

University of Mississippi

eGrove

Electronic Theses and Dissertations

Graduate School

1-1-2019

Cramer type moderate deviations for random fields and mutual information estimation for mixed-pair random variables

Aleksandr Beknazaryan

Follow this and additional works at: <https://egrove.olemiss.edu/etd>



Part of the [Mathematics Commons](#), and the [Statistics and Probability Commons](#)

Recommended Citation

Beknazaryan, Aleksandr, "Cramer type moderate deviations for random fields and mutual information estimation for mixed-pair random variables" (2019). *Electronic Theses and Dissertations*. 1737.
<https://egrove.olemiss.edu/etd/1737>

This Dissertation is brought to you for free and open access by the Graduate School at eGrove. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of eGrove. For more information, please contact egrove@olemiss.edu.

CRAMÉR TYPE MODERATE DEVIATIONS FOR RANDOM FIELDS AND MUTUAL
INFORMATION ESTIMATION FOR MIXED-PAIR RANDOM VARIABLES
DISSERTATION

A Dissertation
presented in partial fulfillment of requirements
for the degree of Doctor of Philosophy
in the Department of Mathematics
The University of Mississippi

by

ALEKSANDR BEKNAZARYAN

August 2019

Copyright Aleksandr Beknazaryan 2019
ALL RIGHTS RESERVED

ABSTRACT

In this dissertation we first study Cramér type moderate deviation for partial sums of random fields by applying the conjugate method. In 1938 Cramér published his results on large deviations of sums of i.i.d. random variables after which a lot of research has been done on establishing Cramér type moderate and large deviation theorems for different types of random variables and for various statistics. In particular, results have been obtained for independent non-identically distributed random variables, for the sum of independent random variables with p -th moment ($p > 2$) and for different types of dependent random variables. In this work we establish Cramér type exact moderate deviation theorem for random fields. We then show that obtained results are applicable to the partial sums of linear random fields with short or long memory and to nonparametric regression with random field errors. We also show that the result for linear random fields can be applied to calculate the tail probability of partial sums of various models such as the autoregressive fractionally integrated moving average FARIMA(p, β, q) processes. The results can also be used to approximate the risk measures such as quantiles and tail conditional expectations of time series or spacial random fields.

We also study the mutual information estimation for mixed-pair random variables. One random variable is discrete and the other one is continuous. We develop a kernel method to estimate the mutual information between two random variables. The estimates enjoy a central limit theorem under some regular conditions on the distributions. The theoretical results are demonstrated by simulation study.

DEDICATION

I dedicate my dissertation work to my loving parents and to my sister Angelika who have always loved me unconditionally, encouraged me to pursue my dreams and on whose constant love and prayer I have relied throughout my journey at the University of Mississippi. I am truly thankful for having you in my life.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere appreciation and deepest gratitude to my advisor Dr. Hailin Sang for his incredible patience, wisdom, helpful suggestions, expert guidance and encouragement throughout my study and research. I have acquired valuable insights through his instructions not only in academic studies but also enthusiasm and rigor in life.

I am also extremely grateful to Dr. Xin Dang for her much-appreciated advice, support and thought-provoking ideas. Her important suggestions and advice are priceless.

Very special thanks to Dr. Gerard Buskes for all his support. The timeless inspiration and encouragement that I got from him helped me a lot during this journey.

I am also very thankful to Dr. Jeremy Clark for his valuable advice on my career and sincere comments and recommendations on my research.

I also would like to thank Dr. Yimin Xiao for his co-operation and generosity which set our joint research project possible as it is till the end.

I am very grateful to Dr. Zhiqu Lu for his valuable time and kindness to serve as my committee member.

My sincere thanks also goes to Dr. Talmage James Reid and Dr. Bing Wei who were always there to help me succeed in my graduate studies.

I am also very grateful to the Graduate School of the University of Mississippi for the financial support.

TABLE OF CONTENTS

ABSTRACT	ii
DEDICATION	iii
ACKNOWLEDGEMENTS	iv
List of Tables	vii
List of Figures	viii
1 INTRODUCTION	1
1.1 MODERATE AND LARGE DEVIATIONS	1
1.2 ENTROPY AND MUTUAL INFORMATION	3
1.3 ENTROPIES IN THE THEORY OF LARGE DEVIATIONS	7
1.4 OVERVIEW	11
1.4.1 Contribution of the dissertation	11
1.4.2 Dissertation Structure	11
2 CRAMÉR TYPE MODERATE DEVIATIONS FOR RANDOM FIELDS	12
2.1 INTRODUCTION	12
2.2 MAIN RESULTS	15
2.3 APPLICATIONS	32
2.3.1 Cramér type moderate deviation for linear random fields	33
2.3.2 Approximation of risk measures	39

2.3.3	Nonparametric regression	40
2.4	CONCLUSION	41
3	ON MUTUAL INFORMATION ESTIMATION FOR MIXED-PAIR RANDOM VARIABLES	43
3.1	INTRODUCTION	43
3.2	MAIN RESULTS	46
3.3	SIMULATION STUDY	53
3.4	CONCLUSION	58
	BIBLIOGRAPHY	60
	VITA	66

LIST OF TABLES

3.3.1 True value of the mutual information and the mean value of the estimates for Pareto and t-distributions.	54
3.3.2 True value of the mutual information and the mean value of the estimates for t-distributions.	57

LIST OF FIGURES

<p>3.3.1 The histograms with kernel density fits of $M = 400$ estimates. Top left: $t(3, 0, 1)$ and $t(12, 0, 1)$. Top right: $t(3, 0, 1)$ and $t(3, 2, 1)$. Bottom left: $t(3, 0, 1)$ and $t(3, 0, 3)$. Bottom right: <i>pareto</i>(1, 2) and <i>pareto</i>(1, 10).</p>	55
<p>3.3.2 The Q-Q plots of $M = 400$ estimates. Top left: $t(3, 0, 1)$ and $t(12, 0, 1)$. Top right: $t(3, 0, 1)$ and $t(3, 2, 1)$. Bottom left: $t(3, 0, 1)$ and $t(3, 0, 3)$. Bottom right: <i>pareto</i>(1, 2) and <i>pareto</i>(1, 10).</p>	56
<p>3.3.3 The histograms and Q-Q plots of $M = 200$ estimates. Left: $t_5(0, I)$ and $t_{25}(0, I)$. Right: $t_5(0, I)$ and $t_5(0, 3)$.</p>	58

1 INTRODUCTION

1.1 MODERATE AND LARGE DEVIATIONS

Let $X_1, X_2, \dots, X_n, \dots$ be a sequence of independent random variables and let $S_n = X_1 + X_2 + \dots + X_n$ be the sum of first n variables. Let $\mu_n = \mathbb{E} S_n$ and $B_n^2 = \text{var}(S_n) = \text{var}(X_1) + \text{var}(X_2) + \dots + \text{var}(X_n)$ be the expected value and the variance of S_n , respectively. The sequence $\{X_n\}$ is said to satisfy the Central Limit Theorem if for any $z_1, z_2 \in \mathbb{R}$

$$\lim_{n \rightarrow \infty} P(z_1 B_n < S_n - \mu_n < z_2 B_n) = \Phi(z_2) - \Phi(z_1),$$

where $\Phi(z)$ is the cumulative distribution function of standard normal distribution. In 1887 Chebyshev presented a wide class of conditions under which the Central Limit Theorem holds. His theorems have then been refined by Markov and got their quite complete forms in the works of Bernstein and Feller. In case z_n unlimitedly grows as $n \rightarrow \infty$ the accuracy of the approximation of $P(S_n - \mu_n > z_n B_n)$ under the conditions of Central Limit Theorem may be quite small. The correction multipliers necessary for increasing the accuracy were first presented in the Cramér's theorem on large deviations ([9]). Let us consider an example that shows how the typical result from this field looks like. Suppose that the random variables $X_1, X_2, \dots, X_n, \dots$ all have expected value equal to zero and variance equal to 1. We then have that $\mu_n = 0$ and $B_n = \sqrt{n}$. Then the probability of

$$S_n \geq z_n \sqrt{n}$$

is equal to $1 - F_n(z_n)$, where $F_n(z)$ is the cumulative distribution function of S_n/\sqrt{n} . For fixed $z_n = z$ we have that

$$\lim_{n \rightarrow \infty} 1 - F_n(z) \rightarrow 1 - \Phi(z).$$

If z_n depends on n and $z_n \rightarrow \infty$ as $n \rightarrow \infty$ then we have that $1 - F_n(z_n) \rightarrow 0$ and $1 - \Phi(z_n) \rightarrow 0$ so that the above formula becomes useless. Thus, in this case we need estimations for the relative accuracy of the approximation, that is, we need to estimate the ratio of $1 - F_n(z_n)$ to $1 - \Phi(z_n)$. In particular, the following natural question arises: under which condition we have that

$$\frac{1 - F_n(z_n)}{1 - \Phi(z_n)} \rightarrow 1 \tag{1.1.1}$$

as $z_n \rightarrow \infty$? This relation holds for any rate of growth of z_n only in the case when all the summands follow the normal distribution. In case the summands are not normal, that relation will hold only in the certain zones that have order not exceeding \sqrt{n} . The "narrowest" zones (those of logarithmic order) are obtained under the condition of existence of certain moments. The extension of the logarithmic range to the order n^a , $a < 1/2$, requires some additional assumptions on the moments and the coincidence of certain number of moments of X_j (that number depends on a) with the corresponding moments of normal distribution. If those assumptions are not satisfied then the expression in the left side of 2.2.15 is described in terms of the Cramér series under the condition that the random variable has moment generating function in a neighborhood of the origin. This condition has been referred to as the Cramér's condition.

Cramér's results on moderate and large deviations have been then refined and developed by [49], [50], [51], [56], [2] and others. Moderate deviation theorems were also obtained for various statistics such as U -statistics (e.g. [41]), L -statistics (e.g. [65]), M -estimators (e.g. [31]) and rank statistics (e.g. [59]). However, the exact moderate deviation for random fields under Cramér's condition has not been well studied. These motivate us to focus on

establishing exact moderate deviation for random fields under Cramér's condition in this dissertation.

1.2 ENTROPY AND MUTUAL INFORMATION

Information theory studies the quantitative laws related to the transferring, storage and processing of information. Information theory focuses on determining the average information transfer rate and solving the problem of maximizing that rate by applying appropriate coding. In order to address these questions, we should first establish a universal quantitative measure of information which should be independent of the specific physical nature of the transmitted messages. When we receive a message about certain event our knowledge of that event changes as we get some information about that event. Note that if the received message concerns a well-known event then, obviously, it does not carry any information. In contrast, if the message concerns a little-known or an unknown event then it carries a lot of information. Thus, the amount of information in a message about a certain event essentially depends on the probability of this event. This is the reason that the probabilistic approach lies in the basis of determining the measure of the amount of information. The measure of the amount of information is based on the concept of entropy. Entropy is a measure of the degree of uncertainty about the state of system (a random variable) X . What does uncertainty mean and how to measure it? Let us consider the following example. Suppose that we have two system: the first system is a die which has 6 states and the second system is a coin that has 2 states. The question is which system's state is harder to predict or, in other words, which system has more uncertainty? The natural answer is that the first system has more uncertainty which shows that the degree of uncertainty of the system depend on the number of its possible states. However, that number is not the unique characteristic of uncertainty. Let us consider two coins, say C_1 and C_2 , both having 2 possible states: tails (T) and heads (H). Suppose that $P(C_1 = T) = P(C_1 = H) = 0.5$ while $P(C_2 = T) = 0.999$ and $P(C_2 = H) = 0.001$. Obviously, the uncertainty of those two systems are different. The

first system has much more uncertainty comparing to the second system which is almost always in the state T . Thus, we see that the degree of uncertainty is also determined by the probabilities of the states of the system. Information theory suggests entropy as a measure of uncertainty. The entropy of a discrete random variable $X \in \mathbb{R}^d$ with the support set \mathcal{X} and probability mass function $p_X(x)$, abbreviated as $p(x)$, is defined to be

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x).$$

Given two discrete random variables X and Y , taking values in \mathcal{X} and \mathcal{Y} , we denote their joint probability distribution as $p_{X,Y}(x, y)$, which is abbreviated as $p(x, y)$, and the conditional probability distribution for the variable y given x as $p_{Y|X}(y|x)$, abbreviated as $p(y|x)$. The conditional entropy $H(Y|X)$ is defined as the entropy of the law $p_{Y|X}(y|x) = p(y|x)$, averaged over x :

$$\begin{aligned} H(Y|X) &= - \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\ &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x) p(y|x) \log p(y|x) \\ &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \end{aligned} \tag{1.2.1}$$

The joint entropy

$$H(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y)$$

of the pair of random variables X and Y can then be written as the entropy of X plus the conditional entropy of Y given X :

$$H(X, Y) = H(X) + H(Y|X).$$

Analogously, the (differential) entropy of a continuous random variable $X \in \mathbb{R}^d$ with probability density function $f(x)$ is defined as

$$H(X) = - \int_{\mathbb{R}^d} f(x) \log f(x) dx$$

and the (differential) conditional entropy of two continuous random variables X and Y with joint probability density function $f(x, y)$ is defined as

$$H(Y|X) = - \int_{x,y} f(x, y) \log f(y|x) dx dy.$$

The joint entropy is again defined to be $H(X, Y) = H(X) + H(Y|X)$.

When we get a message about some system the uncertainty of that system reduces. If everything is known about the system, then there is no point in sending a message. For example, if we receive a message that Paris is the capital of France, then we will not receive any information because we already knew that. But if we get data about unknown system, then we get relevant amount of information and the more uncertain the state of the system is the greater amount of information we will receive. Therefore, the amount of information is measured by a decrease in entropy.

Now suppose that we have two random variables X and Y both of which are either discrete or continuous. The (mutual) information that the variable Y contains about the variable X (and vice versa) is defined to be the amount of reduced entropy of X after observing Y . Analytically this is written as

$$I(X, Y) = H(X) - H(X|Y).$$

Using the formula of joint entropy we can rewrite the mutual information as

$$I(X, Y) = H(X, Y) - H(Y|X) - H(X|Y) = H(Y) - H(Y|X),$$

which shows that the mutual information $I(X, Y)$ measures the reduction in the uncertainty of one of the variables due to the knowledge of the other variable, and is symmetric in X and Y .

Another important concept of information theory that is closely related to the concepts of entropy and mutual information is the so called relative entropy which is also known as the Kullback-Leibler divergence. Given two probability distributions $p(x)$ and $q(x)$ over a discrete random variable X , the relative entropy is defined to be

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

As the definition suggests, the relative entropy is defined only if for all $x \in \mathcal{X}$, $q(x) = 0$ implies $p(x) = 0$ and in that case the convention $0 \log \frac{0}{0} = 0$ implies that the corresponding summand is equal to zero. Similarly, if the probability distributions are given over continuous random variable then the relative entropy is defined to be

$$D(p||q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx,$$

where $p(x)$ and $q(x)$ are the corresponding probability density functions. Relative entropy is a measure of how one probability distribution is different from a second, reference probability distribution. It is not symmetric, namely, $D(p||q) = D(q||p)$ is not necessarily true, however, $D(p||q)$ is always non-negative and it is equal to 0 if and only if the laws p and q are identical. The following formula shows the connection between the mutual information and the relative

entropy.

$$\begin{aligned}
I(X, Y) &= H(X) - H(X|Y) \\
&= - \sum_{x \in \mathcal{X}} p(x) \log p(x) + \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x|y) \\
&= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) + \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x|y) \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x|y)}{p(x)} \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\
&= D(p(x, y) || p(x)p(y)).
\end{aligned} \tag{1.2.2}$$

The estimation of mutual information for the cases when the random variables X and Y are either both discrete or both continuous have been studied by many mathematicians, e.g., [34], [63], [40]; [1], [27], [22], [23] [36], [37], [7].

However, there are only couple of results about the estimation of mutual information of the mixed pair of random variables, where the first random variable is discrete and the second one is continuous. This motivate us to focus on estimating the mutual information of the mixed pair of random variables in this dissertation.

