

Generalised density function estimation using moments and the characteristic function

Gerhard Esterhuizen



Thesis presented in partial fulfilment
of the requirements for the degree of

Master of Science in Electronic Engineering
at the
University of Stellenbosch

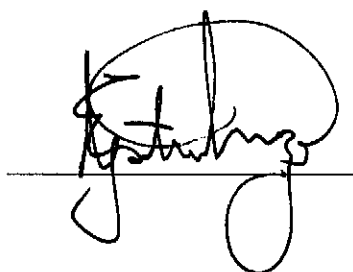
Supervisor: Prof. J.A. du Preez

April 2003

Declaration

I, the undersigned, hereby declare that the work contained in this thesis is my own original work and that I have not previously in its entirety or in part submitted it at any university for a degree.

Signature:

A handwritten signature in black ink, written over a horizontal line. The signature is stylized and appears to be 'K. Stuhm'.

March 2003

Abstract

Probability density functions (PDFs) and cumulative distribution functions (CDFs) play a central role in statistical pattern recognition and verification systems. They allow observations that do not occur according to deterministic rules to be quantified and modelled. An example of such observations would be the voice patterns of a person that is used as input to a biometric security device.

In order to model such non-deterministic observations, a density function estimator is employed to estimate a PDF or CDF from sample data. Although numerous density function estimation techniques exist, all the techniques can be classified into one of two groups, parametric and non-parametric, each with its own characteristic advantages and disadvantages.

In this research, we introduce a novel approach to density function estimation that attempts to combine some of the advantages of both the parametric and non-parametric estimators. This is done by considering density estimation using an abstract approach in which the density function is modelled entirely in terms of its moments or characteristic function. New density function estimation techniques are first developed in theory, after which a number of practical density function estimators are presented.

Experiments are performed in which the performance of the new estimators are compared to two established estimators, namely the Parzen estimator and the Gaussian mixture model (GMM). The comparison is performed in terms of the accuracy, computational requirements and ease of use of the estimators and it is found that the new estimators does combine some of the advantages of the established estimators without the corresponding disadvantages.

Opsomming

Waarskynlikheids digtheidsfunksies (WDFs) en Kumulatiewe distribusiefunksies (KDFs) speel 'n sentrale rol in statistiese patroonherkenning en verifikasie stelsels. Hulle maak dit moontlik om nie-deterministiese observasies te kwantifiseer en te modelleer. Die stempatrone van 'n spreker wat as intree tot 'n biometriese sekuriteits stelsel gegee word, is 'n voorbeeld van so 'n observasie.

Ten einde sulke observasies te modelleer, word 'n digtheidsfunksie afskatter gebruik om die WDF of KDF vanaf data monsters af te skat. Alhoewel daar talryke digtheidsfunksie afskatters bestaan, kan almal in een van twee katagoriee geplaas word, parametries en nie-parametries, elk met hul eie kenmerkende voordele en nadele.

Hierdie werk lê 'n nuwe benadering tot digtheidsfunksie afskatting voor wat die voordele van beide die parametrieses sowel as die nie-parametrieses tegnieke probeer kombineer. Dit word gedoen deur digtheidsfunksie afskatting vanuit 'n abstrakte oogpunt te benader waar die digtheidsfunksie uitsluitlik in terme van sy momente en karakteristieke funksie gemo-delleer word. Nuwe metodes word eers in teorie ondersoek en ontwikkel waarna praktiese tegnieke voorgelê word. Hierdie afskatters het die vermoë om 'n wye verskeidenheid digtheidsfunksies af te skat en is nie net ontwerp om slegs sekere families van digtheidsfunksies optimaal voor te stel nie.

Eksperimente is uitgevoer wat die werkverrigting van die nuwe tegnieke met twee gevestigde tegnieke, naamlik die Parzen afskatter en die Gaussiese mengsel model (GMM), te vergelyk. Die werkverrigting word gemeet in terme van akkuraatheid, vereiste numeriese verwerkingsvermoë en die gemak van gebruik. Daar word bevind dat die nuwe afskatters wel voordele van die gevestigde afskatters kombineer sonder die gepaardgaande nadele.

Acknowledgements

I would like to thank the following people:

- Prof. J.A. Du Preez, my supervisor, for his patient guidance.
- Zelda Weitz for her unfailing love and encouragement.
- My parents and family for their education and support.
- My grandmother and Mr and Mrs Weitz for providing a home away from home.
- Dr Dave Weber and the DSP Lab for providing a creative learning environment.
- Charl Botha for his technical advice.
- Pieter Nel and Koos Hugo for False Bay sailing.

Contents

1	Introduction	1
1.1	Motivation and topicality	1
1.1.1	Density function estimation	2
1.1.2	Our research	3
1.2	Background	5
1.2.1	Random variables	5
1.2.2	Pattern classification	6
1.2.3	Hypothesis tests	7
1.3	Existing techniques	9
1.3.1	Non-parametric estimators	10
1.3.2	Parametric estimators	11
1.3.3	Other techniques	12
1.3.4	Requirements for a new estimator	17
1.4	Objectives	17
1.5	Contributions	18
1.6	Overview of the document	19
2	Novel probability density function estimators	20
2.1	Introduction	20
2.2	Motivation	20
2.3	Definitions and Background	21
2.3.1	Moments	21
2.3.2	Characteristic function	23
2.3.3	Fourier series	25
2.4	Estimators based on moments	26
2.4.1	Motivation	27

2.4.2	A PDF in terms of moments	29
2.4.3	The anti-derivative	37
2.4.4	Numerical integration techniques	38
2.4.5	Fourier series approximation	41
2.4.6	Estimating moments from sample data	44
2.5	Estimators based on the characteristic function	46
2.5.1	Motivation	46
2.5.2	A PDF in terms of a characteristic function	47
2.5.3	Fourier series	51
2.5.4	Limiting case where $N_x \rightarrow \infty$	52
2.6	Conclusions	57
2.6.1	Comparison between characteristic function and moments techniques	58
2.6.2	Comparison with the Parzen estimator and Gaussian mixture model (GMM)	59
3	Novel cumulative distribution function estimators	61
3.1	Introduction	61
3.2	Motivation	62
3.3	Definitions and background	62
3.4	Estimators based on moments	63
3.4.1	A CDF in terms of moments	64
3.4.2	Numerical integration techniques	66
3.4.3	Fourier series approximation	69
3.5	Estimators based on the characteristic function	75
3.5.1	A CDF in terms of a characteristic function	76
3.5.2	Fourier series	79
3.6	Conclusions	81
4	Experimental results	83
4.1	Introduction	83
4.2	Experimental setup	85
4.2.1	Input data	85
4.2.2	Estimation error measure	86
4.2.3	PDF and CDF estimate using moments	87
4.2.4	PDF and CDF estimate using characteristic function	87

4.2.5	Parzen estimator	88
4.2.6	Gaussian Mixture Model	88
4.3	Mean estimation error	88
4.3.1	PDF estimators	88
4.3.2	CDF estimators	97
4.4	Computational requirements	102
4.4.1	PDF Estimators	104
4.4.2	CDF Estimators	106
4.5	Training requirements	107
4.6	Application to speaker verification	109
4.7	Conclusions	111
5	Conclusions and recommendations	114
5.1	Conclusions	114
5.2	Recommendations	116
A	A review of the Fourier transform	120
B	Summary of algorithms	123

List of Tables

4.1	PDF estimators: minimum mean estimation errors, minimum combined mean estimation errors and corresponding combined standard deviations (in brackets) from 100 samples (100× Kullback-Leibler divergence is indicated).	91
4.2	PDF estimators: minimum mean estimation errors, minimum combined mean estimation errors and corresponding combined standard deviations (in brackets) from 1000 samples (100× Kullback-Leibler divergence is indicated).	95
4.3	PDF estimators: the effect of selecting a sub-optimal working point.	96
4.4	CDF estimators: minimum mean estimation errors, minimum combined mean estimation errors and corresponding combined standard deviation (in brackets) from 100 samples (10× integral absolute difference is indicated).	101
4.5	CDF estimators: minimum mean estimation errors, minimum combined mean estimation errors and corresponding combined standard deviation (in brackets) from 1000 samples (10× integral absolute difference is indicated).	102
4.6	Gradient (x100) characterising the relationship between the computation time and the parameter value.	105
4.7	Comparison between GMM and CF technique in a speaker verification application.	111
4.8	Feature matrix for all the estimators.	113
A.1	Selected Fourier transform properties.	122
B.1	PDF estimate from moments using Fourier series.	124
B.2	PDF estimate from samples using Fourier series.	125
B.3	CDF estimate from moments using Fourier series.	126
B.4	CDF estimate from samples using Fourier series.	127

List of Figures

2.1	Successive Taylor series approximations to characteristic function of Gaussian PDF using 5, 14, and 32 terms.	31
2.2	The smoothing effect of a Hamming window on the characteristic function.	34
2.3	Leakage introduced by rectangular windowing of the characteristic function.	36
2.4	Discretisation of the characteristic function.	45
2.5	Reconstructing a PDF directly from data samples, using a triangular windowing function.	50
3.1	Relationship between $f(x)$, $f'(x)$ and $f''(x)$	71
4.1	Mean estimation error: PDF estimate using GMM from 100 samples. . . .	89
4.2	Mean estimation error: PDF estimate using Parzen estimator from 100 samples.	90
4.3	Mean estimation error: PDF estimate using characteristic function from 100 samples.	90
4.4	Mean estimation error: PDF estimate using moments from 100 samples. . .	91
4.5	Typical PDF estimates.	92
4.6	Examples of over-fitted PDF estimates (obtained from 100 samples). . . .	94
4.7	Mean estimation error: CDF estimate using GMM from 100 samples. . . .	97
4.8	Mean estimation error: CDF estimate using Parzen estimator from 100 samples.	98
4.9	Mean estimation error: CDF estimate using characteristic function from 100 samples.	98
4.10	Mean estimation error: CDF estimate using moments from 100 samples. . .	99
4.11	Typical CDF estimates.	100
4.12	Comparison of normalised execution times of PDF estimators: Pentium III 700 MHz.	104

4.13	PDF estimators: estimation error against computational requirements (Parzen is excluded as its computational requirements are too high).	105
4.14	CDF estimators: estimation error against computational requirements (Parzen is excluded as its computational requirements are too high).	107
4.15	Typical class-conditional PDFs of scores corresponding to impostors and known speakers.	110

Mathematical notation

X	Random variable
$f_X(x)$	Probability density function (PDF) of X
$F_X(x)$	Cumulative distribution function (CDF) of X
$\Phi_X(\omega)$	Characteristic function (CF) of X
$\hat{f}_X(x)$	Probability density function (PDF) estimate of X
$\hat{F}_X(x)$	Cumulative distribution function (CDF) estimate of X
$\hat{\Phi}_X(\omega)$	Characteristic function (CF) estimate of X
$\Theta(\omega_u, \omega)$	Frequency domain windowing function
$\theta(\omega_u, x)$	Spatial domain windowing function
$\Delta(\omega)$	Expected value of CF estimate
$\Omega(\omega)$	Variance of CF estimate
x	Outcome of random variable X
x_Δ	Period of function defined on x
x_i	i 'th data sample
ω	Frequency variable
ω_0	Fundamental frequency
ω_u	Upper frequency
m_n	n 'th moment around the origin
N_x	Number of data samples
N_m	Number of moments
N_i	Number of integration intervals
k, m, n	Indices
$\alpha, \beta, \gamma, \lambda, \xi$	Auxiliary symbols
$f(x), f'(x), f''(x)$	Auxiliary functions (spatial domain)
$F(\omega), F'(\omega), F''(\omega)$	Auxiliary functions (frequency domain)

Chapter 1

Introduction

1.1 Motivation and topicality

We have grown accustomed to computers easily performing tasks that we, as humans, consider to be difficult. It is known that a computer can be relied upon to produce the exact same answer each time it performs the same arbitrarily complex numerical calculation, this being due to the extreme deterministic nature of the machine and the problem.

Humans, on the other hand, excel at different kinds of tasks, such as recognising someone we have met before from their voice. This comes so naturally to us that we consider it to be a trivial exercise. However, attempting to duplicate such behaviour using a computer is not a trivial task at all as the problem can only be represented in a non-deterministic fashion. Consider the task of recognising a person from their voice as an example: although a certain phrase spoken by a certain individual never sounds exactly the same as previously, it does contain some characteristics that enable humans (and even some animals) to uniquely identify the person (and the phrase) from their voice. This is sometimes even possible after the sound was severely distorted by, for example, a telephone or recording device.

One way of solving such non-deterministic problems using a computer is to employ some form of statistical model that allows non-deterministic observations to be quantified. An example is found in a speaker verification system which verifies the claimed identity of a person (speaker) based on their voice [1, p. 621]. A recording of the person claiming some identity is presented to the system, which then extracts statistical features from it. These features are then compared to stored values, which are known to represent the claimed person, to ensure that they are within certain specified tolerances. If they are found to be

sufficiently close to the known values, the claimed identity is accepted, else it is rejected. The sensitivity of the system is adjusted in order to minimise a certain cost function (or maximise a utility function), which takes into account the following:

- the probability of the system making certain types of incorrect decisions and
- the practical implications associated with each type of incorrect decision.

Research into this and related fields is far from its infancy and has been going on for many decades. The dream of a machine capable of understanding spoken human language was envisioned as early as the 1950s [1, p. 604]. Although continual and impressive progress has been made ever since then, research in this field is still far from exhausted and only recently have we started encountering everyday consumer products employing these techniques. Examples include biometric security devices, voice portals and electronic devices responding to voice commands.

Research into techniques allowing us to address non-deterministic pattern recognition problems, such as depicted in the example above, is formalised in the field of *statistical pattern recognition*. This allows us to address these problems systematically and within an abstract mathematical framework.

1.1.1 Density function estimation

At the heart of statistical pattern recognition are the probability density function (PDF) [2] and cumulative distribution function (CDF). These functions allow observations that do not follow a deterministic pattern to be characterised and quantified. They are closely related to each other, with the PDF being the derivative of the CDF, and whether a pattern recogniser directly employs a PDF or CDF depends on its implementation details. For the purposes of the following paragraphs, the terms PDF and CDF can be used interchangeably.

In the speaker recognition example presented earlier, a PDF would be used to quantify the probability that a statistical feature takes on values in different ranges of its value space. In doing this, the feature is not associated with a single “correct” value, but is rather specified in terms of how likely it is to attain a value within a certain range. If the values that the feature assume is concentrated in a single dense cluster, the PDF provides information about the location, size and shape of the cluster. Such a single feature that can be represented by a real scalar is referred to as a univariate (or one dimensional) random variable, with its associated PDF being a univariate real function of the random variable.

A PDF can also describe the values that is simultaneously attained by N feature variables, in which case the features under consideration are combined into an N -dimensional feature vector. These underlying features comprising the feature vector are usually related to each other in some way (if not, they could be treated independently from each other as N separate features). A PDF corresponding to such a feature vector, and its underlying random variable, is referred to as multivariate (or N -dimensional).

In order for a practical recogniser to characterise a feature (which represents a random variable), it is required to estimate a PDF from a number of data samples representing possible outcomes of the feature. This task is performed by the density function estimator during the training phase of the system. For example, the speaker verification system introduced above would be presented with a number of voice recordings, obtained in a controlled environment, all from a certain known speaker. Features would then be calculated from these inputs and a PDF would be trained to correspond to this speaker. These PDFs are then later used when verifying the claimed identity of an unknown speaker.

The density function estimator therefore infers characteristics about the underlying density function of a random variable from a number of random samples that represent outcomes of the random variable. Once it is trained, it has the ability to provide the value of the density function at any point within its domain (the sample space). Estimators differ in the way in which they obtain the estimate, the internal representation used to store information about the estimate and the constraints imposed on the density functions that it can accurately approximate. Two important factors that characterise an estimator is its bias and its consistency. Bias refers to whether the expected value of the estimate tends to the actual PDF in the limit where the number of samples from which it is trained tends to infinity. If it does tend to the actual PDF, then the estimator is said to be unbiased, which is a desirable property. Consistency refers to whether the variance of the estimate disappears (tends to zero) in the same limit, which also is a desirable property.

1.1.2 Our research

Established techniques that estimate PDFs from sample data can be broadly characterised according to the basic approach that they follow:

1. Estimate the PDF in terms of some (possibly highly complex) function with a number of free parameters which are selected in order to optimally fit the PDF to the sample (training) data (we include mixture models [3] in this approach).

2. Estimate the PDF in terms of a function or rule that is defined directly on the training samples.

The first approach is known as the parametric approach and the second one known as the non-parametric approach. Although a number of established and respected techniques, each falling into one of these groups, exist, none are without their disadvantages. Furthermore, as each of these two approaches has its own set of unique advantages and disadvantages, we did not attempt to develop another estimator along the lines of one of the above approaches. Instead, we chose to follow a novel approach which resulted in estimators that combine aspects of both of the above approaches. These estimators manage to combine advantages of existing estimators without the disadvantages normally associated with them. The new estimators were based on the following concepts characterising random variables at a very fundamental level:

Moments These are real-valued scalars, characterised by their order, which are often used to characterise aspects of families of PDFs. For example, a Gaussian random variable is characterised by its mean and variance, both of which are moments. Inspired by the *Principle of moments* [4], which states that PDFs encountered in practice are determined entirely in terms of their moments (up to infinite order), we consider ways of estimating arbitrary PDFs using the values of a finite number of moments.

Characteristic function This is a complex-valued function that is related to the PDF through the Fourier transform and which uniquely determines the PDF. Motivation for using this is drawn from the fact that it provides a frequency domain representation of the PDF, which often allows a more efficient approximation to be constructed than one based in the spatial domain. Using this does, however, require a way of estimating a characteristic function from sample data, which is also presented.

As the work is novel, it was decided to limit all research to univariate random variables (and therefore univariate density functions). This is due to the “curse of dimensionality” [3, p. 7] that is associated with multivariate random variables. It is recommended that ways of extending the techniques developed in this thesis to the multivariate case be considered for future research.

1.2 Background

The following sections provide background information that is required to view the rest of the chapter in perspective. An overview of random variables, moments and the characteristic function is provided. This is followed by some examples showing the application of PDFs and CDFs to pattern classification and hypothesis tests which illustrates the central role that PDF and CDF estimators fulfil in pattern recognisers.

1.2.1 Random variables

A real random variable maps all points in a sample space, representing outcomes of a random experiment, onto the real line [2]. When dealing with random variables, we are usually not concerned about the mapping itself, but mostly with the values that the random variable attains. This transforms the problem from the domain of the sample space to the domain of the random variable, allowing the treatment of different types of events (from different sample spaces) in an abstract and consistent fashion without requiring knowledge of the actual sample space. For the remainder of the document, the term “sample space” is used interchangeably to refer to both the domain of the random variable as well as the domain representing the sample space of the actual outcomes (with the exact meaning depending on the context). Important concepts related to random variables, required during the remainder of this document, are now presented in greater detail.

Let X be a real univariate random variable, with x a single outcome (corresponding to some outcome in the sample space). It is fully characterised by its cumulative distribution function (CDF), $F_X(x)$, which is defined as follows:

$$F_X(x) = P\{X \leq x\}. \quad (1.1)$$

We see that the CDF characterises X in a non-deterministic fashion by specifying the probability of the outcomes of X attaining values within a certain range, without limiting a specific outcome in any way (in terms of all possible outcomes). The probability density function (PDF) of X ,

$$f_X(x) = \frac{d}{dx}F_X(x), \quad (1.2)$$

is related to the CDF and provides an indication of the concentration of the outcomes at some point in the sample space.

Moments are scalars, defined entirely by the PDF, that roughly characterise the location

and spread of a random variable with regards to a specific reference point (usually the origin or the mean value). The moments of a univariate random variable are characterised entirely in terms of their order and the point around which they are calculated, with the n 'th order moment around x_0 defined by:

$$\mu'_n(x_0) = \int_{-\infty}^{\infty} (x - x_0)^n f_X(x) dx. \quad (1.3)$$

Some examples of well-known moments are the mean (first moment around the origin) and the variance (second moment around the mean). Parametric density functions are often characterised in terms of their moments (such as the mean, variance, skew and kurtosis).

The characteristic function, $\Phi_X(\omega)$, is a complex-valued function that is also defined entirely in terms of the PDF,

$$\Phi_X(\omega) = \int_{-\infty}^{\infty} f_X(x) e^{j\omega x} dx. \quad (1.4)$$

The above expression shows the characteristic function to be equal to the complex conjugate of the Fourier transform [5, p. 82]. It is usually used analytically to calculate the moments of a PDF as it often provides an easier to use alternative to the expression defining the moments.

1.2.2 Pattern classification

In order to illustrate an application of density functions, we now present a short introduction to pattern classification. Highly detailed accounts of this can be found in Duda and Hart [6], Bishop [3], Fukunaga [7] and Devijver and Kittler [8].

A pattern classification system classifies unlabelled input data as belonging to a class selected from a closed set of candidate classes [9]. We limit ourselves to non-deterministic classifiers that operate on random input data, allowing the use of statistical pattern recognition techniques.

Each class is represented by a probability density function (PDF) that describes outcomes of the data, x , associated with the class. The PDF corresponding to the i 'th class (known as the class-conditional density function) is given by

$$f_X(x | m_i). \quad (1.5)$$

If we also know the prior probability of the i 'th class, $P(m_i)$, we can express the posterior probability of the i 'th class, conditioned on the observation x , using Bayes' Theorem [2, p. 17]:

$$P(m_i | x) = \frac{f_X(x | m_i) P(m_i)}{\sum_j f_X(x | m_j) P(m_j)}. \quad (1.6)$$

This provides an expression for the probability that the i 'th class generated the observed data x . Note that this quantity is automatically normalised so that $\sum_i P(m_i | x) = 1$ (due to the denominator term). If we further assume all misclassifications to carry the same penalty, the observed data is selected as belonging to the class c , the one with the highest posterior probability:

$$c = \underset{i}{\operatorname{argmax}} \{P(m_i | x)\}. \quad (1.7)$$

As the denominator term in the expression defining the posterior probability is the same for all values of i , this classification only depends on the prior probability and the class-conditional density function. If it is further assumed that all classes have equal prior probability, which is a reasonable choice in the absence of any evidence indicating the contrary, the above expression reduces to

$$c = \underset{i}{\operatorname{argmax}} \{f_X(x | m_i)\}, \quad (1.8)$$

which shows that the classification is performed entirely in terms of the class-conditional PDFs. These last two expressions represent classifiers that minimise the expected loss (i.e. the long-term misclassification error) and correspond to the Bayes rule (for minimum loss) [8, p. 24] [7, p. 52]. Although it is likely that they would still misclassify some inputs, they represent the optimal classifier solution for a given set of class-conditional PDFs and prior probabilities.

In both these classifiers, the class-conditional density function played a fundamental role in controlling the operation of the classifier and in ensuring optimal performance. In practice, each of these PDFs would be estimated, from labelled sample data with known class membership, by some PDF estimator. The PDF estimator therefore plays a central role in determining the accuracy of a classifier.

1.2.3 Hypothesis tests

Another application of PDF estimation is found in hypothesis testing. A simple, but complete definition of hypothesis testing is encountered in [10, p. 75]:

Hypothesis testing is the process of inferring from a sample whether or not to accept a certain statement about the population.

The output of a hypothesis test is a binary value, either indicating the acceptance of some hypothesis (called the null hypothesis or H_0) or the rejection of it (which implies acceptance of the alternative hypothesis or H_1). In order to construct a hypothesis test, a test statistic, t , is first selected: this is a random variable, that is defined as a function of a number of observations from the sample space, with a known PDF under the condition that the null hypothesis holds. The space containing the values of the test statistic is then partitioned into two regions by the discriminant function $\delta(t)$, one corresponding to the null hypothesis and one corresponding to alternative hypothesis (called the critical region). Although this also partitions the sample space into two regions, there are good reasons to prefer the use of a test statistic above simply partitioning the sample space directly:

1. By combining multiple, possibly multivariate, observations into a single number, a reduction in the dimensionality of the problem is achieved [11, p. 50]. Ideally, the test statistic summarises only enough information about the observation so that the truth (or falsehood) of the null hypothesis can be established.
2. Often, the PDF of the test statistic is known even though the PDF of the underlying sample space is not [10, p. 31]. An example of this is the creation of a test statistic by summing a large number (typically more than 100) of independent observations all corresponding to the same unknown PDF. According to the central limit theorem [2, p. 118], the test statistic would have a Gaussian (normal) distribution with a mean and variance that can be obtained from the mean and variance of the underlying PDF (describing the sample space). As the mean and variance of the underlying PDF may be estimated without any knowledge of the underlying PDF, this test statistic can be used with any underlying PDF.

In the case of a one-tailed test, the discriminant function simply compares the value of the test statistic to some threshold:

$$\delta(t) = \begin{cases} H_0 : & t > t_a \\ H_1 : & \text{elsewhere,} \end{cases} \quad (1.9)$$

where

$$P(t \leq t_a) = \alpha \quad (1.10)$$

and α is known as the significance level. This level provides a value for the false rejection rate (FRR), which is defined as the probability of rejecting a correct hypothesis. From the definition of the CDF, the threshold, t_a , can be expressed entirely in terms of the CDF of the test statistic:

$$F_T(t_a) = \alpha. \quad (1.11)$$

This allows a hypothesis test with a fixed FRR to be constructed from knowledge of the CDF of the test statistic. It should, however, be noted that the accuracy with which the FRR can be controlled (by selecting α) depends directly on the accuracy of the CDF estimate. Inaccurate estimates would therefore result in values of FRR that could greatly differ from the required significance level.

In classical hypothesis tests [12], the form of the PDF (or CDF) of the test statistic is often known, due to the nature of the problem or the way in which the test statistic is constructed. In applications where this is not the case, the PDF (or CDF) of the test statistic has to be estimated from sample data. Consequently, the accuracy of the hypothesis test is strongly dependent on the accuracy of the CDF estimator.

1.3 Existing techniques

Density function estimators can be broadly classified into two categories: parametric and non-parametric. Although Bishop [3] prefers an additional category, namely *semi-parametric* (which is used to classify mixture densities), we consider the semi-parametric techniques together with the parametric (viewing what Bishop refers to as a parametric estimator simply as a highly constrained parametric estimator). This is done for a simple reason: we investigate existing techniques in order to determine their strengths and weaknesses (in order to attempt improvements). As far as that criteria is concerned, this classification divides the estimators into two broad categories, each with its own distinctive advantages and disadvantages. Also, we do not consider techniques which are highly constrained in terms of the density functions that they can estimate (such as one representing a PDF using only a single Gaussian PDF). *We focus our attention exclusively on techniques that have the ability to (at least in principle) estimate arbitrary PDFs*, although they may have drawbacks in practice. We select a single estimator, that is representative of the class of estimators, from each class of estimators in order to illustrate the philosophy behind each of the two classes of estimators.

In the following sections, we assume X to be a univariate random variable with $\hat{f}_X(x)$



representing its PDF estimate and $\hat{F}_X(x)$ its corresponding CDF estimate. x_i represents the i 'th sample (from a total of N_x samples) drawn from the sample set from which the estimates are obtained.

1.3.1 Non-parametric estimators

Well-known non-parametric density function estimators include the Parzen (a special case of generalised kernel techniques), k Nearest Neighbour and histogram methods [8] [3] [7]. All these estimators utilise the sample data directly when calculating an estimate and provides no way (except for sampling of the estimate itself) of representing the sample data in a more compact fashion. The advantage of this approach is that, as these techniques assume very little about the PDF, they are applicable to a wide range of PDFs. Obtaining an estimate is also easy as it does not involve complicated or iterative parameter estimation techniques. Unfortunately, these techniques suffer from problems relating to computational requirements: as their computational requirements increase with the size of the sample space, they do not scale very well to large sample sizes.

In Chapter 4, experiments were conducted to compare new estimation techniques with established ones. The Parzen estimator was selected as representative of the non-parametric estimators as it provides a good balance between simplicity of use (the k Nearest Neighbour algorithm involves a search which complicates it) and the accuracy of estimates (the histogram is a rather crude technique to use on continuous random variables). A Parzen PDF estimate is obtained by expressing the PDF as a weighted sum of kernel functions, each one centered at the location of a sample:

$$\hat{f}_X(x) = \frac{1}{N_x} \sum_{i=0}^{N_x-1} \varphi(x - x_i), \quad (1.12)$$

where $\varphi(x)$ represents the kernel function used in the approximation. In order for this estimate to represent a valid PDF, each kernel function should also be a valid PDF. From [8] we see that this estimate is biased, as $\hat{f}_X(x)$ tends to $f_X(x)$ convolved with $\varphi(x)$ in the limit where $N_x \rightarrow \infty$. Also, the variance of the estimate tends to zero as $N_x \rightarrow \infty$, proving it to be a consistent estimator (as the estimate improves with an increase in the number of samples). The kernel therefore plays an important role in the accuracy of the estimator as it determines the amount of smoothing that is applied to the estimate: if too little smoothing takes place, the estimate will be over-fitted to the samples (and therefore

seem noisy), while too much smoothing may hide small (but possibly important) features of the PDF.

1.3.2 Parametric estimators

Mixture models present a practical technique that provides a good balance between the non-parametric estimators and the highly constrained traditional parametric estimators. A mixture model approximates a PDF using a weighted sum of density functions:

$$\hat{f}_X(x) = \sum_{i=0}^{M-1} w_i \varphi_i(x), \quad (1.13)$$

where $\varphi_i(x)$ represents the i 'th mixture density, w_i its corresponding weight and M the number of mixtures. Although this expression looks similar to that defining the Parzen estimator, there are two major differences:

1. The shape and location of each mixture density does not correspond directly to a single sample (as the Parzen estimator kernel functions did), but is selected to be optimal according to some criteria. An important implication of this is that the parameters defining the mixture density are not as easily determined as in the case of the Parzen estimator.
2. There are usually significantly fewer mixture components than there are data samples. If the evaluation of each mixture density requires the same computational resources as the evaluation of each Parzen kernel function, the mixture model would be significantly more efficient. The computational requirements of this estimator are also independent of the sample size and only a function of the number of mixtures.

The above two differences also represent, respectively, the biggest disadvantage and advantage that the mixture models have above the non-parametric estimators. A popular choice of density function used in mixture models is the Gaussian density function, which is characterised only by its mean and variance. In order to train the estimator from a sample set, the number of mixtures is first selected (usually using some heuristic rule or with the aid of a clustering algorithm). Parameter values are then calculated so that the resulting mixture PDF estimate shows a good correspondence to the sample set. Training algorithms often select the parameters according to some rule that maximises the likelihood of the mixture estimate with respect to the parameter values.

The *Expectation-Maximisation* (or EM) algorithm [13] [3, p. 65] is an elegant iterative algorithm that allows one to obtain such a maximum likelihood solution. One drawback of this algorithm is that it converges to a maximum likelihood solution corresponding to a local maximum of the likelihood function (i.e the estimate improves up to a certain maximum which may or may not represent the optimal solution). An estimator could therefore, for a certain choice of PDF, sometimes provide highly optimal results, while at other times provide less optimal results. The outcome of a specific training run is highly dependent on the specific sample set and the (often random) initial conditions characterising the training. Estimators employing such iterative training algorithms would generally exhibit higher values of variance in their estimates than estimators that employ closed-form solutions (such as the Parzen estimator).

1.3.3 Other techniques

A number of techniques that address the shortcomings of traditional estimators or problems related to density function estimation are now considered. Preference was given to techniques that exhibited some similarity to those presented in this work (e.g. the use of moments or Fourier series representations) or that addressed problems stated in previous sections. A short overview of the operation of each technique is provided and similarities and differences between these techniques and the new techniques presented in this work are briefly considered.

Some of these techniques represent improved training algorithms or employ parametric estimators using mixtures of flexible basis functions. In other examples, solutions to problems from other branches of engineering and mathematics are applied to the density estimation problem by identifying parallels between these fields.

Improved training algorithms *Bors and Pitas* [14] consider the robust estimation of parameters characterising a Radial Basis Function (RBF) neural network [15, p. 256] that represents a density function using a Gaussian mixture model. The mixture models are usually trained using an iterative algorithm based on second-order statistics (such as the EM algorithm). A drawback of second-order statistics are their sensitivity towards outliers in the training sample set and their large bias when approximating distributions with long tails. They propose the use of robust statistics, based on the median instead of the mean, in order to more accurately estimate the parameters of the mixture density function. Their algorithm (Median RBF)

outperforms estimators based on classical statistics in terms of accuracy when estimating univariate and bivariate density functions. Although this does not reduce the complexity of the mixture model training procedure, it does reduce the bias of the estimate.

Vlassis and Likas [16] also address the training of mixture densities by considering the selection of the optimal number of mixtures to use in a univariate Gaussian mixture model density function. They present a modified training algorithm that dynamically adapts the number of mixture components while iteratively estimating the mixture parameters. Their solution involves the use of the sample kurtosis, which provides a measure of how closely a sample set corresponds to a Gaussian distribution (as the kurtosis of a Gaussian distribution is equal to 0). The mixture model is initialised with a small number of mixtures and the optimal parameter values are estimated using the EM algorithm. Once the training algorithm converges, the weighted kurtosis is computed (as a linear combination of the sample kurtosis of each mixture component). If the absolute value of the weighted kurtosis is found to be too large, implying that at least one significant mixture component is not modelling Gaussian-distributed data, the number of mixtures are increased. This procedure is repeated until the absolute weighted kurtosis attains an acceptably low value. Estimates obtained using this procedure are expected to exhibit lower bias than those obtained by selecting the number of mixtures using a less systematic approach.

