[1] Yacine Aït-Sahalia, Jianqing Fan, and Dacheng Xiu. High-frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association*, 105(492):1504–1517, 2010.

[2] Merve Alanyali, Helen Susannah Moat, and Tobias Preis. Quantifying the relationship between financial news and the stock market. *Scientific reports*, 3:3578, 2013.

[3] Naomi Altman and Christian Leger. Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, 46(2):195–214, 1995.

[4] Pierre-Olivier Amblard and Jean-Marc Brossier. Adaptive estimation of the fourth-order cumulant of a white stochastic process. *Signal processing*, 42(1):37–43, 1995.

[5] George B. Arfken and Hans J. Weber. *Mathematical methods for physicists*. Harcourt/Academic Press, Burlington, MA, fifth edition, 2001.

[6] G. Jogesh Babu, Angelo J. Canty, and Yogendra P. Chaubey. Application of Bernstein polynomials for smooth estimation of a distribution and density function. *J. Statist. Plann. Inference*, 105(2):377–392, 2002.

[7] Jeremy Becnel and Ambar Sengupta. The schwartz space: Tools for quantum mechanics and infinite dimensional analysis. *Mathematics*, 3(2):527–562, 2015.

[8] Janine Bennett, Ray Grout, Philippe Pébay, Diana Roe, and David Thompson. Numerically stable, single-pass, parallel statistics algorithms. In *2009 IEEE International Conference on Cluster Computing and Workshops*, pages 1–8. IEEE, 2009.

[9] Sergei Blinnikov and Richhild Moessner. Expansions for nearly gaussian distributions. *Astronomy and Astrophysics Supplement Series*, 130(1):193–205, 1998.

[10] Michael H Cahill, Diane Lambert, José C Pinheiro, and Don X Sun. Detecting fraud in the real world. In *Handbook of massive data sets*, pages 911–929. Springer, 2002.

[11] John M Chambers, David A James, Diane Lambert, and Scott Vander Wiel. Monitoring networked applications with incremental quantile estimation. *Statistical Science*, pages 463–475, 2006.

[12] Tony F Chan, Gene Howard Golub, and Randall J LeVeque. Updating formulae and a pairwise algorithm for computing sample variances. In *COMPSTAT 1982 5th Symposium held at Toulouse 1982*, pages 30–41. Springer, 1982.

[13] Fei Chen, Diane Lambert, and José C Pinheiro. Incremental quantile estimation for massive tracking. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 516–522. ACM, 2000.

[14] Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. *Mobile Networks and Applications*, 19(2):171–209, 2014.

[15] Zbigniew Ciesielski and Ryszard Zieliński. Polynomial and spline estimators of the distribution function with prescribed accuracy. *Appl. Math. (Warsaw)*, 36(1):1–12, 2009.

[16] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703, 2009.

[17] Christophe Croux and Catherine Dehon. Influence functions of the spearman and kendall correlation measures. *Statistical methods & applications*, 19(4):497–515, 2010.

[18] Harry F Davis. *Fourier series and orthogonal functions*. Courier Corporation, 1989.

[19] Doulaye Dembélé and Gérard Favier. Recursive estimation of fourth-order cumulants with application to identification. *Signal Processing*, 68(2):127–139, 1998.

[20] Luc Devroye and Laszlo Gyorfi. *Nonparametric density estimation: the L1 view*, volume 119. John Wiley & Sons Incorporated, 1985.

[21] SW Dharmadhikari, Kumar Jogdeo, et al. Bounds on moments of certain random variables. *The Annals of Mathematical Statistics*, 40(4):1506–1509, 1969.

[22] Peter J Diggle and Peter Hall. The selection of terms in an orthogonal series density estimator. *Journal of the American Statistical Association*, 81(393):230–233, 1986.

[23] *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.0.17 of 2017-12-22, 2017. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller and B. V. Saunders, eds.