1.3 ENTROPIES IN THE THEORY OF LARGE DEVIATIONS

In previous two parts we presented the main definitions and concepts of two quite broad fields of probability theory and statistics, namely, the theory of moderate and large deviations and the information theory, which are the main topics of this dissertation. In this part we will present the well known Sanov's theorem which can be thought of as one of the bridges that connects those two fields.

For a given sequence X_1, X_2, \dots of independent and identically distributed random variables with mean μ and variance $\sigma^2 < \infty$ let

$$L(\lambda) = \log \mathbb{E} e^{\lambda X_i}$$

be its cumulant generating function and assume that

$$L(\lambda) < \infty \quad \text{for all } \lambda \in \mathbb{R}. \quad (1.3.1)$$

If, as before, we denote by $S_n = X_1 + X_2 + \dots + X_n$ the sum of first n variables then the following theorem holds:

Theorem 1.3.1 (Cramér's theorem). *If the sequence X_1, X_2, \dots satisfies 1.3.1 then for any $x > \mu$ we have*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P} \left\{ \frac{1}{n} S_n > x \right\} = -L^*(x),$$

where L^* given by

$$L^*(x) = \sup_{\lambda \in \mathbb{R}} \{ \lambda x - L(\lambda) \}$$

is the Legendre transform of L .

Note that for the case $X_i \sim \text{Bernoulli}(p)$ we have that $L(\lambda) = \log(pe^\lambda + (1-p))$ and, therefore, for $0 < x < 1$

$$L^*(x) = x \log \frac{x}{p} + (1-x) \log \frac{1-x}{1-p}$$

which is the relative entropy of $(x, 1-x)$ with respect to $(p, 1-p)$.

Theorem 1.3.2 (Moderate deviation principle). *Under the assumptions of Theorem 1.3.2, for any sequence a_n with $\sqrt{n} \ll a_n \ll n$ we have that for all $x > 0$*

$$\lim_{n \rightarrow \infty} \frac{n}{a_n^2} \log \mathbb{P} \left\{ S_n - \mu n \geq x a_n \right\} = -\frac{x^2}{2\sigma^2}.$$

Let us now see how the definition of a function $L^*(x)$ can be generalized to cover the setting where a sequence X_1, X_2, \dots is from some metric space M and we are interested in events of the type $\{X_n \in A\}$ where $A \subset M$ is a Borel set.

Definition 1.3.2. Fix a metric space M . A function $I : M \rightarrow [0, \infty]$ is called

- a rate function if it is lower semicontinuous, which means that the level sets $\{x \in M : I(x) \leq a\}$ are closed for any $a \geq 0$;
- a good rate function if the level sets are compact for any $a \geq 0$.

Definition 1.3.3 (Large deviation principle). A sequence of random variables X_1, X_2, \dots with values in a metric space is said to satisfy a large deviation principle with

- speed $a_n \rightarrow \infty$ and
- rate function I ,

if, for all Borel sets $A \subset M$,

$$\limsup_{n \rightarrow \infty} \frac{1}{a_n} \log \mathbb{P} \left\{ X_n \in A \right\} \leq - \inf_{x \in cl A} I(x),$$

$$\liminf_{n \rightarrow \infty} \frac{1}{a_n} \log \mathbb{P} \left\{ X_n \in A \right\} \geq - \inf_{x \in int A} I(x).$$

Thus, the above Cramér's theorem basically says that $\frac{1}{n}S_n$ satisfies large deviation principle with speed n and good rate function L^* while the moderate deviation principle says that for any sequence $\sqrt{n} \ll a_n \ll n$ the random variables $\frac{S_n - \mu n}{a_n}$ satisfy a large deviation principle with speed a_n^2/n and good rate function $I(x) = \frac{x^2}{2\sigma^2}$.

Now suppose that we are given a sequence X_1, X_2, \dots of independent and identically distributed discrete random variables having finite support \mathcal{X} and we want to find the frequency of a given $x \in \mathcal{X}$ among the first n samples. Then, by applying Cramér's theorem,

for any $0 < a < 1$ and $x \in \mathcal{X}$ we can find the rate of decay of

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i = x\} \geq a\right\}.$$

However, Cramér's theorem cannot be applied in case when we want to find the frequency of more than one symbol, like

$$\left\{\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i = x\} \geq a, \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i = y\} \geq b\right\}, \quad a, b > 0, a + b < 1, x, y \in \mathcal{X}.$$

Such situations can be handled by applying the Sanov's theorem which deals not with the partial sum S_n of first n variables but with their empirical measure

$$\delta_n^X(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i = x\}$$

interpreted as a random element of the space $\mathcal{M}_1(X)$ of probability measures on \mathcal{X} endowed with the metric inherited from the embedding into $\mathbb{R}^{|\mathcal{X}|}$ given by the mapping $\mu \mapsto (\mu(x) : x \in \mathcal{X})$.

Theorem 1.3.3 (Sanov's theorem). *Assume that X_1, X_2, \dots are i.i.d. random variables taking values in a finite set \mathcal{X} and denote by $\mu \in \mathcal{M}_1(X)$ their distribution. Then the empirical measures $\delta_n^X(x)$ satisfy a large deviation principle on the metric space $\mathcal{M}_1(X)$ with speed n and good rate function J given by the relative entropies*

$$J(\nu) = D(\nu||\mu) = \sum_{x \in \mathcal{X}} \nu(x) \log \frac{\nu(x)}{\mu(x)}.$$

More details and proofs of theorems presented in this section can be found in [43]

1.4 OVERVIEW

1.4 Contribution of the dissertation

The contribution of this dissertation is as follows:

- We established exact moderate deviation for random fields under Cramér's condition.
- We applied our results to linear random fields with short or long memory, to non-parametric regression analysis as well as to approximation of the quantiles and tail conditional expectations for the spartial sums of linear random fields
- We gave an estimation for the mutual information of the mixed pair of random variables where one of the variables is discrete and the other one is continuous.
- We conducted simulation study to confirm the theoretical results.

1.4 Dissertation Structure

The structure of this dissertation is organized in the following way. In chapter 2 we obtain Cramér type moderate deviation theorem for random fields and present its applications. In chapter 3 we study the estimation of mutual information for mixed-pair random variables and conduct simulation study to illustrate the theoretical results.

2 CRAMÉR TYPE MODERATE DEVIATIONS FOR RANDOM FIELDS

2.1 INTRODUCTION

In this chapter we study the Cramér type moderate deviations for random fields, in particular linear random fields (often called spatial linear processes in statistics literature) with short or long memory (short or long range dependence). The study of moderate deviation probabilities in non-logarithmic form for independent random variables goes back to 1920s. The first theorem in this field was published by [32] who studied a particular case of the Bernoulli random variables. In his fundamental work, [9] studied the estimation of the tail probability by the standard normal distribution under the condition that the random variable has moment generating function in a neighborhood of the origin (cf. (2.2.1) below). This condition has been referred to as the Cramér condition. Cramér's work was improved by [49] (see also [51], [52]). Their works have stimulated a large amount of research on moderate and large deviations; see below for a brief (and incomplete) review on literature related to this chapter. Nowadays, the area of moderate and large deviation deviations is not only important in probability but also plays an important role in many applied fields, for instance, the premium calculation problem, risk management in insurance (cf.[3]), non-parametric estimation in statistics (see, e.g., [6], [64], [29], [30]) and in network information theory (cf. [38], [39]).

Let X, X_1, X_2, \dots be a sequence of independent and identically distributed (i.i.d.) random variables with mean 0 and variance σ^2 . Let $S_n = \sum_{k=1}^n X_k$ ($n \geq 1$) be the partial

sums. By the central limit theorem,

$$\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |\mathbb{P}(S_n > x\sigma\sqrt{n}) - (1 - \Phi(x))| = 0,$$

where $\Phi(x)$ is the probability distribution of the standard normal random variable. If for a suitable sequence c_n , we have

$$\lim_{n \rightarrow \infty} \sup_{0 \leq x \leq c_n} \left| \frac{\mathbb{P}(S_n > x\sigma\sqrt{n})}{1 - \Phi(x)} - 1 \right| = 0, \quad (2.1.1)$$

or $\mathbb{P}(S_n > x\sigma\sqrt{n}) = (1 - \Phi(x))(1 + o(1))$ uniformly over $x \in [0, c_n]$, then Eq. (2.1.1) is called moderate deviation probability or normal deviation probability for S_n since it can be estimated by the standard normal distribution. We refer to $[0, c_n]$ as a range for the moderate deviation. The most famous result of this kind is the Cramér type moderate deviation. Under Cramér's condition, one has the following Cramér's theorem ([9], [49], [51], [49], p.218; or [52], p.178): If $x \geq 0$ and $x = o(\sqrt{n})$ then

$$\frac{\mathbb{P}(S_n > x\sigma\sqrt{n})}{1 - \Phi(x)} = \exp \left\{ \frac{x^3}{\sqrt{n}} \lambda \left(\frac{x}{\sqrt{n}} \right) \right\} \left[1 + O \left(\frac{x+1}{\sqrt{n}} \right) \right]. \quad (2.1.2)$$

Here $\lambda(z) = \sum_{k=0}^{\infty} c_k z^k$ is a power series with coefficients depending on the cumulants of the random variable X . Eq. (2.1.2) provides more precise approximation than (2.1.1) which holds uniformly on the range $[0, c_n]$ for any $c_n = o(\sqrt{n})$. The moderate deviations under Cramér's condition for independent non-identically distributed random variables were obtained by [13],[49] and [61]. The Cramér type moderate deviation has also been established for the sum of independent random variables with p -th moment, $p > 2$. To name a few, for example, see [56], [44], [45], [42], [60], [2], and [14]. It should be pointed out that the ranges the moderate deviations in these references are smaller (e.g., $c_n = O(\sqrt{\log n})$).

The Cramér type moderate deviations for dependent random variables have also been studied in the literature. [16] and [25] studied the moderate deviation for m -dependent

random variables. [17], [4] studied moderate deviation for mixing processes. [18], [19], [20] and [12] investigated the large and moderate deviations for martingales. [5] established moderate deviation results for linear processes with coefficients satisfying $\sum_{i=1}^{\infty} i|a_i| < \infty$. [67] studied moderate deviations for stationary processes under certain conditions in terms of the physical dependence measure. But it can be verified that the results from [67] can only be applied to linear processes with short memory and their transformations. Recently [47] studied the exact moderate and large deviations for short or long memory linear processes. [57] studied exact moderate and large deviations for linear random fields and applied the moderate result to prove a Davis-Gut law of the iterated logarithm. Nevertheless, in the aforementioned works, the moderate deviations are studied for dependent random variables with p -th moment, $p > 2$. The exact moderate deviation for random fields under Cramér's condition has not been well studied. For example, the optimal range $[0, c_n]$ and the exact rate of convergence in (2.1.1) had been unknown in the random field setting.

The main objective of this part is to establish exact moderate deviation analogous to (2.1.2) for random fields under Cramér's condition. Our main result is Theorem 2.2.1 below, whose proof is based on the conjugate method to change the probability measure as in the classical case (see, e.g., [50], [51]). The extension of this method to the random field setting reveals the deep relationship between the tail probabilities and the properties of the cumulant generating functions of the random variables such as the analytic radius and the bounds, for x within some ranges related to the sum of the variances and the analytic radius of the cumulant generating functions of these random variables. Compared with the results in [57] for linear random fields, Theorem 2.2.1 and 2.3.1 in this chapter provide more precise convergence rate in the moderate deviations and explicit information on the range $[0, c_n]$, which is much bigger than the range in Theorem 2.1 in [57]. In this chapter we use the following notations. For two sequences $\{a_n\}$ and $\{b_n\}$ of real numbers, $a_n \sim b_n$ means $a_n/b_n \rightarrow 1$ as $n \rightarrow \infty$; $a_n \propto b_n$ means that $a_n/b_n \rightarrow C$ as $n \rightarrow \infty$ for some constant $C > 0$; for positive sequences, the notation $a_n \ll b_n$ or $b_n \gg a_n$ means that a_n/b_n is bounded.

For $d, m \in \mathbb{N}$ denote $\Gamma_m^d = [-m, m]^d \cap \mathbb{Z}^d$. Section 2.2 gives the main results. In Section 2.3 we study the application of the main results in linear random fields and nonparametric regression.

2.2 MAIN RESULTS

Let $\{X_{nj}, n \in \mathbb{N}, j \in \mathbb{Z}^d\}$ be a random field with zero means defined on a probability space (Ω, \mathcal{F}, P) . Suppose that for each n , the random variables $X_{nj}, j \in \mathbb{Z}^d$ are independent and satisfy the following Cramér condition: There is a positive constant H_n such that the cumulant generating function

$$L_{nj}(z) = \log \mathbb{E} e^{zX_{nj}} \quad \text{of } X_{nj} \text{ is analytic in } D_n, \quad (2.2.1)$$

where $D_n = \{z \in \mathbb{C} : |z| < H_n\}$ is the disc of radius H_n on the complex plane \mathbb{C} , and \log denotes the principal value of the logarithm so that $L_{nj}(0) = 0$.

Without loss of generality we assume in this section that $\limsup_{n \rightarrow \infty} H_n < \infty$. Within the disc $\{z \in \mathbb{C} : |z| < H_n\}$, L_{nj} can be expanded in a convergent power series

$$L_{nj}(z) = \sum_{k=1}^{\infty} \frac{\gamma_{knj}}{k!} z^k,$$

where γ_{knj} is the cumulant of order k of the random variable X_{nj} . We have that $\gamma_{1nj} = \mathbb{E} X_{nj} = 0$ and $\gamma_{2nj} = \mathbb{E} X_{nj}^2 = \sigma_{nj}^2$. By Taylor's expansion, one can verify that a sufficient condition for (2.2.1) is the following moment condition

$$|\mathbb{E} X_{nj}^m| \leq \frac{m!}{2} \sigma_{nj}^2 H_n^{2-m} \quad \text{for all } m \geq 2.$$

This condition has been used frequently in probability and statistics, see [51] p.55, [28] p.64, [54] p.301, [68] p.164, among others.

Denote

$$S_n = \sum_{j \in \mathbb{Z}^d} X_{nj}, \quad S_{m,n} = \sum_{j \in \Gamma_m^d} X_{nj},$$

$$B_n = \sum_{j \in \mathbb{Z}^d} \sigma_{nj}^2, \quad F_n(x) = P(S_n < x\sqrt{B_n})$$

and assume that S_n is well-defined and $B_n < \infty$ for each $n \in \mathbb{N}$. The following is the main result of this chapter.

Theorem 2.2.1. *Suppose that, for all $n \in \mathbb{N}$ and $j \in \mathbb{Z}^d$, there exist non-negative constants c_{nj} such that*

$$|L_{nj}(z)| \leq c_{nj}, \quad \forall z \in \mathbb{C} \text{ with } |z| < H_n, \quad (2.2.2)$$

and suppose that $B_n H_n^2 \rightarrow \infty$ as $n \rightarrow \infty$, and

$$C_n := \sum_{j \in \mathbb{Z}^d} c_{nj} = O(B_n H_n^2). \quad (2.2.3)$$

If $x \geq 0$ and $x = o(H_n \sqrt{B_n})$, then

$$\frac{1 - F_n(x)}{1 - \Phi(x)} = \exp \left\{ \frac{x^3}{H_n \sqrt{B_n}} \lambda_n \left(\frac{x}{H_n \sqrt{B_n}} \right) \right\} \left(1 + O \left(\frac{x+1}{H_n \sqrt{B_n}} \right) \right), \quad (2.2.4)$$

$$\frac{F_n(-x)}{\Phi(-x)} = \exp \left\{ - \frac{x^3}{H_n \sqrt{B_n}} \lambda_n \left(- \frac{x}{H_n \sqrt{B_n}} \right) \right\} \left(1 + O \left(\frac{x+1}{H_n \sqrt{B_n}} \right) \right), \quad (2.2.5)$$

where

$$\lambda_n(t) = \sum_{k=0}^{\infty} \beta_{kn} t^k$$

is a power series that stays bounded uniformly in n for sufficiently small values of $|t|$ and the coefficients β_{kn} only depend on the cumulants of X_{nj} ($n \in \mathbb{Z}, j \in \mathbb{Z}^d$).