Although these algorithms address some of the concerns associated with mixture models (as they reduce the bias of the estimates and improve robustness) they still do not provide a closed-form solution to the estimate. The possibility of the algorithm converging to a locally optimal solution therefore still exists. Also, both algorithms are biased towards estimating a density function that is a linear combination of Gaussian density functions. Therefore, there are still applications in which a Gaussian mixture model, even with the improved training algorithms, would not represent the optimal choice of estimator.

Spectral density estimation *Pagès-Zamora and Lagunas* [17] present a technique for multivariate density function estimation based on established spectral estimation techniques. They obtain a density function estimate using a memoryless non-linear system (NLS) that expresses the estimate in terms of a Fourier series. The Fourier

series coefficients are inferred from a set of data samples (independently and identically distributed according to some underlying density function) so that the expected mean-squared error between the underlying density function and the estimate is minimised. Parallels are drawn between power spectral density (PSD) estimation and density function estimation: the relationship between the PSD and the autocorrelation function of a random process is similar to the relationship between a density function and its characteristic function (both are Fourier transform relationships). By employing established PSD estimation techniques, a density function estimate is obtained from the empirical characteristic function estimate. An optimal filter is computed that provides an estimate of the density function at a single location in the sample space from a set of training samples. Three methods of obtaining the filters are considered: minimum variance method (MVM), normalised minimum variance method (NMVM) and periodogram method (PM). The accuracy of the new techniques are compared to the histogram method by estimating a bivariate density function from sample data and it is found to outperform it in terms of accuracy.

Bercher and Vignat [18] and *Kay* [19] also follow an approach based on the relationship between PSD estimation and density function estimation. They model the density function as an auto-regressive (AR) process [20] and consider ways in which to estimate the model parameters so as to produce an accurate estimate. It is stated that density function encountered in practice are not likely to be exactly modelled by an AR process and would require a large number of coefficients in order to obtain an accurate estimate. As such long estimates suffer from low stability, regularization techniques are combined with a long AR estimate in order to obtain a stable and accurate estimate. Regularization, however, requires some prior knowledge about the smoothness of the density function. The AR estimator is used to estimate the density function and the entropy of the random variable from sample data and its accuracy is compared to a kernel estimator, a histogram estimator and Vasicek's estimate. It is found to compare favourably with these techniques in terms of accuracy and computational requirements.

Although these techniques express the estimate in terms of a Fourier series, which is similar to some of the new techniques presented in this work, very little consideration is given to the effect that this has on the estimate. Also, no consideration is given to the relationship between an estimate obtained using these PSD techniques and one obtained using an established estimator (such as the Parzen estimator).

Orthogonal series *Silverman* [21] considers a number of density function estimation techniques of which the orthogonal series estimator, originally attributed to *Cencov* in 1962, is one. It operates by expressing a density function in terms of a linear combination of orthogonal basis functions. The weight associated with each basis function is obtained by considering an empirical estimate of the projection of the data samples onto the function. Estimating a density function on a closed interval from a number of samples using a Fourier series is presented as an example. Although the necessity of a frequency domain windowing function is mentioned, no regard is given to the requirements of windowing functions, the effect of truncation of the series and the selection of the fundamental frequency. All these issues are considered in detail in the remainder of this work, when new density function techniques that also employ a Fourier series, are introduced.

Wavelets *Vannucci* [22] consider the estimation of density functions using wavelets. Although these estimators are strictly a special case of orthogonal series estimators, they hold advantages above techniques employing histograms, kernels and (classical) orthogonal series when representing discontinuities and local oscillations. This is due to the property of wavelets that allow functions that are localised in space and in frequency to be accurately approximated. Although consideration is given to the multivariate case, the techniques are only treated in detail for the case of univariate density functions. Experimental results show that linear wavelet estimators are able to accurately estimate smooth density functions while non-linear estimators are best suited towards discontinuities. The accuracy of the estimates are, however, dependent on choice of wavelet family and no single family performs optimally over a range of different density functions.

Another problem with wavelet estimators is selection of the number of wavelets to use in the approximation: too few terms would cause the estimate to neglect details and too many terms would cause over-fitting (the so-called “Dirac disaster”). The latter problem is addressed in a related publication by *Vannucci and Vidakovic* [23]. A technique that penalises the roughness of an estimate is developed and the Fisher information functional is proposed as a measure of roughness.

Similar to the new techniques presented in this work, these estimators attempt to combine the advantages of both the parametric and non-parametric estimators while omitting the disadvantages. The new techniques consider the mathematical relation-

ship between density functions and other quantities characterising them (moments and the characteristic function), while the wavelet approach employs a flexible parametric model. However, both sets of techniques have the ability, in principle, to estimate arbitrary density functions with the minimum of prior knowledge.

Moments *Lindsay, Pilla and Basak* [24] present a technique of estimating a univariate distribution function when the values of a number of moments of a random variable are known. It expresses the distribution function in terms of a mixture distribution function and presents techniques for obtaining optimal values of mixture distribution parameters. The parameters are calculated so that the moments of the resulting distribution function, up to order p , are the same as the moments estimated from the training sample set. A way of optimally selecting the number of moments to use is also presented. Experimental results shows the technique to produce reasonably accurate approximations of linear combinations of chi-square variables when using a mixture of gamma distributions. The selection of basis functions employed in the mixture distribution does however influence the accuracy of the approximation. Prior knowledge of the distribution function being approximated is therefore a requirement for obtaining an optimal estimate.

One of the new techniques presented in this work also computes a distribution function from the values of a number of moments, but using a fundamentally different approach: a distribution estimate is obtained directly in terms of the values of number of moments by using the mathematical relationship between the moments and the distribution function of a random variable.

Inversion integrals *Abate and Whitt* [25] [26] consider applications in operations research and queueing models where probability density and distribution functions are often characterised in terms of transforms (of the density functions). Although some transforms can be inverted using analytical techniques or tabulated formulas, there are many transforms which can only be solved by numerically evaluating the inversion integral. Techniques for evaluating such an integral is considered and it is shown that employing the trapezoidal rule is equivalent to using a Fourier series approximation. The Fourier series coefficients are calculated in terms of the analytical representation characteristic function and a lot of attention is paid to calculation the approximation error bounds. A similar approach is taken by *Witkovský* [27] where they employ an inversion formula that expresses the CDF in terms of an integral

involving the characteristic function. This integral is numerically evaluated in order to compute accurate confidence intervals of a linear combination of Student's t and Fisher-Snedecor's F random variables. Although these techniques of evaluating the inversion integral also considers the relationship between the density function and its characteristic function (similar to some of the new techniques presented in this work), they cannot be directly applied towards density function estimation as they require the characteristic function to be known analytically.

1.3.4 Requirements for a new estimator

From the properties of existing PDF estimation techniques, we can deduce a list of requirements that any new estimator should meet if it is to improve on the established techniques. The following requirements are obtained by considering the combined advantages and disadvantages from both the parametric and the non-parametric estimators:

- Ability to estimate arbitrary PDFs with little or no prior knowledge about their shape.
- Easy to train, preferably employing a closed-form solution. Iterative training schemes are undesirable.
- Computational requirements (for the evaluation of a density height) should preferably be independent from the training sample-size. This would most likely imply a parametric technique.
- Exhibit low variance (when compared to techniques having similar accuracy).
- Consistent (implying that the variance of the estimate decreases as the sample size increases).
- It should be amenable to practical implementation.

1.4 Objectives

Our research objectives were focussed on considering theoretical and practical PDF and CDF estimators, based entirely on moments and the characteristic function, using an abstract mathematical approach. The following research objectives were set:

- Determine the feasibility, suggested by the Principle of Moments, of estimating a PDF and a CDF entirely in terms of the values of a finite number of moments.
- Consider practical techniques of estimating a PDF and a CDF directly in terms of the values of a finite number of moments.
- Determine the feasibility of estimating a PDF and a CDF in terms of a sampled characteristic function.
- Consider practical techniques of estimating a PDF and a CDF directly in terms of a sampled characteristic function, estimated from sample data.
- Consider ways in which the characteristic function can be estimated from sample data.
- Experimentally compare new estimators, that were developed in the course of this work, with established techniques.

1.5 Contributions

The following contributions, published here for the first time (except for a conference publication by the author [28]), is a result of the research that was conducted within the scope of this project:

- Expressions, derived from basic principles, for PDFs and CDFs in terms of a finite number of moments.
- Practical techniques for estimating PDFs and CDFs from a finite set of moments.
- Expressions, derived from basic principles, for PDFs and CDFs in terms of a sampled and windowed characteristic function. This includes a result which presents the Parzen density function estimator from a frequency domain perspective.
- Practical techniques for estimating PDFs and CDFs from a number of samples, by using the characteristic function estimator.
- A consistent estimator for a sampled and windowed characteristic function that operates from sample data.

- A qualitative comparison between these parametric estimators and established parametric and non-parametric estimators.

1.6 Overview of the document

The remainder of the document is organised, as follows, into 4 further chapters:

Chapter 2 The feasibility of estimating a probability density function (PDF) in terms of the values of sample moments is considered and practical techniques are developed to this end. This leads to consideration of the feasibility of estimating a PDF from the characteristic function, also including the development of practical techniques. Section 2.4.5 and Section 2.5.3 provide practical techniques for obtaining a PDF estimate in terms of a finite number of moments and in terms of sample data.

Chapter 3 Cumulative distribution functions (CDFs) are developed in the same fashion as was done for PDFs in the previous chapter. A large part of the theory is inherited from Chapter 2 and this chapter mostly addresses issues specific to CDF estimation from moments and the characteristic function. Again, a number of practical techniques are developed for estimating the CDF using moments of sample data. Section 3.4.3 and Section 3.5.2 provide practical techniques for obtaining a PDF estimate in terms of a finite number of moments and in terms of sample data.

Chapter 4 Experiments that compare the new techniques to established ones (Parzen and GMM) are conducted and the results are shown. Experiments were conducted on synthetic data and an experiment that employed data from a speaker verification system was also conducted. It was found that the new techniques did combine some of the advantages of the established techniques, as was desired. A summary of the characteristics of all the estimators is presented in Section 4.7.

Chapter 5 Conclusions about the work and recommendations for future work are stated.

Chapter 2

Novel probability density function estimators

2.1 Introduction

A novel univariate parameterised density function estimator is now developed. It employs a technique that estimates the probability density function (PDF) in terms of a finite number of sample moments. The feasibility of estimating a PDF in terms of moments is first considered from a theoretical perspective. After the feasibility of this idea is asserted, a number of practical techniques that express a PDF estimate directly in terms of a finite number of moments are then developed.

During the investigation of the feasibility of estimating a PDF from moments, it is noted that the characteristic function plays a central role. The second part of the chapter devotes itself to the estimation of PDFs using the characteristic function, without requiring the use of moments, allowing the estimate to be obtained directly in terms of sample data. The feasibility of this is also first confirmed, after which a practical technique employing a Fourier series is presented. It is compared to the technique employing moments and it is concluded that it is to be preferred above the moments technique in cases where actual sample data is available.

2.2 Motivation

In the previous chapter, two established density function estimators, the GMM and the Parzen estimator, were introduced. We saw that, although both possessed highly desirable

characteristics, each one suffered from some disadvantages that the other one did not. It was stated that it would be desirable to find a density function estimator that combines the advantages of both estimators but omit the disadvantages.

Such an estimator would follow a parametric approach, allowing it to compactly represent the information contained in a sample set, in order to limit the computational requirements. It is also desirable for it to employ some generic parametric model with basis functions that is not based on some specific PDF, allowing it to perform equally well over different PDFs. As the moments of a random variable are often used to define and characterise a random variable, it was decided to attempt the creation of a PDF estimator based on moments (for which a more detailed motivation is found in Section 2.4). The values of a finite number of moments would then replace the sample data representing the PDF and an estimate, that can be evaluated anywhere in the sample domain, would be obtained in terms of these moments.

2.3 Definitions and Background

Some general statistical definitions and mathematical background relating to the PDF estimation techniques presented in this chapter are now presented. Unless otherwise stated, all random variables are assumed to be univariate (with their outcomes taking on scalar values). In-depth explanations of the theory presented here are found in Peebles [2], Kendall and Stuart [4], Stremler [5] and Proakis and Manolakis [20].

2.3.1 Moments

The expected value (or expectation) of a function, $g(x)$, of a random variable, X , with known PDF, $f_X(x)$, is defined by an integral expression:

$$E \{g(X)\} = \int_{-\infty}^{\infty} g(x) f_X(x) dx. \quad (2.1)$$

It is seen that, for a given function $g(x)$, the expected value is only dependent on the PDF. Furthermore, a set of functions, each corresponding to a different $g(x)$, can be used to calculate a set of expected values that characterise the PDF. Through careful selection of the set of functions, each expected value can contain unique information about the PDF that is not contained in any of the other values. This is an important conclusion as it allows a finite number of properties characterising a PDF to be represented by a finite set

of real numbers, making parameterised PDF models possible. An example of this is the classification of a PDF in terms of its mean, variance and skew: these values are referred to as moments and simply correspond to certain choices of $g(x)$.

Moments are characterised by their order, n , and a constant representing the point around which they are calculated, x_0 . In this text, $\mu'_n(x_0)$ refers to the n 'th moment of X around x_0 and is calculated by the integral

$$\mu'_n(x_0) = \int_{-\infty}^{\infty} (x - x_0)^n f_X(x) dx. \quad (2.2)$$

Two special sets of moments are often encountered in practice: first are the moments around the origin, designated by m_n ,

$$\begin{aligned} m_n &= \int_{-\infty}^{\infty} x^n f_X(x) dx \\ &= \mu'_n(0), \end{aligned} \quad (2.3)$$

followed by those calculated around the mean value (referred to as central moments), designated by μ_n . As the mean value of a random variable is simply the first moment around the origin, m_1 , the n 'th central moment is expressed by

$$\begin{aligned} \mu_n &= \int_{-\infty}^{\infty} (x - m_1)^n f_X(x) dx \\ &= \mu'_n(m_1). \end{aligned} \quad (2.4)$$

The importance of the central moments stems from the fact that they are invariant under a coordinate translation and therefore characterise the shape of the distribution around its mean. Characteristics like variance, skew and kurtosis are all central moments and it is evident that the mean cannot be a central moment, as it should change under a coordinate translation. Our interest in moments around the origin becomes apparent when we later consider the characteristic function, which is related to the Fourier transform of the PDF, and can be approximated in terms of moments around the origin.

A moment of order N , $\mu'_N(a)$, calculated around an arbitrary point a , can be represented as a linear combination of a set of moments, up to order N , $\{\mu'_n(b) : 0 \leq n \leq N\}$, calculated

around b (by applying the binomial theorem [29, p. A-1]):

$$\begin{aligned}
\mu'_N(a) &= \mathbb{E}\{(x - a)^N\} = \mathbb{E}\{(x - b + b - a)^N\} \\
&= \mathbb{E}\left\{\sum_{n=0}^N \binom{N}{n} (x - b)^{N-n} (b - a)^n\right\} \\
&= \sum_{n=0}^N \binom{N}{n} (b - a)^n \mathbb{E}\{(x - b)^{N-n}\} \\
&= \sum_{n=0}^N \binom{N}{n} (b - a)^n \mu'_{N-n}(b) .
\end{aligned} \tag{2.5}$$

This allows a set of moments calculated around any arbitrary point, e.g. the mean, to be expressed in terms of a set of moments around the origin. As we can always perform such a transformation, moments around the origin act as an ideal standard reference set by removing the need to remember the point around which a set of moments were calculated.

An important consideration to take into account when calculating moments, is their existence: as moments are calculated as integrals, a moment only exists if its defining integral converges. This is dependent on the function defining the PDF and distribution functions exist that do not possess moments of all orders, [4, p. 55]. The following general rule can prove useful when contemplating the existence of moments: if $\mu'_r(x_0)$ exists, it implies that $\mu'_s(x_0)$ exists for all $s < r$. Conversely, if $\mu'_r(x_0)$ does not exist, it implies that $\mu'_s(x_0)$ does not exist for all $s > r$.

2.3.2 Characteristic function

An alternative way of calculating the moments is by using the characteristic function, $\Phi_X(\omega)$, which is also defined in terms of an expectation [2, p 81],

$$\Phi_X(\omega) = \mathbb{E}\{e^{j\omega X}\}. \tag{2.6}$$

By substituting this into the definition of expectation (Equation 2.1) and using the definition of the Fourier transform (reviewed in Appendix A),

$$\begin{aligned}
\Phi_X(\omega) &= \mathbb{E}\{e^{j\omega X}\} \\
&= \int_{-\infty}^{\infty} f_X(x) e^{j\omega x} dx \\
&= \int_{-\infty}^{\infty} [f_X(x) e^{-j\omega x}]^* dx \\
&= \left[\int_{-\infty}^{\infty} f_X(x) e^{-j\omega x} dx \right]^* \\
&= [\mathcal{F}\{f_X(x)\}]^*,
\end{aligned} \tag{2.7}$$

it is seen (by taking the complex conjugate on both sides of the above expression) that $f_X(x)$ and $\Phi_X^*(\omega)$ constitute a Fourier transform pair where

$$\begin{aligned}
\Phi_X^*(\omega) &= \mathcal{F}\{f_X(x)\} \\
&= \int_{-\infty}^{\infty} f_X(x) e^{-j\omega x} dx
\end{aligned} \tag{2.8}$$

expresses $\Phi_X^*(\omega)$ using the Fourier transform and

$$\begin{aligned}
f_X(x) &= \mathcal{F}^{-1}\{\Phi_X^*(\omega)\} \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi_X^*(\omega) e^{j\omega x} d\omega
\end{aligned} \tag{2.9}$$

expresses $f_X(x)$ using the inverse Fourier transform.

The n 'th moment around the origin is obtained from $\Phi_X(\omega)$ by evaluating its n 'th derivative at the origin [2, p. 81],

$$m_n = \frac{\Phi_X^{(n)}(0)}{j^n}, \tag{2.10}$$

where $\Phi_X^{(n)}(0) = \left. \frac{d^n \Phi_X(\omega)}{d\omega^n} \right|_{\omega=0}$.

As the PDF only attains positive values and has a unity area, it is absolutely integrable:

$$\int_{-\infty}^{\infty} |f_X(x)| dx < \infty. \tag{2.11}$$

Consequently, its Fourier transform, and therefore the characteristic function, *always* exists (as the defining integral always converges). The characteristic function provides a way of calculating the moments of a random variable in situations where the Fourier transform of the PDF and its derivatives (if they exist) are available or can be calculated.

2.3.3 Fourier series

The Fourier series expresses a periodic function $f'(x)$, with period x_Δ , in terms of an infinite weighted sum of harmonically related complex exponential basis functions [5, p. 83], [20, p. 233]:

$$f'(x) = \sum_{k=-\infty}^{\infty} F_k e^{jk\omega_0 x}, \quad (2.12)$$

where F_k is a complex number that represents the k 'th Fourier series coefficient characterising $f'(x)$ and ω_0 represents the fundamental frequency, which is related to the period by

$$\omega_0 = \frac{2\pi}{x_\Delta}. \quad (2.13)$$

All functions $f'(x)$ satisfying the following conditions, known as Dirichlet conditions [20, p. 234], can be expressed in terms of such a Fourier series:

1. $f'(x)$ has a finite number of discontinuities over any x_Δ -wide interval.
2. $f'(x)$ contains a finite number of minima and maxima in any x_Δ -wide interval.
3. $f'(x)$ is absolutely integrable over any x_Δ -wide interval:

$$\int_{\lambda}^{\lambda+x_\Delta} |f'(x)| dx < \infty, \quad (2.14)$$

for any choice of λ .

For functions satisfying these conditions, values of the series coefficients characterising the function are calculated by evaluating an integral expression over a single period of the function (x_Δ):

$$F_k = \frac{1}{x_\Delta} \int_{\lambda}^{\lambda+x_\Delta} f'(x) e^{-jk\omega_0 x} dx, \quad (2.15)$$

As the above result does not depend on a specific choice of periodic interval, the choice of λ is arbitrary. Furthermore, if $f'(x)$ is a real-valued function, the series coefficients always

occur in complex conjugate pairs (with the exception of F_0 which is always a real number):

$$f'(x) \in \mathcal{R} \implies F_k = F_{-k}^*. \quad (2.16)$$

Of particular interest to us is the relationship between the Fourier series representation of $f'(x)$ and the Fourier transform (reviewed in Appendix A) of an aperiodic function, $f(x)$, that is related to $f'(x)$ by the following:

$$f'(x) = \sum_{n=-\infty}^{\infty} f(x - nx_{\Delta}). \quad (2.17)$$

$f'(x)$ is therefore a periodic extension of $f(x)$ with period x_{Δ} . If $F(\omega)$ is the Fourier transform of $f(x)$,

$$F(\omega) = \mathcal{F}\{f(x)\}, \quad (2.18)$$

then F_k is related to $F(\omega)$ by the following:

$$\begin{aligned} F_k &= \frac{1}{x_{\Delta}} F(k\omega_0) \\ &= \frac{1}{x_{\Delta}} F\left(k \frac{2\pi}{x_{\Delta}}\right). \end{aligned} \quad (2.19)$$

This allows the Fourier series coefficients of a periodic extension of a function to be calculated by uniformly sampling and scaling its Fourier transform, with the sampling interval inversely proportional to the period.

2.4 Estimators based on moments

We now investigate the feasibility of estimating a PDF entirely in terms of its moments. We first introduce the concept in principle and justify it using basic mathematical principles. A theoretical investigation follows, with the result being a novel expression for a PDF estimate in terms of a finite number of moments. Finally, a number of practical techniques that allow the PDF estimate to be evaluated anywhere within its domain, are developed from this theory. All these techniques only require the values of a number of moments (which can in turn be estimated from sample data) in order to produce an estimate.

Throughout the entire investigation, potential problems are identified and their possible consequences are acknowledged and addressed.

2.4.1 Motivation

We are accustomed to using moments to describe some aspects of PDFs of interest to us. Examples of this includes the mean, variance, skew, kurtosis or a combination of these which often provide valuable information about a random variable when describing or characterising it. In the case of Gaussian random variables, only the mean and variance are required in order to fully describe the random variable. We therefore feel intuitively attracted to using moments as a way of characterising PDFs and feel that they should provide at least some information allowing a PDF estimate to be constructed.

A second, and more substantiated, motivation for using moments to estimate a PDF stems from the so-called *Principle of Moments*. This is presented by *Kendall and Stuart* in *The Advanced Theory of Statistics* [4, pp. 86-88] where it is argued that PDFs encountered in practice are entirely characterised by their moments and that two distributions having a number of moments in common would bear some resemblance to each other. They furthermore consider the case of having two practical PDFs with the first N_m moments equal to each other and states that, as N_m tends to infinity, the two distributions tend to be identical to each other. From this is then concluded that two distributions, having a number of low-order moments in common, can be expected to be at least approximately equal to each other.

A proof of this (slightly modified from [4]) is obtained by considering a PDF, $f_X(x)$, that is continuous in the interval $x \in [-\frac{x_0}{2}, \frac{x_0}{2}]$, in which most of its area is contained. As a PDF is a non-negative function with unity area, a value of x_0 can always be found that satisfies these conditions. It is then approximated by $\hat{f}_X(x)$ in this interval by using a finite power series:

$$\hat{f}_X(x) = \sum_{n=0}^{N_m-1} a_n x^n \quad , \quad x \in [-\frac{x_0}{2}, \frac{x_0}{2}]. \quad (2.20)$$

The integral-squared error, ε , introduced by the approximation and calculated over $x \in [-\frac{x_0}{2}, \frac{x_0}{2}]$, is written in terms of this power series expansion and the original PDF:

$$\varepsilon = \int_{-\frac{x_0}{2}}^{\frac{x_0}{2}} \left\{ f_X(x) - \sum_{n=0}^{N_m-1} a_n x^n \right\}^2 dx. \quad (2.21)$$

We now find $\hat{f}_X(x)$ as the least-squares approximation of $f_X(x)$ by calculating the set of power series coefficients, $\{a_0, a_1, \dots, a_{N_m-1}\}$, that minimises the integral-squared error, ε .

This is done by equating the partial derivative of the error with respect to each coefficient, $\frac{\partial \varepsilon}{\partial a_i}$, to zero and solving for a_i :

$$\begin{aligned} \frac{\partial \varepsilon}{\partial a_i} &= 2 \int_{-\frac{x_0}{2}}^{\frac{x_0}{2}} \left\{ f_X(x) - \sum_{n=0}^{N_m-1} a_n x^n \right\} x^i dx \\ &= 0. \end{aligned} \quad (2.22)$$

By reorganising the terms, a general solution to this equation is obtained as

$$\int_{-\frac{x_0}{2}}^{\frac{x_0}{2}} x^i f_X(x) dx = \int_{-\frac{x_0}{2}}^{\frac{x_0}{2}} \sum_{n=0}^{N_m-1} a_n x^{n+i} dx, \quad i \in \{0, 1, \dots, N_m - 1\}. \quad (2.23)$$

This represents N_m equations (corresponding to different values of i) that has to be satisfied simultaneously in order for the least squares approximation to be solved. As most of the area of $f_X(x)$ is concentrated within the interval $x \in [-\frac{x_0}{2}, \frac{x_0}{2}]$, the integration limits of the left-hand side expression in the above equation can be changed to $\pm\infty$ without introducing a large error:

$$\int_{-\infty}^{\infty} x^i f_X(x) dx \approx \int_{-\frac{x_0}{2}}^{\frac{x_0}{2}} \sum_{n=0}^{N_m-1} a_n x^{n+i} dx, \quad i \in \{0, 1, \dots, N_m - 1\}. \quad (2.24)$$

The left-hand side is identified as the expression for the i 'th moment of $f_X(x)$ around the origin (m_i), allowing the least-squares approximation of $f_X(x)$ to be expressed entirely in terms of the first N_m moments around the origin:

$$\int_{-\frac{x_0}{2}}^{\frac{x_0}{2}} \sum_{n=0}^{N_m-1} a_n x^{n+i} dx \approx m_i, \quad i \in \{0, 1, \dots, N_m - 1\}. \quad (2.25)$$

The power series coefficients corresponding to the least-squares approximation of $f_X(x)$ are then found by minimising the following expression simultaneously for all $i \in \{0, 1, \dots, N_m - 1\}$:

$$\left| \int_{-\frac{x_0}{2}}^{\frac{x_0}{2}} \sum_{n=0}^{N_m-1} a_n x^{n+i} dx - m_i \right|, \quad i \in \{0, 1, \dots, N_m - 1\}. \quad (2.26)$$

Although the above result does not represent a practical solution to the least-squares approximation (as it is difficult to solve), it is still theoretically significant: an N_m -term least-squares approximation of $f_X(x)$ over the interval $x \in [-\frac{x_0}{2}, \frac{x_0}{2}]$ is seen to be *only*

a function of the first N_m moments of $f_X(x)$. A direct consequence of this is that two distribution functions having the first N_m moments in common, have a similar least squares approximation at least up to order N_m . This *Principle of Moments* prompted further investigation into using moments as a means of completely describing and characterising a PDF.

2.4.2 A PDF in terms of moments

We now use the mathematical relationship between the moments of a PDF and the function describing the PDF to obtain an estimate of the PDF, $\hat{f}_X(x)$, from the values of a number of moments. In order to simplify the equations slightly, we impose the following restrictions:

1. The random variable, X , represented by the PDF, is assumed to be standardised and to therefore have a mean value of zero and unity variance. This is a reasonable assumption, as any random variable, X' , can be standardised as follows, using a simple linear transform:

$$X = \frac{X' - E\{X'\}}{\sqrt{E\{X'^2\} - E\{X'\}^2}} \quad (2.27)$$

This does, however, require that we are able to estimate accurate values of the first two moments of X around the origin ($E\{X'\}$ and $E\{X'^2\}$). As our procedure requires a number of finite moments up to order N_m , this is implied.

2. We know the values of all moments up to order N_m around the *origin*, $\{m_n : 0 \leq n \leq N_m\}$. Equation 2.5 can be used to convert the first moments up to order N_m calculated around an arbitrary point to the required set of $N_m + 1$ moments around the origin.

In Equation 2.10, we expressed moments of X around the origin, m_n , in terms of the characteristic function, $\Phi_X(\omega)$. We now wish to do the inverse of that, and express $\Phi_X(\omega)$ in terms of a finite number of moments. We start by expressing the characteristic function in terms of Taylor polynomials [29, p. 624]:

$$\begin{aligned} \Phi_X(\omega) &= \sum_{n=0}^{N_m-1} \frac{\Phi_X^{(n)}(0)}{n!} \omega^n + R_{N_m}(\omega) \\ &= \hat{\Phi}_X(\omega) + R_{N_m}(\omega), \end{aligned} \quad (2.28)$$

where $\hat{\Phi}_X(\omega)$ is a power series approximation of the characteristic function,

$$\hat{\Phi}_X(\omega) = \sum_{n=0}^{N_m-1} \frac{\Phi_X^{(n)}(0)}{n!} \omega^n, \quad (2.29)$$

and is the same for all random variables having identical values for the first N_m moments. $R_{N_m}(\omega)$ is the *remainder* and indicates the error introduced by this approximation. Equation 2.10 allows us to express $\Phi_X^{(n)}(0)$ in terms of the n 'th moment:

$$\Phi_X^{(n)}(0) = j^n m_n. \quad (2.30)$$

Substituting this into Equation 2.29 allows us to express it in terms of the first N_m moments:

$$\hat{\Phi}_X(\omega) = \sum_{n=0}^{N_m-1} \frac{j^n m_n}{n!} \omega^n. \quad (2.31)$$

Using Equation 2.9, the approximate PDF is expressed in terms of this approximate characteristic function through the Fourier transform relationship:

$$\begin{aligned} \hat{f}_X(x) &= \mathcal{F}^{-1} \left\{ \hat{\Phi}_X^*(\omega) \right\} \\ &= \mathcal{F}^{-1} \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} \omega^n \right\} \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} \omega^n \right\} e^{j\omega x} d\omega. \end{aligned} \quad (2.32)$$

This equation expresses an approximation of the PDF, $\hat{f}_X(x)$, directly in terms of the first N_m moments. As $\hat{\Phi}_X(-\omega) = \hat{\Phi}_X^*(\omega)$ always holds¹, $f_X(x)$ is always a real function (from a property of the Fourier transform), which is expected as a PDF is always a real function. Unfortunately, this integral cannot be evaluated, as the integrand diverges to plus and minus infinity as $\omega \rightarrow \pm\infty$. Regardless of the order of the polynomial, a value of ω always

¹This is seen by noting that $(j^n)^* = (-j)^n$ and $(-x)^n = (-1)^n x^n$ and then comparing $\hat{\Phi}_X(-\omega)$ and $\hat{\Phi}_X^*(\omega)$.

exists for which the highest powered term dominates as ω increases,

$$\begin{aligned} \lim_{\omega \rightarrow \infty} \hat{\Phi}_X(\omega) &= \lim_{\omega \rightarrow \infty} \sum_{n=0}^{N_m-1} \frac{j^n m_n}{n!} \omega^n \\ &= \frac{j^n m_n}{n!} \omega^n \Big|_{n=N_m-1}. \end{aligned} \quad (2.33)$$

Figure 2.1 illustrates the divergence of the approximation of the characteristic function of a Gaussian PDF, corresponding to three different values of N_m .

This is a property of the finite-order polynomial approximation and there is no direct way of avoiding this. It is an undesirable property, as it introduces unwanted high-frequency components into the estimate. These components severely distort the PDF, introducing artifacts such as oscillations and negative values into the estimate. A way of dealing with this is to only reconstruct $\hat{\Phi}_X(\omega)$ within a symmetrical region centered at the origin, for which we know the reconstruction to be well-behaved, and to assume values of zero for the function otherwise. As $\hat{\Phi}_X(\omega)$ is a frequency function that is related to $\hat{f}_X(x)$, this imposes a direct constraint in terms of the energy of the high frequency components that are present in $\hat{f}_X(x)$ and any further estimates using this $\hat{\Phi}_X(\omega)$ would reflect this. Without more detailed knowledge of the actual characteristic function, $\Phi_X(\omega)$, it is impossible to quantify the error introduced by truncating $\hat{\Phi}_X(\omega)$ in this way.

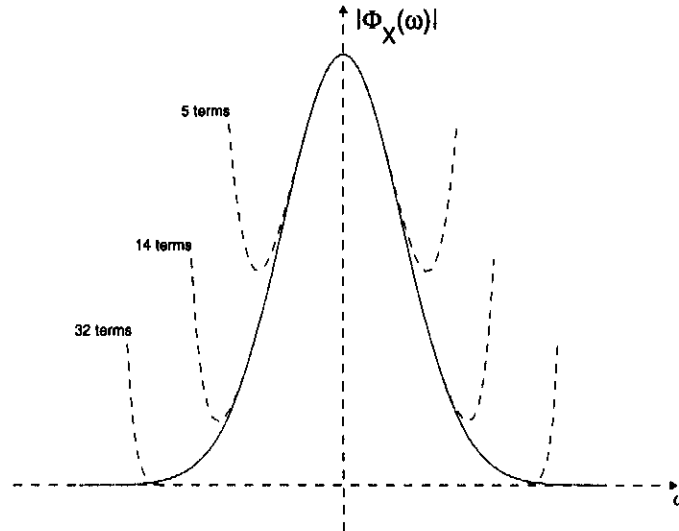


Figure 2.1: Successive Taylor series approximations to characteristic function of Gaussian PDF using 5, 14, and 32 terms.