[24] David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, pages 508–539, 1996.

[25] Michael Falk. Relative deficiency of kernel type estimators of quantiles. *The Annals of Statistics*, pages 261–268, 1984.

[26] Michael Falk et al. Asymptotic normality of the kernel quantile estimator. *The Annals of Statistics*, 13(1):428–433, 1985.

[27] Jean Dickinson Gibbons and Subhabrata Chakraborti. *Nonparametric Statistical Inference*. CRC Press, 2010.

[28] Constantino Goutis and George Casella. Explaining the saddlepoint approximation. *The American Statistician*, 53(3):216–224, 1999.

[29] Włodzimierz Greblicki and Mirosław Pawlak. Hermite series estimates of a probability density and its derivatives. *J. Multivariate Anal.*, 15(2):174–182, 1984.

[30] Włodzimierz Greblicki and Mirosław Pawlak. Pointwise consistency of the Hermite series density estimate. *Statist. Probab. Lett.*, 3(2):65–69, 1985.

[31] Włodzimierz Greblicki and Mirosław Pawlak. *Nonparametric system identification*. Cambridge University Press Cambridge, 2008.

[32] Hugo Lewi Hammer, Anis Yazidi, and Håvard Rue. A new quantile tracking algorithm using a generalized exponentially weighted average of observations. *Applied Intelligence*, 49(4):1406–1420, 2019.

[33] Hugo Lewi Hammer, Anis Yazidi, and Håvard Rue. Tracking of multiple quantiles in dynamically varying data streams. *Pattern Analysis and Applications*, pages 1–13, 2019.

[34] Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

[35] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2011.

[36] Jeffrey D Hart. Data-based choice of the smoothing parameter for a kernel density estimator. *Australian journal of statistics*, 27(1):44–52, 1985.

[37] Chris C Heyde. On a property of the lognormal distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, 25(2):392–393, 1963.

[38] Jonathan Huggins, Ryan P Adams, and Tamara Broderick. Pass-glm: polynomial approximate sufficient statistics for scalable bayesian glm inference. In *Advances in Neural Information Processing Systems*, pages 3611–3621, 2017.

[39] Integral. Truefx. https://www.truefx.com/. Accessed: 2019-05-24.

[40] Raj Jain and Imrich Chlamtac. The p 2 algorithm for dynamic calculation of quantiles and histograms without storing observations. *Communications of the ACM*, 28(10):1076–1085, 1985.

[41] Asma Jmaei, Yousri Slaoui, and Wassima Dellagi. Recursive distribution estimator defined by stochastic approximation method using Bernstein polynomials. *J. Nonparametr. Stat.*, 29(4):792–805, 2017.

[42] Maurice George Kendall, Alan Stuart, and J Keith Ord. *Kendall's advanced theory of statistics. v. 1: Distribution theory*. London (UK) Griffin, 1987.

[43] R Kronmal and Michael Tarter. The estimation of probability densities and cumulatives by fourier series methods. *Journal of the American Statistical Association*, 63(323):925–952, 1968.

[44] Alexandre Leblanc. On estimating distribution functions using Bernstein polynomials. *Ann. Inst. Statist. Math.*, 64(5):919–943, 2012.

[45] E. Liebscher. Hermite series estimators for probability densities. *Metrika*, 37(6):321–343, 1990.

[46] Eckhard Liebscher. Convergence of hermite series density estimators under conditions of weak dependence. *Statistics: A Journal of Theoretical and Applied Statistics*, 31(3):191–214, 1998.

[47] Jacob Loveless, Sasha Stoikov, and Rolf Waeber. Online algorithms in high-frequency trading. *Communications of the ACM*, 56(10):50–56, 2013.

[48] Elizbar A Nadaraya. Some new estimates for distribution functions. *Theory of Probability & Its Applications*, 9(3):497–500, 1964.