Proof. Since $\gamma_{1nj} = 0$, the cumulant generating function $L_{nj}(z)$ of X_{nj} can be written as

$$L_{nj}(z) = \log \mathbb{E} e^{zX_{nj}} = \sum_{k=2}^{\infty} \frac{\gamma_{knj}}{k!} z^k.$$

Cauchy's inequality for the derivatives of analytic functions together with the condition (2.2.2) yields that

$$|\gamma_{knj}| < \frac{k!c_{nj}}{H_n^k}. \quad (2.2.6)$$

By following the conjugate method (cf. Petrov (1965, 1975)), we now introduce an auxiliary sequence of independent random variables $\{\bar{X}_{nj}\}$, $j \in \mathbb{Z}^d$, with the distribution functions

$$\bar{V}_{nj}(x) = e^{-L_{nj}(z)} \int_{-\infty}^x e^{zy} dV_{nj}(y),$$

where $V_{nj}(y) = P(X_{nj} < y)$ and $z \in (-H_n, H_n)$ is a real number whose value will be specified later.

Denote

$$\bar{m}_{nj} = \mathbb{E} \bar{X}_{nj}, \quad \bar{\sigma}_{nj}^2 = \mathbb{E}(\bar{X}_{nj} - \bar{m}_{nj})^2,$$

$$\bar{S}_{m,n} = \sum_{j \in \Gamma_m^d} \bar{X}_{nj}, \quad \bar{S}_n = \sum_{j \in \mathbb{Z}^d} \bar{X}_{nj},$$

$$\bar{M}_n = \sum_{j \in \mathbb{Z}^d} \bar{m}_{nj}, \quad \bar{B}_n = \sum_{j \in \mathbb{Z}^d} \bar{\sigma}_{nj}^2$$

and

$$\bar{F}_n(x) = P(\bar{S}_n < \bar{M}_n + x\sqrt{\bar{B}_n}).$$

Note that, in the above and below, we have suppressed z for simplicity of notations.

We shall see in the later analysis that the quantities \bar{S}_n , \bar{M}_n and \bar{B}_n are well-defined for every n and $z \in \mathbf{R}$ with $|z| < aH_n$, where $a < 1$ is a positive constant which is independent of

n . Throughout the proof we will obtain some estimates holding for the values of z satisfying $|z| < bH_n$, where the positive constant $b < 1$ may vary but is always independent of n . We will then take a to be the smallest one among those constants b . The selection of the constants does not affect the proof since the $z = z_n$ we need in the later analysis has property $z = o(H_n)$.

Also, the change of the order of summation of double series presented in the proof is justified by the absolute convergence of those series in the specified regions.

Step 1: Representation of $P(S_n < x)$ in terms of the conjugate measure

First notice that by equation (2.11) on page 221 of Petrov (1975), for any $m \in \mathbb{N}$, we have

$$P(S_{m,n} < x) = \exp \left\{ \sum_{j \in \Gamma_m^d} L_{nj}(z) \right\} \int_{-\infty}^x e^{-zy} dP(\bar{S}_{m,n} < y). \quad (2.2.7)$$

Note that the condition (2.2.3) implies that $C_n < \infty, n \in \mathbb{N}$. From (2.2.6) it follows that for any w with $|w| < \frac{2}{3}H_n$ and for any $m \in \mathbb{N}$ we have

$$\begin{aligned} \left| \sum_{j \in \Gamma_m^d} L_{nj}(w) \right| &= \left| \sum_{j \in \Gamma_m^d} \sum_{k=2}^{\infty} \frac{\gamma_{knj}}{k!} w^k \right| \\ &\leq \sum_{j \in \Gamma_m^d} \sum_{k=2}^{\infty} \frac{|\gamma_{knj}|}{k!} |w|^k \\ &\leq \sum_{j \in \mathbb{Z}^d} \sum_{k=2}^{\infty} \frac{c_{nj}}{H_n^k} |w|^k \\ &\leq \frac{4}{3} \sum_{j \in \mathbb{Z}^d} c_{nj} = \frac{4}{3} C_n < \infty. \end{aligned} \quad (2.2.8)$$

Therefore, for any v with $|v| < \frac{1}{2}H_n$ and z with $|z| < \frac{1}{6}H_n$,

$$\begin{aligned}
\mathbb{E} \exp\{v\bar{S}_{m,n}\} &= \prod_{j \in \Gamma_m^d} \mathbb{E} \exp\{v\bar{X}_{nj}\} \\
&= \prod_{j \in \Gamma_m^d} \int_{-\infty}^{\infty} e^{vx} d\bar{V}_{nj}(x) = \prod_{j \in \Gamma_m^d} \int_{-\infty}^{\infty} e^{vx} e^{-L_{nj}(z)} e^{zx} dV_{nj}(x) \\
&= \prod_{j \in \Gamma_m^d} e^{-L_{nj}(z)} \int_{-\infty}^{\infty} e^{(v+z)x} dV_{nj}(x) = \prod_{j \in \Gamma_m^d} e^{-L_{nj}(z)} e^{L_{nj}(v+z)} \\
&\rightarrow \exp\left(\sum_{j \in \mathbb{Z}^d} [L_{nj}(v+z) - L_{nj}(z)]\right) < \infty, \text{ as } m \rightarrow \infty.
\end{aligned} \tag{2.2.9}$$

Hence, \bar{S}_n is well-defined and $\bar{S}_{m,n}$ converges to \bar{S}_n in distribution or equivalently in probability or almost surely as $m \rightarrow \infty$.

For the x in $P(S_n < x)$, let $f(y) = \exp\{-zy\}\mathbf{1}\{y < x\}$ and $M > 0$. By Markov's inequality, we have

$$\begin{aligned}
&\mathbb{E} \left\{ f(\bar{S}_{m,n}) \mathbf{1}\{|f(\bar{S}_{m,n})| > M\} \right\} \\
&\leq \mathbb{E} \left\{ \exp\{-z\bar{S}_{m,n}\} \mathbf{1}\{\exp\{-z\bar{S}_{m,n}\} > M\} \right\} \\
&\leq \left[\mathbb{E} \left\{ \exp\{-2z\bar{S}_{m,n}\} \right\} \right]^{\frac{1}{2}} \left[\mathbb{E} \left\{ \mathbf{1}\{\exp\{-z\bar{S}_{m,n}\} > M\} \right\} \right]^{\frac{1}{2}} \\
&\leq \left[\prod_{j \in \Gamma_m^d} e^{-L_{nj}(z)} e^{L_{nj}(-z)} \right]^{\frac{1}{2}} \left[\frac{1}{M} \mathbb{E} \left\{ \exp\{-z\bar{S}_{m,n}\} \right\} \right]^{\frac{1}{2}} \\
&= \frac{1}{\sqrt{M}} \left[\prod_{j \in \Gamma_m^d} e^{-L_{nj}(z)} e^{L_{nj}(-z)} \right]^{\frac{1}{2}} \left[\prod_{j \in \Gamma_m^d} e^{-L_{nj}(z)} e^{L_{nj}(0)} \right]^{\frac{1}{2}}.
\end{aligned}$$

Hence, by (2.2.8) we have that for $|z| < \frac{1}{6}H_n$,

$$\lim_{M \rightarrow \infty} \limsup_{m \rightarrow \infty} \mathbb{E} \left\{ f(\bar{S}_{m,n}) \mathbf{1}\{|f(\bar{S}_{m,n})| > M\} \right\} = 0.$$

Applying Theorem 2.20 from van der Vaart (1998), we have

$$\int_{-\infty}^x e^{-zy} dP(\bar{S}_{m,n} < y) \rightarrow \int_{-\infty}^x e^{-zy} dP(\bar{S}_n < y)$$

as $m \rightarrow \infty$. And taking into account that

$$P(S_{m,n} < x) \rightarrow P(S_n < x)$$

and

$$\exp \left\{ \sum_{j \in \Gamma_m^d} L_{nj}(z) \right\} \rightarrow \exp \left\{ \sum_{j \in \mathbb{Z}^d} L_{nj}(z) \right\}$$

as $m \rightarrow \infty$ we obtain from (2.2.7) that

$$P(S_n < x) = \exp \left\{ \sum_{j \in \mathbb{Z}^d} L_{nj}(z) \right\} \int_{-\infty}^x e^{-zy} dP(\bar{S}_n < y). \quad (2.2.10)$$

Step 2: Properties of the conjugate measure

From the calculation of (2.2.9) it follows that the cumulant generating function $\bar{L}_{nj}(v)$ of the random variable \bar{X}_{nj} exists when $|v|$ is sufficiently small and we have

$$\bar{L}_{nj}(v) = -L_{nj}(z) + L_{nj}(v+z), \quad (2.2.11)$$

$j \in \mathbb{Z}^d$. Denoting by $\bar{\gamma}_{knj}$ the cumulant of order k of the random variable \bar{X}_{nj} , we obtain

$$\bar{\gamma}_{knj} = \left[\frac{d^k \bar{L}_{nj}(v)}{dv^k} \right]_{v=0} = \frac{d^k L_{nj}(z)}{dz^k}.$$

Setting $k = 1$ and $k = 2$ we find that

$$\bar{m}_{nj} = \frac{dL_{nj}(z)}{dz} = \sum_{\ell=2}^{\infty} \frac{\gamma_{\ell nj}}{(\ell-1)!} z^{\ell-1}, \quad (2.2.12)$$

and

$$\bar{\sigma}_{nj}^2 = \frac{d^2 L_{nj}(z)}{dz^2} = \sum_{\ell=2}^{\infty} \frac{\gamma_{\ell nj}}{(\ell-2)!} z^{\ell-2}. \quad (2.2.13)$$

Hence, for $|z| < \frac{1}{2}H_n$, (2.2.12) implies

$$\begin{aligned} |\bar{M}_n| &= \left| \sum_{j \in \mathbb{Z}^d} \bar{m}_{nj} \right| = \left| \sum_{j \in \mathbb{Z}^d} \sum_{k=2}^{\infty} \frac{\gamma_{knj}}{(k-1)!} z^{k-1} \right| \\ &\leq \sum_{j \in \mathbb{Z}^d} \sum_{k=2}^{\infty} \frac{k! c_{nj}}{H_n^k} \frac{|z|^{k-1}}{(k-1)!} \leq \frac{3}{H_n} \sum_{j \in \mathbb{Z}^d} c_{nj} = \frac{3C_n}{H_n}, \end{aligned} \quad (2.2.14)$$

which means that \bar{M}_n is well-defined and, as a function of $z \in \mathbb{C}$, is analytic in $|z| < \frac{1}{2}H_n$.

Also, without loss of generality, we assume that

$$\limsup_n \frac{C_n}{B_n H_n^2} \leq 1. \quad (2.2.15)$$

By the definition of \bar{M}_n and (2.2.12), we have

$$\begin{aligned} \bar{M}_n &= z \sum_{j \in \mathbb{Z}^d} \gamma_{2nj} + \sum_{j \in \mathbb{Z}^d} \sum_{k=3}^{\infty} \frac{\gamma_{knj}}{(k-1)!} z^{k-1} \\ &= z B_n + \sum_{j \in \mathbb{Z}^d} \sum_{k=3}^{\infty} \frac{\gamma_{knj}}{(k-1)!} z^{k-1}. \end{aligned} \quad (2.2.16)$$

It follows from (2.2.6) that

$$\begin{aligned} \left| \sum_{k=3}^{\infty} \frac{\gamma_{knj}}{(k-1)!} z^{k-1} \right| &\leq |z| \sum_{k=3}^{\infty} \frac{k! c_{nj}}{H_n^k} \frac{|z|^{k-2}}{(k-1)!} \\ &= \frac{|z| c_{nj}}{H_n^2} \sum_{k=3}^{\infty} k \left| \frac{z}{H_n} \right|^{k-2} \leq \frac{|z| c_{nj}}{2H_n^2} \end{aligned}$$

for $|z| < b_1 H_n$ and a suitable positive constant $b_1 < 1$ which is independent of j and n . This together with (2.2.16) implies that for $|z| < b_1 H_n$

$$|z| \left(B_n - \frac{C_n}{2H_n^2} \right) \leq |\overline{M}_n| \leq |z| \left(B_n + \frac{C_n}{2H_n^2} \right).$$

Taking into account the condition (2.2.15), we get that

$$\overline{M}_n \propto |z| B_n. \quad (2.2.17)$$

Moreover, (2.2.16) implies that for $|z| < \frac{1}{2} H_n$,

$$\begin{aligned} |\overline{M}_n - z B_n| &\leq \sum_{j \in \mathbb{Z}^d} \sum_{k=3}^{\infty} \frac{k! c_{nj}}{H_n^k} \frac{|z|^{k-1}}{(k-1)!} \\ &\leq \frac{|z|^2}{H_n^3} \sum_{j \in \mathbb{Z}^d} \sum_{k=3}^{\infty} k c_{nj} \frac{|z|^{k-3}}{H_n^{k-3}} \leq \frac{8|z|^2 C_n}{H_n^3}. \end{aligned} \quad (2.2.18)$$

Also, by the definition of \overline{B}_n and (2.2.13), we have

$$\begin{aligned} \overline{B}_n &= \sum_{j \in \mathbb{Z}^d} \gamma_{2nj} + \sum_{j \in \mathbb{Z}^d} \sum_{k=3}^{\infty} \frac{\gamma_{knj}}{(k-2)!} z^{k-2} \\ &= B_n + \sum_{j \in \mathbb{Z}^d} \sum_{k=3}^{\infty} \frac{\gamma_{knj}}{(k-2)!} z^{k-2}. \end{aligned} \quad (2.2.19)$$

It follows from (2.2.6) that

$$\left| \sum_{k=3}^{\infty} \frac{\gamma_{knj}}{(k-2)!} z^{k-2} \right| \leq \sum_{k=3}^{\infty} \frac{k! c_{nj}}{H_n^k} \frac{|z|^{k-2}}{(k-2)!} \leq \frac{c_{nj}}{2H_n^2}$$

for $|z| < b_2 H_n$ and a suitable positive constant $b_2 < 1$ which is independent of j and n . This together with (2.2.19) implies that for $|z| < b_2 H_n$, \overline{B}_n is well-defined and

$$B_n - \frac{C_n}{2H_n^2} \leq |\overline{B}_n| \leq B_n + \frac{C_n}{2H_n^2}.$$

Condition (2.2.15) then implies that

$$\overline{B}_n \propto B_n. \quad (2.2.20)$$

Furthermore, (2.2.19) and (2.2.6) imply that for $|z| < \frac{1}{2}H_n$,

$$\begin{aligned} \left| \overline{B}_n - B_n \right| &\leq \sum_{j \in \mathbb{Z}^d} \sum_{k=3}^{\infty} \frac{k! c_{nj}}{H_n^k} \frac{|z|^{k-2}}{(k-2)!} \\ &\leq \frac{|z|}{H_n^3} \sum_{j \in \mathbb{Z}^d} \sum_{k=3}^{\infty} k(k-1) c_{nj} \frac{|z|^{k-3}}{H_n^{k-3}} \\ &\leq \frac{28|z|C_n}{H_n^3}. \end{aligned} \quad (2.2.21)$$

Step 3: Selection of z

Let $z = z_n$ be the real solution of the equation

$$x = \frac{\overline{M}_n}{\sqrt{B_n}}, \quad (2.2.22)$$

and let

$$t = t_n = \frac{x}{H_n \sqrt{B_n}}. \quad (2.2.23)$$

Then

$$t = \frac{\overline{M}_n}{H_n B_n} = \frac{1}{H_n B_n} \sum_{j \in \mathbb{Z}^d} \sum_{k=2}^{\infty} \frac{\gamma_{knj}}{(k-1)!} z^{k-1}. \quad (2.2.24)$$

By (2.2.14) we know that $\frac{\overline{M}_n}{H_n B_n}$ is analytic in a disc $|z| < \frac{1}{2}H_n$ and

$$\left| \frac{\overline{M}_n}{H_n B_n} \right| \leq \frac{3C_n}{H_n^2 B_n}$$

in that disc. It follows from Bloch's theorem (see, e.g., Privalov (1984), page 256) that (2.2.24) has a real solution which can be written as

$$z = \sum_{m=1}^{\infty} a_{mn} t^m \quad (2.2.25)$$

for

$$|t| < \left(\sqrt{\frac{1}{2} + \frac{3C_n}{H_n^2 B_n}} - \sqrt{\frac{3C_n}{H_n^2 B_n}} \right)^2.$$

Moreover, the absolute value of that sum in (2.2.25) is less than $\frac{1}{2}H_n$. Condition (2.2.3) implies that there exists a disc with center at $t = 0$ and radius R that does not depend on n within which the series on the right side of (2.2.25) converges.