However, the structure of the Fourier transform does provide us with valuable information as to the type of approximation errors we can expect as high frequency components play the most important role in the realisation of entities comprising the micro structure of the PDF. Information regarding macroscopic structure, on the other hand, are mostly contained in the lower frequency components. Elimination of high frequencies would therefore imply smoothing of the PDF estimate, with the smaller features being affected most. This becomes apparent when we consider the mathematical effect that multiplication of the spectrum by a rectangular windowing function, which is exactly what this proposed truncation does, has on the PDF. The rectangular windowing function is defined as follows:

$$\Theta_{rect}(\omega_u, \omega) = \begin{cases} 1 & |\omega| \leq \omega_u \\ 0 & |\omega| > \omega_u, \end{cases} \quad (2.34)$$

where ω_u is referred to as the upper frequency. The convolution property of the Fourier transform states that multiplication in the frequency domain amounts to convolution in the spatial domain, i.e.

$$\begin{aligned} \mathcal{F}^{-1}\left\{\Phi_1(\omega)\Phi_2(\omega)\right\} &= \int_{-\infty}^{\infty} f_1(\xi)f_2(x-\xi)d\xi \\ &= f_1(x) \star f_2(x), \end{aligned} \quad (2.35)$$

where $\Phi_n(\omega) = \mathcal{F}\{f_n(x)\}$ and $f_1(x) \star f_2(x)$ denotes the convolution of $f_1(x)$ and $f_2(x)$.

Substituting the characteristic function and the windowing function for $\Phi_1(\omega)$ and $\Phi_2(\omega)$ respectively in Equation 2.35, shows the PDF to be equal to the inverse Fourier transform of the characteristic function convolved with the inverse Fourier transform of

the windowing function, $\frac{\sin(\omega_u x)}{\pi x}$:

$$\begin{aligned}
\hat{f}_X(x) &= \frac{1}{2\pi} \int_{-\omega_u}^{\omega_u} \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} w^n \right\} e^{jwx} dw \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \Theta_{rect}(\omega_u, \omega) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} w^n \right\} e^{jwx} dw \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \Theta_{rect}(\omega_u, \omega) \hat{\Phi}_X^*(\omega) e^{jwx} dw \\
&= \mathcal{F}^{-1} \left\{ \Theta_{rect}(\omega_u, \omega) \hat{\Phi}_X^*(\omega) \right\} \\
&= \mathcal{F}^{-1} \left\{ \Theta_{rect}(\omega_u, \omega) \right\} \star \mathcal{F}^{-1} \left\{ \hat{\Phi}_X^*(\omega) \right\} \\
&= \frac{\sin \omega_u x}{\pi x} \star \mathcal{F}^{-1} \left\{ \hat{\Phi}_X^*(\omega) \right\}.
\end{aligned} \tag{2.36}$$

A higher value of ω_u results in a narrower function applied in the convolution, therefore causing less smoothing. A rectangular window is not the only option and other windows, for instance Hamming, Hanning or Bartlett (triangular) windows can also be employed. A general windowing function, $\Theta(\omega_u, \omega)$, is incorporated into Equation 2.32, resulting in a PDF estimate that contains no energy above the upper frequency (ω_u):

$$\hat{f}_X(x) = \frac{1}{2\pi} \int_{-\omega_u}^{\omega_u} \Theta(\omega_u, \omega) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} w^n \right\} e^{jwx} dw. \tag{2.37}$$

This allows an appropriate choice of windowing function to completely mask the effect of the polynomial divergence. The general windowing function depicted above, $\Theta(\omega_u, \omega)$, is an even real-valued function attaining the value of zero outside of a symmetrical range centered at the origin, $|\omega| > \omega_u$:

$$\Theta(\omega_u, \omega) = \begin{cases} \Theta'(\omega_u, |\omega|) & ; \quad |\omega| \leq \omega_u \\ 0 & ; \quad |\omega| > \omega_u, \end{cases} \tag{2.38}$$

and its inverse Fourier transform, $\theta(\omega_u, x)$, is also a real-valued even function:

$$\theta(\omega_u, x) = \mathcal{F}^{-1} \left\{ \Theta(\omega_u, \omega) \right\}. \tag{2.39}$$

Figure 2.2 illustrates the smoothing effect that the truncation of the characteristic function by a Hamming window has on the density function approximation. It features the reconstruction of a PDF, consisting of a mixture of a Gaussian and a uniform density function, from a truncated characteristic function. As ω_u is selected to have a low value, the smoothing significantly degrades the uniform density function, which requires high-frequency components to realise its discontinuities.

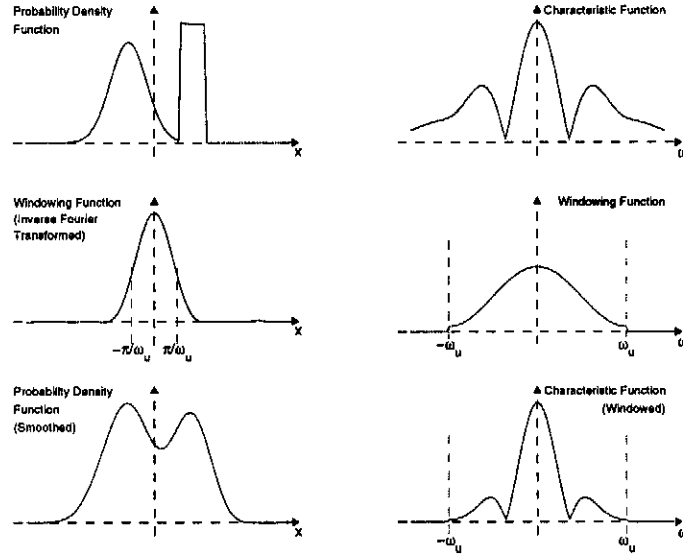


Figure 2.2: The smoothing effect of a Hamming window on the characteristic function.

In principle we are not overly concerned about this smoothing effect, as it aids the estimator in generalising across the sample data by averaging (smoothing) over it. Even if we had an expression for the actual characteristic function describing the sample data, we would still prefer to generate an estimate with suppressed high frequency components. This is because the PDF reconstructed using the entire spectrum would reflect the exact distribution of the *observed* sample set, and would therefore be over-specialised. One of the reasons for using density functions is for their generalisation abilities in obtaining a more general picture of the underlying process than the raw data provides. Smoothing does, however, become a problem if the highest frequency component is so low that important discriminating features and characteristics of the PDF are obscured.

Elimination of the high frequencies unfortunately also has a second, and more problematic, side-effect, namely leakage [20, pp. 623-629]. As the windowing is performed in the *frequency domain*, leakage is observed in the *spatial domain*, with the effects visible in

the reconstructed PDF. Although this is the inverse of the scenario involving time-domain windowing, popularly encountered in signal processing, the same principles hold and similar problems are therefore still experienced. Our concern stems from the fact that a PDF may *never* attain negative values, as this is undefined and wreaks havoc with any method that considers density functions in the logarithmic domain. Doing this would also cause the estimate to violate one of the principle properties of a valid PDF. The inverse Fourier transformed windowing function, $\theta(\omega_u, x)$ in Equation 2.39, can attain negative values, for example when using rectangular, Hamming or Hanning windowing functions. As this is convolved with a PDF that is a non-negative function, the resulting function is therefore almost guaranteed to *also* attain negative values. Leakage has a further tendency to introduce ripple into a density function at regions of constant probability density. This is due to windowing functions exhibiting (damped) oscillatory behaviour in the spatial domain. Regions of low and near-constant probability density, expected far (> 5 standard deviations) away from the mean value, is mostly affected by this.

We address this problem by taking the absolute value of $\hat{f}_X(x)$ in Equation 2.37 as the value of the PDF estimate,

$$\hat{f}_X(x) = \left| \frac{1}{2\pi} \int_{-\omega_u}^{\omega_u} \Theta(\omega_u, \omega) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} \omega^n \right\} e^{j\omega x} d\omega \right|. \quad (2.40)$$

Figure 2.3 illustrates the effect of leakage introduced by rectangular windowing of the characteristic function. This should be compared with Figure 2.2, which used the same upper frequency, ω_u , and PDF, but which displayed much less leakage. On the other hand, the rectangular windowing function results in significantly less smoothing due to the narrower main lobe. There is therefore always a trade-off between the main lobe width (principally responsible for smoothing) and the normalised side lobe magnitudes (principally responsible for ripple) [20, p. 436].

Some solutions to $\hat{f}_X(x)$, obtained by evaluating the integral in Equation 2.40, are only accurate over a subset of their domain surrounding the origin (the reasons for this is discussed in detail in the following sections where it is encountered). In such an event the definition of $\hat{f}_X(x)$ is further amended so that it attains a value of zero outside the interval, $|x| \leq \frac{x\Delta}{2}$:

$$\hat{f}_X(x) = \begin{cases} \left| \frac{1}{2\pi} \int_{-\omega_u}^{\omega_u} \Theta(\omega_u, \omega) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} \omega^n \right\} e^{j\omega x} d\omega \right| & ; \quad |x| \leq \frac{x\Delta}{2} \\ 0 & ; \quad |x| > \frac{x\Delta}{2}. \end{cases} \quad (2.41)$$

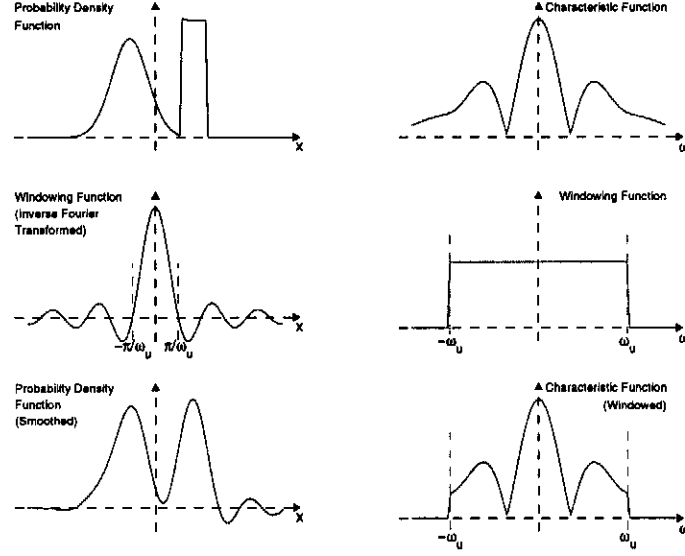


Figure 2.3: Leakage introduced by rectangular windowing of the characteristic function.

In performing the truncation illustrated above, a part of the sample space is effectively neglected and all data that may have been observed in that part is simply disregarded. Justification for this action is found in the Bienaymé – Tchebycheff inequality [4, p. 88]. It provides an upper bound for the probability of observing data outside a symmetrical interval surrounding the mean, μ_X , for a specified variance, σ_X^2 :

$$P\{|X - \mu_X| \geq \lambda \sigma_X\} \leq \frac{1}{\lambda^2}. \quad (2.42)$$

Recalling that X is standardised, we consider the inequality with $\mu_X = 0$ and $\sigma_X = 1$ to obtain an upper bound for the error introduced by truncating $\hat{f}_X(x)$, to the range $x \in [-\frac{x_\Delta}{2}, \frac{x_\Delta}{2}]$:

$$\varepsilon = \left(\frac{2}{x_\Delta} \right)^2. \quad (2.43)$$

This error is interpreted as the probability of assigning a zero probability density to a sample that may have a non-zero probability density. In other words, it indicates the fraction of the sample space (located in the tails of the PDF) that is effectively ignored in this approximation. Larger values of x_Δ produces better results and, depending on the application, it is recommended to use a value of at least 5 times the standard deviation.

We now have the ability to express a PDF estimate mathematically in terms of a number of sample moments, using Equation 2.41. The expression allows the introduction of an arbitrary frequency-domain windowing function, which determines the overall level of detail of the approximation and requires the selection of a value determining the size of an interval over which the estimate is calculated.

In the following sections, we devote our attention towards practical ways of solving this integral expression. The objective is to obtain one or more closed-form solutions that expresses $\hat{f}_X(x)$ in terms of a number of elementary mathematical functions of a finite number of moments. A number of symbols from Equation 2.41 are used throughout and retain the meanings they had there: $\hat{f}_X(x)$, $\Theta(\omega_u, \omega)$, N_m , m_N , ω_u , x_Δ , ω and x . Additionally, ω_0 indicates the fundamental frequency, applicable to techniques employing a discrete frequency approximation.

2.4.3 The anti-derivative

We start by attempting to solve the integral directly using the anti-derivative. By interchanging the order of summation and integration in Equation 2.40, assuming a rectangular windowing function for $\Theta(\omega_u, \omega)$ and expanding the complex exponential into a real sine and cosine term, we obtain the following expression:

$$\begin{aligned}\hat{f}_X(x) &= \left| \frac{1}{2\pi} \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} \left\{ \int_{-\omega_u}^{\omega_u} \omega^n e^{j\omega x} d\omega \right\} \right| \\ &= \left| \frac{1}{2\pi} \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} \left\{ \int_{-\omega_u}^{\omega_u} \omega^n \cos(\omega x) d\omega + j \int_{-\omega_u}^{\omega_u} \omega^n \sin(\omega x) d\omega \right\} \right|.\end{aligned}\tag{2.44}$$

This expresses $\hat{f}_X(x)$ in terms of a weighted sum of basis functions, where the n 'th basis function is the inverse Fourier transform of ω^n multiplied by a rectangular windowing function. Evaluating this expression requires that the anti-derivatives of the integrands be evaluated:

$$\begin{aligned}\int \omega^n \cos(\omega x) d\omega &= \sum_{k=0}^n k! \binom{n}{k} \frac{\omega^{n-k}}{x^{k+1}} \sin(\omega x + \frac{k\pi}{2}) \\ \int \omega^n \sin(\omega x) d\omega &= - \sum_{k=0}^n k! \binom{n}{k} \frac{\omega^{n-k}}{x^{k+1}} \cos(\omega x + \frac{k\pi}{2}).\end{aligned}\tag{2.45}$$

Evaluating the anti-derivative for values of x close to the origin proves to be problematic, due to the divergent nature of the terms. All but one term (corresponding to $k = 0$) tends to $\pm\infty$ in the limit as $x \rightarrow 0$. This results in numerical instability as attempts are made to evaluate $\hat{f}_X(x)$ near the origin.

The anti-derivative also poses another problem: a higher-order Taylor polynomial approximation to the characteristic function would usually be preferable to a lower-order one, as it provides a better approximation further away from the origin. In this case, increasing the order of the approximation, n , also increases the maximum value of the arguments to the factorial and combinatory term. This introduces additional numerical instability into the calculations. Unfortunately, this is also not solved by transforming the problem to the logarithmic domain, as it requires the accurate summation of a number terms of differing order of magnitude.

This leads to a conflicting requirement, as a higher number of moments is preferred in order to improve the approximation to the characteristic function, while it leads to increased numerical instability when evaluating the anti-derivative. It is therefore not recommended that the anti-derivative be used to evaluate the integral and more practical alternatives are now considered.

2.4.4 Numerical integration techniques

Employing a numerical integration technique eliminates the problems associated with the anti-derivative. The restriction on the maximum number of moments vanishes, thereby allowing a reconstruction exhibiting a high level of detail. It also solves the problem of the value of the approximation at small values of x (i.e. close to the origin), as the integral itself is defined there, although the anti-derivative is not. A further advantage is the ability to seamlessly integrate almost any windowing function, without adding complexity. The only requirement is the ability to obtain values of the windowing function sampled at regular intervals in its domain, while using the anti-derivative required analytical introduction of the windowing function expression into the solution. Careful use of windowing functions, as well as the ability to obtain an estimate rich in high frequencies, contains the error introduced by taking the absolute value in Equation 2.40.

Numerical integration techniques partition the domain of integration into a finite number of intervals, N_i , and then approximate the integral over each interval in terms of the value of the integrand at the interval boundaries. These approximate integrals are then accumulated to approximate the original integral. Techniques differ in the way that the

integration domain is partitioned as well as the way in which the integral is approximated in terms of the boundary values.

We can, however, obtain a general expression that facilitates the evaluation of Equation 2.40 by means of different numerical integration techniques. This requires the following changes to the integral expression:

1. The continuous integration over the interval $\omega \in [-\omega_u, \omega_u]$ is replaced by an $(N_i + 1)$ -term discrete sum. This requires that all occurrences of the integration variable, ω , be replaced by a discrete variable, $\{\gamma_k : 0 \leq k \leq N_i\}$.
2. A weight, β_k is added to each term comprising the sum. This accounts for the specific way in which the technique combines the values at the interval boundaries to estimate the integrand over the interval.
3. A scaling term, α , is added to account for loss of the differential ($d\omega$) and also to balance any scaling factors introduced by the β_k factors.

Expressions for α , β_k and γ_k are dependent on the details of the numerical integration technique that is employed. Certain techniques may even impose restrictions on the values that N_i may attain. The changes proposed above are applied to Equation 2.40 and the resulting expression is simplified, resulting in a general expression:

$$\begin{aligned}
 \hat{f}_X(x) &= \left| \frac{1}{2\pi} \int_{-\omega_u}^{\omega_u} \Theta(\omega_u, \omega) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} \omega^n \right\} e^{j\omega x} d\omega \right| \\
 &= \left| \frac{1}{2\pi} \alpha \sum_{k=0}^{N_i} \beta_k \Theta(\omega_u, \gamma_k) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} (\gamma_k)^n \right\} e^{j\gamma_k x} \right| \\
 &= \left| \sum_{k=0}^{N_i} \beta_k \frac{\alpha}{2\pi} \Theta(\omega_u, \gamma_k) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} (\gamma_k)^n \right\} e^{j\gamma_k x} \right|.
 \end{aligned} \tag{2.46}$$

By factoring out the part of the expression that is *not* a function of x ,

$$\varphi_k = \beta_k \frac{\alpha}{2\pi} \Theta(\omega_u, \gamma_k) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} (\gamma_k)^n \right\}, \tag{2.47}$$

$\hat{f}_X(x)$ can be expressed in a form resembling a Fourier series:

$$\hat{f}_X(x) = \left| \sum_{k=0}^{N_i} \varphi_k e^{j\gamma_k x} \right|. \quad (2.48)$$

In order to use the above equations one only needs to determine suitable expressions for α , β_k and γ_k corresponding to the desired integration technique. This is now performed for a popular integration technique.

Simpson's rule [29] is a numerical integration technique that uses parabolas to approximate the areas under the curve being integrated. The integration domain, $[-\omega_u, \omega_u]$ is partitioned by an *odd* number of equidistant points, into N_i bins. In order to apply this rule to the solution of the integral, define the unknown expressions as follows:

$$\begin{aligned} \alpha &= \frac{2\omega_u}{3N_i} \\ \beta_k &= \begin{cases} 1 & ; \quad k = 0, N_i \\ 3 - (-1)^k & ; \quad k = 1, 2, 3, \dots, N_i - 1 \end{cases} \\ \gamma_k &= \frac{2k - N_i}{N_i} \omega_u. \end{aligned} \quad (2.49)$$

Considering the expression for α leads to an interesting observation of major practical importance: Equation 2.48 expresses $\hat{f}_X(x)$ in terms of a finite sum of a number of weighed complex exponential functions. If the frequencies of the exponential terms, γ_k , are harmonically related, the sum constitutes a Fourier series [5, p. 83] resulting in $\hat{f}_X(x)$ being a periodic function (with the period being the inversely proportional to the fundamental frequency). When employing Simpson's rule, γ_k is always an integer multiple of $\frac{2\omega_u}{N_i}$, and it is therefore expected that $\hat{f}_X(x)$ would be periodic with period $x_0 = \frac{\pi N_i}{\omega_u}$. This is verified

by combining Equation 2.48 and Equation 2.49 and substituting $x = x + \frac{2\pi N_i}{\omega_u}$:

$$\begin{aligned}
\hat{f}_X\left(x + \frac{2\pi N_i}{\omega_u}\right) &= \left| \sum_{k=0}^{N_i} \varphi_k e^{j\gamma_k\left(x + \frac{2\pi N_i}{\omega_u}\right)} \right| \\
&= \left| \sum_{k=0}^{N_i} \varphi_k e^{j\gamma_k x} e^{j\left(\frac{2k-N_i}{N_i}\omega_u\right)\left(\frac{2\pi N_i}{\omega_u}\right)} \right| \\
&= \left| \sum_{k=0}^{N_i} \varphi_k e^{j\gamma_k x} e^{j2\pi(2k-N_i)} \right| \\
&= \left| \sum_{k=0}^{N_i} \varphi_k e^{j\gamma_k x} \right| \\
&= \hat{f}_X(x).
\end{aligned} \tag{2.50}$$

Care should therefore be taken to guard against artifacts introduced by the numerical integration procedures. Periodicity, illustrated above, is encountered when using any technique that samples the integrand uniformly and symmetrically around the origin, i.e. if γ_k can be expressed as $\gamma_k = k\omega_0$, with $k = 0, 1, 2, \dots$. In this case, the estimate is only valid over an interval defined in terms of the fundamental frequency, ω_0 , and the PDF is assumed to be zero outside the interval $|x| \leq \frac{\pi}{\omega_0}$:

$$\hat{f}_X(x) = \begin{cases} \left| \sum_{k=0}^{N_i} \varphi_k e^{j\gamma_k x} \right| & ; \quad |x| \leq \frac{\pi}{\omega_0} \\ 0 & ; \quad |x| > \frac{\pi}{\omega_0}. \end{cases} \tag{2.51}$$

2.4.5 Fourier series approximation

A solution that expresses $\hat{f}_X(x)$ as a sum of harmonic complex exponential functions is now considered. This is accomplished by sampling the characteristic function at discrete intervals in the frequency domain, and using this discretised function to obtain the PDF estimate. As the resulting function is periodic, we only define $\hat{f}_X(x)$ to be equal to the periodic function over a single period, and to be zero everywhere else.

Let $f(x)$ be a function that is defined in terms of the integral in Equation 2.40 and let $F(\omega)$ be its Fourier transform:

$$f(x) = \frac{1}{2\pi} \int_{-\omega_u}^{\omega_u} \Theta(\omega_u, \omega) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} \omega^n \right\} e^{j\omega x} d\omega. \tag{2.52}$$

As $\Theta(\omega_u, \omega)$ is equal to zero outside the interval $|\omega| \leq \omega_u$, the integration bounds can be changed to $\pm\infty$ without changing the value of the integral. The integral is then identified as an inverse Fourier transform, allowing $F(\omega)$ to be isolated:

$$\begin{aligned}
f(x) &= \frac{1}{2\pi} \int_{-\omega_u}^{\omega_u} \Theta(\omega_u, \omega) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} \omega^n \right\} e^{j\omega x} d\omega \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \Theta(\omega_u, \omega) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} \omega^n \right\} e^{j\omega x} d\omega \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{j\omega x} d\omega, \\
&= \mathcal{F}^{-1}\{F(\omega)\}
\end{aligned} \tag{2.53}$$

with

$$F(\omega) = \Theta(\omega_u, \omega) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} \omega^n \right\}. \tag{2.54}$$

Let $f'(x)$ be a periodic extension of $f(x)$ with a period of x_Δ :

$$f'(x) = \sum_{k=-\infty}^{\infty} f(x - kx_\Delta). \tag{2.55}$$

From the relationship between the Fourier series and the Fourier transform (Section 2.3.3), $f'(x)$ can be expressed in terms of a Fourier series, with the series coefficients calculated in terms of $F(\omega)$, the Fourier transform of $f(x)$:

$$f'(x) = \sum_{k=-\infty}^{\infty} F_k e^{j\frac{2\pi k}{x_\Delta} x}, \tag{2.56}$$

with

$$F_k = \frac{1}{x_\Delta} F\left(\frac{2\pi k}{x_\Delta}\right). \tag{2.57}$$

Furthermore, as $F(\omega)$ is zero outside the interval $|\omega| \leq \omega_u$, the Fourier series only contains

a finite number of terms:

$$\begin{aligned} f'(x) &= \sum_{k=-\infty}^{\infty} F_k e^{j \frac{2\pi k}{x_\Delta} x} \\ &= \sum_{k=-\alpha}^{\alpha} F_k e^{j \frac{2\pi k}{x_\Delta} x} \end{aligned} \quad (2.58)$$

where

$$\alpha = \left\{ \max(k) : \frac{2\pi k}{x_\Delta} \leq \omega_u, k \in \mathcal{N} \right\}. \quad (2.59)$$

If x_Δ is selected so that $f(x) \approx 0$ if $|x| > \frac{x_\Delta}{2}$, then $f(x) \approx f'(x)$ in the interval $|x| \leq \frac{x_\Delta}{2}$. If $f(x)$ represents a PDF estimate, such a value of x_Δ can always be found (with the help of the Bienaymé – Tchebycheff inequality), as a PDF has finite area. We can therefore express $f(x)$ approximately in terms of a Fourier series as follows (by substituting the definitions of $F(\omega)$ and F_k):

$$\begin{aligned} f(x) &\approx f'(x) \\ &= \sum_{k=-\alpha}^{\alpha} F_k e^{j \frac{2\pi k}{x_\Delta} x} \\ &= \sum_{k=-\alpha}^{\alpha} \frac{1}{x_\Delta} F\left(\frac{2\pi k}{x_\Delta}\right) e^{j \frac{2\pi k}{x_\Delta} x} \\ &= \sum_{k=-\alpha}^{\alpha} \frac{1}{x_\Delta} \Theta\left(\omega_u, \frac{2\pi k}{x_\Delta}\right) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} \left(\frac{2\pi k}{x_\Delta}\right)^n \right\} e^{j \frac{2\pi k}{x_\Delta} x}. \end{aligned} \quad (2.60)$$

The above approximates $f(x)$ in terms of a finite Fourier series that only involves a finite number of moments and a windowing function. By substituting this approximation for the integral expression in Equation 2.40, we express the PDF estimate in terms of a Fourier series that only involves a finite number of moments and a windowing function:

$$\begin{aligned} \hat{f}_X(x) &= \left| \sum_{k=-\alpha}^{\alpha} \frac{1}{x_\Delta} \Theta\left(\omega_u, \frac{2\pi k}{x_\Delta}\right) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} \left(\frac{2\pi k}{x_\Delta}\right)^n \right\} e^{j \frac{2\pi k}{x_\Delta} x} \right| \\ &= \left| \sum_{k=-\alpha}^{\alpha} \beta_k e^{j \frac{2\pi k}{x_\Delta} x} \right|. \end{aligned} \quad (2.61)$$

$$\beta_k = \frac{1}{x_\Delta} \Theta(\omega_u, \frac{2\pi k}{x_\Delta}) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} \left(\frac{2\pi k}{x_\Delta} \right)^n \right\}. \quad (2.62)$$

This represents $\hat{f}_X(x)$ as a periodic function. As a PDF always has unity area and can therefore never be periodic, the resulting periodic function cannot be directly used as a PDF. An aperiodic PDF is obtained by defining the PDF to be equal to the above function in the interval $|x| \leq \frac{x_\Delta}{2}$ and zero outside of it:

$$\hat{f}_X(x) = \begin{cases} \left| \sum_{k=-\alpha}^{\alpha} \beta_k e^{j \frac{2\pi k}{x_\Delta} x} \right| & ; \quad |x| \leq \frac{x_\Delta}{2} \\ 0 & ; \quad |x| > \frac{x_\Delta}{2}. \end{cases} \quad (2.63)$$

Obtaining a PDF estimate in terms of a Fourier series therefore required the following steps:

1. Sampling the characteristic function to express the PDF estimate in terms of an (infinite) Fourier series. This causes the PDF estimate to be periodically extended and requires correct selection of x_Δ to minimise the distortion caused by this.
2. Windowing of the characteristic function (by multiplying the characteristic function estimate by a windowing function, $\Theta(\omega_u, \omega)$). This introduces smoothing into the PDF estimate as it is convolved with the inverse Fourier transformed windowing function.
3. Truncation of the PDF estimate to zero outside its first period.

Figure 2.4 illustrates the discretisation of the characteristic function and the relationship between the period of the PDF and the sampling frequency of $\hat{\Phi}_X(\omega)$.

2.4.6 Estimating moments from sample data

Although the moments played a central role in the calculations in the preceding section, no mention was made as to how they were obtained. The methods for calculating approximations to density functions are applied directly when the values of the moments are known beforehand or can be calculated for a finite number of moments. If, however, only sample data are available, the moments have to be obtained by estimating it from the sample data. These estimates are then substituted for values of the moments in all the

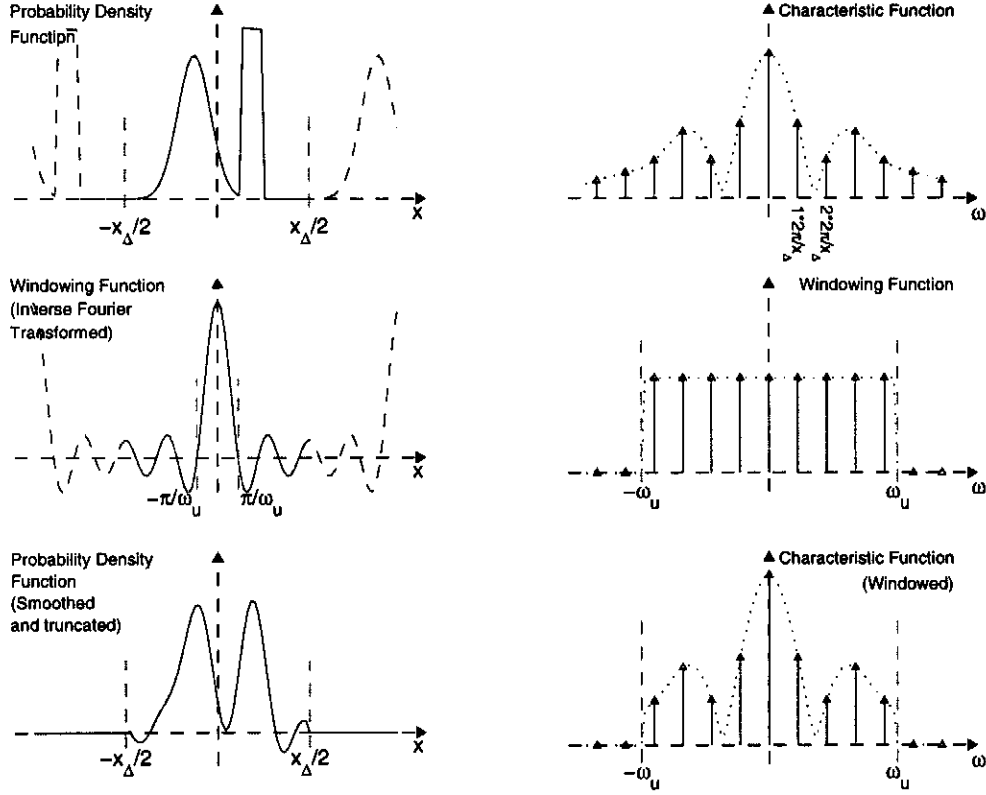


Figure 2.4: Discretisation of the characteristic function.

preceding expressions. Assume that the sample set consists of N_x samples, with the i 'th sample indicated by x_i . An estimate for the n 'th order moment is then obtained by

$$\hat{m}_n = \frac{1}{N_x} \sum_{i=0}^{N_x-1} x_i^n. \quad (2.64)$$

In order to establish the validity of this estimation, we consider the mean value and variance of the estimate in terms of the actual (population) moments. This is obtained from [4, p. 229]:

$$\mathbb{E}\{\hat{m}_n\} = m_n. \quad (2.65)$$

$$\mathbb{E}\{(\hat{m}_n - m_n)^2\} = \frac{1}{N_x} (m_{2n} - m_n^2). \quad (2.66)$$

As the mean value of the estimate is equal to the value being estimated, this represents an unbiased estimate [30, p. 119]. The variance vanishes in the limit as the number of

samples, N_x , tends to infinity, which is a sufficient condition for the estimate to be weakly consistent. These properties of the estimator, unbiasedness and consistency, are desirable and causes the accuracy of the estimate to increase as the sample size is increased.

2.5 Estimators based on the characteristic function

We now present a PDF estimator that follows a direct approach to estimating a PDF from sample data, without using moments at all. It is presented in this section as it is a direct result of the previous estimator that expresses a PDF in terms of its moments. It builds on the theory presented in the previous sections and follows a similar line of reasoning. As such, most of the theory from the previous sections are not repeated here although the following works relies on it.

The biggest advantage of this new technique is that it eliminates the complications associated with the Taylor polynomial approximation of the characteristic function. In order to remove the dependence on the Taylor polynomial, a technique that directly estimates the characteristic function from the sample data is proposed and investigated. The PDF estimator derived from this characteristic function estimator is then shown to be identical to the Parzen estimator. We therefore manage to arrive at the Parzen estimator by using a frequency domain approach (which is contrary to the more popular approach that is based in the spatial domain). From this, a practical PDF estimator that uses a Fourier series and represents a parametric approximation to the Parzen estimator is developed.

As the new estimator operates directly from sample data, it is recommended that it be used in situations where sample data is available and the moments-based one be used if the moments present the only available information.

2.5.1 Motivation

In the previous sections we estimated a PDF from the values of a finite number of moments. Although our motivation for this course of action was taken from the Principle of Moments and all the techniques presented expressed the PDF estimate in terms of the moments, the characteristic function played a central role throughout: the PDF estimate was always obtained directly from the characteristic function, which was estimated from the moments using a Taylor polynomial approximation.

Unfortunately the Taylor approximation held a number of disadvantages, most notably being the large number of moments required to obtain an accurate PDF estimate (which is

impractical). It also diverges when not evaluated close to the origin, thereby complicating implementations and limiting the amount of detail that an estimate can contain.