[49] Valeriy Naumov and Olli Martikainen. Exponentially weighted simultaneous estimation of several quantiles. *World Academy of Science, Engineering and Technology*, 8:563–568, 2007.

[50] Peter M Neely. Comparison of several algorithms for computation of means, standard deviations and correlation coefficients. *Communications of the ACM*, 9(7):496–499, 1966.

[51] Jurg Ott and Richard A Kronmal. Some classification procedures for multivariate binary data using orthogonal functions. *Journal of the American Statistical Association*, 71(354):391–399, 1976.

[52] Philippe Pébay, Timothy B Terriberry, Hemanth Kolla, and Janine Bennett. Numerically stable, scalable formulas for parallel and online computation of higher-order multivariate central moments with arbitrary weights. *Computational Statistics*, 31(4):1305–1325, 2016.

[53] A. Pepelyshev, E. Rafajłowicz, and A. Steland. Estimation of the quantile function using Bernstein-Durrmeyer polynomials. *J. Nonparametr. Stat.*, 26(1):1–20, 2014.

[54] Jouni Puuronen and Aapo Hyvärinen. Hermite polynomials and measures of non-gaussianity. In *Artificial Neural Networks and Machine Learning–ICANN 2011*, pages 205–212. Springer, 2011.

[55] Kimmo EE Raatikainen. Simultaneous estimation of several percentiles. *Simulation*, 49(4):159–163, 1987.

[56] Kimmo EE Raatikainen. Sequential procedure for simultaneous estimation of several percentiles. *Trans. Society for Computer Simulation*, 1:21–44, 1990.

[57] R.-D. Reiss. Nonparametric estimation of smooth distribution functions. *Scand. J. Statist.*, 8(2):116–119, 1981.

[58] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[59] Stuart C. Schwartz. Estimation of probability density by an orthogonal series. *Ann. Math. Statist.*, 38:1261–1265, 1967.

[60] Simon J Sheather and James Stephen Marron. Kernel quantile estimators. *Journal of the American Statistical Association*, 85(410):410–416, 1990.

[61] Michael Stephanou and Melvin Varughese. On the properties of hermite series based distribution function estimators. *Metrika*, 2020.

[62] Michael Stephanou, Melvin Varughese, and Iain Macdonald. Sequential quantiles via Hermite series density estimation. *Electron. J. Stat.*, 11(1):570–607, 2017, arXiv preprint arXiv:1507.05073v2.

[63] Gábor Szego. *Orthogonal polynomials*. American Mathematical Society, Providence, R.I., fourth edition, 1975. American Mathematical Society, Colloquium Publications, Vol. XXIII.

[64] Luke Tierney. A space-efficient recursive procedure for estimating a quantile of an unknown distribution. *SIAM Journal on Scientific and Statistical Computing*, 4(4):706–711, 1983.

[65] Nitya Tiwari and Prem C Pandey. A technique with low memory and computational requirements for dynamic tracking of quantiles. *Journal of Signal Processing Systems*, 91(5):411–422, 2019.

[66] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[67] Gilbert G. Walter. Properties of Hermite series estimation of probability density. *Ann. Statist.*, 5(6):1258–1264, 1977.

[68] G. S. Watson and M. R. Leadbetter. Hazard analysis. II. *Sankhyā Ser. A*, 26:101–116, 1964.

[69] BP Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420, 1962.

[70] Max Welling. Robust series expansions for probability density estimation. *Report—California Institute of Technology, Computational Vision Lab*, page 22, 1999.

[71] DHD West. Updating mean and variance estimates: An improved method. *Communications of the ACM*, 22(9):532–535, 1979.

[72] Wei Xiao. An online algorithm for nonparametric correlations. *arXiv preprint arXiv:1712.01521*, 2017.

[73] Anis Yazidi and Hugo Hammer. Multiplicative update methods for incremental quantile estimation. *IEEE transactions on cybernetics*, 49(3):746–756, 2017.