It can be checked from (2.2.24) and (2.2.25) that

$$a_{1n} = H_n \quad \text{and} \quad a_{2n} = -\frac{H_n^2}{2B_n} \sum_{j \in \mathbb{Z}^d} \gamma_{3nj}. \quad (2.2.26)$$

Cauchy's inequality implies that for every $m \in \mathbb{N}$,

$$|a_{mn}| \leq \frac{H_n}{2R^m}.$$

Therefore, as $t \rightarrow 0$, $a_{1n}t$ becomes the dominant term of the series in (2.2.25). Hence, for sufficiently large n we have

$$\frac{1}{2}tH_n \leq z \leq 2tH_n, \quad z = o(H_n)$$

and taking into account (2.2.23) we get

$$\frac{x}{2\sqrt{B_n}} \leq z \leq \frac{2x}{\sqrt{B_n}}. \quad (2.2.27)$$

It follows from (2.2.8) and (2.2.14) that for $z < \frac{1}{2}H_n$,

$$\left| z\bar{M}_n - \sum_{j \in \mathbb{Z}^d} L_{nj}(z) \right| \leq \frac{3|z|}{H_n} C_n + \frac{4}{3} C_n < 3C_n.$$

For the solution z of the equation (2.2.22) we also have

$$\begin{aligned} z\bar{M}_n - \sum_{j \in \mathbb{Z}^d} L_{nj}(z) &= \sum_{j \in \mathbb{Z}^d} \sum_{k=2}^{\infty} \frac{\gamma_{knj}}{(k-1)!} z^k - \sum_{j \in \mathbb{Z}^d} \sum_{k=2}^{\infty} \frac{\gamma_{knj}}{k!} z^k \\ &= \sum_{j \in \mathbb{Z}^d} \sum_{k=2}^{\infty} \frac{(k-1)\gamma_{knj}}{k!} \left(\sum_{m=1}^{\infty} a_{mn} t^m \right)^k \\ &:= \sum_{j \in \mathbb{Z}^d} \frac{\gamma_{2nj}}{2} a_{1n}^2 t^2 - \sum_{k=3}^{\infty} b_{kn} t^k \tag{2.2.28} \\ &= \frac{H_n^2 B_n t^2}{2} - H_n^2 B_n t^3 \sum_{k=3}^{\infty} \frac{b_{kn}}{H_n^2 B_n} t^{k-3} \\ &= \frac{H_n^2 B_n t^2}{2} - H_n^2 B_n t^3 \lambda_n(t), \end{aligned}$$

where $\lambda_n(t) = \sum_{k=0}^{\infty} \beta_{kn} t^k$ with $\beta_{kn} = b_{(k+3)n} (H_n^2 B_n)^{-1}$.

Recall that the series $\sum_{m=1}^{\infty} a_{mn} t^m$ converges in the disc centered at $t = 0$ with radius $R > 0$ that does not depend on n , and the absolute value of this sum is less than $\frac{1}{2}H_n$. We see from (2.2.28) that the function $\lambda_n(t)$ is obtained by the substitution of $\sum_{m=1}^{\infty} a_{mn} t^m$ in a series that converges on the interval $(-\frac{1}{2}H_n, \frac{1}{2}H_n)$. It follows from Cauchy's inequality that

$$|\beta_{kn}| \leq \frac{3C_n}{H_n^2 B_n R^{k+3}} \leq \frac{3}{R^{k+3}}, \quad k \geq 0,$$

which means that for $|t| < \frac{1}{2}R$, $\lambda_n(t)$ stays bounded uniformly in n . In particular, by (2.2.26) and (2.2.28), we have $\beta_{0n} = \frac{H_n}{6B_n} \sum_{j \in \mathbb{Z}^d} \gamma_{3nj}$.

From now on we will assume that z is the unique real solution of the equation (2.2.22).

Step 4: The case $0 \leq x \leq 1$

Now we prove the theorem for the case $0 \leq x \leq 1$ using the method presented in Petrov and Robinson (2006). Throughout the proof, C denotes a positive constant which may vary from line to line, but is independent of j, n and z . If $f_n(s)$ is the characteristic function of $S_n/\sqrt{B_n}$ we then have that for $|s| < H_n\sqrt{B_n}/2$

$$\begin{aligned} f_n(s) &= \int_{-\infty}^{\infty} e^{isu} dP(S_n \leq u\sqrt{B_n}) \\ &= \int_{-\infty}^{\infty} e^{isy/\sqrt{B_n}} dP(S_n \leq y) \\ &= \exp \left\{ \sum_{j \in \mathbb{Z}^d} L_{nj}(is/\sqrt{B_n}) \right\}. \end{aligned}$$

Then

$$\begin{aligned} \log f_n(s) &= \sum_{j \in \mathbb{Z}^d} L_{nj}(is/\sqrt{B_n}) = \sum_{j \in \mathbb{Z}^d} \sum_{k=2}^{\infty} \frac{\gamma_{knj}}{k!} (is/\sqrt{B_n})^k \\ &= - \sum_{j \in \mathbb{Z}^d} \frac{\gamma_{2nj}}{2} s^2/B_n + \sum_{j \in \mathbb{Z}^d} \sum_{k=3}^{\infty} \frac{\gamma_{knj}}{k!} (is/\sqrt{B_n})^k = -s^2/2 + \sum_{j \in \mathbb{Z}^d} \sum_{k=3}^{\infty} \frac{\gamma_{knj}}{k!} (is/\sqrt{B_n})^k. \end{aligned}$$

Thus, using (2.2.6) we get that for $|s| < \delta H_n\sqrt{B_n}/2$, with $0 < \delta < 1$,

$$|\log f_n(s) + s^2/2| \leq \sum_{j \in \mathbb{Z}^d} \sum_{k=3}^{\infty} c_{nj} \left(\frac{|s|}{H_n\sqrt{B_n}} \right)^k \leq C_n \left(\frac{|s|}{H_n\sqrt{B_n}} \right)^3 (1 - \delta)^{-1}$$

Then, for appropriate choice of δ we have that

$$|f_n(s) - e^{-s^2/2}| < C \frac{e^{-s^2/4}|s|^3 C_n}{H_n^3 \sqrt{B_n}^3} < C \frac{e^{-s^2/4}|s|^3}{H_n \sqrt{B_n}},$$

for $|s| < \delta H_n \sqrt{B_n}/2$. Now applying Theorem 5.1 from Petrov (1995) with $b = 1/\pi$ and $T = \delta H_n \sqrt{B_n}/2$ we get that

$$\sup_x |F_n(x) - \Phi(x)| < \frac{C}{H_n \sqrt{B_n}}. \quad (2.2.29)$$

Since $0 \leq x \leq 1$, $B_n H_n^2 \rightarrow \infty$ as $n \rightarrow \infty$, and $\lambda_n\left(\frac{x}{H_n \sqrt{B_n}}\right)$ is bounded uniformly in n , we have

$$\exp\left\{\frac{x^3}{H_n \sqrt{B_n}} \lambda_n\left(\frac{x}{H_n \sqrt{B_n}}\right)\right\} = 1 + O(H_n^{-1} B_n^{-1/2}).$$

Together with condition (2.2.3), to have (2.2.4) in the case $0 \leq x \leq 1$, it is sufficient to show

$$\frac{1 - F_n(x)}{1 - \Phi(x)} = 1 + O\left(\frac{C}{H_n \sqrt{B_n}}\right),$$

which is given by (2.2.29), since $1/2 \leq \Phi(x) \leq \Phi(1)$ for $0 \leq x \leq 1$.

So we will limit the proof of the theorem to the case $x > 1$, $x = o(H_n \sqrt{B_n})$.

Step 5: The case $x > 1$, $x = o(H_n \sqrt{B_n})$

Making a change of variables $y \rightsquigarrow \bar{M}_n + y\sqrt{B_n}$ and applying (2.2.22), we can rewrite (2.2.10) as

$$\begin{aligned} 1 - F_n(x) &= \exp\left\{-z\bar{M}_n + \sum_{j \in \mathbb{Z}^d} L_{nj}(z)\right\} \int_{(x\sqrt{B_n} - \bar{M}_n)/\sqrt{B_n}}^{\infty} \exp\left\{-zy\sqrt{B_n}\right\} d\bar{F}_n(y) \\ &= \exp\left\{-z\bar{M}_n + \sum_{j \in \mathbb{Z}^d} L_{nj}(z)\right\} \int_0^{\infty} \exp\left\{-zy\sqrt{B_n}\right\} d\bar{F}_n(y). \end{aligned} \quad (2.2.30)$$

Denote $r_n(x) = \bar{F}_n(x) - \Phi(x)$ and we show that for sufficiently large n

$$\sup_x |r_n(x)| \leq \frac{C}{H_n \sqrt{B_n}}. \quad (2.2.31)$$

Let $\bar{f}_n(s)$ be the characteristic function of $(\bar{S}_n - \bar{M}_n)/\sqrt{\bar{B}_n}$. We then have that

$$\begin{aligned}
\bar{f}_n(s) &= \int_{-\infty}^{\infty} e^{isu} dP(\bar{S}_n \leq u\sqrt{\bar{B}_n} + \bar{M}_n) \\
&= \int_{-\infty}^{\infty} e^{is(y-\bar{M}_n)/\sqrt{\bar{B}_n}} dP(\bar{S}_n \leq y) \\
&= \exp \left\{ -is\bar{M}_n/\sqrt{\bar{B}_n} - \sum_{j \in \mathbb{Z}^d} L_{nj}(z) \right\} \int_{-\infty}^{\infty} e^{(z+is/\sqrt{\bar{B}_n})y} dP(S_n \leq y) \\
&= \exp \left\{ -is\bar{M}_n/\sqrt{\bar{B}_n} - \sum_{j \in \mathbb{Z}^d} L_{nj}(z) + \sum_{j \in \mathbb{Z}^d} L_{nj}(z + is/\sqrt{\bar{B}_n}) \right\}.
\end{aligned}$$

Then by (2.2.11) for $|z| < \frac{1}{2}H_n$ and $|s| < H_n\sqrt{\bar{B}_n}/6$ we have that

$$\begin{aligned}
\log \bar{f}_n(s) &= -is\bar{M}_n/\sqrt{\bar{B}_n} + \sum_{j \in \mathbb{Z}^d} \bar{L}_{nj}(is/\sqrt{\bar{B}_n}) \\
&= -\frac{1}{2}s^2 + \frac{1}{6}(is/\sqrt{\bar{B}_n})^3 \left[\frac{d^3 \sum_{j \in \mathbb{Z}^d} \bar{L}_{nj}(y)}{dy^3} \right]_{y=\theta is/\sqrt{\bar{B}_n}},
\end{aligned}$$

where $0 \leq |\theta| \leq 1$. For $|z| < \frac{1}{2}H_n$ and $|s| < \delta H_n\sqrt{\bar{B}_n}/6$, with $0 < \delta < 1$, we have that

$$\begin{aligned}
&\left| \left[\frac{d^3 \sum_{j \in \mathbb{Z}^d} \bar{L}_{nj}(y)}{dy^3} \right]_{y=\theta is/\sqrt{\bar{B}_n}} \right| = \left| \left[\frac{d^3}{dy^3} \sum_{j \in \mathbb{Z}^d} \sum_{k=1}^{\infty} \frac{\bar{\gamma}_{knj}}{k!} y^k \right]_{y=\theta is/\sqrt{\bar{B}_n}} \right| \\
&= \left| \sum_{j \in \mathbb{Z}^d} \sum_{k=3}^{\infty} \frac{\bar{\gamma}_{knj}}{(k-3)!} (\theta is/\sqrt{\bar{B}_n})^{k-3} \right| \\
&\leq \sum_{j \in \mathbb{Z}^d} \sum_{k=3}^{\infty} k(k-1)(k-2) \frac{c_{nj}}{(H_n/2)^k} \left(s/\sqrt{\bar{B}_n} \right)^{k-3} = \frac{48C_n}{H_n^3} \left(1 - \frac{s/\sqrt{\bar{B}_n}}{H_n/2} \right)^{-4} \\
&\leq \frac{48C_n}{H_n^3} (1-\delta)^{-4}.
\end{aligned}$$

Thus,

$$|\log \bar{f}_n(s) + s^2/2| < \frac{8|s|^3 C_n}{H_n^3 \sqrt{\bar{B}_n}} (1-\delta)^{-4}.$$

Then, for appropriate choice of δ we have that

$$|\bar{f}_n(s) - e^{-s^2/2}| < C \frac{e^{-s^2/4}|s|^3 C_n}{H_n^3 \sqrt{B_n}^3} < C \frac{e^{-s^2/4}|s|^3}{H_n \sqrt{B_n}}$$

for $|s| < \delta H_n \sqrt{B_n}/6$. Now applying (2.2.20) and Theorem 5.1 from Petrov (1995) with $b = 1/\pi$ and $T = \delta H_n \sqrt{B_n}/6$, we have (2.2.31).

By (2.2.31) we have

$$\begin{aligned} \int_0^\infty \exp\left\{-zy\sqrt{B_n}\right\} d\bar{F}_n(y) &= \frac{1}{\sqrt{2\pi}} \int_0^\infty \exp\left\{-zy\sqrt{B_n} - \frac{y^2}{2}\right\} dy - r_n(0) \\ &\quad + z\sqrt{B_n} \int_0^\infty r_n(y) \exp\left\{-zy\sqrt{B_n}\right\} dy \quad (2.2.32) \\ &= \frac{1}{\sqrt{2\pi}} \int_0^\infty \exp\left\{-zy\sqrt{B_n} - \frac{y^2}{2}\right\} dy + \alpha_n, \end{aligned}$$

where $|\alpha_n| \leq \frac{C}{H_n \sqrt{B_n}}$.