If it is, however, assumed that sample data is available, alternative techniques of estimating the characteristic function may be considered. These characteristic function estimators can then be combined with the theory presented in the previous sections to create a PDF estimator that operates directly from sample data.

2.5.2 A PDF in terms of a characteristic function

Peebles [2, p 81] defines the characteristic function of X in terms of the expected value of a complex exponential function of X :

$$\Phi_X(\omega) = E\{e^{j\omega X}\}. \quad (2.67)$$

Obtaining an expression for $E\{e^{j\omega X}\}$ requires that an expression for the PDF of X , $f_X(x)$ be known. Since we are trying to estimate it, we unfortunately do not possess such an expression. We do, however, assume to be in possession of a number of IID (independently and identically distributed) samples corresponding to the distribution $f_X(x)$. An approximation of $\Phi_X(\omega)$ is obtained by estimating $E\{e^{j\omega X}\}$ from the samples as follows (where x_i denotes the i 'th sample and N_x the number of samples):

$$\hat{\Phi}'_X(\omega) = \frac{1}{N_x} \sum_{i=0}^{N_x-1} e^{j\omega x_i}. \quad (2.68)$$

Substituting this expression into Equation 2.9 expresses an approximation of the PDF in terms of the inverse Fourier transform of this approximate characteristic function:

$$\begin{aligned} \hat{f}_X(x) &= \mathcal{F}^{-1}\{\hat{\Phi}'^*_X(\omega)\} \\ &= \mathcal{F}^{-1}\left\{\frac{1}{N_x} \sum_{i=0}^{N_x-1} e^{-j\omega x_i}\right\} \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left\{\frac{1}{N_x} \sum_{i=0}^{N_x-1} e^{-j\omega x_i}\right\} e^{j\omega x} d\omega. \end{aligned} \quad (2.69)$$

Exchanging the order of summation and integration in the above equation, and evaluating the resulting inverse Fourier integral (by using a transform pair from Appendix A) produces an interesting result: the PDF is realised as a sum of impulse (dirac-delta) functions, each

one located at the position of a data sample,

$$\begin{aligned}
\hat{f}_X(x) &= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \left\{ \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-jwx_i} e^{jwx} dw \right\} \\
&= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \mathcal{F}^{-1} \left\{ e^{-jwx_i} \right\} \\
&= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \delta(x - x_i).
\end{aligned} \tag{2.70}$$

$\hat{f}_X(x)$ therefore provides an *exact* representation of the sample data, without introducing any assumptions about its distribution or performing any modelling and is therefore of limited practical use. This issue was addressed in Section 2.4.2 when smoothing of the PDF, introduced by the truncation of the high frequency components of the characteristic function, was discussed. It was concluded that smoothing of a PDF that *exactly* represents a set of sample data is required to aid generalisation.

In order to smooth the PDF, we introduce a windowing function to suppress the high frequency components of the characteristic function. Let $\Theta(\omega_u, \omega)$ denote the real frequency domain windowing function, which attains a value of zero outside the interval $|\omega| \leq \omega_u$, and $\theta(\omega_u, x)$ its inverse Fourier transform, also a real function. We apply this windowing function by incorporating it into the characteristic function estimator, thereby obtaining a new estimate:

$$\begin{aligned}
\hat{\Phi}_X(\omega) &= \Theta(\omega_u, \omega) \hat{\Phi}'_X(\omega) \\
&= \Theta(\omega_u, \omega) \frac{1}{N_x} \sum_{i=0}^{N_x-1} e^{jwx_i}.
\end{aligned} \tag{2.71}$$

A new PDF estimate is now obtained by taking the inverse Fourier transform of the complex conjugate of this estimate. The convolution/multiplication duality of the Fourier transform, discussed in Section 2.4.2, is utilised to express the PDF in terms of the windowing function and the data samples:

$$\begin{aligned}
\hat{f}_X(x) &= \mathcal{F}^{-1}\left\{\hat{\Phi}_X^*(\omega)\right\} \\
&= \mathcal{F}^{-1}\left\{\Theta(\omega_u, \omega) \hat{\Phi}_X'^*(\omega)\right\} \\
&= \mathcal{F}^{-1}\left\{\Theta(\omega_u, \omega)\right\} \star \mathcal{F}^{-1}\left\{\hat{\Phi}_X'^*(\omega)\right\} \\
&= \theta(\omega_u, x) \star \frac{1}{N_x} \sum_{i=0}^{N_x-1} \delta(x - x_i) \\
&= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \theta(\omega_u, x - x_i).
\end{aligned} \tag{2.72}$$

This function is seen to be similar to Equation 2.70, but with the impulse functions replaced by the inverse Fourier transformed windowing function, $\theta(\omega_u, x)$. This causes the presence of a data sample at some location to be spread out into the neighbourhood of the sample, instead of being concentrated at a single point. A consequence of this is that each sample contributes globally to the estimate, instead of having a localised effect, which allows the estimate to generalise over more than one sample set.

By comparing Equation 2.72 to the Parzen density function estimator [6, p. 88] [8, p. 425], shows them to be identical. The major difference between this estimate and the Parzen estimate is that this one was *derived from a purely frequency domain perspective*, while the Parzen estimator is specified entirely in the spatial domain. The Parzen estimator is a biased and consistent estimator, with its bias determined by selection of the kernel function, $\theta(\omega_u, x)$. It will be seen later, when considering the limiting properties of the characteristic function estimator in Equation 2.68, that Equation 2.72 also represents a biased and consistent estimator, that is in all aspects *identical* to the Parzen estimator.

Figure 2.5 illustrates the operation of the estimator in Equation 2.72 by reconstructing a PDF using only a small number of data samples (10). The PDF is the mixture of a Gaussian and a uniform density that was introduced in the previous section. Two cases are illustrated, one using a low upper frequency (ω_u) and the other using a higher frequency. In the case of the high value of ω_u , the presence of the windowing function centered at each data sample is clearly visible, and the estimate severely differs from the actual PDF. This effect decreases as ω_u is lowered and more smoothing is introduced, allowing a fair approximation to be produced. It is seen that this approximation deviates most from the actual PDF at places of sharp discontinuity. This is expected as this is where high frequency components are required most.

This example employed a triangular windowing function, $\Theta_{\Delta}(\omega_u, \omega)$, with width $2\omega_u$. An important property of this function is that its Inverse Fourier transform, $\theta_{\Delta}(\omega_u, \omega)$, *never* attains a negative value, ensuring that the resulting estimate is a non-negative function (which is an important property of a PDF):

$$\Theta_{\Delta}(\omega_u, \omega) = \begin{cases} \frac{-|\omega|}{\omega_u} + 1 & ; \quad |\omega| \leq \omega_u \\ 0 & ; \quad |\omega| > \omega_u \end{cases} \quad (2.73)$$

$$\theta_{\Delta}(\omega_u, \omega) = \frac{2 \sin^2(\frac{1}{2}\omega_u x)}{\pi \omega_u x^2}. \quad (2.74)$$

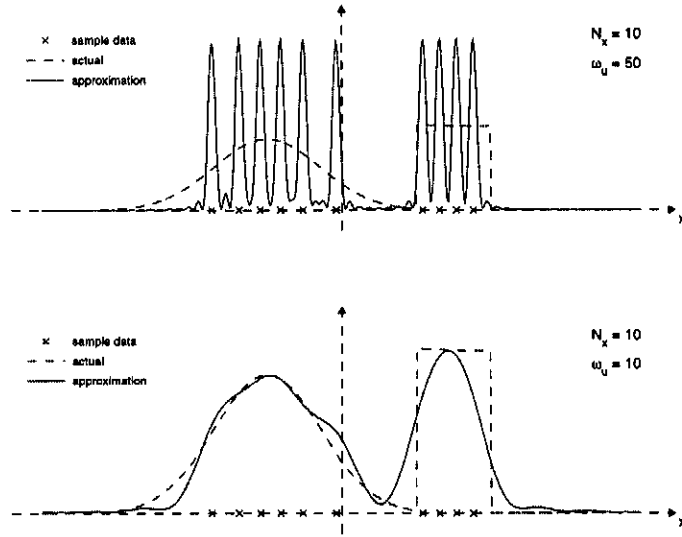


Figure 2.5: Reconstructing a PDF directly from data samples, using a triangular windowing function.

Unfortunately, there is no guarantee against the estimate, $\hat{f}_X(x)$, not attaining negative values in general, as it is dependent on the choice of windowing function. The absolute value of Equation 2.72 is therefore taken as the PDF estimate in order to ensure that it only takes on non-negative values:

$$\hat{f}_X(x) = \left| \frac{1}{N_x} \sum_{i=0}^{N_x-1} \theta(\omega_u, x - x_i) \right|. \quad (2.75)$$

This provides a conceptually simple expression for the PDF estimate and expresses it directly in terms of the inverse Fourier transformed windowing function and the values of

a number of data samples. Unfortunately, it does not provide a parametric representation of the PDF over and above that provided by the raw data samples. Next, we consider an alternative solution that infers a set of parametric constants that acts as parameters, thereby replacing the sample data. This has the advantage of providing a more compact representation of the sample data, thereby reducing computational requirements.

2.5.3 Fourier series

The technique presented in this section is similar to the one presented in Section 2.4.5: the characteristic function is sampled at a number of equidistant points in order to produce a Fourier series solution that expresses a PDF estimate in terms of a number of complex exponential basis functions.

An alternative way of representing the PDF in terms of the data samples and the windowing function is obtained by expanding the inverse Fourier transform in Equation 2.72, instead of invoking the convolution/multiplication duality. Again the absolute value is taken to ensure that the PDF does not attain negative values:

$$\begin{aligned}\hat{f}_X(x) &= \left| \mathcal{F}^{-1} \left\{ \Theta(\omega_u, \omega) \hat{\Phi}_X^*(\omega) \right\} \right| \\ &= \left| \frac{1}{2\pi} \int_{-\omega_u}^{\omega_u} \left\{ \Theta(\omega_u, \omega) \frac{1}{N_x} \sum_{i=0}^{N_x-1} e^{-j\omega x_i} \right\} e^{j\omega x} d\omega \right|.\end{aligned}\tag{2.76}$$

This provides us with an expression for the PDF estimate, expressed directly in terms of the sample data and some windowing function that smoothes over the observed data. In order to construct a parametric estimator, the argument of the inverse Fourier transform in the above equation is sampled at a number of equally spaced intervals, each $\frac{2\pi}{x_\Delta}$ apart. This changes the integral into a summation and adds a scaling constant, as the inverse Fourier transform becomes a Fourier series:

$$\begin{aligned}\hat{f}_X(x) &= \left| \frac{1}{x_\Delta} \sum_{k=-\alpha}^{\alpha} \Theta(\omega_u, \frac{2\pi k}{x_\Delta}) \frac{1}{N_x} \sum_{i=0}^{N_x-1} e^{-j\frac{2\pi k}{x_\Delta} x_i} e^{j\frac{2\pi k}{x_\Delta} x} \right| \\ &= \left| \sum_{k=-\alpha}^{\alpha} \frac{1}{x_\Delta} \Theta(\omega_u, \frac{2\pi k}{x_\Delta}) \frac{1}{N_x} \sum_{i=0}^{N_x-1} e^{-j\frac{2\pi k}{x_\Delta} x_i} e^{j\frac{2\pi k}{x_\Delta} x} \right| \\ \hat{f}_X(x) &= \left| \sum_{k=-\alpha}^{\alpha} \beta_k e^{j\frac{2\pi k}{x_\Delta} x} \right|,\end{aligned}\tag{2.77}$$

$$\alpha = \{\max(k) : \frac{2\pi k}{x_\Delta} \leq \omega_u, k \in \mathcal{N}\}, \quad (2.78)$$

$$\beta_k = \frac{1}{x_\Delta N_x} \Theta(\omega_u, \frac{2\pi k}{x_\Delta}) \sum_{i=0}^{N_x-1} e^{-j \frac{2\pi k}{x_\Delta} x_i}. \quad (2.79)$$

It was noted earlier that a Fourier series always represents a periodic function. In the above expression for $\hat{f}_X(x)$, the resulting PDF estimate has a period of x_Δ (which is therefore related to the sampling interval $\frac{2\pi}{x_\Delta}$) and represents a periodically extended version of the estimate in Equation 2.72. The distortion introduced by the periodic extension can be limited by selecting x_Δ to be large enough. The resulting function is again truncated to a symmetrical single-period interval centered at the origin (which is where most of its area is contained as it has a zero mean):

$$\hat{f}_X(x) = \begin{cases} \left| \sum_{k=-\alpha}^{\alpha} \beta_k e^{j \frac{2\pi k}{x_\Delta} x} \right| & ; \quad |x| \leq \frac{x_\Delta}{2} \\ 0 & ; \quad |x| > \frac{x_\Delta}{2}. \end{cases} \quad (2.80)$$

This technique is identical to the one presented Section 2.4.5, with the only difference being the way in which the values of the characteristic function are obtained: where it was previously expressed as a function of a number of moments and a windowing function, it is now a function of a number of observed data samples and a windowing function.

2.5.4 Limiting case where $N_x \rightarrow \infty$

We now consider the behaviour of the characteristic function (CF) estimator presented in Equation 2.71 in the limiting case where the number of samples, N_x , tends to infinity. This proves it to be a biased and consistent estimator of the characteristic function and PDF. Let X_i be a random variable representing the value of the i 'th sample that is used in the characteristic function estimate. It is assumed that all X_i are independent random variables with PDF $f_X(x)$ and characteristic function $\Phi_X(\omega)$ (which is what we are trying to estimate). The characteristic function estimator is then also represented as a random complex-valued function, $\Delta(\omega)$, in terms of these random variables representing the outcomes of the sample data:

$$\Delta(\omega) = \frac{1}{N_x} \sum_{i=0}^{N_x-1} \Theta(\omega_u, \omega) e^{j\omega X_i}. \quad (2.81)$$

We now find the expected value of this random function (representing the expected value of the estimator) as the product of the windowing function and the actual characteristic function:

$$\begin{aligned}
E\{\Delta(\omega)\} &= E\left\{\frac{1}{N_x} \sum_{i=0}^{N_x-1} \Theta(\omega_u, \omega) e^{j\omega X_i}\right\} \\
&= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \Theta(\omega_u, \omega) E\left\{e^{j\omega X_i}\right\} \\
&= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \Theta(\omega_u, \omega) \Phi_X(\omega) \\
&= \Theta(\omega_u, \omega) \Phi_X(\omega).
\end{aligned} \tag{2.82}$$

As the expected value of the estimate is not equal to the actual CF, we conclude that the estimator is biased. The PDF estimate that is associated with this expected CF estimate corresponds to a smoothed version of the actual PDF, as the actual PDF is convolved with the inverse Fourier transformed windowing function, $\theta(\omega_u, x)$:

$$\begin{aligned}
\hat{f}_X(x) &= \mathcal{F}^{-1}\left\{\Theta(\omega_u, \omega) \Phi_X^*(\omega)\right\} \\
&= \mathcal{F}^{-1}\left\{\Theta(\omega_u, \omega)\right\} \star \mathcal{F}^{-1}\left\{\Phi_X^*(\omega)\right\} \\
&= \theta(\omega_u, x) \star f_X(x).
\end{aligned} \tag{2.83}$$

We next obtain the variance of the CF estimate, $\Omega(\omega)$, in order to characterise its behaviour in the limit as $N_x \rightarrow \infty$:

$$\begin{aligned}
\Omega(\omega) &= E \left\{ \left| \Delta(\omega) - \Theta(\omega_u, \omega) \Phi_X(\omega) \right|^2 \right\} \\
&= E \left\{ \left[\Delta(\omega) - \Theta(\omega_u, \omega) \Phi_X(\omega) \right] \left[\Delta^*(\omega) - \Theta^*(\omega_u, \omega) \Phi_X^*(\omega) \right] \right\} \\
&= E \left\{ \Delta(\omega) \Delta^*(\omega) - \Theta(\omega_u, \omega) \Phi_X(\omega) \Delta^*(\omega) - \Delta(\omega) \Theta^*(\omega_u, \omega) \Phi_X^*(\omega) \right. \\
&\quad \left. + \Theta(\omega_u, \omega) \Phi_X(\omega) \Theta^*(\omega_u, \omega) \Phi_X^*(\omega) \right\} \\
&= E \left\{ \Delta(\omega) \Delta^*(\omega) - \Theta(\omega_u, \omega) \Phi_X(\omega) \Delta^*(\omega) - \Delta(\omega) \Theta^*(\omega_u, \omega) \Phi_X^*(\omega) \right. \\
&\quad \left. + \left| \Theta(\omega_u, \omega) \Phi_X(\omega) \right|^2 \right\} \\
&= E \left\{ \Delta(\omega) \Delta^*(\omega) \right\} - E \left\{ \Theta(\omega_u, \omega) \Phi_X(\omega) \Delta^*(\omega) \right\} - E \left\{ \Delta(\omega) \Theta^*(\omega_u, \omega) \Phi_X^*(\omega) \right\} \\
&\quad + E \left\{ \left| \Theta(\omega_u, \omega) \Phi_X(\omega) \right|^2 \right\} \\
&= E \left\{ \Delta(\omega) \Delta^*(\omega) \right\} - \Theta(\omega_u, \omega) \Phi_X(\omega) \Theta^*(\omega_u, \omega) \Phi_X^*(\omega) - \Theta(\omega_u, \omega) \Phi_X(\omega) \Theta^*(\omega_u, \omega) \Phi_X^*(\omega) \\
&\quad + \left| \Theta(\omega_u, \omega) \Phi_X(\omega) \right|^2 \\
&= E \left\{ \Delta(\omega) \Delta^*(\omega) \right\} - \left| \Theta(\omega_u, \omega) \Phi_X(\omega) \right|^2 - \left| \Theta(\omega_u, \omega) \Phi_X(\omega) \right|^2 \\
&\quad + \left| \Theta(\omega_u, \omega) \Phi_X(\omega) \right|^2 \\
&= E \left\{ \Delta(\omega) \Delta^*(\omega) \right\} - \left| \Theta(\omega_u, \omega) \Phi_X(\omega) \right|^2.
\end{aligned}
\tag{2.84}$$

To simplify the process, we evaluate the argument of the first term (involving the expectation) separately:

$$\begin{aligned}
\Delta(\omega)\Delta^*(\omega) &= \left[\frac{1}{N_x} \sum_{i=0}^{N_x-1} \Theta(\omega_u, \omega) e^{j\omega X_i} \right] \left[\frac{1}{N_x} \sum_{k=0}^{N_x-1} \Theta(\omega_u, \omega) e^{-j\omega X_k} \right] \\
&= \frac{\Theta(\omega_u, \omega)^2}{N_x^2} \sum_{i=0}^{N_x-1} \sum_{k=0}^{N_x-1} e^{j\omega X_i} e^{-j\omega X_k} \\
&= \frac{\Theta(\omega_u, \omega)^2}{N_x^2} \sum_{i=0}^{N_x-1} \left[e^{j\omega X_i} e^{-j\omega X_i} + \sum_{\substack{k=0 \\ k \neq i}}^{N_x-1} e^{j\omega X_i} e^{-j\omega X_k} \right] \\
&= \frac{\Theta(\omega_u, \omega)^2}{N_x^2} \sum_{i=0}^{N_x-1} \left[1 + \sum_{\substack{k=0 \\ k \neq i}}^{N_x-1} e^{j\omega X_i} e^{-j\omega X_k} \right],
\end{aligned} \tag{2.85}$$

after which the expected value is then calculated:

$$\begin{aligned}
\mathbb{E} \left\{ \Delta(\omega)\Delta^*(\omega) \right\} &= \mathbb{E} \left\{ \frac{\Theta(\omega_u, \omega)^2}{N_x^2} \sum_{i=0}^{N_x-1} \left[1 + \sum_{\substack{k=0 \\ k \neq i}}^{N_x-1} e^{j\omega X_i} e^{-j\omega X_k} \right] \right\} \\
&= \frac{\Theta(\omega_u, \omega)^2}{N_x^2} \sum_{i=0}^{N_x-1} \left[1 + \sum_{\substack{k=0 \\ k \neq i}}^{N_x-1} \mathbb{E} \left\{ e^{j\omega X_i} e^{-j\omega X_k} \right\} \right].
\end{aligned} \tag{2.86}$$

As the two random variables, X_i and X_k , are independent, the single expectation above can be factored into two separate expectation operations:

$$\begin{aligned}
\mathbb{E}\left\{\Delta(\omega)\Delta^*(\omega)\right\} &= \frac{\Theta(\omega_u, \omega)^2}{N_x^2} \sum_{i=0}^{N_x-1} \left[1 + \sum_{\substack{k=0 \\ k \neq i}}^{N_x-1} \mathbb{E}\left\{e^{j\omega X_i}\right\} \mathbb{E}\left\{e^{-j\omega X_k}\right\}\right] \\
&= \frac{\Theta(\omega_u, \omega)^2}{N_x^2} \sum_{i=0}^{N_x-1} \left[1 + \sum_{\substack{k=0 \\ k \neq i}}^{N_x-1} \Phi_X(\omega)\Phi_X^*(\omega)\right] \\
&= \frac{\Theta(\omega_u, \omega)^2}{N_x^2} \sum_{i=0}^{N_x-1} \left[1 + \sum_{\substack{k=0 \\ k \neq i}}^{N_x-1} |\Phi_X(\omega)|^2\right] \tag{2.87} \\
&= \frac{\Theta(\omega_u, \omega)^2}{N_x^2} N_x \left[1 + (N_x - 1) |\Phi_X(\omega)|^2\right] \\
&= \frac{\Theta(\omega_u, \omega)^2}{N_x} + \frac{\Theta(\omega_u, \omega)^2}{N_x} N_x |\Phi_X(\omega)|^2 - \frac{\Theta(\omega_u, \omega)^2}{N_x} |\Phi_X(\omega)|^2 \\
&= \frac{\Theta(\omega_u, \omega)^2}{N_x} - \frac{|\Theta(\omega_u, \omega)\Phi_X(\omega)|^2}{N_x} + |\Theta(\omega_u, \omega)\Phi_X(\omega)|^2.
\end{aligned}$$

This is substituted back into Equation 2.84 to obtain an expression for the variance of the estimate in terms of the actual characteristic function, the windowing function and the number of samples:

$$\begin{aligned}
\Omega(\omega) &= \frac{\Theta(\omega_u, \omega)^2}{N_x} - \frac{|\Theta(\omega_u, \omega)\Phi_X(\omega)|^2}{N_x} + |\Theta(\omega_u, \omega)\Phi_X(\omega)|^2 - |\Theta(\omega_u, \omega)\Phi_X(\omega)|^2 \\
&= \frac{\Theta(\omega_u, \omega)^2}{N_x} - \frac{|\Theta(\omega_u, \omega)\Phi_X(\omega)|^2}{N_x} \tag{2.88} \\
&= \frac{\Theta(\omega_u, \omega)^2}{N_x} \left[1 - |\Phi_X(\omega)|^2\right].
\end{aligned}$$

The variance therefore vanishes as the number of samples (N_x) tends to infinity. This makes the characteristic function estimator consistent and implies that an increase in the number of samples, from which the estimator is trained, always leads to a decrease in the expected estimation error (for a given windowing function and PDF).

2.6 Conclusions

Estimating a probability density function (PDF) only in terms of only a finite number of moments was found to be entirely feasible in theory as well as in practice (Section 2.4). This is possible due to the Fourier relationship existing between the characteristic function and the PDF as well as the ability to approximate the characteristic function from a number of moments using a Taylor series. From a theoretical perspective, an integral expression (Equation 2.40) was derived that expresses a PDF estimate in terms of a finite number of moments and a function that controls the accuracy trade-offs involved in the estimate (the windowing function). Three practical techniques of evaluating this integral expression were considered in detail:

Anti-derivative This evaluated the integral by considering the anti-derivative of the integrand (Section 2.4.3). It suffers from problems regarding accuracy and sensitivity towards the sample data distribution and its use is therefore not recommended.

Numerical integration These techniques evaluated the integral using some numerical integration technique (Section 2.4.4) and was developed to be applicable to a wide range of numerical integration techniques. Although it is recommended above the anti-derivative technique, care should still be taken as certain choices of parameters can introduce artifacts in the estimate.

Fourier series The characteristic function estimate (which is obtained in terms of a number of moments) is uniformly sampled and then used to express the PDF estimate in terms of a Fourier series (Section 2.4.5). As a Fourier series is best suited to representing a periodic function and a PDF can never be periodic, care had to be taken to ensure a robust technique that produces valid results. Fortunately, all challenges in this regard were addressed and successfully resolved, resulting in a practical technique with favourable characteristics. This technique should be preferred above the other ones that were considered.

Estimating a density function using only the characteristic function (Section 2.5), without involving moments at all, was also considered and was found to be feasible in theory as well as in practice. A characteristic function estimator that operates directly from sample data was first presented (Equation 2.71), after which a theoretical relationship between this estimator and the PDF estimate was established (Equation 2.72). A surprising result was that this estimator was found to be *identical* to the Parzen estimator (even in terms

of its bias and consistency), although it was derived from a frequency domain perspective contrary to the Parzen estimator that is defined in the spatial domain. In order to allow practical application of this estimator, a technique involving the Fourier series was developed (Section 2.5.3). As this technique has favourable characteristics making it well suited to practical application, alternatives involving the anti-derivative or numerical integration techniques were not considered.

2.6.1 Comparison between characteristic function and moments techniques

Estimating the density function directly using the characteristic function holds numerous advantages above the techniques that employ moments. Omitting the Taylor series approximation of the characteristic function eliminates a prime source of errors: a higher-order Taylor polynomial is generally preferred as it provides a better approximation than a lower-order one. But, a higher-order approximation is also a function of higher-order moments, which introduces complications if the moments are estimated from sample data (as opposed to accurate values being available from some other source). From Equation 2.66 it is seen that the variance of the estimate of the n 'th order central moment, \hat{m}_n , is a function of the moment of order $2n$. This causes the variance to rapidly increase as the order is increased, weakening the approximation. The robustness of the moment estimator also decreases rapidly as the order is increased due to the sensitivity towards outlier data.

Another problem encountered with the Taylor polynomial stems from the the fact that the accuracy of the reconstructed function decreases as the function is evaluated further away from the origin, due to the polynomial diverging. This was addressed by only evaluating the characteristic function over a subset of its domain (by introducing a windowing function), which also decreases the accuracy of the estimate as it introduced smoothing into the estimate.

None of these complications arise when estimating the characteristic function directly from the sample data, and no trade-off exists. There is no lower limit on the amount of smoothing that the estimator imposes and an artificial limit would be required in practice, to prevent over-specialisation of the estimator. Consequently, it is recommended that the Fourier series PDF estimator based on the characteristic function always be preferred above the moments-based estimator in the presence of sample data. If the only available data is in the form of the values of a number of moments, it is recommended that the Fourier series PDF estimator based on moments be used.

2.6.2 Comparison with the Parzen estimator and Gaussian mixture model (GMM)

It is insightful to compare these new techniques based on moments and the characteristic function with two established techniques, namely the Parzen estimator and the Gaussian mixture model (GMM). Both the Parzen estimator and the GMM are specialised cases of mixture models, in which the density function is represented by a number of weighted density functions. They differ in the way in which each one determines the parameters defining the component density functions and their corresponding weights from a training dataset.

The Parzen estimator places a mixture density component at the location of each data sample in the training dataset. All components are identical and get assigned the same weight that is inversely proportional to the training set size (and therefore number of components). Although this estimator is desirable in terms of accuracy and ease of training (it simply stores the data), it suffers from complexity problems as its computational complexity is proportional to the size of the training dataset.

The GMM, on the other hand, is able to compactly represent a dataset using a (relatively) small number of mixture components. It achieves this by iteratively optimising a cost function so as to provide an optimal fit between the estimate and the training dataset, using a fixed number of mixture components. This has the ability to combine the effect of a cluster of closely-spaced data samples, thereby removing the one-to-one mapping between data samples and mixture components that characterised the Parzen estimator. Consequently, it can achieve accuracy that is similar to the Parzen estimator, but with a large reduction in the number of mixture components, as redundancies in the training set are ignored. Its largest disadvantage does, however, also lie in its training algorithm: the iterative training procedure is prone to converging to a solution that is only locally optimal, yielding inaccurate estimates and increasing the overall variance of the estimator.

The techniques presented in this chapter combine favourable characteristics from both these estimators by taking a frequency domain approach to density estimation. In terms of training requirements, they employ a closed-form training algorithm that calculates the values of a number of parameters directly in terms of the values of a number of data samples or moments (without requiring any optimisation or iteration). In this regard they are similar to the Parzen estimator as they have a predictable and simple training procedure (that the GMM lacks). Their similarities to the GMM stems from the fact that they are able to represent the information in the training dataset using a fixed number

of parameters, independent from the training set size. This provides them with good scalability (which was lacking in the Parzen estimator).

The end result is a family of PDF estimators, derived using a frequency domain perspective, that combine the scalability of the GMM with the predictability of the Parzen estimator.

Chapter 3

Novel cumulative distribution function estimators

3.1 Introduction

Cumulative density functions (CDFs) are closely related to probability density functions (PDFs), with the latter simply being the derivative of the former. Although it is possible to obtain a CDF from a PDF by using some numerical integration technique, this does not always present a practical or sufficiently accurate solution. In this chapter we investigate the possibility of estimating cumulative density functions from moments and sample data in the same way that was done for PDFs in the previous chapter. The object is to again establish the theoretical feasibility of such an approach as well as the creation of techniques that can be implemented and applied in practice.

Due to the similarities between the PDF and CDF, a large part of the theory developed in the previous chapter is inherited by this chapter without repetition. Consequently, the work in this chapter relies on the theory presented in the previous chapter, with only additional theory relating exclusively to CDF estimation included here.

This chapter follows a similar layout to the previous one, where the techniques based on moments are first presented, followed by those based on the characteristic function. For both techniques, the problem is first approached from a purely theoretical point of view after which attention is paid to the development of techniques that can be used in practice. As was the case with the PDF estimators, two estimators employing Fourier series techniques are developed.

3.2 Motivation

CDF estimators based on existing PDF estimators (such as Parzen and GMM) also suffer from the same problems as their PDF counterparts. Inspired by the success achieved in designing PDF estimators using a frequency domain approach, we wish to apply the same principles to the investigation of CDF estimators. As PDFs and CDFs are mathematically very closely related, it is natural to expect that, by applying the same principles used to develop PDF estimators, similar CDF estimators can also be developed.

From our experience with PDF estimators, we have a prior indication of the problems associated with different approaches, allowing them to be addressed in advance. Also, a great deal of the theory introduced in the previous chapter is equally applicable to CDF estimation, as it was developed within a more general framework than was required for PDF estimation only. This should allow us to develop CDF estimators being comparable to the PDF estimators from the previous chapter with a minimum of additional work.

3.3 Definitions and background

The techniques presented in this chapter extends those presented in Section 2.4 and Section 2.5. The cumulative distribution function, $F_X(x)$, is defined in terms of a probability function, which can be expressed in terms of the PDF, $f_X(x)$:

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= \int_{-\infty}^x f_X(\lambda) d\lambda. \end{aligned} \tag{3.1}$$

A property of the Fourier transform allows us to express the Fourier transform of the definite integral of a function in terms of the Fourier transform of the function and the frequency variable [5]:

$$\mathcal{F}\left\{\int_{-\infty}^x f(\lambda) d\lambda\right\} = \frac{1}{j\omega} F(\omega) + \pi F(0) \delta(\omega) \tag{3.2}$$

where $F(\omega) = \mathcal{F}\{f(x)\}$.

The above two equations are combined, after taking the Fourier transform on both sides of the first equation, to relate the Fourier transform of the CDF to the Fourier transform

of the PDF:

$$\begin{aligned}\mathcal{F}\{F_X(x)\} &= \mathcal{F}\left\{\int_{-\infty}^x f_X(\lambda)d\lambda\right\} \\ &= \frac{1}{j\omega}\mathcal{F}\{f_X(x)\} + \pi\mathcal{F}\{f_X(x)\}\Big|_{\omega=0}\delta(\omega).\end{aligned}\tag{3.3}$$

Recalling the relationship between the PDF and the characteristic function and noting that $\Phi_X^*(0) = 1$ (from a property of the characteristic function [2, p. 81]), allows us to obtain a relationship between the CDF and the characteristic function:

$$\begin{aligned}\mathcal{F}\{F_X(x)\} &= \frac{1}{j\omega}\mathcal{F}\{f_X(x)\} + \pi\mathcal{F}\{f_X(x)\}\Big|_{\omega=0}\delta(\omega) \\ &= \frac{1}{j\omega}\Phi_X^*(\omega) + \pi\Phi_X^*(\omega)\Big|_{\omega=0}\delta(\omega) \\ &= \frac{1}{j\omega}\Phi_X^*(\omega) + \pi\Phi_X^*(0)\delta(\omega) \\ &= \frac{1}{j\omega}\Phi_X^*(\omega) + \pi\delta(\omega).\end{aligned}\tag{3.4}$$

Taking the inverse Fourier transform on both sides of the above equation allows it to be simplified and enables us to express the CDF directly in terms of the characteristic function and a constant term:

$$\begin{aligned}F_X(x) &= \mathcal{F}^{-1}\left\{\frac{1}{j\omega}\Phi_X^*(\omega)\right\} + \mathcal{F}^{-1}\left\{\pi\delta(\omega)\right\} \\ &= \mathcal{F}^{-1}\left\{\frac{1}{j\omega}\Phi_X^*(\omega)\right\} + \frac{1}{2}.\end{aligned}\tag{3.5}$$

3.4 Estimators based on moments

The above expression allows us to obtain a CDF estimate entirely in terms of a characteristic function estimate. Using similar reasoning to that followed in Section 2.4, an integral expression for the CDF in terms of the values of a finite number of moments are derived.

Two practical techniques of evaluating this integral are then considered: one using a numerical integration technique and one using a Fourier series. Unlike the case of the PDF estimator, no technique involving the anti-derivative is considered, as it was deemed impractical. The derivation of the numerical integration technique is only slightly more complicated than that done for the PDF estimator, while the Fourier series technique proves

to be a great deal more complicated. This is due to the impossibility of approximating a CDF directly using a periodic function, as it is a monotonic function by definition.

3.4.1 A CDF in terms of moments

We now consider the relationship between the CDF and the moments of a random variable, X , and use that to derive an expression for the CDF directly in terms of a finite number of moments.

Consider Equation 3.5: using a Taylor series expansion, the characteristic function is expressed as a polynomial in terms of the moments of the random variable (Equation 2.31),

$$\begin{aligned} F_X(x) &= \mathcal{F}^{-1} \left\{ \frac{1}{j\omega} \Phi_X^*(\omega) \right\} + \frac{1}{2} \\ &= \mathcal{F}^{-1} \left\{ \frac{1}{j\omega} \sum_{n=0}^{\infty} \frac{m_n}{j^n n!} \omega^n \right\} + \frac{1}{2}. \end{aligned} \quad (3.6)$$

The $\frac{1}{j\omega}$ term is multiplied by this polynomial, thereby resulting in another polynomial of lower order, allowing the CDF to be expressed in terms of an infinite number of moments,

$$F_X(x) = \mathcal{F}^{-1} \left\{ \sum_{n=0}^{\infty} \frac{m_n}{j^{n+1} n!} \omega^{n-1} \right\} + \frac{1}{2}. \quad (3.7)$$

Note that this polynomial expresses the CDF identically (and is therefore not an approximation), but requires moments of infinite order, which presents problems in practice. In order to use this expression in a practical estimator, the series expansion is truncated to a finite number of terms (N_m), resulting in an expression for a CDF *estimate* that only requires a finite number of moments (also N_m),

$$\hat{F}_X(x) = \mathcal{F}^{-1} \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^{n+1} n!} \omega^{n-1} \right\} + \frac{1}{2}. \quad (3.8)$$

This expression is undefined as the Fourier transform integral fails to converge due to the divergent polynomial series. This problem was also encountered when we expressed the PDF in terms of a finite number of moments. Multiplying the series by a windowing function with finite support, $\Theta(\omega_u, \omega)$, remedies this problem, as it forces the argument of

the Fourier transform to zero outside the interval $|\omega| \leq \omega_u$,

$$\hat{F}_X(x) = \mathcal{F}^{-1} \left\{ \Theta(\omega_u, \omega) \sum_{n=0}^{N_m-1} \frac{m_n}{j^{n+1}n!} \omega^{n-1} \right\} + \frac{1}{2}, \quad (3.9)$$

where

$$\Theta(\omega_u, \omega) = 0 \quad \text{if} \quad |\omega| > \omega_u. \quad (3.10)$$

By expanding the inverse Fourier transform, a CDF estimate is expressed entirely in terms of a finite number of moments and a windowing function:

$$\hat{F}_X(x) = \frac{1}{2\pi} \int_{-\omega_u}^{\omega_u} \Theta(\omega_u, \omega) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^{n+1}n!} \omega^{n-1} \right\} e^{j\omega x} d\omega + \frac{1}{2}. \quad (3.11)$$

In order to consider the effect of the windowing function on this approximation, we compare the expression for the CDF estimate in Equation 3.9 with the one representing the actual CDF in Equation 3.7. As the power series in these expressions have identical coefficients and only differ in their number of terms, a value of ω_u can be found so that they are approximately equal to each other in the interval $|\omega| \leq \omega_u$,

$$\sum_{n=0}^{N_m-1} \frac{m_n}{j^{n+1}n!} \omega^{n-1} \approx \sum_{n=0}^{\infty} \frac{m_n}{j^{n+1}n!} \omega^{n-1} \quad , \quad |\omega| \leq \omega_u. \quad (3.12)$$

Furthermore, as the windowing function in Equation 3.9 attains a value of zero outside the interval $|\omega| \leq \omega_u$, the power series can be modified to contain an infinite number of terms, leading only to the introduction of a small error (from the approximate equality in the previous expression):

$$\begin{aligned} \hat{F}_X(x) &= \mathcal{F}^{-1} \left\{ \Theta(\omega_u, \omega) \sum_{n=0}^{N_m-1} \frac{m_n}{j^{n+1}n!} \omega^{n-1} \right\} + \frac{1}{2}, \\ &\approx \mathcal{F}^{-1} \left\{ \Theta(\omega_u, \omega) \sum_{n=0}^{\infty} \frac{m_n}{j^{n+1}n!} \omega^{n-1} \right\} + \frac{1}{2}. \end{aligned} \quad (3.13)$$

By now applying the convolution/multiplication property of the Fourier transform, the

above can be expressed as follows:

$$\begin{aligned}
\hat{F}_X(x) &\approx \mathcal{F}^{-1} \left\{ \Theta(\omega_u, \omega) \sum_{n=0}^{\infty} \frac{m_n}{j^{n+1} n!} \omega^{n-1} \right\} + \frac{1}{2} \\
&= \mathcal{F}^{-1} \left\{ \Theta(\omega_u, \omega) \right\} \star \mathcal{F}^{-1} \left\{ \sum_{n=0}^{\infty} \frac{m_n}{j^{n+1} n!} \omega^{n-1} \right\} + \frac{1}{2} \\
&= \theta(\omega_u, x) \star \mathcal{F}^{-1} \left\{ \sum_{n=0}^{\infty} \frac{m_n}{j^{n+1} n!} \omega^{n-1} \right\} + \frac{1}{2}
\end{aligned} \tag{3.14}$$

By comparing this to the expression for the CDF (Equation 3.7), it is seen that this CDF estimate represents a smoothed version of the actual CDF, as it is convolved with the inverse Fourier transformed windowing function, $\theta(\omega_u, x)$. This convolution also introduces a complication in the form of leakage, thereby placing a restriction on the windowing function: if the CDF estimate is to be a monotonic function, the convolution kernel, $\theta(\omega_u, x)$, should be non-negative. In the case of the PDF estimator, this was addressed by simply taking the absolute value of the estimate. Unfortunately, we cannot do the same in this situation as the same problem manifests itself in terms of the estimate becoming non-monotonic (as opposed to non-negative). It is therefore recommended that a Bartlett (triangular) window always be employed by the CDF estimators.

We next direct our attention to ways of solving the integral in Equation 3.11 (representing a CDF estimate in terms of a number of moments) in order to obtain the value of the CDF at any point in its domain, when presented with a number of moments. Direct evaluation using the anti-derivative, even when using a rectangular windowing function, is not feasible and was therefore not attempted. This was noted earlier when a similar integral was encountered when the PDF was estimated in terms of a number of moments. We do, however, consider a numerical integration technique as well as one based on the Fourier series.

3.4.2 Numerical integration techniques

Evaluating Equation 3.11 using a numerical integration technique is similar to the procedure followed to calculate the PDF estimate using a numerical integration technique. Care should, however, be taken as the integrand possesses a singularity at the origin. This is due to the ω term in the denominator, corresponding to $n = 0$. Fortunately, the integrand is continuous in the limit as $\omega \rightarrow 0$ and exists (as the Dirichlet Integral exists, [31, p. 394]).

In order to apply a numerical integration technique, the value of the integrand should therefore be defined in terms of its limiting value at the origin, in the case where $n = 0$. This is best accomplished by evaluating the integral of the term corresponding to $n = 0$ separately from the rest:

$$\hat{F}_X(x) = \kappa + \frac{1}{2\pi} \int_{-\omega_u}^{\omega_u} \Theta(\omega_u, \omega) \left\{ \sum_{n=1}^{N_m-1} \frac{m_n}{j^{n+1}n!} \omega^{n-1} \right\} e^{j\omega x} d\omega + \frac{1}{2}, \quad (3.15)$$

with

$$\begin{aligned} \kappa &= \frac{1}{2\pi} \int_{-\omega_u}^{\omega_u} \Theta(\omega_u, \omega) \left\{ \frac{m_0}{j^1 0!} \omega^{-1} \right\} e^{j\omega x} d\omega \\ &= -j \frac{1}{2\pi} \int_{-\omega_u}^{\omega_u} \Theta(\omega_u, \omega) \frac{e^{j\omega x}}{\omega} d\omega \\ &= \frac{1}{2\pi} \int_{-\omega_u}^{\omega_u} \Theta(\omega_u, \omega) \frac{\sin \omega x}{\omega} d\omega. \end{aligned} \quad (3.16)$$

This results in two integrals, one real and one complex, of which only one contains a singularity in the integrand that requires special consideration. The integral expression for κ is approximated using a numerical integration technique by partitioning the integration domain, $[-\omega_u, \omega_u]$, into N_i intervals, replacing the integral by a discrete summation and introducing scaling factors. As this results in the integrand being evaluated at a number of points in its domain, it is assigned the limiting value at the origin, $\omega = 0$, in order to account for the singularity:

$$\begin{aligned} \kappa &= \frac{1}{2\pi} \alpha \sum_{k=0}^{N_i} \beta_k \Theta(\omega_u, \gamma_k) \frac{\sin \gamma_k x}{\gamma_k} \\ &= \sum_{k=0}^{N_i} \beta_k \frac{\alpha}{2\pi} \Theta(\omega_u, \gamma_k) \frac{\sin \gamma_k x}{\gamma_k} \\ &= \sum_{k=0}^{N_i} \varphi'_k, \end{aligned} \quad (3.17)$$

with

$$\varphi'_k = \begin{cases} \beta_k \frac{\alpha}{2\pi} \Theta(\omega_u, \gamma_k) \frac{\sin \gamma_k x}{\gamma_k} & ; \quad \gamma_k \neq 0 \\ \beta_k \frac{\alpha}{2\pi} \Theta(\omega_u, \gamma_k) x & ; \quad \gamma_k = 0. \end{cases} \quad (3.18)$$

The remaining integral, from Equation 3.15, is also approximated by a numerical inte-

gration technique, using the same method as above, and the result is combined with the above result. As this integrand contains no singularities it can simply be evaluated anywhere in its domain. This expresses the CDF, $\hat{F}_X(x)$, directly in terms of a finite number of sample moments:

$$\begin{aligned}
\hat{F}_X(x) &= \kappa + \frac{1}{2\pi} \alpha \sum_{k=0}^{N_i} \beta_k \Theta(\omega_u, \gamma_k) \left\{ \sum_{n=1}^{N_m-1} \frac{m_n}{j^{n+1} n!} (\gamma_k)^{n-1} \right\} e^{j\gamma_k x} + \frac{1}{2} \\
&= \kappa + \sum_{k=0}^{N_i} \beta_k \frac{\alpha}{2\pi} \Theta(\omega_u, \gamma_k) \left\{ \sum_{n=1}^{N_m-1} \frac{m_n}{j^{n+1} n!} (\gamma_k)^{n-1} \right\} e^{j\gamma_k x} + \frac{1}{2} \\
&= \kappa + \sum_{k=0}^{N_i} \varphi_k e^{j\gamma_k x} + \frac{1}{2} \\
&= \sum_{k=0}^{N_i} \varphi'_k + \sum_{k=0}^{N_i} \varphi_k e^{j\gamma_k x} + \frac{1}{2},
\end{aligned} \tag{3.19}$$

with

$$\varphi_k = \beta_k \frac{\alpha}{2\pi} \Theta(\omega_u, \gamma_k) \left\{ \sum_{n=1}^{N_m-1} \frac{m_n}{j^{n+1} n!} (\gamma_k)^{n-1} \right\}. \tag{3.20}$$

Note that the discretisation of the continuous integrand can introduce unwanted artifacts into the resulting expression for $\hat{F}_X(x)$. Any technique that samples the integrand uniformly and symmetrically (with respect to the origin), results in $\hat{F}_X(x)$ being a periodic function. This is always the case when γ_k only attains constant-scaled integral values, i.e. $\gamma_k = k\omega_0$ where ω_0 acts as the fundamental frequency. The period of $\hat{F}_X(x)$ is then equal to $x_0 = \frac{2\pi}{\omega_0}$. In order to account for this, the function is only calculated within an interval corresponding to half its first period, $|x| \leq \frac{\pi}{\omega_0}$ and is assigned the theoretical limiting values attained at $\pm\infty$ (0 and 1) outside the interval:

$$\hat{F}_X(x) = \begin{cases} 0 & ; \quad x < -\frac{\pi}{\omega_0} \\ \left| \sum_{k=0}^{N_i} \varphi'_k + \sum_{k=0}^{N_i} \varphi_k e^{j\gamma_k x} + \frac{1}{2} \right| & ; \quad |x| \leq \frac{\pi}{\omega_0} \\ 1 & ; \quad x > \frac{\pi}{\omega_0} \end{cases} \tag{3.21}$$

In order to apply Simpson's rule, N_i is selected to be an even number and α, β_k and γ_k

are obtained by the following expressions (identical to Equation 2.49):

$$\begin{aligned}\alpha &= \frac{2\omega_u}{3N_i} \\ \beta_k &= \begin{cases} 1 & ; \quad k = 0, N_i \\ 3 - (-1)^k & ; \quad k = 1, 2, 3, \dots, N_i - 1 \end{cases} \\ \gamma_k &= \frac{2k - N_i}{N_i} \omega_u\end{aligned}\tag{3.22}$$

As $\gamma_k = k \frac{2\omega_u}{N_i}$, the resulting function is periodic and the fundamental period is identified to be $\omega_0 = \frac{2\omega_u}{N_i}$. The resulting CDF, $\hat{F}_X(x)$, should therefore only be calculated over the period $|x| \leq \frac{\pi N_i}{\omega_u}$.

3.4.3 Fourier series approximation

We now obtain an expression for $\hat{F}_X(x)$ in terms of a Fourier series that involves a finite number of moments. Although this is similar to techniques developed in Section 2.4.5 and Section 2.5.3, complications arise due to the frequency (ω) term present in the denominator. A general result is first obtained and this is then applied to the problem at hand to obtain the desired expression.

Assume $f(x)$ to be a continuous real finite-energy function with most of its energy concentrated in an x_Δ -wide interval surrounding the origin and $F(\omega)$, a complex continuous function, to be its Fourier transform:

$$\begin{aligned}|f(x)|_{|x| \leq \frac{x_\Delta}{2}} &\gg |f(x)|_{|x| > \frac{x_\Delta}{2}} \approx 0 \\ \int_{|x| \leq \frac{x_\Delta}{2}} f(x) dx &\gg \int_{|x| > \frac{x_\Delta}{2}} f(x) dx \approx 0 \\ F(\omega) &= \mathcal{F}\{f(x)\}.\end{aligned}\tag{3.23}$$

Obtain the periodic function $f'(x)$ by periodically extending $f(x)$ with a period of x_Δ ,

$$f'(x) = \sum_{k=-\infty}^{\infty} f(x - kx_\Delta).\tag{3.24}$$

Note that $f'(x) \approx f(x)$ in the interval $|x| \leq \frac{x_\Delta}{2}$, due to the above constraints placed on $f(x)$. This periodic extension of $f(x)$ can now be expressed as a Fourier series, with series

coefficients related to $F(\omega)$, the Fourier transform of $f(x)$:

$$f'(x) = \frac{1}{x_\Delta} \sum_{k=-\infty}^{\infty} F\left(\frac{2\pi}{x_\Delta}\right) e^{\frac{j2\pi k}{x_\Delta} x}. \quad (3.25)$$

This Fourier series is used to obtain, $F'(\omega)$, the Fourier *transform* of $f'(x)$, in terms of an infinite sum of weighted impulse functions, with their weights obtained from the above Fourier series coefficients:

$$F'(\omega) = \frac{2\pi}{x_\Delta} \sum_{k=-\infty}^{\infty} F\left(\frac{2\pi}{x_\Delta}\right) \delta\left(\omega - \frac{2\pi k}{x_\Delta}\right) \quad (3.26)$$

In order to confirm that this expression represents the Fourier transform of $f'(x)$, its inverse Fourier transform should be taken and the sifting property of the impulse function applied. $F'(\omega)$ therefore consists of an infinite sum of impulse functions located at integer multiples of the fundamental frequency $\omega_0 = \frac{2\pi}{x_\Delta}$. The impulse function located at the origin of $F'(\omega)$, corresponding to $k = 0$ in the above expression, is only responsible for the mean value of $f'(x)$, calculated over an integral number of periods. Removing it therefore only results in the removal of a constant term from $f'(x)$, without having any other effect on its overall shape or periodicity. It is isolated from the above expression, thereby introducing another function, $F''(\omega)$, representing the Fourier transform of $f'(x)$ without the constant term:

$$F'(\omega) = F''(\omega) + \frac{2\pi}{x_\Delta} F(0) \delta(\omega), \quad (3.27)$$

where

$$F''(\omega) = \frac{2\pi}{x_\Delta} \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} F\left(\frac{2\pi k}{x_\Delta}\right) \delta\left(\omega - \frac{2\pi k}{x_\Delta}\right) \quad (3.28)$$

and

$$F''(0) = 0. \quad (3.29)$$

$F''(\omega)$ therefore represents the Fourier transform of a continuous, x_Δ -periodic, *zero-mean* function, $f''(x)$, which is identical to $f'(x)$, except for a constant term. Taking the inverse Fourier transform on both sides of the above expression expresses $f'(x)$ in terms of $f''(x)$

and a constant term:

$$\begin{aligned} f'(x) &= \mathcal{F}^{-1}\{F''(\omega)\} + \mathcal{F}^{-1}\left\{\frac{2\pi}{x_\Delta}F(0)\delta(\omega)\right\} \\ &= f''(x) + \frac{F(0)}{x_\Delta}. \end{aligned} \quad (3.30)$$

A graphical representation of $f(x)$, $f'(x)$ and $f''(x)$, their Fourier transforms and their relationships are presented in Figure 3.1.

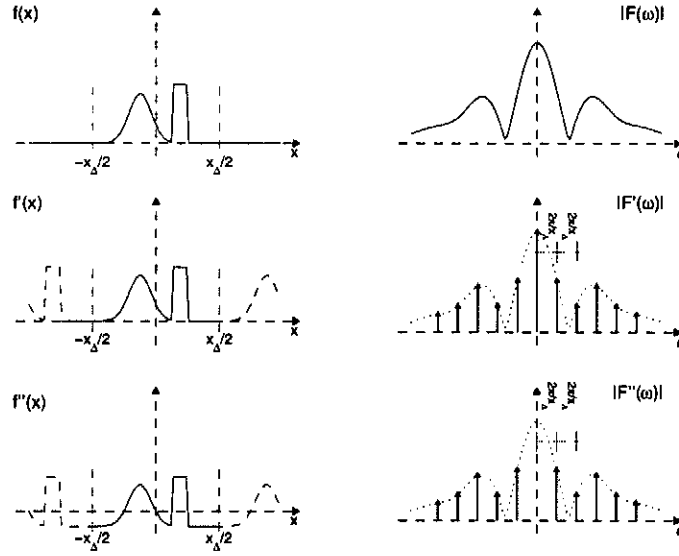


Figure 3.1: Relationship between $f(x)$, $f'(x)$ and $f''(x)$

We next consider expressions involving integrals of the functions introduced above. It was noted earlier that $f(x) \approx f'(x)$ in the interval $|x| \leq \frac{x_\Delta}{2}$. This allows us to conditionally equate definite integrals over these two functions in an interval near the origin:

$$\int_{-\frac{x_\Delta}{2}}^x f(\xi) d\xi \approx \int_{-\frac{x_\Delta}{2}}^x f'(\xi) d\xi \quad ; \quad |x| \leq \frac{x_\Delta}{2}. \quad (3.31)$$

Recall that constraints were also placed on the integral of $f(x)$, requiring it to have most of its area concentrated in the x_Δ -wide interval centered at the origin. As a consequence of this, the lower bound of the integral involving $f(x)$ in the above expression can be changed from $-\frac{x_\Delta}{2}$ to $-\infty$ with only slight effect, as $\int_{-\infty}^{-\frac{x_\Delta}{2}} f(\xi) d\xi \ll \int_{-\frac{x_\Delta}{2}}^{\frac{x_\Delta}{2}} f(\xi) d\xi$. This is done and the right-hand side terms are successively substituted from above expressions until the

integral of $f'(x)$ is expressed approximately in terms of the integral of $f''(x)$:

$$\begin{aligned}
\int_{-\infty}^x f(\xi) d\xi &\approx \int_{-\frac{x_{\Delta}}{2}}^x f(\xi) d\xi && ; \quad |x| \leq \frac{x_{\Delta}}{2} \\
&\approx \int_{-\frac{x_{\Delta}}{2}}^x f'(\xi) d\xi && ; \quad |x| \leq \frac{x_{\Delta}}{2} \\
&= \int_{-\frac{x_{\Delta}}{2}}^x \left\{ f''(\xi) + \frac{F(0)}{x_{\Delta}} \right\} d\xi && ; \quad |x| \leq \frac{x_{\Delta}}{2} \\
&= \int_{-\frac{x_{\Delta}}{2}}^x f''(\xi) d\xi + F(0) \left\{ \frac{x}{x_{\Delta}} + \frac{1}{2} \right\} && ; \quad |x| \leq \frac{x_{\Delta}}{2} \quad (3.32) \\
&= \int_{-\infty}^x f''(\xi) d\xi - \int_{-\infty}^{-\frac{x_{\Delta}}{2}} f''(\xi) d\xi + F(0) \left\{ \frac{x}{x_{\Delta}} + \frac{1}{2} \right\} && ; \quad |x| \leq \frac{x_{\Delta}}{2} \\
&= g''(x) - g''(-\frac{x_{\Delta}}{2}) + F(0) \left\{ \frac{x}{x_{\Delta}} + \frac{1}{2} \right\} && ; \quad |x| \leq \frac{x_{\Delta}}{2} \\
&= g''(\xi) \Big|_{\xi=-\frac{x_{\Delta}}{2}}^x + F(0) \left\{ \frac{x}{x_{\Delta}} + \frac{1}{2} \right\} && ; \quad |x| \leq \frac{x_{\Delta}}{2},
\end{aligned}$$

where

$$g''(x) = \int_{-\infty}^x f''(\xi) d\xi. \quad (3.33)$$

We now proceed to express $g''(x)$ in terms of $F(\omega)$, the Fourier transform of $f(x)$. Consider the above integral defining $g''(x)$: a property of the Fourier transform allows an improper integral of this form to be expressed in terms of the Fourier transform of the integrand and a term involving the mean value of the integrand. As the mean value of $f''(x)$ is zero, $g''(x)$ can be expressed entirely in terms of $F''(\omega)$, the Fourier transform of $f''(x)$, which was defined earlier:

$$\begin{aligned}
g''(x) &= \int_{-\infty}^x f''(\xi) d\xi \\
&= \mathcal{F}^{-1} \left\{ \frac{1}{j\omega} F''(\omega) + \pi F''(0) \delta(\omega) \right\} \\
&= \mathcal{F}^{-1} \left\{ \frac{1}{j\omega} F''(\omega) \right\} \\
&= \mathcal{F}^{-1} \left\{ \frac{1}{j\omega} \frac{2\pi}{x_{\Delta}} \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} F\left(\frac{2\pi k}{x_{\Delta}}\right) \delta\left(\omega - \frac{2\pi k}{x_{\Delta}}\right) \right\}.
\end{aligned} \quad (3.34)$$

By moving the Fourier transform into the summation and evaluating the resulting expres-

sion (by expanding the Fourier transform into its defining integral and employing the sifting property of the impulse function), we express $g''(x)$ in terms of an infinite sum involving $F(\omega)$:

$$\begin{aligned}
g''(x) &= \mathcal{F}^{-1} \left\{ \frac{1}{j\omega} \frac{2\pi}{x_\Delta} \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} F\left(\frac{2\pi k}{x_\Delta}\right) \delta\left(\omega - \frac{2\pi k}{x_\Delta}\right) \right\} \\
&= \frac{2\pi}{jx_\Delta} \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} F\left(\frac{2\pi k}{x_\Delta}\right) \mathcal{F}^{-1} \left\{ \frac{1}{\omega} \delta\left(\omega - \frac{2\pi k}{x_\Delta}\right) \right\} \\
&= \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} \frac{1}{2\pi jk} F\left(\frac{2\pi k}{x_\Delta}\right) e^{j\frac{2\pi k}{x_\Delta}x}.
\end{aligned} \tag{3.35}$$

The above expression represents $g''(x)$ as a Fourier series and is substituted back into Equation 3.32:

$$\begin{aligned}
\int_{-\infty}^x f(\xi) d\xi &\approx g''(\xi) \Big|_{\xi=-\frac{x_\Delta}{2}}^x + F(0) \left\{ \frac{x}{x_\Delta} + \frac{1}{2} \right\} \quad ; \quad |x| \leq \frac{x_\Delta}{2} \\
&= \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} \frac{1}{2\pi jk} F\left(\frac{2\pi k}{x_\Delta}\right) e^{j\frac{2\pi k}{x_\Delta}\xi} \Big|_{\xi=-\frac{x_\Delta}{2}}^x + F(0) \left\{ \frac{x}{x_\Delta} + \frac{1}{2} \right\} \quad ; \quad |x| \leq \frac{x_\Delta}{2}.
\end{aligned} \tag{3.36}$$

This provides the result we are seeking: it expresses the integral of a function with finite support directly in terms of a sum involving its Fourier transform and, implicitly, its mean value. In order to apply it to CDF estimation, we let $f(x)$ represent a PDF with finite support, implying a mean value, $F(0)$, of unity. As most of the area under a PDF is concentrated within a few standard deviations from the mean value, this restriction is not unreasonable and should produce a fair approximation. The integral then represents a CDF estimate, $\hat{F}_X(x)$, which is expressed in terms of the Fourier transform of the PDF,

the conjugate characteristic function $\Phi_X^*(\omega)$:

$$\begin{aligned}
\hat{F}_X(x) &= \int_{-\infty}^x f_X(\xi) d\xi \\
&\approx \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} \frac{1}{2\pi j k} \Phi_X^*\left(\frac{2\pi k}{x_\Delta}\right) e^{j \frac{2\pi k}{x_\Delta} \xi} \Big|_{\xi=-\frac{x_\Delta}{2}}^x + \Phi_X^*(0) \left\{ \frac{x}{x_\Delta} + \frac{1}{2} \right\} \quad ; \quad |x| \leq \frac{x_\Delta}{2} \\
&= \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} \frac{1}{2\pi j k} \Phi_X^*\left(\frac{2\pi k}{x_\Delta}\right) e^{j \frac{2\pi k}{x_\Delta} \xi} \Big|_{\xi=-\frac{x_\Delta}{2}}^x + \frac{x}{x_\Delta} + \frac{1}{2} \quad ; \quad |x| \leq \frac{x_\Delta}{2}.
\end{aligned} \tag{3.37}$$

Implementing an infinite sum such as the one in the above expression is not feasible in practice. Multiplying $\Phi_X^*(\omega)$ with a windowing function, $\Theta(\omega_u, \omega)$, ensures its value to be zero outside the interval $|\omega| \leq \omega_u$ and allows the sum to be truncated to a finite number of terms. $\Theta(\omega_u, \omega)$ should be a real symmetrical function attaining a maximum value of 1 at the origin and should have a non-negative inverse Fourier transform. Substituting this new frequency function, $\Theta(\omega_u, \omega)\Phi_X^*(\omega)$, for the characteristic function in the above expression has two effects on $f_X(x)$: it is smoothed and it is spread out. The smoothing does not pose a problem, but care should be taken, as excessive leakage, caused by spreading out $f_X(x)$, can result in a weakened estimate. Fortunately, a large enough value of x_Δ can always be selected so that the leakage is negligible. Due to the periodicity of the resulting $\hat{F}_X(x)$, it should never be evaluated outside the interval $|x| \leq \frac{x_\Delta}{2}$, but rather assigned its limiting values of 0 and 1. We express the CDF estimate as a finite sum involving the characteristic function and windowing function:

$$\hat{F}_X(x) = \sum_{\substack{k=-\alpha \\ k \neq 0}}^{\alpha} \frac{1}{2\pi j k} \Theta\left(\omega_u, \frac{2\pi k}{x_\Delta}\right) \Phi_X^*\left(\frac{2\pi k}{x_\Delta}\right) e^{j \frac{2\pi k}{x_\Delta} \xi} \Big|_{\xi=-\frac{x_\Delta}{2}}^x + \frac{x}{x_\Delta} + \frac{1}{2} \quad ; \quad |x| \leq \frac{x_\Delta}{2} \tag{3.38}$$

$$\alpha = \left\{ \max(k) : \frac{2\pi k}{x_\Delta} \leq \omega_u, \quad k \in \mathcal{N} \right\}.$$

In order to obtain a CDF estimate in terms of a number of moments, an expression for the characteristic function estimate in terms of N_m moments is substituted for $\Phi_X^*(\omega)$ in the

above expression:

$$\begin{aligned} \hat{F}_X(x) = \sum_{\substack{k=-\alpha \\ k \neq 0}}^{\alpha} \frac{1}{2\pi jk} \Theta(\omega_u, \frac{2\pi k}{x_\Delta}) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} \left(\frac{2\pi k}{x_\Delta} \right)^n \right\} e^{j \frac{2\pi k}{x_\Delta} \xi} \Big|_{\xi=-\frac{x_\Delta}{2}}^x \\ + \frac{x}{x_\Delta} + \frac{1}{2} \quad ; \quad |x| \leq \frac{x_\Delta}{2}. \end{aligned} \quad (3.39)$$

Finally, common terms are factored out to simplify the estimator and the estimate is assigned limiting values outside the interval $|x| \leq \frac{x_\Delta}{2}$:

$$\hat{F}_X(x) = \begin{cases} 0 & ; \quad x < -\frac{x_\Delta}{2} \\ \sum_{k=-\alpha}^{\alpha} \beta_k e^{j \frac{2\pi k}{x_\Delta} x} + \frac{x}{x_\Delta} & ; \quad |x| \leq \frac{x_\Delta}{2} \\ 1 & ; \quad x > \frac{x_\Delta}{2}, \end{cases} \quad (3.40)$$

with

$$\beta_k = \begin{cases} \frac{1}{2\pi jk} \Theta(\omega_u, \frac{2\pi k}{x_\Delta}) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} \left(\frac{2\pi k}{x_\Delta} \right)^n \right\} & ; \quad k \neq 0 \\ \frac{1}{2} - \sum_{\substack{n=-\alpha \\ n \neq 0}}^{\alpha} (-1)^n \beta_n & ; \quad k = 0. \end{cases} \quad (3.41)$$

This represents a practical estimator that expresses a CDF estimate directly in terms of a number of finite moments, using a Fourier series. Care should be taken when determining ω_u , specifying the highest frequency that is used in the estimate (which determines the number of frequency components): as the characteristic function is expressed as a Taylor series (in terms of a number of moments), it diverges above a certain frequency, thereby placing a limit on the maximum value of ω_u . Unfortunately the value of this maximum is highly dependent on the actual CDF being estimated and implementations should therefore determine ω_u dynamically by testing for divergence of the Taylor series.

3.5 Estimators based on the characteristic function

Using a similar approach to that followed in Section 2.5, we now consider ways of estimating a CDF directly from sample data using a simple characteristic function estimator. Some of the theory presented in the previous sections was developed within a more general

framework than was required at that stage, allowing it to be reused in this section.

In the chapter dealing with PDF estimators, we found that direct estimation of the characteristic function from sample data held numerous advantages above techniques employing moments. This was due to complications introduced by estimation of the characteristic function from moments (using a Taylor series). As the same complications were encountered with the CDF estimator employing moments, we now consider a method of estimating the CDF directly from sample data. This results in a more usable alternative to the moments-based technique in cases where sample data is available.

Estimation of a CDF using a characteristic function estimator that operates directly from sample data, thereby eliminating the use of moments, is first considered. It is shown that the proposed characteristic function estimator represents the Parzen estimator identically (in that it estimates a CDF by placing a kernel function at the location of each data sample). A practical technique of estimating a CDF using a Fourier series is then presented. This technique operates directly from sample data and represents a parametric approximation to the Parzen CDF estimate. It is recommended that it be used instead of the moments technique in applications where actual sample data is available.