Denote

$$I_1 = \int_0^\infty \exp\left\{-zy\sqrt{B_n} - \frac{y^2}{2}\right\} dy = \psi(z\sqrt{B_n})$$

and

$$I_2 = \int_0^\infty \exp\left\{-\frac{\bar{M}_n}{\sqrt{B_n}} - \frac{y^2}{2}\right\} dy = \psi(\bar{M}_n B_n^{-\frac{1}{2}}),$$

where

$$\psi(x) = \frac{1 - \Phi(x)}{\Phi'(x)} = e^{\frac{x^2}{2}} \int_x^\infty e^{-\frac{t^2}{2}} dt$$

is the Mills ratio which is known to satisfy

$$\frac{x}{x^2 + 1} < \psi(x) < \frac{1}{x},$$

for all $x > 0$. Hence, by (2.2.27) and (2.2.20) we obtain

$$\begin{aligned}
\frac{\alpha_n}{xI_1} &= \frac{\alpha_n z \sqrt{\overline{B}_n}}{x} + \frac{\alpha_n}{xz \sqrt{\overline{B}_n}} \\
&\leq C \left(\frac{z \sqrt{\overline{B}_n}}{x H_n \sqrt{\overline{B}_n}} + \frac{1}{H_n \sqrt{\overline{B}_n} x z \sqrt{\overline{B}_n}} \right) \\
&\leq C \left(\frac{1}{H_n \sqrt{\overline{B}_n}} + \frac{1}{H_n \sqrt{\overline{B}_n} x^2} \right) \\
&\leq \frac{C}{H_n \sqrt{\overline{B}_n}}.
\end{aligned}$$

Hence,

$$\alpha_n = I_1 O\left(\frac{x}{H_n \sqrt{\overline{B}_n}}\right). \quad (2.2.33)$$

For every $y_1 < y_2$ we have that $\psi(y_2) - \psi(y_1) = (y_2 - y_1)\psi'(u)$, where $y_1 < u < y_2$. As for $u > 0$, $|\psi'(u)| < u^{-2}$, then using (2.2.3), (2.2.27), (2.2.17), (2.2.18), (2.2.20) and (2.2.21) we get that

$$\begin{aligned}
|I_2 - I_1| &= \left| \psi'(u) \right| \left| \overline{M}_n B_n^{-\frac{1}{2}} - z \sqrt{\overline{B}_n} \right| \\
&\leq \frac{1}{u^2 \sqrt{\overline{B}_n}} \left| \overline{M}_n - z \sqrt{\overline{B}_n} \sqrt{\overline{B}_n} \right| \\
&\leq \frac{1}{u^2 \sqrt{\overline{B}_n}} \left(\left| \overline{M}_n - z B_n \right| + \left| z B_n - z \sqrt{\overline{B}_n} \sqrt{\overline{B}_n} \right| \right) \\
&\leq \frac{C}{(\frac{1}{4}x)^2 \sqrt{\overline{B}_n}} \left(\frac{z^2 C_n}{H_n^3} + z \sqrt{\overline{B}_n} \left| \sqrt{\overline{B}_n} - \sqrt{\overline{B}_n} \right| \right) \\
&\leq \frac{C}{x^2 \sqrt{\overline{B}_n}} \left(\frac{x^2 C_n}{B_n H_n^3} + \frac{x |B_n - \overline{B}_n|}{\sqrt{\overline{B}_n} + \sqrt{\overline{B}_n}} \right) \\
&\leq \frac{C}{x^2 \sqrt{\overline{B}_n}} \left(\frac{x^2 C_n}{B_n H_n^3} + \frac{x z C_n}{H_n^3 \sqrt{\overline{B}_n}} \right) \\
&\leq \frac{C}{x^2 \sqrt{\overline{B}_n}} \left(\frac{x^2 C_n}{B_n H_n^3} + \frac{x^2 C_n}{H_n^3 B_n} \right) = \frac{C C_n}{B_n^{\frac{3}{2}} H_n^3} \\
&\leq \frac{C}{H_n \sqrt{\overline{B}_n}}.
\end{aligned}$$

Hence,

$$\frac{|I_2 - I_1|}{xI_2} \leq \frac{C}{xH_n\sqrt{B_n}\psi(\overline{M}_n B_n^{-\frac{1}{2}})} = \frac{C}{xH_n\sqrt{B_n}\psi(x)} < \frac{C}{xH_n\sqrt{B_n}} \frac{x^2 + 1}{x} < \frac{C}{H_n\sqrt{B_n}},$$

which means that

$$I_1 = I_2 \left(1 + O\left(\frac{x}{H_n\sqrt{B_n}}\right) \right). \quad (2.2.34)$$

Finally, combining (2.2.30), (2.2.22), (2.2.23), (2.2.28), (2.2.32) and (2.2.33) we get

$$\begin{aligned} 1 - F_n(x) &= \exp \left\{ -\frac{H_n^2 B_n t^2}{2} + H_n^2 B_n t^3 \lambda_n(t) \right\} \int_0^\infty \exp \left\{ -zy\sqrt{B_n} \right\} d\overline{F}_n(y) \\ &= \exp \left\{ -\frac{x^2}{2} + \frac{x^3}{H_n\sqrt{B_n}} \lambda_n\left(\frac{x}{H_n\sqrt{B_n}}\right) \right\} \left(\frac{1}{\sqrt{2\pi}} I_1 + \alpha_n \right) \\ &= \exp \left\{ -\frac{x^2}{2} + \frac{x^3}{H_n\sqrt{B_n}} \lambda_n\left(\frac{x}{H_n\sqrt{B_n}}\right) \right\} \frac{1}{\sqrt{2\pi}} I_1 \left(1 + O\left(\frac{x}{H_n\sqrt{B_n}}\right) \right). \end{aligned}$$

By (2.2.34) and the fact that $I_2 = \psi(x)$, we see that

$$\frac{1 - F_n(x)}{1 - \Phi(x)} = \exp \left\{ \frac{x^3}{H_n\sqrt{B_n}} \lambda_n\left(\frac{x}{H_n\sqrt{B_n}}\right) \right\} \left(1 + O\left(\frac{x}{H_n\sqrt{B_n}}\right) \right).$$

This proves (2.2.4). The proof of (2.2.5) follows a same pattern and is omitted. \square

For the rest of the chapter, we only state the results for $x \geq 0$. Since $\lambda_n(t) = \sum_{k=0}^\infty \beta_{kn} t^k$ stays bounded uniformly in n for sufficiently small values of $|t|$ and $\beta_{0n} = \frac{H_n}{6B_n} \sum_{j \in \mathbb{Z}^d} \gamma_{3nj}$ from the proof of Theorem 2.2.1, we have the following corollary:

Corollary 2.2.2. *Assume the conditions of Theorem 2.2.1 hold. Then for $x \geq 0$ with $x = O\left((H_n\sqrt{B_n})^{1/3}\right)$ we have*

$$\frac{1 - F_n(x)}{1 - \Phi(x)} = \exp \left\{ \frac{x^3}{6B_n^{3/2}} \sum_{j \in \mathbb{Z}^d} \gamma_{3nj} \right\} \left(1 + O\left(\frac{x+1}{H_n\sqrt{B_n}}\right) \right).$$

Notice that $\frac{x^3}{6B_n^{3/2}} \sum_{j \in \mathbb{Z}^d} \gamma_{3nj} = O(1)$ under the condition $x = O\left((H_n \sqrt{B_n})^{1/3}\right)$. Also taking into the account the fact that for $x > 0$

$$1 - \Phi(x) < \frac{e^{-x^2/2}}{x\sqrt{2\pi}},$$

we obtain the following corollaries:

Corollary 2.2.3. *Under the conditions of Theorem 2.2.1, we have that for $x \geq 0$ with $x = O\left((H_n \sqrt{B_n})^{1/3}\right)$,*

$$1 - F_n(x) = \left(1 - \Phi(x)\right) \exp\left\{\frac{x^3}{6B_n^{3/2}} \sum_{j \in \mathbb{Z}^d} \gamma_{3nj}\right\} + O\left(\frac{e^{-x^2/2}}{H_n \sqrt{B_n}}\right).$$

Corollary 2.2.4. *Assume the conditions of Theorem 2.2.1 and $\sum_{j \in \mathbb{Z}^d} \gamma_{3nj} = 0$ for all $n \in \mathbb{N}$. Then for $x \geq 0$ with $x = O\left((H_n \sqrt{B_n})^{1/3}\right)$, we have*

$$F_n(x) - \Phi(x) = O\left(\frac{e^{-x^2/2}}{H_n \sqrt{B_n}}\right).$$

Also as $1 - \Phi(x) \sim \frac{1}{x\sqrt{2\pi}} e^{-x^2/2}$, as $x \rightarrow \infty$, we have

Corollary 2.2.5. *Under the conditions of Theorem 2.2.1, if $x \rightarrow \infty$, $x = o(H_n \sqrt{B_n})$, then*

$$\frac{F_n(x + \frac{c}{x}) - F_n(x)}{1 - F_n(x)} \rightarrow 1 - e^{-c}$$

for every positive constant c .

2.3 APPLICATIONS

In this section, we provide some applications of the main result in Section 2.2. First, we derive a moderate deviation result for linear random fields with short or long memory;

then we apply this result to risk measures and apply a same argument to study nonparametric regression.

2.3 Cramér type moderate deviation for linear random fields

Let $X = \{X_j, j \in \mathbb{Z}^d\}$ be a linear random field defined on a probability space (Ω, \mathcal{F}, P) by

$$X_j = \sum_{i \in \mathbb{Z}^d} a_i \varepsilon_{j-i}, \quad j \in \mathbb{Z}^d,$$

where the innovations $\varepsilon_i, i \in \mathbb{Z}^d$, are i.i.d. random variables with mean zero and finite variances σ^2 , and where $\{a_i, i \in \mathbb{Z}^d\}$ is a sequence of real numbers that satisfy $\sum_{i \in \mathbb{Z}^d} a_i^2 < \infty$.

Linear random fields have been studied extensively in probability and statistics. We refer to Sang and Xiao (2018) for a brief review on studies in limit theorems, large and moderate deviations for linear random fields and to Koul et al (2016), Lahiri and Robinson (2016) and the reference therein for recent developments in statistics.

By applying Theorem 2.2.1 in Section 2.2, we establish the following moderate deviation result for linear random fields with short or long memory, under Cramér's condition on the innovations $\varepsilon_i, i \in \mathbb{Z}^d$. Compared with the moderate deviation results in Sang and Xiao (2018), our Theorem 2.3.1 below gives more precise convergence rate which holds on much wider range for x .

Suppose that there is a disc centered at $z = 0$ within which the cumulant generating function $L(z) = L_{\varepsilon_i}(z) = \log \mathbb{E} e^{z\varepsilon_i}$ of ε_i is analytic and can be expanded in a convergent power series

$$L(z) = \sum_{k=1}^{\infty} \frac{\gamma_k}{k!} z^k,$$

where γ_k is the cumulant of order k of the random variables $\varepsilon_i, i \in \mathbb{Z}^d$. We have that $\gamma_1 = \mathbb{E} \varepsilon_i = 0$ and $\gamma_2 = \mathbb{E} \varepsilon_i^2 = \sigma^2, i \in \mathbb{Z}^d$.

We write

$$S_n = \sum_{j \in \Gamma_n^d} X_j = \sum_{j \in \mathbb{Z}^d} b_{nj} \varepsilon_j, \quad (2.3.1)$$

where $b_{nj} = \sum_{i \in \Gamma_n^d} a_{i-j}$. In the setting of Section 2.2, we have $X_{nj} = b_{nj} \varepsilon_j$, $j \in \mathbb{Z}^d$. Then it can be verified that for all $n \geq 1$ and $j \in \mathbb{Z}^d$, X_{nj} satisfy condition (2.2.1) for suitably chosen H_n . In the notation of Section 2.2, we have

$$B_n = \sigma^2 \sum_{j \in \mathbb{Z}^d} b_{nj}^2, \quad F_n(x) = P(S_n < x \sqrt{B_n}).$$

Hence, we can apply Theorem 2.2.1 to prove the following theorem. Here, as usual (see, e.g., [58]), we shall say that the function $l(\cdot) : [1, \infty) \rightarrow \mathbb{R}$ is a slowly varying function (at infinity) if l is a real-valued, positive and measurable function on $[1, \infty]$ and

$$\lim_{x \rightarrow +\infty} \frac{l(\lambda x)}{l(x)} = 1$$

for every $\lambda > 0$.

Theorem 2.3.1. *Assume that the linear random field $X = \{X_j, j \in \mathbb{Z}^d\}$ has short memory, i.e.,*

$$A := \sum_{i \in \mathbb{Z}^d} |a_i| < \infty, \quad a := \sum_{i \in \mathbb{Z}^d} a_i \neq 0, \quad (2.3.2)$$

or long memory with coefficients

$$a_i = l(|i|) b(i/|i|) |i|^{-\alpha}, \quad i \in \mathbb{Z}^d, |i| \neq 0, \quad (2.3.3)$$

where $\alpha \in (d/2, d)$ is a constant, $l(\cdot) : [1, \infty) \rightarrow \mathbb{R}$ is a slowly varying function at infinity and $b(\cdot)$ is a continuous function defined on the unit sphere \mathbb{S}_{d-1} . Suppose that there exist

positive constants H and C such that

$$|L(z)| < C \tag{2.3.4}$$

in the disc $|z| < H$. Then for all $x \geq 0$ with $x = o(n^{d/2})$, we have

$$\frac{1 - F_n(x)}{1 - \Phi(x)} = \exp \left\{ \frac{x^3}{n^{d/2}} \lambda_n \left(\frac{x}{n^{d/2}} \right) \right\} \left(1 + O \left(\frac{x+1}{n^{d/2}} \right) \right), \tag{2.3.5}$$

where

$$\lambda_n(t) = \sum_{k=0}^{\infty} \beta_{kn} t^k$$

is a power series that stays bounded uniformly in n for sufficiently small values of $|t|$ and the coefficients β_{kn} only depend on the cumulants of ε_i and on the coefficients a_i of the linear random field.

Proof. Since $\gamma_1 = 0$, we see that the cumulant generating function $L_{nj}(z)$ of the random variable $b_{nj}\varepsilon_j, j \in \mathbb{Z}^d$, is given by

$$L_{nj}(z) = \log \mathbb{E} e^{z b_{nj} \varepsilon_j} = \sum_{k=2}^{\infty} \frac{\gamma_k b_{nj}^k}{k!} z^k.$$

Cauchy's inequality for the derivatives of analytic functions together with the condition (2.3.4) yields that

$$|\gamma_k| < \frac{k!C}{H^k}. \tag{2.3.6}$$

Denote $M_n = \max_{j \in \mathbb{Z}^d} |b_{nj}|$. Then by (2.3.6), for any H_n with $0 < H_n \leq \frac{H}{2M_n}$ and for any z with $|z| < H_n$ we have

$$\begin{aligned}
|L_{nj}(z)| &\leq \sum_{k=2}^{\infty} \frac{|\gamma_k| |b_{nj}|^k}{k!} |z|^k \leq C \sum_{k=2}^{\infty} \frac{|b_{nj} H_n|^k}{H^k} \\
&= \frac{C}{H} \frac{b_{nj}^2 H_n^2}{H - |b_{nj} H_n|} \leq \frac{2C b_{nj}^2 H_n^2}{H^2}.
\end{aligned}$$

Hence,

$$C_n = \sum_{j \in \mathbb{Z}^d} \frac{2C b_{nj}^2 H_n^2}{H^2} = \frac{2C B_n H_n^2}{\sigma^2 H^2}.$$

Then by Theorem 2.2.1, if $B_n H_n^2 \rightarrow \infty$ as $n \rightarrow \infty$, we have

$$\frac{1 - F_n(x)}{1 - \Phi(x)} = \exp \left\{ \frac{x^3}{H_n \sqrt{B_n}} \lambda_n \left(\frac{x}{H_n \sqrt{B_n}} \right) \right\} \left(1 + O \left(\frac{x+1}{H_n \sqrt{B_n}} \right) \right) \quad (2.3.7)$$

for $x \geq 0, x = o(H_n \sqrt{B_n})$.

If the linear random field has long memory then we have that (see Surgailis (1982), Theorem 2) $B_n \propto n^{3d-2\alpha} l^2(n)$. As the function $b(\cdot)$ is bounded, then for $j \in \Gamma_n^d$ we have

$$\begin{aligned}
|b_{nj}| &\leq C_1 \sum_{i \in \Gamma_n^d} l(|i-j|) |i-j|^{-\alpha} \\
&\leq C_1 \sum_{k=1}^{2dn} k^{d-1} l(k) k^{-\alpha} \propto n^{d-\alpha} l(n),
\end{aligned}$$

where we have used the fact (see Bingham et al. (1987) or Seneta (1976)) that for a slowly varying function $l(x)$ defined on $[1, \infty)$ and for any $\theta > -1$,

$$\int_1^x y^\theta l(y) dy \sim \frac{x^{\theta+1} l(x)}{\theta+1}, \quad \text{as } x \rightarrow \infty.$$

It follows from the definition of a_i in (2.3.3) that (for sufficiently large n) $M_n = \max_{j \in \mathbb{Z}^d} |b_{nj}|$ is attained at some $j \in \Gamma_n^d$. Hence, $M_n = O(n^{d-\alpha} l(n))$. We take $H_n \propto n^{-d+\alpha} l^{-1}(n)$ which

yields

$$H_n \sqrt{B_n} \propto n^{d/2}.$$

Then the result follows from (2.3.7).

If the linear random field has short memory, i.e., $A := \sum_{i \in \mathbb{Z}^d} |a_i| < \infty$, $a := \sum_{i \in \mathbb{Z}^d} a_i \neq 0$, we can take $M_n = A$ and $H_n = \frac{H}{2A}$. Moreover, we also have

$$\sum_{j \in \mathbb{Z}^d} |b_{nj}| \leq \sum_{j \in \mathbb{Z}^d} \sum_{i \in \Gamma_n^d} |a_{i-j}| = (2n+1)^d \sum_{i \in \mathbb{Z}^d} |a_i| = A(2n+1)^d$$

and

$$\sum_{j \in \mathbb{Z}^d} |b_{nj}| \geq \left| \sum_{j \in \mathbb{Z}^d} \sum_{i \in \Gamma_n^d} a_{i-j} \right| = (2n+1)^d \left| \sum_{i \in \mathbb{Z}^d} a_i \right| = |a|(2n+1)^d,$$

which means that $\sum_{j \in \mathbb{Z}^d} |b_{nj}| \propto n^d$.