3.5.1 A CDF in terms of a characteristic function

From Equation 3.5 we obtain an expression for a CDF estimate, $\hat{F}_X(x)$, in terms of a characteristic function estimate, $\hat{\Phi}_X(\omega)$:

$$\hat{F}_X(x) = \mathcal{F}^{-1} \left\{ \frac{1}{j\omega} \hat{\Phi}_X^*(\omega) \right\} + \frac{1}{2}. \quad (3.42)$$

By using the characteristic function estimator introduced in Section 2.5, the characteristic function estimate is expressed directly in terms of N_x data samples (where x_i refers to the i 'th data sample):

$$\hat{\Phi}_X(\omega) = \frac{1}{N_x} \sum_{i=0}^{N_x-1} e^{j\omega x_i}. \quad (3.43)$$

Combining these last two equations allows us to express a CDF estimate in terms of a number of data samples:

$$\begin{aligned}
\hat{F}_X(x) &= \mathcal{F}^{-1} \left\{ \frac{1}{j\omega} \frac{1}{N_x} \sum_{i=0}^{N_x-1} e^{-j\omega x_i} \right\} + \frac{1}{2} \\
&= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \mathcal{F}^{-1} \left\{ \frac{1}{j\omega} e^{-j\omega x_i} \right\} + \frac{1}{2} \\
&= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \left[\frac{1}{2} \operatorname{sgn}(x - x_i) \right] + \frac{1}{2} \\
&= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \left[\frac{1}{2} \operatorname{sgn}(x - x_i) + \frac{1}{2} \right] \\
&= \frac{1}{N_x} \sum_{i=0}^{N_x-1} u(x - x_i),
\end{aligned} \tag{3.44}$$

where $\operatorname{sgn}(x)$ represents the signum (sign) function, which is defined [5, p. 90] as:

$$\operatorname{sgn}(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0, \end{cases} \tag{3.45}$$

with its Fourier transform given by

$$\begin{aligned}
\mathcal{F}\{\operatorname{sgn}(x)\} &= \frac{2}{j\omega} \\
\mathcal{F}\{\operatorname{sgn}(x - x_i)\} &= \frac{2}{j\omega} e^{-j\omega x_i}.
\end{aligned} \tag{3.46}$$

$u(x)$ is the unit step function, which is defined in terms of the signum function,

$$\begin{aligned}
u(x) &= \frac{1}{2} + \frac{1}{2} \operatorname{sgn}(x) \\
&= \begin{cases} 0 & x < 0 \\ \frac{1}{2} & x = 0 \\ 1 & x > 0. \end{cases}
\end{aligned} \tag{3.47}$$

Compare this expression for the CDF estimate to Equation 2.70 which expressed the PDF in terms of a number of impulse functions: they are similar except for the choice of kernel function, the one employed by the CDF estimator being the integral of the one employed by the PDF estimator.

Equation 3.44 therefore represents the CDF counterpart of Equation 2.70 and, as in the case of the PDF estimator, has limited practical application as it does not provide any additional information above that provided by the raw data samples. In order to improve the estimator, a real windowing function with finite support, $\Theta(\omega_u, \omega)$, is introduced into the characteristic function estimator and a new CDF estimate is obtained (by interchanging the order of the summation and the inverse Fourier transform and using the multiplication/convolution property of the Fourier transform):

$$\begin{aligned}
\hat{F}_X(x) &= \mathcal{F}^{-1} \left\{ \frac{\Theta(\omega_u, \omega)}{j\omega} \frac{1}{N_x} \sum_{i=0}^{N_x-1} e^{-j\omega x_i} \right\} + \frac{1}{2} \\
&= \mathcal{F}^{-1} \left\{ \frac{1}{N_x} \sum_{i=0}^{N_x-1} \left[\frac{\Theta(\omega_u, \omega)}{j\omega} e^{-j\omega x_i} \right] \right\} + \frac{1}{2} \\
&= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \left[\mathcal{F}^{-1} \left\{ \frac{\Theta(\omega_u, \omega)}{j\omega} e^{-j\omega x_i} \right\} \right] + \frac{1}{2} \\
&= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \left[\mathcal{F}^{-1} \left\{ \Theta(\omega_u, \omega) \right\} \star \mathcal{F}^{-1} \left\{ \frac{1}{j\omega} e^{-j\omega x_i} \right\} \right] + \frac{1}{2} \\
&= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \left[\theta(\omega_u, x) \star \frac{1}{2} \text{sgn}(x - x_i) \right] + \frac{1}{2} \\
&= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \left[\int_{-\infty}^{\infty} \theta(\omega_u, \lambda) \frac{1}{2} \text{sgn}(x - x_i - \lambda) d\lambda \right] + \frac{1}{2}.
\end{aligned} \tag{3.48}$$

In order to evaluate this convolution integral, it is first separated into two integrals, one corresponding to positive values of the signum function and one to negative values. The fact that $\theta(\omega_u, x)$ has unit area (as $\Theta(\omega_u, \omega)$, which represents its Fourier transform, was constrained to attain a value of unity at the origin) is then used to express the second integral in terms of the first integral. After simplifying, the constant term is then moved

into the summation and the expression is again simplified:

$$\begin{aligned}
\hat{F}_X(x) &= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \left[\int_{-\infty}^{\infty} \theta(\omega_u, \lambda) \frac{1}{2} \operatorname{sgn}(x - x_i - \lambda) d\lambda \right] + \frac{1}{2} \\
&= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \left[\frac{1}{2} \int_{-\infty}^{x-x_i} \theta(\omega_u, \lambda) d\lambda - \frac{1}{2} \int_{x-x_i}^{\infty} \theta(\omega_u, \lambda) d\lambda \right] + \frac{1}{2} \\
&= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \left[\frac{1}{2} \int_{-\infty}^{x-x_i} \theta(\omega_u, \lambda) d\lambda - \frac{1}{2} \left\{ 1 - \int_{-\infty}^{x-x_i} \theta(\omega_u, \lambda) d\lambda \right\} \right] + \frac{1}{2} \\
&= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \left[\int_{-\infty}^{x-x_i} \theta(\omega_u, \lambda) d\lambda - \frac{1}{2} \right] + \frac{1}{2} \\
&= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \left[\int_{-\infty}^{x-x_i} \theta(\omega_u, \lambda) d\lambda - \frac{1}{2} + \frac{1}{2} \right] \\
&= \frac{1}{N_x} \sum_{i=0}^{N_x-1} \int_{-\infty}^{x-x_i} \theta(\omega_u, \lambda) d\lambda.
\end{aligned} \tag{3.49}$$

This expresses the CDF estimate in terms of a number of kernel functions that are obtained by integrating the inverse Fourier transformed windowing function, $\theta(\omega_u, x)$, and is the CDF counterpart of Equation 2.72 (the Parzen PDF estimator). It is seen that, in order for the CDF estimate to be monotonic, $\theta(\omega_u, x)$ should be a non-negative function. This restriction was not required by the PDF estimator. Consequently, windowing functions that qualifies for use in the PDF estimator does not necessarily qualify for use in the CDF estimator.

3.5.2 Fourier series

In Section 2.5.3, we presented a PDF estimator that expressed a PDF in terms of sample data using a Fourier series. Using the same approach, a CDF estimator is now presented that combines the characteristic function estimator considered in the previous section (and introduced in the previous chapter) with theory developed in order to estimate a CDF directly from sample data.

Equation 3.38, reproduced here, expresses an approximation of the CDF, $\hat{F}_X(x)$, in terms of the characteristic function, $\Phi_X(\omega)$, and a windowing function, $\Theta(\omega_u, \omega)$, which

are both uniformly sampled:

$$\begin{aligned}\hat{F}_X(x) &= \int_{-\infty}^x \hat{f}_X(\xi) d\xi \\ &= \sum_{\substack{k=-\alpha \\ k \neq 0}}^{\alpha} \frac{1}{2\pi jk} \Theta(\omega_u, \frac{2\pi k}{x_\Delta}) \Phi_X^*(\frac{2\pi k}{x_\Delta}) e^{j\frac{2\pi k}{x_\Delta} \xi} \Big|_{\xi=-\frac{x_\Delta}{2}}^x + \frac{x}{x_\Delta} + \frac{1}{2} \quad ; \quad |x| \leq \frac{x_\Delta}{2}\end{aligned}\quad (3.50)$$

$$\alpha = \{ \max(k) : \frac{2\pi k}{x_\Delta} \leq \omega_u, \quad k \in \mathcal{N} \}.$$

This is combined with the characteristic function estimator presented in Equation 3.43 by substituting the characteristic function estimate for the characteristic function in the above expression. The result is an expression for $\hat{F}_X(x)$ directly in terms of a number of data samples:

$$\begin{aligned}\hat{F}_X(x) &= \sum_{\substack{k=-\alpha \\ k \neq 0}}^{\alpha} \frac{1}{2\pi jk} \Theta(\omega_u, \frac{2\pi k}{x_\Delta}) \left\{ \hat{\Phi}_X^*(\omega) \right\} e^{j\frac{2\pi k}{x_\Delta} \xi} \Big|_{\xi=-\frac{x_\Delta}{2}}^x + \frac{x}{x_\Delta} + \frac{1}{2} \quad ; \quad |x| \leq \frac{x_\Delta}{2} \\ &= \sum_{\substack{k=-\alpha \\ k \neq 0}}^{\alpha} \frac{1}{2\pi jk} \Theta(\omega_u, \frac{2\pi k}{x_\Delta}) \left\{ \frac{1}{N_x} \sum_{i=0}^{N_x-1} e^{-j\frac{2\pi k}{x_\Delta} x_i} \right\} e^{j\frac{2\pi k}{x_\Delta} \xi} \Big|_{\xi=-\frac{x_\Delta}{2}}^x + \frac{x}{x_\Delta} + \frac{1}{2} \quad ; \quad |x| \leq \frac{x_\Delta}{2}.\end{aligned}\quad (3.51)$$

Common terms are factored out, all the constant terms are consolidated¹ and the estimate is assigned limiting values outside the interval $|x| \leq \frac{x_\Delta}{2}$:

$$\hat{F}_X(x) = \begin{cases} 0 & ; \quad x < -\frac{x_\Delta}{2} \\ \sum_{k=-\alpha}^{\alpha} \beta_k e^{j\frac{2\pi k}{x_\Delta} x} + \frac{x}{x_\Delta} & ; \quad |x| \leq \frac{x_\Delta}{2} \\ 1 & ; \quad x > \frac{x_\Delta}{2}, \end{cases}\quad (3.52)$$

¹ β_0 is obtained by expressing the sum corresponding to $\xi = -\frac{x_\Delta}{2}$ in terms of the other β_k and noting that $e^{j\frac{2\pi k}{x_\Delta} \xi} \Big|_{\xi=-\frac{x_\Delta}{2}} = (-1)^k$.

with

$$\beta_k = \begin{cases} \frac{1}{2\pi jk} \Theta(\omega_u, \frac{2\pi k}{x_\Delta}) \left\{ \frac{1}{N_x} \sum_{i=0}^{N_x-1} e^{-j \frac{2\pi k}{x_\Delta} x_i} \right\} & ; \quad k \neq 0 \\ \frac{1}{2} - \sum_{\substack{n=-\alpha \\ n \neq 0}}^{\alpha} (-1)^n \beta_n & ; \quad k = 0. \end{cases} \quad (3.53)$$

This represents a practical CDF estimator that operates from sample data only and overcomes the problems associated with the techniques employing moments. It should be preferred above the moments-based techniques in applications where sample data is available. Due to the additional constraints placed on the windowing function (when compared to the PDF estimators), the CDF estimator presented above should normally require a higher value of ω_u than the corresponding PDF estimator would in order to obtain an estimate of comparable accuracy (this being due to the restricted windowing functions causing more smoothing).

3.6 Conclusions

Although a large part of the theory required by this chapter was inherited from the previous chapter, estimating a cumulative density function (CDF) using moments and the characteristic function presented its own challenges.

Estimating a CDF only in terms of a finite number of moments was found to be feasible in theory as well as in practice (Section 3.4) and an integral expression for the CDF estimate, involving the values of a finite number of moments, was obtained (Equation 3.11). As evaluating the integral using the anti-derivative did not prove to be a desirable course of action in the previous chapter (dealing with PDF estimation), it was not attempted here. Instead only a numerical integration technique and a Fourier series technique were considered. The former technique was developed in a generic fashion and can be applied to different numerical integration techniques (Section 3.4.2).

However, developing a CDF estimator employing a Fourier series solution proved to be more challenging. This was due to the asymmetry of a CDF and the fact that it does not have finite support, combined with the Fourier series being optimal for representing periodic functions or functions with finite support. Fortunately all these challenges were addressed and successfully resolved and the result is a practical technique that allows a CDF estimate (involving a Fourier series) to be obtained in terms of a finite number of

moments (Section 3.4.3). As the numerical integration technique is more difficult to apply than the Fourier series one, it is recommended that the Fourier series technique be preferred in practice. The numerical integration technique should, however, be used in applications where the approximations introduced by the Fourier series technique are undesirable.

Extending the techniques to make use of the characteristic function estimator presented in the previous chapter only required the combination of the moments technique with results from the previous chapter. The theoretical result was an integral expression for a CDF estimate in terms of sample data only, without involving moments (Equation 3.49). This produced a CDF estimate that was identical to the result produced by the Parzen estimator, but derived using a frequency domain perspective. From this, a practical technique involving the Fourier series, that has the ability to estimate the CDF using sample data only, was developed. This technique can be viewed as a parametric approximation to the Parzen estimator and it combines the desirable accuracy and ease of use of the Parzen estimator with the desirable computational complexity of a parametric technique. In the presence of sample data, this technique should be preferred above the one employing moments, with the moments technique reserved for applications where only moments data is available.

The result is a family of scalable CDF estimators that are simple to train and have the ability to either operate directly from sample data or from the values of a number of moments only.

Chapter 4

Experimental results

4.1 Introduction

In the previous chapters, two types of estimators were introduced for both probability density functions (PDFs) and cumulative density functions (CDFs): the first type estimated these functions from a set of moments characterising a random variable while the second type utilised the characteristic function, which it estimated from a set of data samples. Although the theory behind each type of PDF and CDF estimator was extensively developed, a number of practical techniques (or algorithms) that allowed the theory to be implemented in practice were also presented.

One such technique provided an estimate in the form of a Fourier series expansion and was derived for both the PDF and CDF estimators. Software implementations were created for the following four Fourier series techniques in order to evaluate the performance of the techniques and also to verify the validity of the theory from which it was derived:

1. PDF estimate using moments, Section 2.4.5.
2. PDF estimate from sample data using the characteristic function, Section 2.5.3.
3. CDF estimate using moments, Section 3.4.3.
4. CDF estimate from sample data using the characteristic function, Section 3.5.2.

As the theory presented in the previous chapters is novel and derived from basic principles, experiments were designed to provide a quantitative and critical evaluation of the theory. The object was to verify the correctness of the theory presented in the previous

chapters and to obtain an indication of the performance that could be expected from estimators based on this theory. In order to provide a broad perspective on the performance of the estimators, a comparison was drawn between the performance of the techniques listed above and two well-known and established density function estimators, namely the Gaussian Mixture Model (GMM) and the Parzen estimator. Most of the experiments were conducted on synthetic data, which allowed comparison of the techniques under controlled conditions. Each estimator's performance can be characterised in a number of different ways, with the importance of a certain aspect of its performance dependent on the details of the actual application in which the estimator is deployed. Experiments were designed to compare the following characteristics of the estimators:

Accuracy This indicates the extent to which the PDF (or CDF) estimate corresponds to the actual PDF (or CDF) from which the samples, on which the estimate is based, was drawn. It is quantified by some distance or error measure that indicates the amount by which the estimate differs from the actual function.

Computational requirements This indicates the amount of computing resources that is required to train and evaluate the estimate. As an estimator with lower computational complexity requires a shorter time to perform a certain task than one with higher complexity, experiments were designed in which the time taken to perform certain estimates were measured for each estimator. These figures provide important information when selecting a technique to be used in situations where execution speed is an important factor, e.g. a real-time system or a problem that involves the processing of large amounts of data.

Sensitivity to parameters Each estimator possesses a number of degrees of freedom or parameters that has to be specified by the user of the estimator. As unsuitable choices of parameters can produce undesirable results, it is advantageous to find a range of "safe" parameter values which provide adequate (and in some cases optimal) performance over a range of conditions. The experiments involving the accuracy and computational complexity were all performed over ranges of parameter values in order to determine choices of parameter values that optimise some aspect of each estimator or causes it to perform satisfactorily over a range of conditions.

Training requirements Each estimator collects information, mostly in the form of statistics, from the training set in order to estimate a PDF or CDF. Details of this training

procedure differ between the estimators and is an important consideration in applications where the estimate is often updated.

At the end of the chapter, one of the estimators is compared to the GMM by constructing a practical example of a classifier that can function as part of a speaker verification application.

4.2 Experimental setup

4.2.1 Input data

In order to evaluate the density function estimators, random data was used as inputs to the experiments and synthetic data corresponding to five different PDFs were generated by a computer. PDFs with strong characteristics were selected, thereby emphasising the strengths and weaknesses of each estimator with regards to certain PDF characteristics encountered in practice. Furthermore, as each PDF possessed unique characteristics, it allowed a controlled test to be performed over a range of operating conditions. Input data corresponding to the following density functions were used in the experiments:

normal A symmetrical, continuous and unimodal density function defined by a Gaussian density function.

uniform A symmetrical, unimodal density function containing sharp discontinuities.

skew An asymmetrical continuous and unimodal density function defined by a Rayleigh density function.

bimodal A symmetrical, continuous and bimodal density function consisting of two translated Gaussian density functions.

mixture An asymmetrical, discontinuous and bimodal density function defined by a mixture between a Gaussian and a uniform density function.

In order to be able to provide a fair comparison between the performance of each estimator on different PDFs, the above density functions were all normalised to have zero means and unity variance (or values very close to these).

Sample data corresponding to each of the above PDFs was generated by independently drawing samples from a uniform or Gaussian random process and then applying the necessary transforms to obtain the desired distribution. A number of datasets (20), each

containing the same number of samples, were generated for each PDF. This allowed statistics about the performance of each estimator to be collected by repeating an experiment over all the datasets corresponding to a certain PDF.

4.2.2 Estimation error measure

For each PDF estimator, an estimation error was calculated as the Kullback-Leibler [3, p 59] distance between the actual PDF and the PDF estimate:

$$\varepsilon_P = - \int_{-\infty}^{\infty} f_X(x) \ln \frac{\hat{f}_X(x)}{f_X(x)} dx, \quad (4.1)$$

where $f_X(x)$ denotes the actual PDF and $\hat{f}_X(x)$ an estimate. This is an asymmetrical distance that is only equal to zero if $f_X(x)$ and $\hat{f}_X(x)$ are identical. The above integral was evaluated using a numerical method (the trapezoid rule [29]) which evaluated the integrand over 1000 points uniformly spaced in the interval $[-8, 8]$.

For each estimator and PDF combination a number of trials were performed, each time calculating ε_P over a number of datasets corresponding to the same PDF. The mean and variance of the estimation error was then calculated, providing a realistic indication of the ability of a particular estimator to estimate a particular PDF with certain characteristics. This operation was repeated over all PDFs and for each estimator.

A similar process was followed to evaluate the CDF estimators, except that a different distance measure was employed. In this case, the distance was defined as the integral absolute difference between the actual CDF and the CDF estimate:

$$\varepsilon_C = - \int_{-\infty}^{\infty} |F_X(x) - \hat{F}_X(x)| dx, \quad (4.2)$$

where $F_X(x)$ denotes the actual CDF and $\hat{F}_X(x)$ an estimate. This error was also integrated over 1000 steps using the trapezoid rule in the interval $[-8, 8]$. As with the PDF estimates, the mean and variance of the estimation error was calculated for all combinations of estimators and PDFs over a number of trials.

These error rates were calculated over a range of parameter values, characterising each estimator, in order to illustrate the performance of each estimator over a subset of its operating range. The ranges were selected to include optimal working points or to show asymptotic behaviour (where applicable) and were limited to values that could typically

be used in practice. Evaluating the estimators over a range of operating conditions also ensured a fair comparison between estimators as it prevented a selection of parameters that only favoured some estimators.

Details of each estimator as well as the parameters that were varied and their ranges are now provided.

4.2.3 PDF and CDF estimate using moments

The expressions for the PDF and CDF estimators are given by Equations 2.62-2.63 and Equations 3.40-3.41 respectively. Values for the adjustable parameters were selected as follows, with the number of frequency components being varied over a range of values:

- N_m : All estimates employed 100 moments, as working with higher order moments is impractical due to numerical instability and the high variance of the estimators.
- α : The number of frequency components used in the approximation was varied in the range $[4, 30]$.
- x_Δ : The reconstruction period was selected to be equal to 16.
- $\Theta(\omega_u, \omega)$: A Hamming window was selected for the PDF estimator as it resulted in less smoothing than the Bartlett (triangular) window did (due to the narrower convolution kernel). This improved the accuracy of the approximation, most notably in the tail regions where smoothing often introduces large errors. The CDF estimate did, however, employ a Bartlett window due to additional constraints on the windowing function (that the convolution kernel should be non-negative) in order to ensure that the CDF estimate is monotonic.

4.2.4 PDF and CDF estimate using characteristic function

The expressions for the PDF and CDF estimators are given by Equations 2.79-2.80 and Equations 3.52-3.53 respectively. Values for the parameters were selected as follows, with the number of frequency components being varied over a range of values:

- α : The number of frequency components used in the approximation was varied in the range $[2, 100]$.
- x_Δ : The reconstruction period was selected to be equal to 16.

- $\Theta(\omega_u, \omega)$: A Hamming window was selected for the PDF estimator and a Bartlett (triangular) window for the CDF estimator (using the same motivation that was used for the moments technique).

4.2.5 Parzen estimator

A Gaussian density function was used as kernel for the PDF estimator and a Gaussian cumulative density function was used for the CDF estimator. Each kernel only has a single parameter, corresponding to the variance of the Gaussian density function, which was varied in the range $[1, 100]^{-1}$.

4.2.6 Gaussian Mixture Model

The PDF was approximated using a mixture of Gaussian density functions. The mixture density parameters were estimated from data using the iterative Expectation-Maximisation (EM) algorithm. The CDF was approximated from the same mixture parameters that was used for the PDF estimate, but using Gaussian *cumulative* density functions as bases. The number of mixtures was varied in the range $[1, 20]$.

4.3 Mean estimation error

A number of graphs showing the mean estimation error plotted against values of the operational parameter which were varied are now presented for each estimator. Each result was obtained by performing 20 trials in which PDF estimates were obtained from 100 data samples.

A total of 8 graphs are shown, one corresponding to each PDF estimator and one corresponding to each CDF estimator. Each graph contains 5 plots of the mean estimation error, each one corresponding to one of the input data PDFs (normal, uniform, skew, bimodal, mixture). The minimum mean estimation error is also indicated individually for each PDF plot and represents the optimal working point for a particular PDF and estimator combination when estimating a PDF from 100 samples.

4.3.1 PDF estimators

Graphs showing the estimation errors corresponding to the 4 different PDF estimators are shown in Figure 4.1 (GMM), Figure 4.2 (Parzen), Figure 4.3 (characteristic function)

and Figure 4.4 (moments). Plots of typical PDF estimates, obtained from 100 samples corresponding to the Gaussian/uniform mixture PDF, are shown in Figure 4.5.

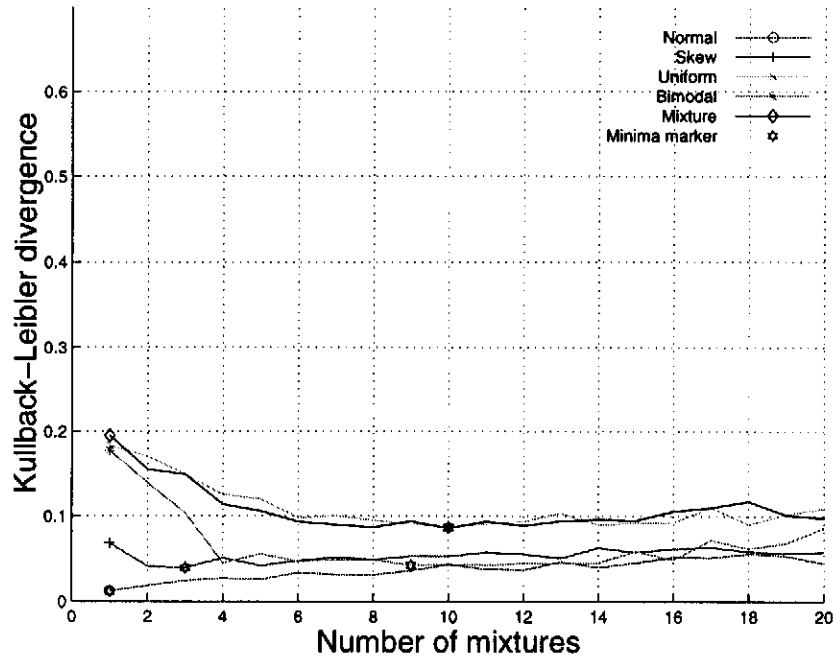


Figure 4.1: Mean estimation error: PDF estimate using GMM from 100 samples.

All 4 estimators exhibited similar performance when evaluated in terms of the relative performance on datasets corresponding to different PDFs, values of the minimum errors and the grouping and location of the optimal working points. Table 4.1 summarises the minimum mean estimation errors for each technique/PDF combination. A combined mean estimation error was also calculated by averaging the mean estimation errors corresponding to the 5 different types of PDF. The minimum value of this combined error is indicated in the last column of the table (along with its corresponding combined standard deviation in brackets). The minimum error in each column is indicated using bold typeface.

From this table, we note that all the estimators are sensitive to the actual PDF and that no single one performs optimally on all the PDFs. By considering the optimal working points, we note that all the estimators show an affinity towards smooth and continuous PDFs and require more parameters or degrees of freedom to optimally estimate a PDF containing discontinuities. For cases where no prior knowledge exists about the actual PDF, it is desirable to find a working point that provides optimal performance over more than one PDF. Such a point is found by considering the parameter value at which the

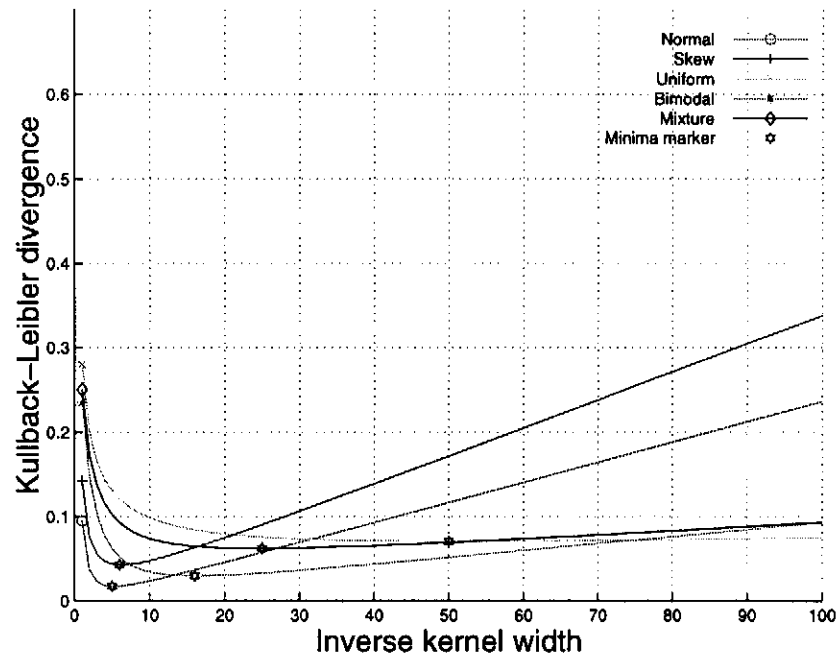


Figure 4.2: Mean estimation error: PDF estimate using Parzen estimator from 100 samples.

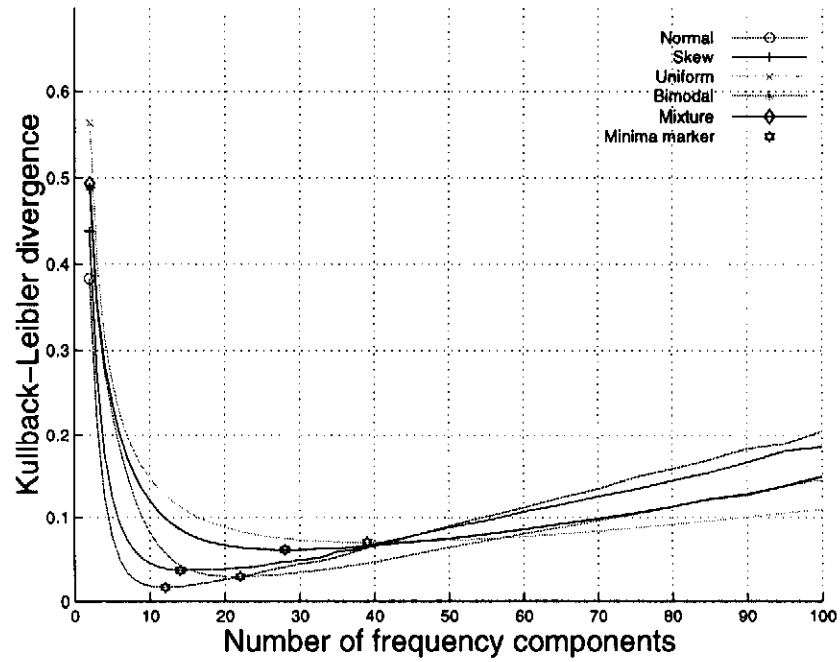


Figure 4.3: Mean estimation error: PDF estimate using characteristic function from 100 samples.

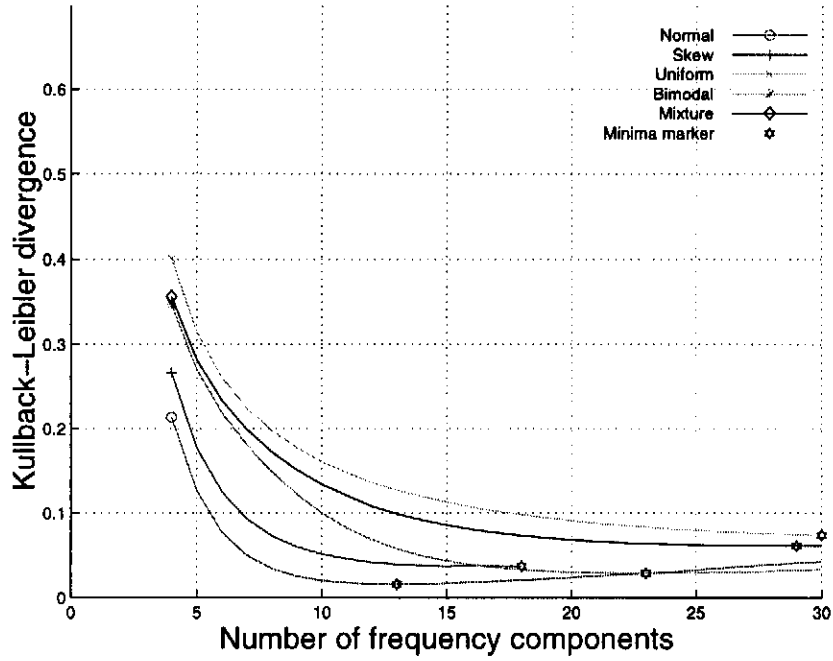


Figure 4.4: Mean estimation error: PDF estimate using moments from 100 samples.

	Normal	Uniform	Skew	Bimodal	Mixture	Combined
Moments	1.60	9.77	3.69	3.29	7.32	5.23 (1.36)
CF	1.60	7.02	3.70	2.93	6.11	4.90 (1.43)
GMM	1.22	8.71	3.95	4.26	8.60	6.23 (3.11)
Parzen	1.70	7.05	4.28	2.99	6.21	5.47 (1.72)

Table 4.1: PDF estimators: minimum mean estimation errors, minimum combined mean estimation errors and corresponding combined standard deviations (in brackets) from 100 samples ($100 \times$ Kullback-Leibler divergence is indicated).

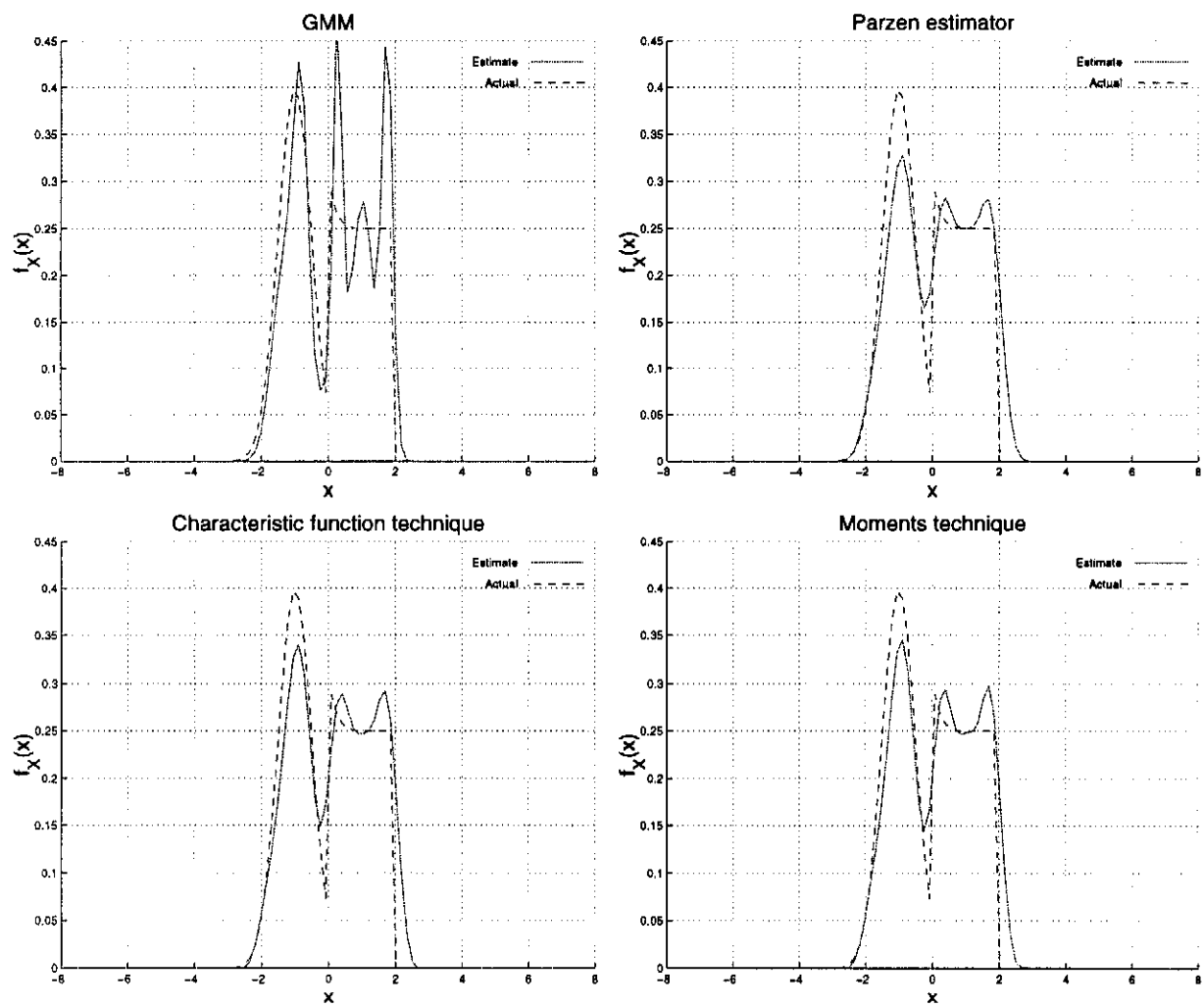


Figure 4.5: Typical PDF estimates.

combined estimation error of each estimator attains a minimum.