As for all $n \in \mathbb{N}$ we have that $|b_{nj}| \leq A$ by the definition of A , then

$$\sum_{j \in \mathbb{Z}^d} b_{nj}^2 \leq A \sum_{j \in \mathbb{Z}^d} |b_{nj}| \leq A^2(2n+1)^d.$$

On the other hand, for $j \in \Gamma_{\lfloor n/2 \rfloor}^d$ we have that $|b_{nj}| > |a|/2$ for sufficiently large n . Hence,

$$\sum_{j \in \mathbb{Z}^d} b_{nj}^2 \geq \sum_{j \in \Gamma_{\lfloor n/2 \rfloor}^d} b_{nj}^2 \geq \frac{a^2}{4} (2 \lfloor n/2 \rfloor + 1)^d.$$

Thus, $\sum_{j \in \mathbb{Z}^d} b_{nj}^2 \propto n^d$ and the result follows from (2.3.7). □

To the best of our knowledge, Theorem 2.3.1 is the first result that gives the exact tail probability for partial sums of random fields with dependence structure under the Cramér condition.

Due to its preciseness, Theorem 2.3.1 can be applied to evaluate the performance of approximation of the distribution of linear random fields by truncation. We often use the random variable $X_j^m = \sum_{i \in \Gamma_m^d} a_i \varepsilon_{j-i}$ with finite terms to approximate the linear random field

$X_j = \sum_{i \in \mathbb{Z}^d} a_i \varepsilon_{j-i}$ in practice. For example, the moving average with finite terms $MA(m)$ is applied to approximate the linear process (moving average with infinite terms). In this case, Theorem 2.3.1 also applies to the partial sum $S_n^m = \sum_{j \in \Gamma_n^d} X_j^m = \sum_{j \in \mathbb{Z}^d} b_{nj}^m \varepsilon_j$. Here only finite terms b_{nj}^m are non-zero. Denote

$$B_n^m = \sigma^2 \sum_{j \in \mathbb{Z}^d} (b_{nj}^m)^2, \quad F_n^m(x) = P(S_n^m < x \sqrt{B_n^m}).$$

Then for all $x \geq 0$ with $x = o(n^{d/2})$, we have

$$\frac{1 - F_n^m(x)}{1 - \Phi(x)} = \exp \left\{ \frac{x^3}{n^{d/2}} \lambda_n^m \left(\frac{x}{n^{d/2}} \right) \right\} \left(1 + O \left(\frac{x+1}{n^{d/2}} \right) \right),$$

where

$$\lambda_n^m(t) = \sum_{k=0}^{\infty} \beta_{kn}^m t^k,$$

and where the coefficients β_{kn}^m have similar definition as β_{kn} . To see the difference between the two tail probabilities of the partial sums, we have

$$\begin{aligned} \frac{1 - F_n(x)}{1 - F_n^m(x)} &= \exp \left\{ \frac{x^3}{n^{d/2}} \left[\lambda_n \left(\frac{x}{n^{d/2}} \right) - \lambda_n^m \left(\frac{x}{n^{d/2}} \right) \right] \right\} \left(1 + O \left(\frac{x+1}{n^{d/2}} \right) \right) \\ &= \exp \left\{ \frac{x^3}{n^{d/2}} \left[\beta_{0n} - \beta_{0n}^m + \sum_{k=1}^{\infty} (\beta_{kn} - \beta_{kn}^m) \left(\frac{x}{n^{d/2}} \right)^k \right] \right\} \left(1 + O \left(\frac{x+1}{n^{d/2}} \right) \right), \end{aligned}$$

here as in the proof of Theorem 2.3.1, we take $M_n = \max_{j \in \mathbb{Z}^d} |b_{nj}|$, $H_n = \frac{H}{2M_n}$, $M_n^m = \max_{j \in \mathbb{Z}^d} |b_{nj}^m|$, $H_n^m = \frac{H}{2M_n^m}$,

$$\beta_{0n} = \frac{H_n}{6B_n} \sum_{j \in \mathbb{Z}^d} \gamma_{3nj} = \frac{H\gamma_3}{12M_n B_n} \sum_{j \in \mathbb{Z}^d} (b_{nj})^3,$$

$$\beta_{0n}^m = \frac{H_n^m}{6B_n^m} \sum_{j \in \mathbb{Z}^d} \gamma_{3nj}^m = \frac{H\gamma_3}{12M_n^m B_n^m} \sum_{j \in \mathbb{Z}^d} (b_{nj}^m)^3.$$

If $\gamma_3 \neq 0$, $\frac{1-F_n(x)}{1-F_n^m(x)}$ is dominated by $\exp\left\{\frac{x^3}{n^{d/2}}(\beta_{0n} - \beta_{0n}^m)\right\}$. If $\gamma_3 = 0$, then $\beta_{0n} = \beta_{0n}^m = 0$ and $\frac{1-F_n(x)}{1-F_n^m(x)}$ can be dominated by $\exp\left\{\frac{x^4}{n^d}(\beta_{1n} - \beta_{1n}^m)\right\}$ which depends on whether $\gamma_4 = 0$. In general, Theorem 2.3.1 can be applied to evaluate whether the truncated version X_j^m is a good approximation to X_j in terms of the ratio $\frac{1-F_n(x)}{1-F_n^m(x)}$ for x in different ranges which depends on the property of the innovation ε and the sequence $\{a_i, i \in \mathbb{Z}^d\}$.

Theorem 2.3.1 can be applied to calculate the tail probability of the partial sum of some well-known dependent models. For example, the autoregressive fractionally integrated moving average FARIMA(p, β, q) processes in one dimensional case introduced by Granger and Joyeux (1980) and Hosking (1981), which is defined as

$$\phi(B)X_n = \theta(B)(1 - B)^{-\beta}\varepsilon_n.$$

Here p, q are nonnegative integers, $\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p$ is the AR polynomial and $\theta(z) = 1 + \theta_1 z + \dots + \theta_q z^q$ is the MA polynomial. Under the conditions that $\phi(z)$ and $\theta(z)$ have no common zeros, the zeros of $\phi(\cdot)$ lie outside the closed unit disk and $-1/2 < \beta < 1/2$, the FARIMA(p, β, q) process has linear process form $X_n = \sum_{i=0}^{\infty} a_i \varepsilon_{n-i}$, $n \in \mathbb{N}$, with $a_i = \frac{\theta(1)}{\phi(1)} \frac{i^{\beta-1}}{\Gamma(\beta)} + O(i^{-1})$. Here $\Gamma(\cdot)$ is the gamma function.

2.3 Approximation of risk measures

Theorem 2.3.1 can be applied to approximate the risk measures such as quantiles and tail conditional expectations for the partial sums S_n in (2.3.1) of linear random field $X = \{X_j, j \in \mathbb{Z}^d\}$. Given the tail probability $\alpha \in (0, 1)$, let $Q_{\alpha, n}$ be the upper α -th quantile of S_n . Namely $P(S_n \geq Q_{\alpha, n}) = \alpha$. By Theorem 2.3.1, for all $x \geq 0$ with $x = o(n^{d/2})$,

$$P(S_n > x\sqrt{B_n}) = \exp\left\{\frac{x^3}{n^{d/2}}\lambda_n\left(\frac{x}{n^{d/2}}\right)\right\}(1 - \Phi(x))(1 + o(1)).$$

We approximate $Q_{\alpha,n}$ by $x_\alpha\sqrt{B_n}$, where $x = x_\alpha = o(n^{d/2})$ can be solved numerically from the equation

$$\exp\left\{\frac{x^3}{n^{d/2}}\lambda_n\left(\frac{x}{n^{d/2}}\right)\right\}(1 - \Phi(x)) = \alpha.$$

The tail conditional expectation is computed as

$$\begin{aligned} E(S_n|S_n \geq Q_{\alpha,n}) &= \frac{Q_{\alpha,n}P(S_n \geq Q_{\alpha,n}) + \int_{Q_{\alpha,n}}^\infty P(S_n \geq w)dw}{P(S_n \geq Q_{\alpha,n})} \\ &= Q_{\alpha,n} + \frac{\sqrt{B_n}}{\alpha} \int_{Q_{\alpha,n}/\sqrt{B_n}}^\infty \exp\left\{\frac{y^3}{n^{d/2}}\lambda_n\left(\frac{y}{n^{d/2}}\right)\right\}(1 - \Phi(y))dy, \end{aligned}$$

which can be solved numerically. The quantile and tail conditional expectation, which are also called value at risk (VaR) or expected shortfall (ES) in finance and risk theory, are important measures to model the extremal behavior of random variables in practice. The precise moderate deviation results in this article provide a vehicle in the computation of these two measures of time series or spacial random fields. See Peligrad et al (2014a) for a brief review of VaR and ES in the literature and a study of them when a linear process has p -th moment ($p > 2$) or has a regularly varying tail with exponent $t > 2$.

2.3 Nonparametric regression

Consider the following regression model

$$Y_{n,j} = g(z_{n,j}) + X_{n,j}, \quad j \in \Gamma_n^d,$$

where g is a bounded continuous function on \mathbb{R}^m , $z_{n,j}$'s are the fixed design points over $\Gamma_n^d \subseteq \mathbb{Z}^d$ with values in a compact subset of \mathbb{R}^m , and $X_{n,j} = \sum_{i \in \mathbb{Z}^d} a_i \varepsilon_{n,j-i}$ is a linear random field over \mathbb{Z}^d , where the i.i.d. innovations $\varepsilon_{n,i}$ satisfy the same conditions as in Subsection 2.3.1. Regression models with independent or weakly dependent random field

errors have been well-studied in the literature, see, e.g., El Machkoui (2007), El Machkouri and Stoica (2010), Hallin et al (2004).

The kernel regression estimator for the function g on the basis of sample pairs $(z_{n,j}, Y_{n,j})$, $j \in \Gamma_n^d$, is

$$g_n(z) = \sum_{j \in \Gamma_n^d} w_{n,j}(z) Y_{n,j},$$

where the weight functions $w_{n,j}(\cdot)$'s on \mathbb{R}^m have form

$$w_{n,j}(z) = \frac{K\left(\frac{z - z_{n,j}}{h_n}\right)}{\sum_{i \in \Gamma_n^d} K\left(\frac{z - z_{n,i}}{h_n}\right)}.$$

Here $K : \mathbb{R}^m \rightarrow \mathbb{R}^+$ is a kernel function and h_n is a sequence of bandwidths which goes to zero as $n \rightarrow \infty$. Notice that the weight functions satisfy the condition $\sum_{j \in \Gamma_n^d} w_{n,j}(z) = 1$.

For fixed $z \in \mathbf{R}^m$, let $S_n := g_n(z) - \mathbb{E}g_n(z)$. Then it can be written as

$$S_n = \sum_{j \in \Gamma_n^d} w_{n,j}(z) X_{n,j} = \sum_{j \in \mathbb{Z}^d} b_{n,j} \varepsilon_{n,j},$$

where $b_{n,j} = \sum_{i \in \Gamma_n^d} w_{n,i}(z) a_{i-j}$. Let $B_n = \sigma^2 \sum_{j \in \mathbb{Z}^d} b_{n,j}^2$, $M_n = \max_{j \in \mathbb{Z}^d} |b_{n,j}|$. Assume that the innovations $\varepsilon_{n,i}$ satisfy the Cramér's condition (2.2.1) with $H_n \propto M_n^{-1}$. By the same analysis as in the proof of Theorem 2.3.1, if $B_n H_n^2 \rightarrow \infty$ as $n \rightarrow \infty$, $x \geq 0$, $x = o(H_n \sqrt{B_n})$, we derive a moderate deviation result for $S_n = g_n(z) - \mathbb{E}g_n(z)$ that is similar to (2.3.7). This result can be applied to quantify the convergence rate of $S_n = g_n(z) - \mathbb{E}g_n(z) \rightarrow 0$, as $n \rightarrow \infty$.

2.4 CONCLUSION

Under the condition proposed by Cramér we have obtained moderate deviation theorem for random fields. The indices of elements of the considered random fields belong to $\mathbb{N} \times \mathbb{Z}^d$. Obtained result is consistent with the classical Cramér -Petrov theorem for the partial sum of a sequence of independent and identically distributed (iid) random variables and is also applicable to wide class of non iid random variables. In particular, applying the

main theorem 2.2.1 we have obtained moderate deviation theorem 2.3.1 for linear random fields with short or long memory. Theorem 2.3.1 can further be applied to calculate the tail probability of the partial sum of certain dependent models such as the autoregressive fractionally integrated moving average FARIMA(p, β, q) processes. Theorem 2.3.1 can be also applied to approximate the risk measures such as quantiles and tail conditional expectations of time series or spacial random fields.

3 ON MUTUAL INFORMATION ESTIMATION FOR MIXED-PAIR RANDOM VARIABLES

3.1 INTRODUCTION

The entropy of a discrete random variable $X \in \mathbb{R}^d$ with countable support $\{x_1, x_2, \dots\}$ and $p_i = \mathbb{P}(X = x_i)$ is defined to be

$$H(X) = - \sum_i p_i \log p_i,$$

and the (differential) entropy of a continuous random variable $Y \in \mathbb{R}^d$ with probability density function $f(y)$ is defined as

$$H(Y) = - \int_{\mathbb{R}^d} f(y) \log f(y) dy.$$

If $d \geq 2$, $H(X)$ or $H(Y)$ is also called the joint entropy of the components in X or Y . Entropy is a measure of distribution uncertainty and naturally it has application in the fields of information theory, statistical classification, pattern recognition and so on.

From the definition of entropy it follows that the entropy of a discrete random variable X having probability mass function $p(x)$ is the expected value of $-\log p(X) = \log \frac{1}{p(X)}$. Analogously, the (differential) entropy of a continuous random variable Y with probability density function $f(y)$ is the expected value of $-\log f(X) = \log \frac{1}{f(X)}$. We will use this interpretation of entropy later on to define the corresponding estimator for a mutual information. Before that let us first recall the definition and some basic properties of mutual information.

Let P_X, P_Y be probability measures on some arbitrary measure spaces \mathcal{X} and \mathcal{Y} respectively. Let P_{XY} be the joint probability measure on the space $\mathcal{X} \times \mathcal{Y}$. If P_{XY} is absolutely continuous with respect to the product measure $P_X \times P_Y$, let $\frac{dP_{XY}}{d(P_X \times P_Y)}$ be the Radon-Nikodym derivative. Then the general definition of the mutual information (e.g., [15]) is given by

$$I(X, Y) = \int_{\mathcal{X} \times \mathcal{Y}} dP_{XY} \log \frac{dP_{XY}}{d(P_X \times P_Y)}. \quad (3.1.1)$$

If two random variables X and Y are either both discrete or both continuous then the mutual information of X and Y can be expressed in terms of entropies as

$$I(X, Y) = H(X) + H(Y) - H(X, Y), \quad (3.1.2)$$

where

$$H(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y)$$

is the joint entropy of X and Y with $p(x, y)$ being the joint density function of X and Y . Given the above definitions of entropy and mutual information it is possible to derive certain very natural properties of them. In particular, for random variables X and Y , which are either both discrete or both of them are continuous, we have that

- $H(X) \geq 0$ if X has discrete distribution;
- $H(X + c) = H(X)$;
- $H(aX) = H(X) + \log|a|$ if X is continuous;
- $H(aX) = H(X)$ if X is discrete, $a \neq 0$;
- $H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$
- $I(X, Y) = I(Y, X) = H(Y) - H(Y|X)$;
- $I(X, X) = H(X)$;

- $I(X, Y) \geq 0$ with equality if and only if X and Y are independent;
- Among PDFs $f(x)$ with support $[a, b]$, the maximum-entropy distribution is the uniform distribution;
- The Gaussian distribution has maximum entropy relative to all distributions over $(-\infty, \infty)$ with finite mean μ and finite variance σ^2 .

Property $H(X, Y) = H(X) + H(Y|X)$ is called the chain rule of entropy.

However, in practice and application, we often need to work on a mixture of continuous and discrete random variables. There are several ways for the mixture. 1). One random variable X is discrete and the other random variable Y is continuous; 2). A random variable Z has both discrete and continuous components, i.e., $Z = X$ with probability p and $Z = Y$ with probability $1 - p$, where $0 < p < 1$, X is a discrete random variable and Y is a continuous random variable; 3). a random vector with each dimension component being discrete, continuous or mixture as in 2).

In [46], the authors extend the definition of the joint entropy for the first case mixture, i.e., for the pair of random variables, where the first random variable is discrete and the second one is continuous. Our goal is to study the mutual information for that case and provide the estimation of the mutual information from a given i.i.d. sample $\{X_i, Y_i\}_{i=1}^N$.

In [15], the authors applied the k -nearest neighbor method to estimate the Radon-Nikodym derivative and, therefore, to estimate the mutual information for all three mixed cases. In the literature, if the random variables X and Y are either both discrete or both continuous, the estimation of mutual information is usually performed by the estimation of the three entropies in (3.1.2). The estimation of a differential entropy has been well studied. An incomplete list of the related research includes the nearest-neighbor estimator [34], [63], [40]; the kernel estimator [1], [27], [22], [23] and the orthogonal projection estimator [36], [37]. [7] studied the plug-in entropy estimator for the finite value discrete case and obtained the mean, the variance and the central limit theorem of this estimator. [66] studied the

coverage-adjusted entropy estimator with unobserved values for the infinite value discrete case.

3.2 MAIN RESULTS

Consider a random vector $Z = (X, Y)$. We call Z a mixed-pair if $X \in \mathbb{R}$ is a discrete random variable with countable support $\mathcal{X} = \{x_1, x_2, \dots\}$ while $Y \in \mathbb{R}^d$ is a continuous random variable. Observe that $Z = (X, Y)$ induces measures $\{\mu_1, \mu_2, \dots\}$ that are absolutely continuous with respect to the Lebesgue measure, where $\mu_i(A) = \mathbb{P}(X = x_i, Y \in A)$, for every Borel set A in \mathbb{R}^d . There exists a non-negative function $g(x, y)$ with $h(x) := \int_{\mathbb{R}^d} g(x, y) dy$ be the probability mass function on \mathcal{X} and $f(y) := \sum_i g_i(y)$ be the marginal density function of Y . Here, $g_i(y) = g(x_i, y)$, $i \in \mathbb{N}$. In particular, denote $p_i = h(x_i)$, $i \in \mathbb{N}$. We have that

$$f_i(y) = \frac{1}{p_i} g_i(y)$$

is the probability density function of Y conditioned on $X = x_i$. In [46], the authors gave the following regulation of mixed-pair and then defined the joint entropy of a mixed-pair.

Definition 3.2.1. (*Good mixed-pair*). *A mixed-pair random variables $Z = (X, Y)$ is called good if the following condition is satisfied:*

$$\int_{\mathcal{X} \times \mathbb{R}^d} |g(x, y) \log g(x, y)| dx dy = \sum_i \int_{\mathbb{R}^d} |g_i(y) \log g_i(y)| dy < \infty.$$

Essentially, we have a good mixed-pair random variables when restricted to any of the X values, the conditional differential entropy of Y is well-defined.

Definition 3.2.2. (*Entropy of a mixed-pair*). *The entropy of a good mixed-pair random variable is defined by*

$$H(Z) = - \int_{\mathcal{X} \times \mathbb{R}^d} g(x, y) \log g(x, y) dx dy = - \sum_i \int_{\mathbb{R}^d} g_i(y) \log g_i(y) dy.$$

As $g_i(y) = p_i f_i(y)$ then we have that

$$\begin{aligned}
H(Z) &= - \sum_i \int_{\mathbb{R}^d} g_i(y) \log g_i(y) dy \\
&= - \sum_i \int_{\mathbb{R}^d} p_i f_i(y) \log p_i f_i(y) dy \\
&= - \sum_i p_i \log p_i \int_{\mathbb{R}^d} f_i(y) dy - \sum_i p_i \int_{\mathbb{R}^d} f_i(y) \log f_i(y) dy \\
&= - \sum_i p_i \log p_i - \sum_i p_i \int_{\mathbb{R}^d} f_i(y) \log f_i(y) dy \\
&= H(X) + \sum_i p_i H(Y|X = x_i).
\end{aligned} \tag{3.2.3}$$

We take the convention $\log 0 = 0$ and $\log 0/0 = 0$. From the general formula of the mutual information (3.1.1), we get that

$$\begin{aligned}
I(X, Y) &= \int_{\mathcal{X} \times \mathbb{R}^d} g(x, y) \log \frac{g(x, y) dx dy}{h(x) f(y) dx dy} dx dy \\
&= \sum_i \int_{\mathbb{R}^d} g_i(y) \log \frac{g_i(y)}{p_i f(y)} dy \\
&= \sum_i \int_{\mathbb{R}^d} g_i(y) \log g_i(y) dy - \sum_i \int_{\mathbb{R}^d} g_i(y) \log p_i dy - \sum_i \int_{\mathbb{R}^d} g_i(y) \log f(y) dy \\
&= \sum_i \int_{\mathbb{R}^d} p_i f_i(y) \log [p_i f_i(y)] dy - \sum_i p_i \log p_i \int_{\mathbb{R}^d} f_i(y) dy - \int_{\mathbb{R}^d} f(y) \log f(y) dy \\
&= \sum_i p_i \log p_i \int_{\mathbb{R}^d} f_i(y) dy + \sum_i p_i \int_{\mathbb{R}^d} f_i(y) \log f_i(y) dy - \sum_i p_i \log p_i - \int_{\mathbb{R}^d} f(y) \log f(y) dy \\
&= -H(Z) + H(X) + H(Y) = H(Y) - \sum_i p_i H(Y|X = x_i) := H(Y) - \sum_i I_i.
\end{aligned} \tag{3.2.4}$$

Let $(X, Y), (X_1, Y_1), \dots, (X_N, Y_N)$ be a random sample drawn from a mixed distribution with discrete component having support $\{0, 1, \dots, m\}$, and let $p_i = \mathbb{P}(X = i)$, $0 \leq i \leq m$ with $0 < p_i < 1, \sum p_i = 1$. Also suppose that the continuous component has pdf

$f(y)$. Denote $\hat{p}_i = \sum_{k=1}^N \mathbb{I}(X_k = i)/N$, $0 \leq i \leq m$, and let

$$\begin{aligned} \bar{I}_i &= -\hat{p}_i \left[N\hat{p}_i \right]^{-1} \sum_{k=1}^N \mathbb{I}(X_k = i) \log f_i(Y_k) \\ &= -N^{-1} \sum_{k=1}^N \mathbb{I}(X_k = i) \log f_i(Y_k) \end{aligned} \tag{3.2.5}$$

and

$$\bar{H}(Y) = -N^{-1} \sum_{k=1}^N \log f(Y_k) \tag{3.2.6}$$

be the estimators of $I_i = p_i H(Y|X = i)$, $0 \leq i \leq m$, and $H(Y)$ respectively, where $f_i(y)$ is the probability density function of Y conditioned on $X = i$, $0 \leq i \leq m$. Denote $a = (1, -1, \dots, -1)^\top$. Let Σ be the covariance matrix of $(\log f(Y), \mathbb{I}(X = 0) \log f_0(Y), \dots, \mathbb{I}(X = m) \log f_m(Y))^\top$.

Theorem 3.2.1. $a^\top \Sigma a > 0$ if and only if X and Y are dependent. For the estimator

$$\bar{I}(X, Y) = \bar{H} - \sum_{i=0}^m \bar{I}_i \tag{3.2.7}$$

of $I(X, Y)$ we have that

$$\sqrt{N}(\bar{I}(X, Y) - I(X, Y)) \rightarrow N(0, a^\top \Sigma a) \tag{3.2.8}$$

given that X and Y are dependent. Furthermore, the variance $a^\top \Sigma a$ can be calculated by

$$\begin{aligned}
a^\top \Sigma a &= \text{var}(\log f(Y)) + \sum_{i=0}^m p_i E_i[\log f_i(Y)]^2 - \sum_{i=0}^m p_i^2 (E_i[\log f_i(Y)])^2 \\
&\quad - 2 \sum_{i=0}^m p_i [E_i \log f_i(Y) \log f(Y) - E_i \log f_i(Y) E \log f(Y)] \\
&\quad - 2 \sum_{0 \leq i < j \leq m} p_i p_j [E_i \log f_i(Y)] [E_j \log f_j(Y)],
\end{aligned} \tag{3.2.9}$$

where E_i is the conditional expectation of Y given $X = i$, $0 \leq i \leq m$.

Proof. First of all, $a^\top \Sigma a \geq 0$ since Σ is the variance covariance matrix. If $a^\top \Sigma a = 0$ then

$$\text{var} \left(\log f(Y) - \sum_{i=0}^m \mathbb{I}(X = i) \log f_i(Y) \right) = a^\top \Sigma a = 0$$

and $\log f(Y) - \sum_{i=0}^m \mathbb{I}(X = i) \log f_i(Y) \equiv C$ for some constant C . But

$$\log f(Y) - \sum_{i=0}^m \mathbb{I}(X = i) \log f_i(Y) = \sum_{i=0}^m \mathbb{I}(X = i) \log \frac{f(Y)}{f_i(Y)}.$$

Hence $\log \frac{f(Y)}{f_i(Y)} \equiv C$. Then $f_i(y) = cf(y)$ for some constant $c > 0$ and for all $0 \leq i \leq m$.

But $f(y) = \sum_{i=0}^m p_i f_i(y) = cf(y) \sum_{i=0}^m p_i = cf(y)$. Hence, $c \equiv 1$ and $f_i(y) = f(y)$ for all $0 \leq i \leq m$. Then X and Y are independent. On the other hand, if X and Y are independent, then $f_i(y) = f(y)$ for all $0 \leq i \leq m$. Therefore, $\log f(Y) - \sum_{i=0}^m \mathbb{I}(X = i) \log f_i(Y) = 0$ and $a^\top \Sigma a = 0$. Hence, $a^\top \Sigma a = 0$ if and only if X and Y are independent.

Notice that the vector $(\bar{H}(Y), \bar{I}_0, \dots, \bar{I}_m)^\top$ is the sample mean of a sequence of i.i.d. random vectors

$$\{(\log f(Y_k), \mathbb{I}(X_k = 0) \log f_0(Y_k), \dots, \mathbb{I}(X_k = m) \log f_m(Y_k))^\top\}_{k=1}^N$$

with mean $(H(Y), I_0, \dots, I_m)^\top$. Then, by central limit theorem, we have

$$\sqrt{N} \left(\begin{pmatrix} \bar{H} \\ \bar{I}_0 \\ \vdots \\ \bar{I}_m \end{pmatrix} - \begin{pmatrix} H \\ I_0 \\ \vdots \\ I_m \end{pmatrix} \right) \rightarrow N(\bar{0}, \Sigma),$$

and, given $a^\top \Sigma a > 0$, we have (3.2.8). By the formula for variance decomposition, we have

$$\begin{aligned} & \text{var}(\mathbb{I}(X = i) \log f_i(Y)) \\ &= E\{\text{var}[\mathbb{I}(X = i) \log f_i(Y)|X]\} + \text{var}\{E[\mathbb{I}(X = i) \log f_i(Y)|X]\} \\ &= E\{\mathbb{I}(X = i) \text{var}[\log f_i(Y)|X]\} + \text{var}\{\mathbb{I}(X = i) E[\log f_i(Y)|X]\} \\ &= E\left\{\mathbb{I}(X = i) \sum_{j=0}^m \text{var}_j(\log f_j(Y)) \mathbb{I}(X = j)\right\} \\ & \quad + \text{var}\left\{\mathbb{I}(X = i) \sum_{j=0}^m E_j(\log f_j(Y)) \mathbb{I}(X = j)\right\} \\ &= \text{var}_i[\log f_i(Y)] E\{\mathbb{I}(X = i)\} + (E_i[\log f_i(Y)])^2 \text{var}\{\mathbb{I}(X = i)\} \\ &= p_i \text{var}_i[\log f_i(Y)] + (p_i - p_i^2) (E_i[\log f_i(Y)])^2 \\ &= p_i E_i[\log f_i(Y)]^2 - p_i^2 (E_i[\log f_i(Y)])^2, \end{aligned} \tag{3.2.10}$$

$0 \leq i \leq m$. Here var_i is the conditional variance of Y when $X = i$, $0 \leq i \leq m$. By similar calculation,

$$\begin{aligned} & \text{Cov}\left(\mathbb{I}(X = i) \log f_i(Y), \mathbb{I}(X = j) \log f_j(Y)\right) \\ &= -p_i p_j [E_i \log f_i(Y)] [E_j \log f_j(Y)], \end{aligned} \tag{3.2.11}$$

for all $0 \leq i < j \leq m$, and

$$\begin{aligned} & \text{Cov} \left(\mathbb{I}(X = i) \log f_i(Y), \log f(Y) \right) \\ &= p_i [E_i \log f_i(Y) \log f(Y) - E_i \log f_i(Y) E \log f(Y)]. \end{aligned} \quad (3.2.12)$$

Thus, the covariance matrix Σ of $(\log f(Y), \mathbb{I}(X = 0) \log f_0(Y), \dots, \mathbb{I}(X = m) \log f_m(Y))^\top$ and therefore $a^\top \Sigma a$ can be calculated by the above calculation (3.2.10)-(3.2.12). We then have (3.2.9). \square

We consider the case when the random variables X and Y are dependent. Note that in this case $a^\top \Sigma a > 0$ and we have (3.2.8). However, $\bar{I}(X, Y)$ is not a practical estimator since the density functions involved are not known.

Now let $K(\cdot)$ be a kernel function in \mathbb{R}^d and let h be the bandwidth. Then

$$\hat{f}_{ik}(y) = \left\{ (N\hat{p}_i - 1)h^d \right\}^{-1} \sum_{j \neq k} \mathbb{I}(X_j = i) K\{(y - Y_j)/h\}$$

are the “leave-one-out” estimators of the functions f_i , $0 \leq i \leq m$, and

$$\hat{I}_i = -N^{-1} \sum_{k=1}^N \mathbb{I}(X_k = i) \log \hat{f}_{ik}(Y_k) \quad (3.2.13)$$

are estimators of $I_i = p_i H(Y|X = i)$, $0 \leq i \leq m$. Also

$$\hat{H} = -N^{-1} \sum_{k=1}^N \log \hat{f}_k(Y_k) \quad (3.2.14)$$

is an estimator of $H(Y)$, where

$$\begin{aligned}
\hat{f}_k(y) &= \left\{ (N-1)h^d \right\}^{-1} \sum_{j \neq k} K\{(y - Y_j)/h\} \\
&= \left\{ (N-1)h^d \right\}^{-1} \sum_{j \neq k} \left[\sum_{i=0}^m \mathbb{I}(X_k = i) \right] K\{(y - Y_j)/h\} \\
&= \sum_{i=0}^m \frac{N\hat{p}_i - 1}{N-1} \hat{f}_{ik}(y).
\end{aligned} \tag{3.2.15}$$

Theorem 3.2.2. *Assume that the tails of f_0, \dots, f_m are decreasing like $|x|^{-\alpha_0}, \dots, |x|^{-\alpha_m}$, respectively, as $|x| \rightarrow \infty$. Also assume that the kernel function has appropriately heavy tails as in [22]. If $h = o(N^{-1/8})$ and $\alpha_0, \dots, \alpha_m$ are all greater than $7/3$ in the case $d = 1$, greater than 6 in the case $d = 2$ and greater than 15 in the case $d = 3$, then for the estimator*

$$\hat{I}(X, Y) = \hat{H} - \sum_{i=0}^m \hat{I}_i, \tag{3.2.16}$$

we have

$$\sqrt{N}(\hat{I}(X, Y) - I(X, Y)) \rightarrow N(0, a^\top \Sigma a). \tag{3.2.17}$$

Proof. Under the conditions in the theorem, applying the formula (3.1) or (3.2) from [23], we have

$$\hat{H} = \bar{H} + o(N^{-1/2}), \quad \hat{I}_0 = \bar{I}_0 + o(N^{-1/2}), \dots, \quad \hat{I}_m = \bar{I}_m + o(N^{-1/2}).$$

Together with Theorem 3.2.1, we have (3.2.17). □

We may take the probability density function of Student- t distribution with proper degree of freedom instead of the normal density function as the kernel function. On the other hand, if X and Y are independent then $I(X, Y) = \bar{I}(X, Y) = 0$ and we have that $\hat{I}(X, Y) = o(N^{-1/2})$.