It is interesting to note that performance of the estimators are very similar when the combined error is considered, with the difference between the smallest and largest error being within 30% of the smallest error. This is quite remarkable, considering the difference in the underlying approaches followed by the different techniques. Of special interest is the fact that the two estimators presented in the previous chapters outperform the two established estimators on average. In terms of the standard deviations (indicated in brackets in the last column), the new estimators also exhibited the two lowest values, providing assurance that the reduced mean error does not come at a cost in terms of the variance of the estimates. Also note the high standard deviation exhibited by the GMM, compared to the other estimators. This can be ascribed to the iterative training algorithm that is prone to being caught in local minima (of the loss function) and is sensitive to the initial conditions of the training procedure, causing it to sometimes produce inaccurate estimates. This highlights a major advantage that the other techniques share above the GMM: they all provide closed-form solutions to the estimates that do not require iteration or careful setup of initial conditions.

This close match in overall performance instills confidence in the fact that the theory presented in the previous chapters is correct as well as applicable in practice (over a range of operating conditions). It is unlikely that these results would be obtained if the theory had been incorrect or if relevant facts had been omitted from consideration.

Furthermore, the fact that the two new estimators outperform the established estimators for this specific selection of PDFs and number of samples indicates that there are situations where these new estimators are better suited to the problem at hand than the established ones (when considering performance in terms of mean approximation error).

For the PDF estimator using moments, there is a restriction on the maximum value of the number of frequency components (which is the parameter that was varied during the experiment) that can be used in the approximation. This is due to the fact that each frequency component represents a value sampled from a characteristic function approximation, that was generated using a Taylor series expansion. This function starts to diverge at a certain frequency, which is dependent on the actual PDF. This explains the reason for the premature termination of some of the plots in Figure 4.4. An implication of this is that the number of frequency components should not be a fixed number in practice, but rather be adapted to the actual characteristic function approximation. Alternatively, if a fixed number of components are required, a conservative (low) number of frequency

components should be selected. Unfortunately, imposing a maximum value on the number of frequency components causes a certain minimum amount of smoothing to always take place. Although this can be beneficial in the case of sparse sample data, as it prevents over-fitting to the dataset, one would expect the performance of this estimator to show a less marked improvement, when compared to the other estimators, as more data samples are added. This restriction makes the technique using moments impractical to use in situations where other methods (such as the one using the characteristic function) can be used and it should only be used in situations where the only available data are the values of a number of moments. Nonetheless, we have demonstrated that under such conditions accurate estimates can still be obtained.

Both the Parzen estimator and the characteristic function technique showed tendencies of over-fitting to the data for choices of parameters beyond the optimal points. In both cases this was due to a decrease in the amount of smoothing applied to the estimate, allowing loss of continuity between neighbouring samples. This effect is most pronounced when the sample data sparsely populates the sample space. Examples of over-fitting by the Parzen estimator and characteristic function technique are seen in Figure 4.6. In particular, note an interesting resemblance between the Parzen estimate and the characteristic function estimate in the part of the estimate representing the uniform density: what might have been mistaken for the Gibbs phenomenon [20, p. 259] in the characteristic function approach (as it uses a Fourier series expansion and the estimate resembles an oscillation) is actually just a symptom of over-fitting.

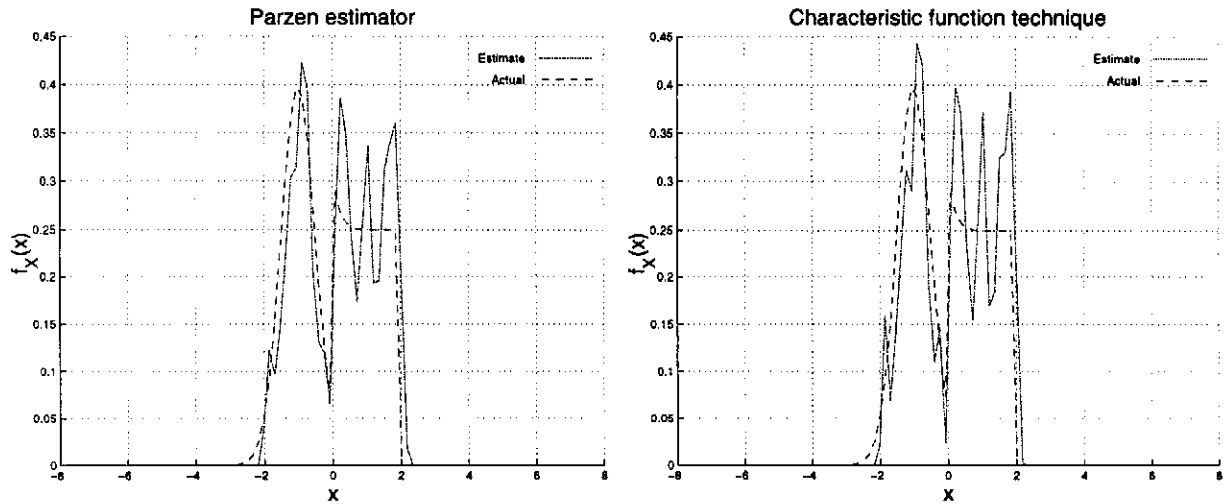


Figure 4.6: Examples of over-fitted PDF estimates (obtained from 100 samples).

In order to determine the influence that the number of samples from which the PDF is estimated has on the performance of each technique, the experiments were repeated using 1000 input samples. Doing this also allows us to verify the general applicability of conclusions drawn from the results based on 100 samples.

Each experiment (corresponding to a combination of estimator and PDF) was again repeated over 20 trials and the mean approximation errors and standard deviations were calculated. A combined mean error and standard deviation was also again calculated for each PDF by averaging the mean estimation error and standard deviation over the 5 different types of PDF. Generally, the results followed a similar pattern to the experiments using 100 data samples, but with an overall reduction in the values of the approximation errors. The graphs of the mean approximation errors corresponding to 1000 input samples are not shown, as it does not provide any particular additional insight. However, a summary of the minimum mean estimation errors are provided in Table 4.2.

	Normal	Uniform	Skew	Bimodal	Mixture	Combined
Moments	0.53	10.46	2.12	3.39	7.45	4.79 (0.33)
CF	0.41	2.34	0.85	0.55	2.11	1.66 (0.29)
GMM	0.09	5.65	0.78	0.65	3.91	2.32 (0.97)
Parzen	0.45	3.18	1.19	0.59	2.38	2.04 (0.41)

Table 4.2: PDF estimators: minimum mean estimation errors, minimum combined mean estimation errors and corresponding combined standard deviations (in brackets) from 1000 samples (100× Kullback-Leibler divergence is indicated).

All the estimators, with the exception of the moments technique showed a large improvement (between 60 and 70 percent) in terms of their average error rates. This was predicted earlier when it was noted that the estimates obtained using the moments technique (which improved by less than 10 percent) suffered from extreme bias due to the limit on the maximum number of frequency components that may be used. Using a larger number of samples did, however, improve the variance of the estimator on the same scale as it did for the other estimators (between 65 and 80 percent).

Also of interest, is the performance of the estimators on 1000 samples, when using values of the optimal operating points determined from 100 samples. This provides an indication of the sensitivity of the estimator to the number of samples. Ideally, an estimator would have a single optimal operating point regardless of the number of samples used in the estimate. Unfortunately this is not a very realistic expectation. It is, however, still feasible to require that when using the optimal working point corresponding to 100 input samples

to obtain an estimate from 1000 samples, that

- the error would not be significantly worse than at the optimal working point corresponding to 1000 samples, and
- that the error would not be worse than that obtained from 100 samples.

Table 4.3 contains values of estimation errors corresponding to optimal and sub-optimal working points. Each column contains the minimum combined mean estimation errors (and corresponding standard deviations) that was obtained by estimating the PDFs from either 100 or 1000 samples using a choice of parameters corresponding to a working point that is optimally suited for an estimate from either 100 or 1000 samples. Estimates were obtained for all 5 PDFs and the mean estimation errors and standard deviations were calculated over the results corresponding to 20 datasets (the first two columns were therefore copied from the last columns of Table 4.1 and Table 4.2).

	Working point optimal for					
	100 samples		1000 samples		100 samples	
	Estimate obtained from					
	100 samples		1000 samples		1000 samples	
Moments	5.23	(1.36)	4.79	(0.33)	N/A	
CF	4.90	(1.43)	1.66	(0.29)	2.80	(0.26)
GMM	6.23	(3.11)	2.32	(0.97)	8.26	(3.43)
Parzen	5.47	(1.72)	2.04	(0.41)	3.28	(0.28)

Table 4.3: PDF estimators: the effect of selecting a sub-optimal working point.

Both the characteristic function technique and Parzen estimator showed an improvement in accuracy when using 1000 samples than when using 100 samples (when performing estimates at the optimal working point corresponding to 100 samples). The accuracy of estimates obtained from 1000 samples using the optimal working point for 100 samples also showed acceptable values when compared to the optimal values corresponding to 1000 samples. This allows us to conclude that operating points corresponding to a number of samples can be used when estimating a PDF from a larger number of samples, without undue loss of accuracy, in the cases of the characteristic function technique and Parzen estimator (at least in some situations).

Unfortunately, the results show this not to be the case for the GMM and moments technique. In the case of the moments technique, the optimal working point corresponding

to 100 samples could not be used to estimate the PDF from 1000 samples as it required too high frequency components. Therefore, the number of parameters should always be selected based on the actual sample data at hand and never use a fixed number.

The GMM actually showed a degradation in accuracy (past the accuracy obtained from 100 samples), when using a sub-optimal working point to determine an estimate from 1000 samples. The GMM therefore seems to be more sensitive towards choices of parameter values than the Parzen and characteristic function techniques.

4.3.2 CDF estimators

Graphs showing the estimation errors corresponding to the 4 different CDF estimators are shown in Figure 4.7 (GMM), Figure 4.8 (Parzen), Figure 4.9 (characteristic function) and Figure 4.10 (moments). Plots of typical CDF estimates, obtained from 100 samples corresponding to the Gaussian/uniform mixture PDF, are shown in Figure 4.11.

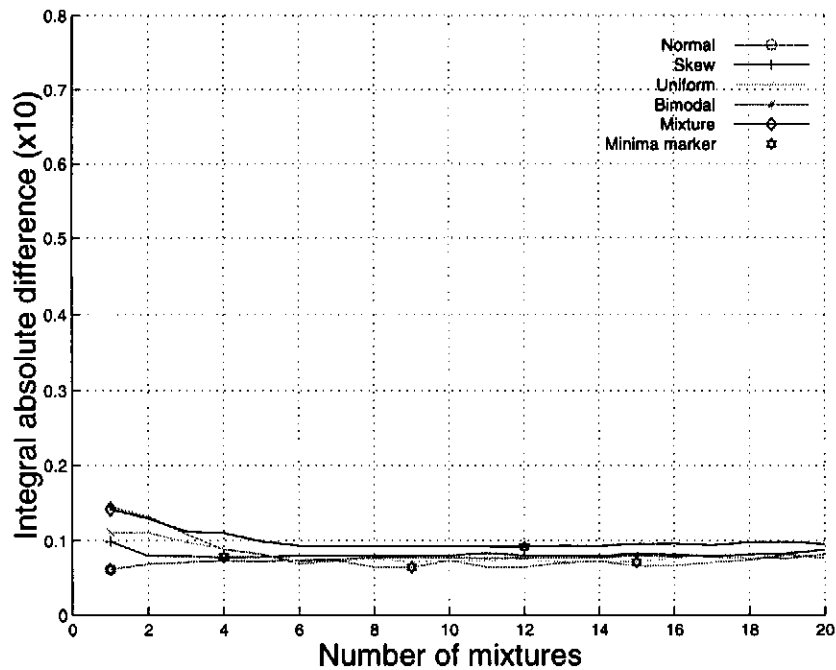


Figure 4.7: Mean estimation error: CDF estimate using GMM from 100 samples.

Compared to the PDF estimators, all the CDF estimators exhibited a general apathy towards the choice of parameters as well as the actual PDF being estimated. For all the estimators except the GMM, the plots corresponding to the different PDFs, are barely

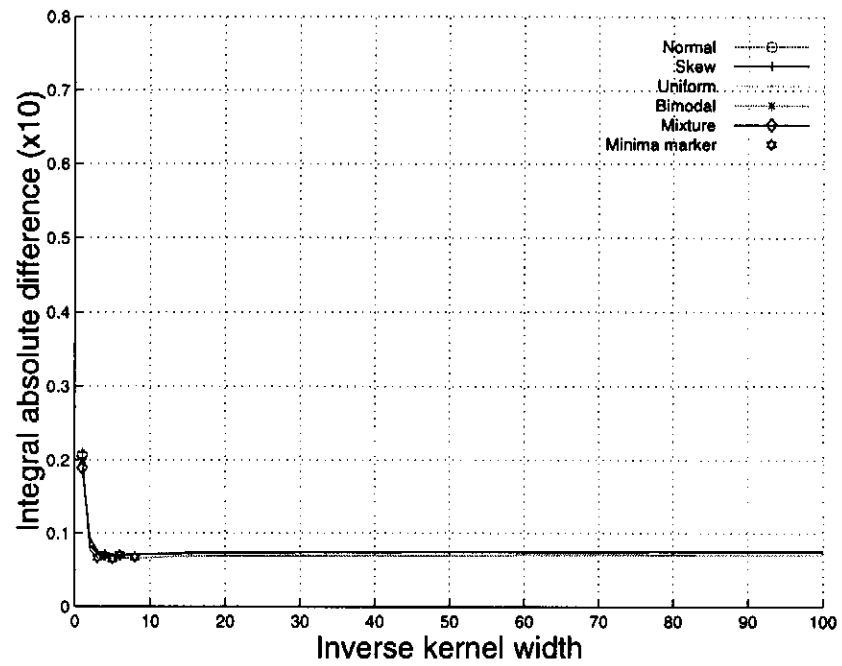


Figure 4.8: Mean estimation error: CDF estimate using Parzen estimator from 100 samples.

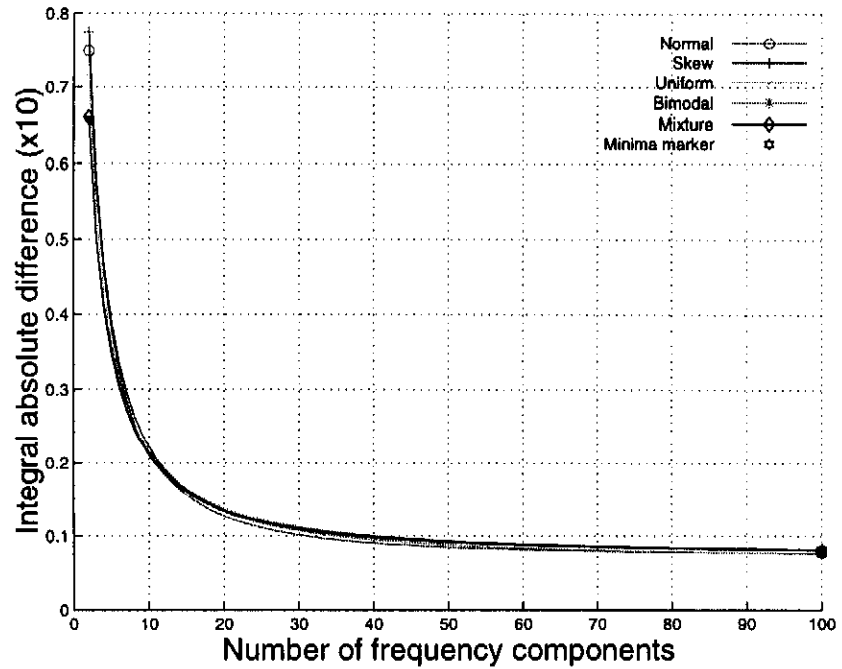


Figure 4.9: Mean estimation error: CDF estimate using characteristic function from 100 samples.

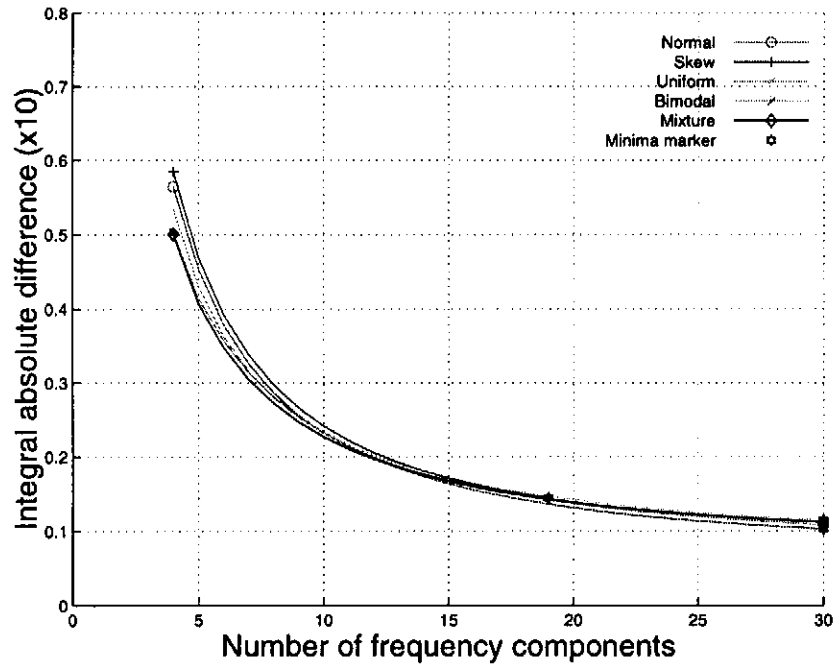


Figure 4.10: Mean estimation error: CDF estimate using moments from 100 samples.

discernible from each other. Furthermore, all the estimators with the exception of the moments technique, reached a point after which the graphs essentially became flat, thereby removing nearly all the dependence of the error on the value of the parameter.

These estimators therefore had a lower limit, independent from the type of CDF being estimated, on their mean approximation errors. This guards against over-fitting, to which some of the PDF estimators (notably the Parzen estimator and characteristic function technique) were prone, as a wide range of parameter values produce near-optimal results. However, in the case of the moments technique, the graph does not flatten out completely over the range of parameter values that was tested. It is therefore expected for this estimator to show a higher minimum approximation error as it never reaches its optimal point.¹ This is confirmed by the error rates shown in Table 4.4, which shows the estimation error corresponding to the different estimators for different density functions, as well as the combined error (obtained by averaging the mean estimation error over the 5 types of PDFs).

From this table we again see that all the estimators, with the exception of the moments

¹Increasing the number of frequency components would not improve the situation, as this would require an increase in the number of moments used in the characteristic function estimate. This is not recommended due to the impracticality of estimating high order moments.

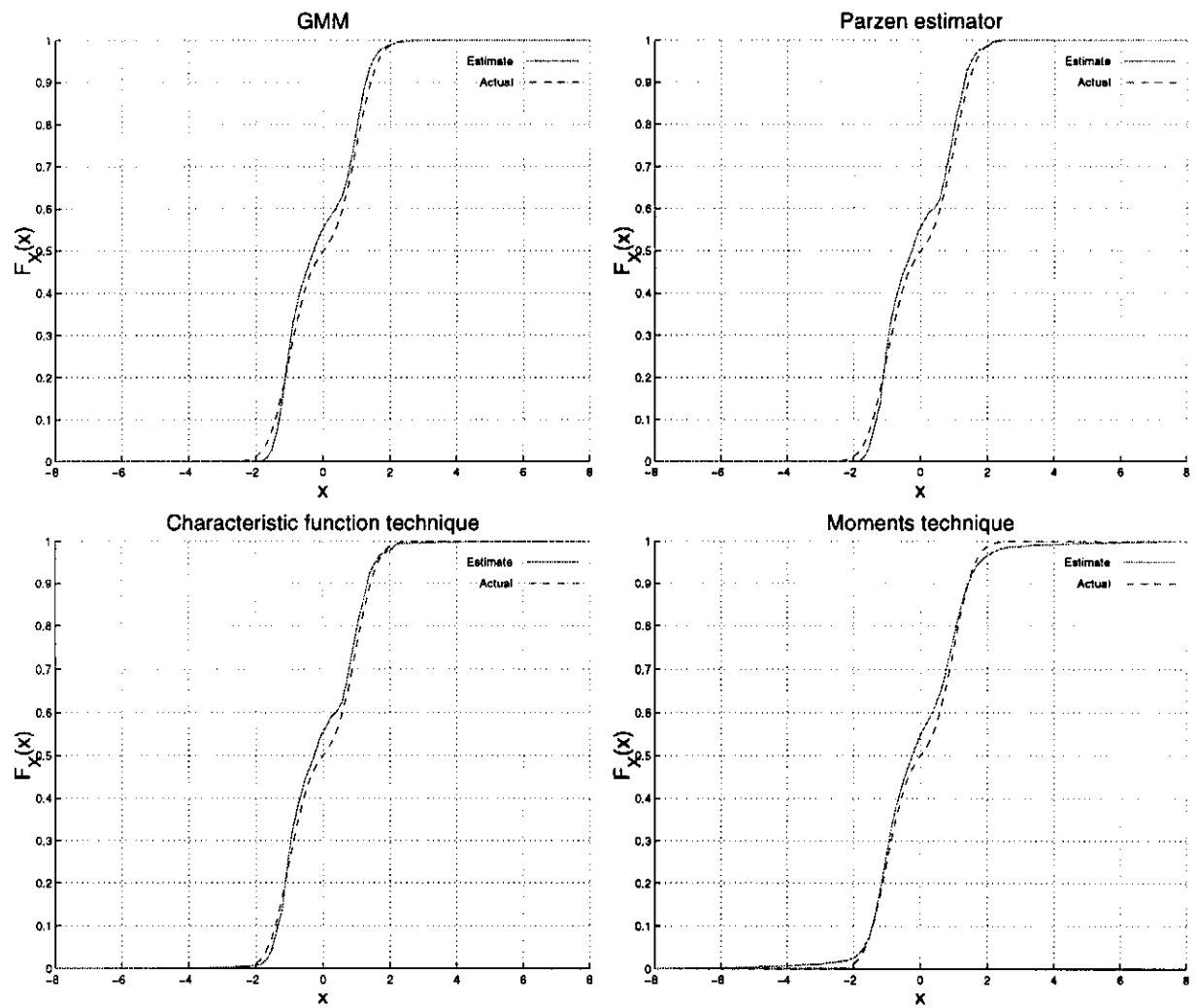


Figure 4.11: Typical CDF estimates.

technique, perform similarly in terms of their minimum approximation errors. Also, by comparing the graphs we note that the characteristic function and moments techniques exhibit similar performance. Based on these observations, we draw the conclusion that these new CDF estimators, presented in previous chapters, do perform correctly and produce estimates with accuracy that is comparable to established estimators. Use of the moments technique is, however, discouraged in applications where another estimator (such as the characteristic function technique) can be used.

	Normal	Uniform	Skew	Bimodal	Mixture	Combined
Moments	1.37	1.47	1.45	1.42	1.43	1.43 (0.29)
CF	0.77	0.80	0.81	0.76	0.82	0.79 (0.29)
GMM	0.63	0.72	0.78	0.66	0.92	0.78 (0.38)
Parzen	0.66	0.67	0.69	0.65	0.70	0.68 (0.31)

Table 4.4: CDF estimators: minimum mean estimation errors, minimum combined mean estimation errors and corresponding combined standard deviation (in brackets) from 100 samples ($10\times$ integral absolute difference is indicated).

The Parzen estimator outperforms all the estimators, except the GMM in the case of a single Gaussian PDF. As the Parzen estimator provides the least parametric (i.e. corresponding to a reduction in information when compared to the raw data samples) estimate it has the ability to provide the best fit to the data, thereby ensuring the highest accuracy. This should be contrasted to the case of the PDF estimators, where a certain amount of generalisation (smoothing) helped to reduce the problem of over-fitting (which plagued the Parzen estimator for certain choices of parameter and sample set size).

Recall that the characteristic function and moments techniques used a different frequency domain windowing function for the CDF estimate than for the PDF estimate. The one that was selected for the CDF estimate was required to have a non-negative inverse Fourier transform (which was not required for the PDF estimator). This choice resulted in a windowing function which applied more smoothing to the estimate than the PDF estimator did. Consequently, we expect it to perform worse when compared to the other estimators. This constraint on windowing function would, unfortunately, also be present in practice.

Table 4.5 shows the minimum mean and minimum average mean approximation errors corresponding to 1000 input samples. The error rates of the GMM, Parzen estimator and characteristic function technique show a reduction in the error rate, when compared to the results corresponding to 100 input samples, while the moments technique shows virtually

no change. However, as the standard deviation of all the estimators decreased with an increase in the number of samples, we can conclude that all the estimators show consistent behaviour, thereby deriving benefit from an increase in sample size.

	Normal	Uniform	Skew	Bimodal	Mixture	Combined
Moments	1.43	1.37	1.51	1.40	1.34	1.41 (0.09)
CF	0.34	0.38	0.37	0.36	0.34	0.36 (0.10)
GMM	0.20	0.30	0.23	0.23	0.26	0.25 (0.10)
Parzen	0.23	0.23	0.23	0.22	0.21	0.23 (0.09)

Table 4.5: CDF estimators: minimum mean estimation errors, minimum combined mean estimation errors and corresponding combined standard deviation (in brackets) from 1000 samples ($10\times$ integral absolute difference is indicated).

4.4 Computational requirements

When selecting an algorithm for use in an application where computational resources are expensive, such as a real-time or embedded application, the computational requirements are an important consideration. In this section we compare the computational requirements of the four estimators considered in the previous section: GMM, Parzen, characteristic function and moments technique. The aim is again to find out how the new estimators, introduced in the previous chapters, measure up against some established estimators. We assume throughout that the algorithms would be implemented as software targeted for a modern and ubiquitous microprocessor architecture (and therefore do not consider the performance implications on dedicated or unusual hardware architectures). This is done to ensure that any results obtained here reflect practical usage considerations and that they can be implemented in real-life systems with the minimum of effort. Furthermore, we measure the amount of computational power required by an estimator in terms of the time required to perform a certain task as this is one of the chief constraints in a real-time system.

Determining the amount of time to execute a certain algorithm on a specified modern computing platform theoretically proves to be a daunting task. This is due to the nature of modern microprocessor architectures and the large difference in performance that can be extracted by even seemingly meaningless optimisations. It was therefore decided to determine the computational requirements of each algorithm empirically by constructing

software benchmark applications. By drawing on the similarities in the operations that different estimators perform, the task was narrowed down to the construction of a set of benchmarks that allowed a comparison between all of the estimators. The following assumptions were made:

1. As all the CDF estimators considered here are nearly identical to their PDF counterparts, their comparative performance can be considered to be the same as that of their respective PDF estimators. This removed the need to benchmark any of the CDF estimators.
2. Furthermore, as the PDF and CDF estimators for the characteristic function and moments techniques are identical (as they both employ Fourier series and only differ in the way that the series coefficients are calculated), they are assumed to have identical computational requirements.

Three benchmark implementations were therefore constructed: one implementing the GMM PDF estimator, one implementing the Parzen PDF estimator and one implementing the characteristic function technique PDF estimator (which also represented the moments technique). *Each benchmark application measured the time required to evaluate a PDF estimate at a fixed number of positions in the sample space.*

All benchmarks were written in the C programming language and the execution time was then measured on a number of different hardware platforms. The tests were executed on the Linux operating system and the programs were compiled using the GNU C Compiler at its highest optimisation level (level 6).

In order to determine the influence that selectable parameters had on the computational requirements of each estimator, all the benchmarks were repeated over a range of parameters. The parameter which was varied, as well as the range over which it was varied, being again specific to each estimator:

GMM The number of mixtures was varied in the range [5, 100].

Parzen The number of training data samples was varied in the range [5, 100].

Characteristic function / moments The number of Fourier series coefficients was varied in the range [5, 100].

Note that the parameter that was varied in the case of the Parzen estimator is not the same one that was varied in the experiments involving the approximation error.

4.4.1 PDF Estimators

Figure 4.12 contains graphs of the normalised execution times of the different estimators plotted against their parameter values. These results were obtained by executing the benchmarks on an Intel Pentium III 700 MHz computer and normalising it so that the execution time of the characteristic function and moment estimators are equal to 1 at its highest parameter value.

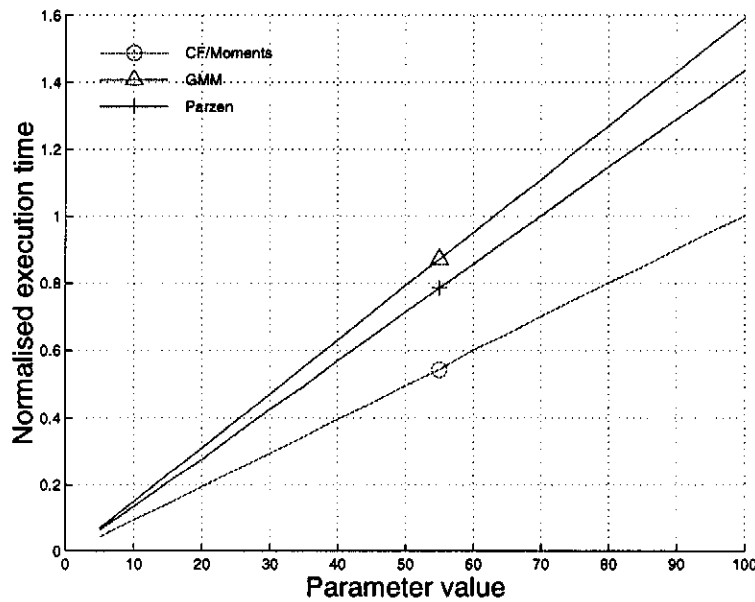


Figure 4.12: Comparison of normalised execution times of PDF estimators: Pentium III 700 MHz.

On their own, these execution times do not provide an adequate basis for comparison between the estimators, as the different parameters are completely unrelated to each other and cannot be compared. It is, however, possible to use this information to compare the speed/accuracy trade-offs of some of the techniques, thereby providing us with a practically relevant comparison.

In the case of the GMM, characteristic function and moments techniques, the same parameters that were varied during the accuracy trials were also varied during the computational requirement trials. This allows a direct comparison between the accuracy of these estimators and their computational requirements: by mapping between the parameter value and the computational requirements and between the parameter value and the estimation error, the relationship between the estimation error and computational require-

ments are obtained. Furthermore, as the mapping between the parameter value and the computational requirements (represented by the execution time) are linear, from Figure 4.12, it is easily calculated from the gradients of the graphs.

The same benchmarks were also repeated on two other hardware platforms (an Intel Pentium II 233 MHz and an Intel Pentium IV 1700 MHz), and the results again normalised. The gradients of the graphs of the computation time vs parameter values for each estimator / hardware platform are tabulated in Table 4.6. Combining these values with the results of the previous section allows us to generate graphs of the mean approximation error (in terms of the Kullback-Leibler divergence) against the computational requirements (in terms of the normalised execution time). Figure 4.13 shows these graphs comparing the GMM and the characteristic function and moments techniques, for estimates that were derived from 100 samples as well as for those derived from 1000 samples. The mean estimation error averaged over all PDFs were used as an indication of the approximation error.

	Pentium II 233	Pentium III 700	Pentium IV 1700
Moments / CF	1.000	1.000	1.000
GMM	1.558	1.585	4.505
Parzen	1.373	1.430	3.441

Table 4.6: Gradient (x100) characterising the relationship between the computation time and the parameter value.

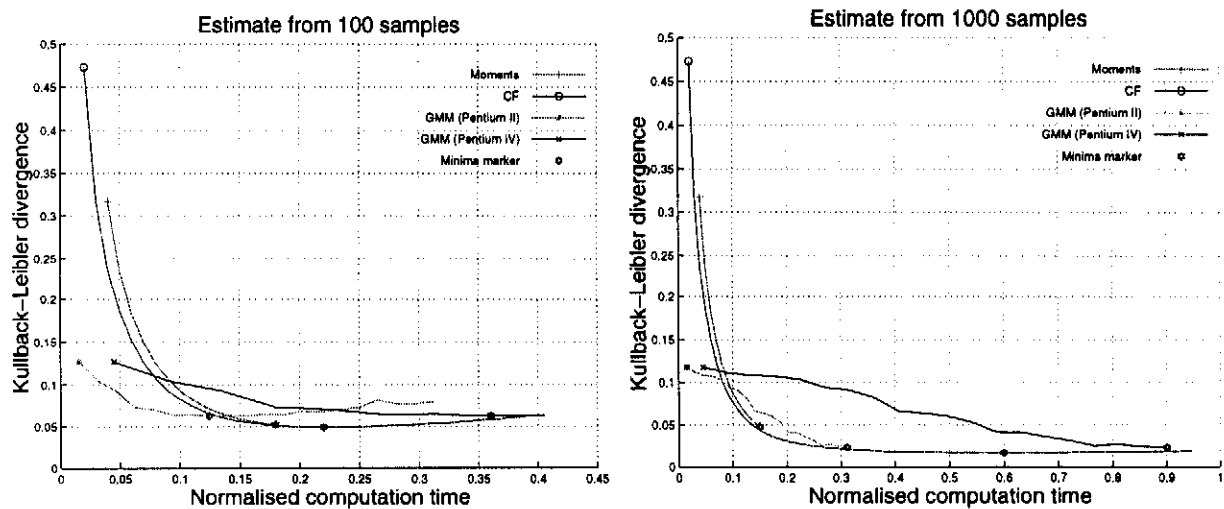


Figure 4.13: PDF estimators: estimation error against computational requirements (Parzen is excluded as its computational requirements are too high).

From the graph we observe that the higher accuracy of the characteristic function and moments technique estimators come at a price as they require more computational resources. Note, however, that the graph corresponding to the GMM does not intersect the other graphs close to the optimal working points of the other estimators. Consequently, the new techniques allow us to improve on the estimates produced by the GMM in applications where the computational requirements allow it.

The optimal choice of estimator is also highly dependent on the hardware platform and the number of samples from which the estimate is obtained (with the difference between the GMM and the other estimators being most pronounced on the Pentium IV using 1000 samples). The locations at which the graphs representing the GMM intersect with those representing the characteristic function and moments techniques corresponded to a number of mixtures between 2 and 5 and to between 8 and 13 frequency components.

Results corresponding to the Parzen estimator are excluded from the graphs as its performance is not comparable to that of the other estimators over the range of values that was considered. As this estimator uses all the sample data directly to produce an estimate, the computational resources it requires is proportional to the size of the training set. Consequently, it requires a significantly larger amount of computing power than the other estimators when obtaining estimates from large sample sets. For the case of 100 samples, the Parzen estimator requires between 5 and 10 times (dependent on the hardware architecture) the computing power that the characteristic function and moments techniques require. It is therefore recommended to only use the Parzen estimator in applications where the sample size is not larger than the number of parameters required by a parametric estimator to produce an adequate estimate.

4.4.2 CDF Estimators

Figure 4.14 contains graphs comparing the estimation error against the computational requirements for different CDF estimators. The mean estimation error averaged over all PDFs was again used as a measure of the approximation error and the normalised computing time used to quantify the computational requirements.

From these results we conclude that from a performance perspective that is quantified wholly in terms of the approximation accuracy and computational requirements, the GMM CDF estimator outperforms all the other estimators. In order for the characteristic function technique to show the same accuracy as the GMM requires a large increase in the computation time. This degradation in performance, when compared to the results

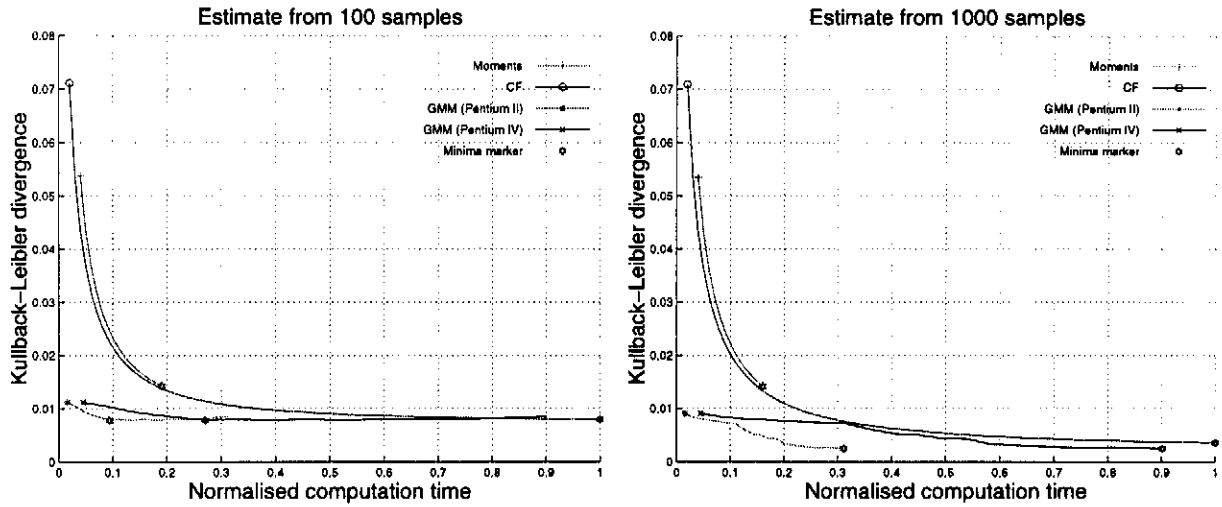


Figure 4.14: CDF estimators: estimation error against computational requirements (Parzen is excluded as its computational requirements are too high).

obtained for the PDF estimators, is again explained by the choice of frequency domain windowing function that was employed in the characteristic function and moments techniques. Note, however, that the characteristic function and moments techniques still hold some other advantages above the GMM in terms of their simplified training procedure.

When used as a CDF estimator, the Parzen estimator suffers from the same disadvantages that was encountered with it as a PDF estimator. As its performance is also proportional to the size of the sample set, it should only be selected above the GMM in applications where the sample set is smaller than the number of mixtures that is required to provide an estimate of adequate accuracy.

4.5 Training requirements

Ease of training is an important consideration in applications where the density function often changes during the operation of the system, requiring the estimator to adapt to the new density function. In these situations, the training requirements of the estimators should also be given some consideration (along with accuracy and computational requirements). Although we do not wish to compare the training algorithms of the different techniques in detail (as the selection of a training algorithm is dependent on the specific application of the technique), we briefly present some general considerations to be kept in mind when training each of the estimators.

Parzen Of all the estimators considered so far, the Parzen estimator requires the least amount of training. Although a Parzen estimate can be obtained without any training by simply using a predefined kernel function, this could produce poor results (in terms of approximation error). This was seen in situations where estimates were obtained using parameter values that were not close to the optimal working points. In order to improve this situation, it is recommended that the kernel employed by the estimator be matched to the dataset by considering the expected distance between data samples and ensuring that the kernel width is compatible with this value. Doing this need not be a complex operation and sensible values can be obtained by only considering the standard deviation and size of the sample set. Even if this is done, the Parzen estimator still requires the least amount of training.

Characteristic function technique The estimator based on the characteristic function technique allows the parameters (Fourier series coefficients) to be directly trained from the sample data using the expressions in Equation 2.79 and Equation 3.53. Both operations involve the evaluation of a number of complex exponential functions, defined over the sample set. The complexity of these operations are linear with respect to the sample set size and the number of parameters and training algorithms provide a closed-form solution to the parameter values that does not require any iterative training procedure. Furthermore, the training procedure is deterministic and does not depend on the selection of initial conditions.

Moments technique Training the estimator using the moments technique requires a two-step approach: first the moments are estimated from the sample data, after which the Fourier series coefficients are estimated using the values of the moments. Care should be taken to ensure that the expression for the characteristic function, which is approximated as a Taylor series with coefficients that is expressed in terms of the values of the moments, does not diverge.

GMM A popular way of training GMMs is to use the Expectation-Maximisation algorithm [3, p.65]. This is an iterative training algorithm that is prone to complications involving singularities and convergence towards local optima. Knowing in advance how many iterations would be required to produce adequate results is also difficult due to its sensitivity towards initial conditions. This makes it difficult to design a general training procedure for use in practice: if the training is prematurely terminated it results in an increased approximation error and variance, while leaving the

procedure until the estimate attains sufficient accuracy might violate some real-time constraint.

The GMM training algorithm does not perform particularly well when the size of the sample set is not much larger than the number of parameters (by at least an order of magnitude). If this is not ensured, singularities can occur if a mixture component represents only a few isolated samples unless proper care is taken (usually in the form of some heuristic rules).

4.6 Application to speaker verification

An experiment was conducted to evaluate the performance of one of the estimators on data obtained from a speaker verification application. This application either accepts or rejects the claimed identity of a person that presents speech to the system. It functions by calculating a score (in the form of a likelihood) from a sequence of feature vectors extracted from the speech and a model corresponding to the claimed identity of the person.

It is expected that scores corresponding to correct claimed identities would generally be concentrated at different values than those corresponding to impostors. Two gaussian mixture models (GMMs) are then trained to represent class-conditional density functions: one corresponding to known speakers and one corresponding to impostors. When presented with speech from an unknown speaker claiming some identity, the score is first calculated and then classified as belonging to either the impostors or the known speakers. The experiment involved replacing the GMM by the PDF estimator based on the characteristic function (CF) and then comparing the performance of the new classifier with the one employing the GMM.

The input data consisted of 60000 data samples, corresponding to scores, that were labelled as either belonging to impostors or to speakers with correctly claimed identities. This dataset was divided into a training set (consisting of 6000 samples) and a test set (consisting of the remaining 54000 samples). From the labelled training set, two models were trained, one corresponding to the impostors and one to the known speakers. Estimates of the class-conditional PDFs obtained from the PDF estimator using the CF technique is shown in Figure 4.15. All the data in the test set were then classified using these models (according to Equation 1.6) as either belonging to the impostors or the known speakers.

In order to compare the performance of the classifier using the new estimator to the one using the GMM, the McNemar test was employed: this test allows one to compare whether

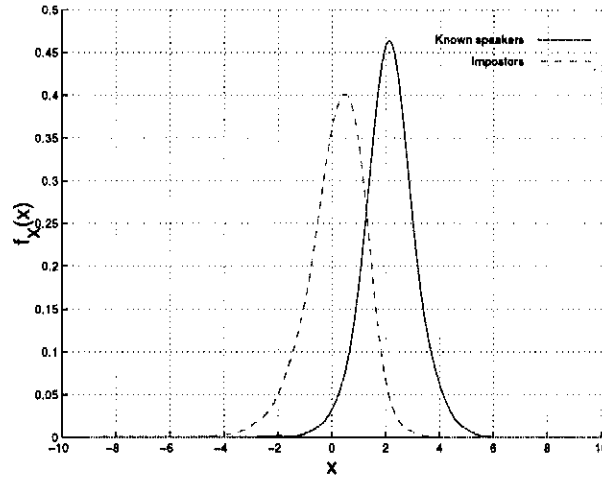


Figure 4.15: Typical class-conditional PDFs of scores corresponding to impostors and known speakers.

differences in accuracy between two classifiers are significant or can simply be ascribed to chance (as neither classifier is perfect and therefore bound to make random errors). It only considers results where the two classifiers differ in correctness (with the one providing a correct verdict and the other one an incorrect verdict). If both estimators are equally accurate, it is expected that the probability of estimator A being correct and estimator B being wrong would be the same as the probability of B being correct and A being wrong. This is the same as saying that the number of results where A is correct and B is wrong (x) is binomially distributed [2]:

$$P(x) = \sum_{k=0}^N \binom{N}{k} p^k (1-p)^{N-k}, \quad (4.3)$$

with N (the number of trials) being equal to the total number of results where A and B differ and $p = 0.5$. This allows us to construct a two-tailed test against which we can test the null hypothesis stating that the two classifiers are equally accurate.

Values for x and N were calculated over a number of trials, corresponding to different parameter values characterising the estimators, and the McNemar test (with a significance level of 5%) was performed to determine under which circumstances one estimator outperformed another one. The results of these trials are summarised in Table 4.7.

It is seen that, when using 8 frequency components or more, the CF estimator shows little difference in performance from the GMM. The fact that it is easier to train and exhibits

CF: Number of frequency components	GMM: Number of Mixtures									
	1	2	3	4	5	6	7	8	9	10
2	E	E	E	E	E	E	E	E	E	E
4	E	E	E	E	E	E	E	E	E	E
6	x	x	E	x	E	E	E	x	E	E
8	x	x	x	x	x	x	x	x	x	x
10	x	x	x	x	x	E	E	E	x	x
12	N	x	x	x	x	E	x	x	x	x
14	N	x	x	x	x	x	x	x	x	x
16	N	x	x	x	x	x	x	x	x	x
18	N	x	x	x	x	x	x	x	x	x
20	N	x	x	x	E	x	x	x	x	x

E: GMM better N: CF better x: Equal

Table 4.7: Comparison between GMM and CF technique in a speaker verification application.

less variance than the GMM (from earlier experiments) makes it a suitable replacement for the GMM in this application.

4.7 Conclusions

We will now present a comparison between the characteristics of the estimators, based on a number of practically relevant criteria, from information which was obtained in the previous sections. This serves as a summary of this chapter and may be consulted when selecting an estimator for a specific application.

The comparison is presented in Table 4.8, with features of the estimators in the rows and the different estimators in the columns. All features are positive (advantageous) and a mark in a cell indicates that the corresponding estimator possesses a certain feature (when compared to the other estimators). Characteristic features, which are not associated with at least one of the GMM or the Parzen estimators but which are associated with one of other estimators, are indicated in bold. This indicates the features upon which the new estimators improve on at least one of the old estimators. The following features are tabulated:

Easy to train An unsupervised training algorithm exists that is not overly complicated, in terms of the ease of implementation, when compared to the algorithm that evaluates the PDF or CDF. The training should also be able to proceed in an unsupervised fashion without requiring external intervention and should provide near optimal results over a range of input conditions (density functions).

Compact data representation The estimator provides a more compact representation of the sample data than the raw data does after the training phase. This would imply that the estimator is of a parametric nature.

Accurate PDF estimate The estimator produces accurate PDF estimates.

Accurate CDF estimate The estimator produces accurate CDF estimates.

Low variance PDF estimate The PDF estimates exhibit low variance.

Low variance CDF estimate The CDF estimates exhibit low variance.

Low computational requirements The estimator requires low computational overhead to produce an accurate estimate.

Moments only The estimator is capable of producing an estimate from the values of a number of moments only.

Easy to select parameters It is possible to select values of parameters that provide acceptable results over a wide range of operating conditions. This also implies that the estimator is not overly sensitive to the values of the parameters.

Resistant to PDF over-fitting The estimator is resistant to over-fitting when estimating a PDF.

Resistant to CDF over-fitting The estimator is resistant to over-fitting when estimating a CDF.

Suitable for small datasets The estimator is suitable for use on datasets containing less than 100 samples.

Suitable for medium datasets The estimator is suitable for use on datasets containing between 100 and 1000 data samples.

Suitable for large datasets The estimator is suitable for use on datasets containing more than 1000 data samples.

	GMM	Parzen	CF	Moments
Easy to train		✓	✓	
Compact data representation	✓		✓	✓
Accurate PDF estimate	✓	✓	✓	✓
Accurate CDF estimate	✓	✓	✓	✓
Low variance PDF estimate		✓	✓	✓
Low variance CDF estimate	✓	✓	✓	✓
Low computational requirements	✓		✓	✓
Easy to select parameters		✓	✓	✓
Moments only				✓
Resistant to PDF over-fitting	✓			✓
Resistant to CDF over-fitting	✓	✓	✓	✓
Suitable for small datasets		✓	✓	✓
Suitable for medium datasets	✓		✓	✓
Suitable for large datasets	✓		✓	✓

Table 4.8: Feature matrix for all the estimators.

The experiments attempted to provide an objective view of the capabilities and strong and weak points of each estimator. It was shown that no single estimator is capable of performing all tasks equally well and some are simply suited better to a specific task than any other estimator. Nonetheless, from the table it is evident that the new estimators are able to combine positive features from both the Parzen estimator and the GMM. This is heartening as it allows it to be used in situations where neither one of the established estimators would provide adequate performance, either in terms of accuracy, computational requirements or training complexity.

We therefore conclude that the theory presented in the previous chapters are correct and present theoretical and practical contributions to the field of density function estimation.

Chapter 5

Conclusions and recommendations

5.1 Conclusions

This work presents contributions to the field of density function estimation by considering a novel approach to density function estimation from the following perspectives:

Theoretical study The estimation of density functions (PDFs and CDFs) using only moments or the characteristic function was shown to be feasible in theory. The problem of density function estimation was considered from a frequency domain perspective which produced positive results. Expressions for density function estimates, with desirable characteristics, were obtained entirely in terms of a finite number of moments (Section 2.4.2 and Section 3.4.1) and in terms of a characteristic function estimate from sample data (Section 2.5.2 and Section 3.5.1). Both types of estimators produced an estimate that represented a smoothed version of the actual density function (to which the moments or characteristic function corresponded), with the exact amount of smoothing being controlled by the choice of a windowing function. Even though the level of detail that the moments-based estimator is able to attain is limited by the impracticalities of estimating high order moments from sample data, it was still found capable of providing results comparable to existing estimators in terms of accuracy. The estimators based on the characteristic function estimate was shown to be identical to the Parzen estimator, although the derivation was done from a frequency-domain perspective. This estimator also overcame the limits associated with the accuracy of the moments-based technique, thereby presenting a more practically feasible estimator.

Practical estimators A number of ways were considered in which the theory can be applied to the creation of new practical estimators that employ moments or the characteristic function. Four new parametric techniques that employ a Fourier series were developed, allowing the PDF and CDF to be estimated from both moments and sample data (using the characteristic function estimator). All these techniques represented estimators that are easy to train from sample data and have the ability to approximate arbitrary density functions. As they are parametric techniques, they also provide a compact representation of the sample data and have the ability to generalise over the sample set. The amount of smoothing employed by each estimator is controlled by the selection of a windowing (or kernel) function. Furthermore, the techniques based on the characteristic function estimator represent a parametric approximation to the Parzen estimator, thereby eliminating one of its greatest disadvantages (high computational requirements due to non-parametric nature) while retaining its major advantages (the ability to estimate arbitrary PDFs using a closed-form estimator that does not require an iterative training procedure).

Experimental validation The validity of the theory as well as the new estimation techniques were established by comparing the performance of the new estimators to two established ones, one representing parametric estimators (the Gaussian mixture model) and the other representing non-parametric estimators (the Parzen estimator). The comparison was done in terms of accuracy, computational requirements and training requirements as well as other factors such as ease of implementation. It was found that the new estimators exhibited performance that compared favourably with that of the established ones. Some of the new estimators also outperformed the established estimators in terms of accuracy, computational requirements or ease of use under certain conditions, thereby showing that there are applications in which they represent better choices than the established estimators that were considered. From this it was concluded that the theory developed during this work is indeed correct and that the new estimators can be used in practice to complement existing estimators.

By comparing the above to the statement of our research objectives (Section 1.4) it is seen that all the objectives have been met and some even superseded (due to the successful application of the theory and the positive experimental results).

5.2 Recommendations

As a lot of the work presented in this thesis is novel, there is potential for further research along similar paths. It is recommended that research be continued on one of the following:

1. The basis functions from which the estimates employing the characteristic function is constructed, is determined by the characteristic function estimator presented in Equation 2.71. By experimenting with different estimators, different basis functions may be employed. The objective would be to find basis functions more suitable to PDF and CDF estimation than the sinusoidal ones employed by the current estimator.
2. All the estimators presented in this work featured a frequency domain windowing function with finite support, $\Theta(\omega_u, \omega)$. This function determined the degree of smoothing that the estimator applied and also allowed the estimate to be represented using a parametric representation (as it limited the free parameters to a finite number). By considering alternative windowing functions than those presented in the previous chapters, estimates which are more optimal in terms of the estimation error or computational requirements may be found.
3. Extending this work to multivariate random variables would allow it to be applied to a much wider range of problems. Due to the “curse of dimensionality”, this generalisation would not necessarily be a trivial exercise. In order to construct practical techniques for use in higher dimensions, it should be attempted to construct techniques with complexity that is linear in the number of dimensions (instead of the exponential relationship predicted by the “curse”). It is expected that work in this regard would concentrate on the selection of suitable windowing functions and the normalisation of the sample data, as this could allow the estimate to be compactly represented in the frequency domain, thereby breaking the “curse”.

Bibliography

- [1] J. R. Deller Jr, J. G. Proakis, and J. H. Hansen, *Discrete-Time Processing of Speech Signals*. Prentice Hall, 1993.
- [2] P. Z. Peebles Jr, *Probability, Random Variables and Random Signal Principles*. McGraw-Hill, third ed., 1993.
- [3] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.
- [4] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, vol. 1. Charles Griffin & Company Limited, third ed., 1947.
- [5] F. G. Stremler, *Introduction to Communication Systems*. Addison-Wesley Publishing Company, third ed., 1990.
- [6] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.
- [7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Morgan Kaufmann, second ed., 1990.
- [8] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Prentice Hall, 1982.
- [9] M. Golfarelli, M. D., and M. D., "On the error-reject trade-off in biometric verification systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 786–796, July 1997.
- [10] W. J. Conover, *Practical Nonparametric Statistics*. John Wiley and Sons, 1980.
- [11] H. Neave and P. Worthington, *Distribution-Free Tests*. Unwin Hyman Ltd, 1988.

- [12] E. L. Lehmann, *Testing Statistical Hypothesis*. John Wiley and Sons, 1959.
- [13] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, pp. 47–60, Nov. 1996.
- [14] A. Bors and I. Pitas, "Robust estimation for radial basis functions," in *IEEE Workshop on Neural Networks for Signal Processing*, pp. 105–114, 1994.
- [15] S. Haykin, *Neural Networks: A Comprehensive Foundation*. second ed., 1999.
- [16] N. Vlassis and A. Likas, "A kurtosis-based dynamic approach to Gaussian mixture modeling," *IEEE Transactions on Systems, Man and Cybernetics, Part A*, vol. 29, pp. 393–399, July 1999.
- [17] A. Pagès-Zamora and M. A. Lagunas, "Joint probability density function estimation by spectral estimate methods," in *ICASSP*, vol. 5, pp. 2936–2939, 1996.
- [18] J.-F. Bercher and C. Vignat, "Estimating the entropy of a signal with applications," *IEEE Transactions on Signal Processing*, vol. 48, pp. 1687–1694, June 2000.
- [19] S. Kay, "Model-based probability density function estimation," *IEEE Signal Processing Letters*, vol. 5, pp. 318–320, Dec. 1998.
- [20] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*. Prentice Hall, third ed., 1996.
- [21] B. Silverman, "Density estimation for statistics and data analysis," *Monographs on Statistics and Applied Probability*, vol. 26, 1986.
- [22] M. Vannucci, "Nonparametric density estimation using wavelets," Tech. Rep. 95-26, Duke University, USA, 1995.
- [23] M. Vannucci and B. Vidakovic, "Preventing the Dirac disaster: wavelets based density estimation," Tech. Rep. 95-27, Duke University, USA, 1995.
- [24] B. G. Lindsay, R. S. Pilla, and P. Basak, "Moment-based approximations of distributions using mixtures: theory and applications," *Annals of the Institute of Statistical Mathematics*, vol. 52, pp. 215–230, June 2000.
- [25] J. Abate and W. Whitt, "Numerical inversion of probability generating functions," *Operations Research Letters*, pp. 245–251, 1992.

- [26] J. Abate and W. Whitt, "The fourier-series method for inverting transforms of probability distributions," *Queueing Systems*, pp. 5–88, 1992.
- [27] V. Witkovský, "On the exact computation of the density and of the quantiles of linear combinations of t and F random variables," *Journal of Statistical Planning and Inference*, pp. 1–13, 2001.
- [28] G. Esterhuizen and J. A. Du Preez, "Efficient frequency domain univariate Parzen approximation and applications," in *IEEE Africon*, pp. 287–292, 2002.
- [29] R. L. Finney and G. B. Thomas Jr, *Calculus*. Addison-Wesley Publishing Company, 1990.
- [30] V. Barnett, *Comparitive Statistical Inference*. John Wiley and Sons, 1973.
- [31] A. Jeffrey, *Complex Analysis and Applications*. CRC Press, 1992.
- [32] G. B. Folland, *Fourier Analysis and its Applications*. Wadsworth and Brooks/Cole Mathematical Series, 1992.

Appendix A

A review of the Fourier transform

A short overview of the Fourier transform is now presented. More detailed accounts are found in Folland [32], Stremler [5] and Proakis and Manolakis [20].

Let $f(x)$ be real- or complex-valued function defined on the real line. Its Fourier transform, denoted by $F(\omega)$, is a complex-valued function that is defined by

$$F(\omega) = \int_{-\infty}^{\infty} f(x)e^{-j\omega x}dx, \quad (\text{A.1})$$

where ω is a real scalar representing radial frequency. The Dirichlet conditions provide a set of sufficient conditions that guarantee existence of the Fourier transform:

1. $f(x)$ has a finite number of discontinuities in any finite interval.
2. $f(x)$ has a finite number of minima and maxima in any finite interval.
3. $f(x)$ is absolutely integrable:

$$\int_{-\infty}^{\infty} |f(x)|dx < \infty. \quad (\text{A.2})$$

If these conditions are met, $f(x)$ can be reconstructed from its Fourier transform using the inverse Fourier transform that is defined by

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)e^{j\omega x}d\omega. \quad (\text{A.3})$$

The Fourier transform and inverse Fourier transform operations are denoted using a

cursive letter F ,

$$\begin{aligned} F(\omega) &= \mathcal{F}\{f(x)\}, \\ f(x) &= \mathcal{F}^{-1}\{F(\omega)\} \end{aligned} \tag{A.4}$$

and $f(x)$ and $F(\omega)$ constitute a Fourier transform pair, which is denoted using a double arrow,

$$f(x) \leftrightarrow F(\omega). \tag{A.5}$$

It is customary to refer to $F(\omega)$ as the frequency domain representation of $f(x)$ and to $f(x)$ as the time or spatial domain representation of $F(\omega)$ (depending on the definition of x).

A number of Fourier transform properties allow operations in the one domain to be related to operations in the other. Table A.1 contains a list of some of the properties that are often encountered:

Operation	$f(x)$	$F(\omega)$
Linearity	$k_1 f_1(x) + k_2 f_2(x)$	$k_1 F_1(\omega) + k_2 F_2(\omega)$
Translation	$f(x - x_0)$ $f(x) e^{j\omega_0 x}$	$F(\omega) e^{-j\omega x_0}$ $F(\omega - \omega_0)$
Conjugation	$f^*(x)$ $f^*(-x)$	$F^*(-\omega)$ $F^*(\omega)$
Duality	$F(x)$	$2\pi f(-\omega)$
Convolution	$\int_{-\infty}^{\infty} f_1(\lambda) f_2(x - \lambda) d\lambda$ $f_1(x) f_2(x)$	$F_1(\omega) F_2(\omega)$ $\frac{1}{2\pi} \int_{-\infty}^{\infty} F_1(\xi) F_2(\omega - \xi) d\xi$
Integration	$\int_{-\infty}^x f(\lambda) d\lambda$ $\pi f(0) \delta(x) - \frac{f(x)}{jx}$	$\pi F(0) \delta(\omega) + \frac{F(\omega)}{j\omega}$ $\int_{-\infty}^{\omega} F(\xi) d\xi$
Differentiation	$\frac{d^n}{dx^n} f(x)$ $(-jt)^n f(x)$	$(j\omega)^n F(\omega)$ $\frac{d^n}{d\omega^n} F(\omega)$

Table A.1: Selected Fourier transform properties.

Appendix B

Summary of algorithms

A summary of the algorithms describing the following four practical estimators are now presented:

- Table B.1: PDF from moments (Section 2.4.5).
- Table B.2: PDF from sample data (Section 2.5.3).
- Table B.3: CDF from moments (Section 3.4.3).
- Table B.4: PDF from sample data (Section 3.5.2).

Inputs:

Values of the first N_m moments: $S = \{m_0, m_1, \dots, m_{N_m-1}\}$.

Known parameters:

$\Theta(\omega_u, \omega)$: real windowing function with $\Theta(\omega_u, \omega) = 0$ if $|\omega| > \omega_u$.

x_Δ : width of interval over which estimate attains non-zero values.

K : number of parameters.

Training procedure:

Compute the mean: $\mu_x = m_1$.

Compute the parameters:

$$\gamma_k = \frac{1}{x_\Delta} \Theta\left(\frac{2\pi[K-1]}{x_\Delta}, \frac{2\pi k}{x_\Delta}\right) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} \left(\frac{2\pi k}{x_\Delta}\right)^n \right\}, \quad k = 0, 1, \dots, K-1.$$

Estimate:

$$\hat{f}_X(x) = \begin{cases} \left| \sum_{k=-(K-1)}^{K-1} \gamma_k e^{j \frac{2\pi k}{x_\Delta} x} \right| & ; \quad |x - \mu_x| \leq \frac{x_\Delta}{2} \\ 0 & ; \quad |x - \mu_x| > \frac{x_\Delta}{2}. \end{cases}$$

Table B.1: PDF estimate from moments using Fourier series.

Inputs:

Values of N_x samples: $S = \{x_0, x_1, \dots, x_{N_x-1}\}$.

Known parameters:

$\Theta(\omega_u, \omega)$: real windowing function with $\Theta(\omega_u, \omega) = 0$ if $|\omega| > \omega_u$.

x_Δ : width of interval over which estimate attains non-zero values.

K : number of parameters.

Training procedure:

Compute the sample mean: $\mu_x = \frac{1}{N_x} \sum_{i=0}^{N_x-1} x_i$.

Compute the parameters:

$$\gamma_k = \frac{1}{x_\Delta N_x} \Theta\left(\frac{2\pi[K-1]}{x_\Delta}, \frac{2\pi k}{x_\Delta}\right) \left\{ \sum_{n=0}^{N_x-1} e^{-j\frac{2\pi k}{x_\Delta} x_i} \right\}, \quad k = 0, 1, \dots, K-1.$$

Estimate:

$$\hat{f}_X(x) = \begin{cases} \left| \sum_{k=-(K-1)}^{K-1} \gamma_k e^{j\frac{2\pi k}{x_\Delta} x} \right| & ; \quad |x - \mu_x| \leq \frac{x_\Delta}{2} \\ 0 & ; \quad |x - \mu_x| > \frac{x_\Delta}{2}. \end{cases}$$

Table B.2: PDF estimate from samples using Fourier series.

Inputs:

Values of the first N_m moments: $S = \{m_0, m_1, \dots, m_{N_m-1}\}$.

Known parameters:

$\Theta(\omega_u, \omega)$: real windowing function with $\Theta(\omega_u, \omega) = 0$ if $|\omega| > \omega_u$.

x_Δ : width of interval over which estimate attains non-terminal values.

K : number of parameters.

Training procedure:

Compute the mean: $\mu_x = m_1$.

Compute the parameters:

$$\gamma_k = \begin{cases} \frac{1}{2\pi jk} \Theta\left(\frac{2\pi[K-1]}{x_\Delta}, \frac{2\pi k}{x_\Delta}\right) \left\{ \sum_{n=0}^{N_m-1} \frac{m_n}{j^n n!} \left(\frac{2\pi k}{x_\Delta}\right)^n \right\} & ; \quad k = 1, 2, \dots, K-1 \\ \frac{1}{2} - \sum_{\substack{n=-(K-1) \\ n \neq 0}}^{K-1} (-1)^n \gamma_n & ; \quad k = 0. \end{cases}$$

Estimate:

$$\hat{F}_X(x) = \begin{cases} 0 & ; \quad x - \mu_x < \frac{-x_\Delta}{2} \\ \sum_{k=-(K-1)}^{K-1} \gamma_k e^{j \frac{2\pi k}{x_\Delta} x} + \frac{x}{x_\Delta} & ; \quad |x - \mu_x| \leq \frac{x_\Delta}{2} \\ 1 & ; \quad x - \mu_x > \frac{x_\Delta}{2}. \end{cases}$$

Table B.3: CDF estimate from moments using Fourier series.

Inputs:

Values of N_x samples: $S = \{x_0, x_1, \dots, x_{N_x-1}\}$.

Known parameters:

$\Theta(\omega_u, \omega)$: real windowing function with $\Theta(\omega_u, \omega) = 0$ if $|\omega| > \omega_u$.

x_Δ : width of interval over which estimate attains non-zero values.

K : number of parameters.

Training procedure:

Compute the sample mean: $\mu_x = \frac{1}{N_x} \sum_{i=0}^{N_x-1} x_i$.

Compute the parameters:

$$\gamma_k = \begin{cases} \frac{1}{2\pi jk} \Theta\left(\frac{2\pi[K-1]}{x_\Delta}, \frac{2\pi k}{x_\Delta}\right) \left\{ \frac{1}{N_x} \sum_{i=0}^{N_x-1} e^{-j\frac{2\pi k}{x_\Delta} x_i} \right\} & ; \quad k = 1, 2, \dots, K-1 \\ \frac{1}{2} - \sum_{\substack{n=-(K-1) \\ n \neq 0}}^{K-1} (-1)^n \gamma_n & ; \quad k = 0. \end{cases}$$

Estimate:

$$\hat{F}_X(x) = \begin{cases} 0 & ; \quad x - \mu_x < \frac{-x_\Delta}{2} \\ \sum_{k=-(K-1)}^{K-1} \gamma_k e^{j\frac{2\pi k}{x_\Delta} x} + \frac{x}{x_\Delta} & ; \quad |x - \mu_x| \leq \frac{x_\Delta}{2} \\ 1 & ; \quad x - \mu_x > \frac{x_\Delta}{2}. \end{cases}$$

Table B.4: CDF estimate from samples using Fourier series.