3.3 SIMULATION STUDY

In this section we conduct a simulation study with $m = 1$, i.e., the random variable X takes two possible values 0 and 1, to confirm the main results stated in (3.2.17) for the kernel mutual information estimation of good mixed-pairs. First we study some one dimensional examples. Let $t(\nu, \mu, \sigma)$ be the Student t distribution with degree of freedom ν , location parameter μ and scale parameter σ and let $pareto(x_m, \alpha)$ be the Pareto distribution with density function $f(x) = \alpha x_m^\alpha x^{-(\alpha+1)} \mathbb{I}(x \geq x_m)$. We study the mixture for the following four cases: 1). $t(3, 0, 1)$ and $t(12, 0, 1)$; 2). $t(3, 0, 1)$ and $t(3, 2, 1)$; 3). $t(3, 0, 1)$ and $t(3, 0, 3)$; 4). $pareto(1, 2)$ and $pareto(1, 10)$. For each case, $p_0 = 0.3$ for the first distribution and $p_1 = 0.7$ for the second distribution.

The second row of Table 3.3.1 lists the mathematica calculation of the mutual information (MI) as stated in (3.2.4) for each case. The third row of Table 3.3.1 gives the average of 400 estimates based on formula (3.2.16). For each estimate, we use the probability density function of the Student t distribution with degree of freedom 3, i.e. $t(3, 0, 1)$, as the kernel function. We also have simulation study with kernel functions satisfying the conditions in the main results and obtained similar results. We take $h = N^{-1/5}$ as the bandwidth for the first three cases and $h = N^{-1/5}/24$ for the last case. The data size for each estimate is $N = 50,000$ in each case. The Pareto distributions $pareto(1, 2)$ and $pareto(1, 10)$ have very dense area on the right of 1. This is the reason that we take a relatively small bandwidth for this case. To apply the kernel method in estimation, one should select an optimal bandwidth based on some criteria, for example, to minimize the mean squared error. It is interesting to investigate the bandwidth selection problem from both theoretical and application viewpoints. However, it seems that the study in this direction is very difficult. We leave it as an open question for future study. It is clear that the average of the estimates matches the true value of mutual information.

We apply mathematica to calculate the covariance matrix Σ of

$$(\log f(Y), \mathbb{I}(X = 0) \log f_0(Y), \mathbb{I}(X = 1) \log f_1(Y))^T$$

and, therefore, the value of $a^T \Sigma a$ for each case. The values of $a^T \Sigma a$ are 0.02189236, 0.3092179, 0.1540501 and 0.2748102 respectively for the four cases. The fourth row of Table 3.3.1 lists the values of $(a^T \Sigma a / N)^{1/2}$ which serves as the asymptotic approximation of the standard deviation of the estimator $\hat{I}(X, Y)$ in the central limit theorem (3.2.17). The last row gives the sample standard deviation from $M = 400$ estimates. These two values also have good match.

mixture	$t(3, 0, 1)$	$t(3, 0, 1)$	$t(3, 0, 1)$	$pareto(1, 2)$
	$t(12, 0, 1)$	$t(3, 2, 1)$	$t(3, 0, 3)$	$pareto(1, 10)$
MI	0.011819	0.20023	0.102063	0.201123
mean of estimates	0.01167391	0.1991132	0.1014199	0.2010447
$(a^T \Sigma a / N)^{1/2}$	0.0006617	0.0025	0.0018	0.0023
sample sd	0.0006616724	0.002345997	0.001819982	0.002349275

Table 3.3.1: True value of the mutual information and the mean value of the estimates for Pareto and t-distributions.

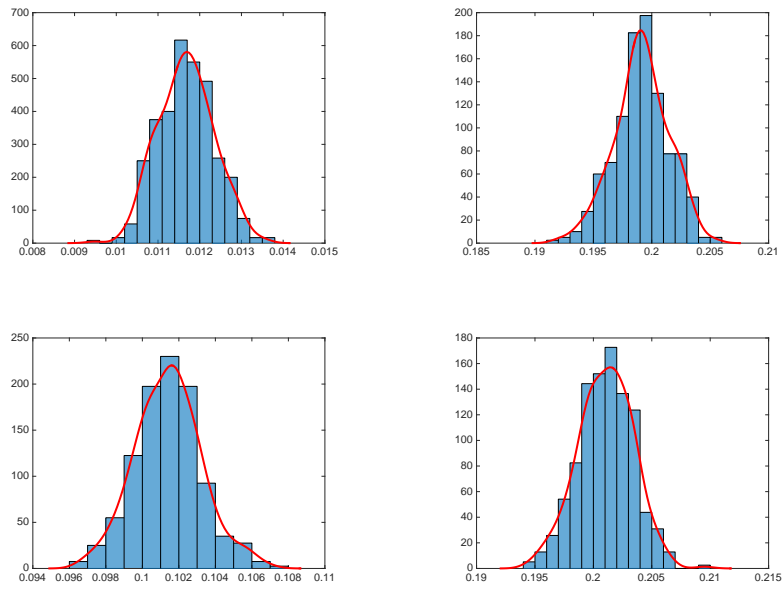


Figure 3.3.1: The histograms with kernel density fits of $M = 400$ estimates. Top left: $t(3, 0, 1)$ and $t(12, 0, 1)$. Top right: $t(3, 0, 1)$ and $t(3, 2, 1)$. Bottom left: $t(3, 0, 1)$ and $t(3, 0, 3)$. Bottom right: $\text{pareto}(1, 2)$ and $\text{pareto}(1, 10)$.

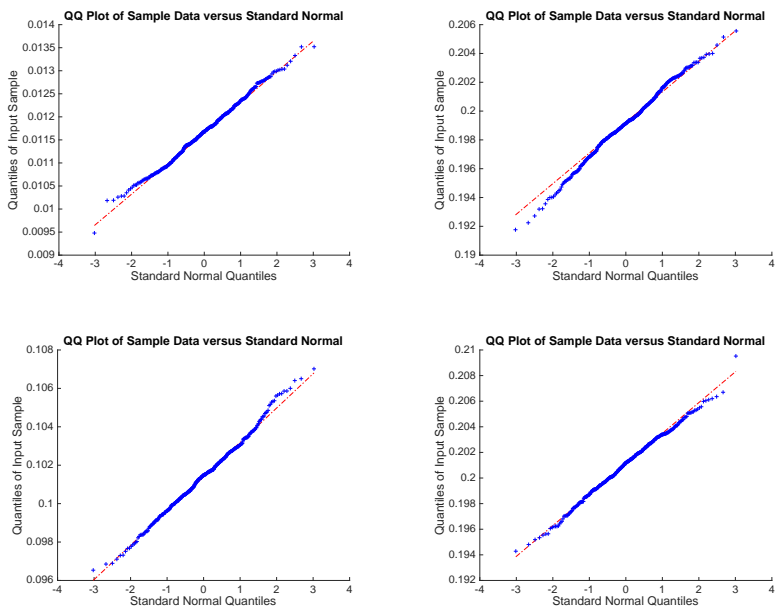


Figure 3.3.2: The Q-Q plots of $M = 400$ estimates. Top left: $t(3, 0, 1)$ and $t(12, 0, 1)$. Top right: $t(3, 0, 1)$ and $t(3, 2, 1)$. Bottom left: $t(3, 0, 1)$ and $t(3, 0, 3)$. Bottom right: $pareto(1, 2)$ and $pareto(1, 10)$.

Figure 3.3.1 and 3.3.2 show the histograms with kernel density fits and normal Q-Q plots of 400 estimates for each case. It is clear that the values of $\hat{I}(X, Y)$ follow a normal distribution.

We study two examples in the two dimensional case. Let $t_\nu(\mu, \Sigma_0)$ be the two dimensional Student t distribution with degree of freedom ν , mean μ and shape matrix Σ_0 . We study the mixture in two cases: 1). $t_5(0, I)$ and $t_{25}(0, I)$; 2). $t_5(0, I)$ and $t_5(0, 3I)$. Here I is the identity matrix. For each case, $p_0 = 0.3$ for the first distribution and $p_1 = 0.7$ for the second distribution. Table 3.3.2 summarizes 200 estimates of the mutual information with $h = N^{-1/5}$ and sample size $N = 50,000$ for each estimate. We take $t_3(0, I)$ as the kernel function. Same as the one dimensional case, we apply mathematica to calculate the

true value of MI and $(a^\top \Sigma a/N)^{1/2}$ which is given in formula (3.2.9). Figure 3.3.3 shows the histograms with kernel density fits and normal Q-Q plots of 200 estimates for each example. It is clear that the values of $\hat{I}(X, Y)$ also follow a normal distribution in the two dimensional case. In summary, the simulation study confirms the central limit theorem as stated in (3.2.17).

mixture	$t_5(0, I)$ $t_{25}(0, I)$	$t_5(0, I)$ $t_5(0, 3I)$
MI	0.01158	0.202516
mean of estimates	0.0112381	0.2022715
$(a^\top \Sigma a/N)^{1/2}$	0.0006577826	0.002312909
sample sd	0.0008356947	0.002315134

Table 3.3.2: True value of the mutual information and the mean value of the estimates for t-distributions.

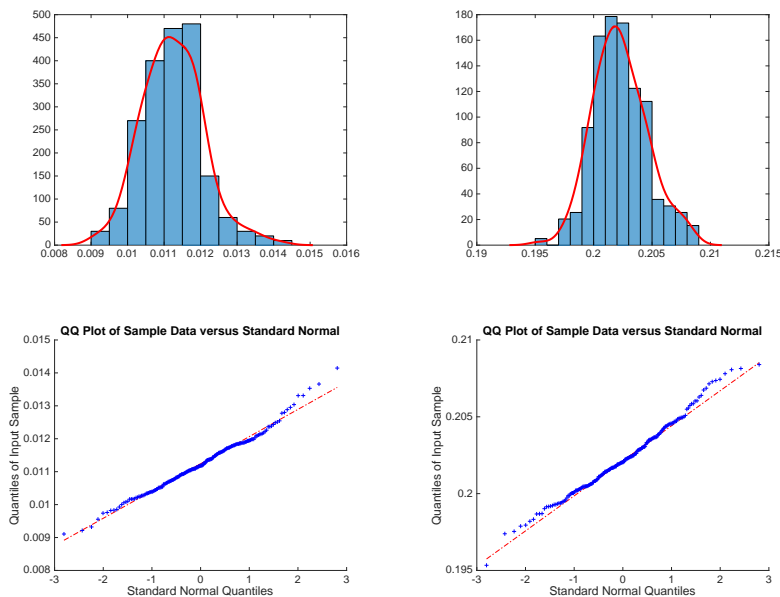


Figure 3.3.3: The histograms and Q-Q plots of $M = 200$ estimates. Left: $t_5(0, I)$ and $t_{25}(0, I)$. Right: $t_5(0, I)$ and $t_5(0, 3)$.

3.4 CONCLUSION

We considered a pair of random variables one of which is discrete and the other one is continuous. In classical case, that is, in case when both random variables are of same type, either discrete or continuous, the estimation of mutual information of two random variables is usually done by applying the so-called 3H formula which allows to represent the mutual information in terms of 3 entropies. Therefore, in order to estimate the mutual information for the mixed-pair random variables we start from the most general formula for mutual information and prove that a formula analogous to the classical 3H formula also holds for this case and it involves an entropy of the so called good mixed-pair random variables introduced in [46]. Then, after representing the mutual information of mixed-pair random variables in terms of entropies we constructed estimators for each of those entropies and, thus, we obtained an estimator $\bar{I}(X, Y)$ for the mutual information $I(X, Y)$

given by 3.2.7, We have then shown that the Central Limit Theorem holds for the estimator $\bar{I}(X, Y)$ and we have also derived the explicit formula for the variance of the limiting normal distribution. Next we obtained an estimator $\hat{I}(X, Y)$ of $I(X, Y)$ given by 3.2.16. Compared to the estimator $\bar{I}(X, Y)$ the estimator $\hat{I}(X, Y)$ is much more practical in the sense that it only depends on the given data and on the choice of the kernel function. Then, under the assumption that the tails of conditional density functions $f_i(y)$ of $Y|X = x_i$ are decreasing sufficiently fast at infinity and that the kernel function has appropriately heavy tail we have shown that the estimator $\hat{I}(X, Y)$ also enjoys Central Limit Theorem. Finally, we conducted simulation study with discrete component X taking the values 0 and 1 with probabilities 0.3 and 0.7 respectively, and with continuous component Y following t -distribution and Pareto distribution with different parameters. Our kernel function is the density function of t -distribution with 3 degrees of freedom. Then, as the tables 3.3.1 and 3.3.2 demonstrate the obtained values closely match the corresponding theoretical true values.

BIBLIOGRAPHY

Bibliography

- [1] Ahmad, I. A. and Lin, P. E. 1976. A nonparametric estimation of the entropy for absolutely continuous distributions. *IEEE Trans. Information Theory*. **22**, 372-375.
- [2] Amosova, N. N., 1979. On probabilities of moderate deviations for sums of independent random variables. *Teor. Veroyatn. Primen.* **24**, 858–865.
- [3] Asmussen, S. and Albrecher, H., 2010. *Ruin Probabilities*. World Scientific, Hackensack, NJ.
- [4] Babu, G. J. and Singh, K., 1978a. Probabilities of moderate deviations for some stationary strong-mixing processes. *Sankhyā Ser. A* **40**, 38–43.
- [5] Babu, G. J. and Singh, K., 1978b. On probabilities of moderate deviations for dependent processes. *Sankhyā Ser. A* **40**, 28–37.
- [6] Bahadur, R. and Rao, R. R., 1960. On deviations of the sample mean. *Ann. Math. Statist.* **31**, 1015–1027.
- [7] Basharin, G. P. 1959. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability and Its Applications*. **4**, 333-336.
- [8] Bingham, N. H., Goldie, C. M. and Teugels, J. L., 1987. *Regular Variation*. Cambridge University Press, Cambridge, UK.
- [9] Cramér H., 1938. *Sur un nouveau théorème-limite de la théorie des probabilités*, Actual. Sci. et Ind., Paris, 736.
- [10] El Machkouri, M., 2007. Nonparametric regression estimation for random fields in a fixed-design. *Stat. Inference Stoch. Process.* **10**, 29–47.
- [11] El Machkouri, M. and Stoica, R., 2010. Asymptotic normality of kernel estimates in a regression model for random fields. *J. Nonparametr. Stat.* **22**, 955–971.
- [12] Fan, X., Grama, I. G. and Liu, Q., 2013. Cramér large deviation expansions for martingales under Bernstein’s condition. *Stoch. Process. Appl.* **123**, 3919–3942.
- [13] Feller W., 1943. Generalization of a probability limit theorem of Cramér. *Trans. Amer. Math. Soc.* **54**, 361–372.
- [14] Frolov, A. N., 2005. On probabilities of moderate deviations of sums for independent random variables. *J. Math. Sci. (New York)* **127**, 1787–1796.

- [15] Gao, W., Kannan, S., Oh, S. and Viswanath, P. 2017. Estimating mutual information for discrete-continuous mixtures. *Advances in Neural Information Processing Systems*. 5988-5999.
- [16] Ghosh, M., 1974. Probabilities of moderate deviations under m-dependence. *Canad. J. Statist.* **2**, 157–168.
- [17] Ghosh, M. and Babu, G. J., 1977. Probabilities of moderate deviations for some stationary ϕ -mixing processes. *Ann. Probab.* **5**, 222–234.
- [18] Grama, I. G., 1997. On moderate deviations for martingales. *Ann. Probab.* **25**, 152–183.
- [19] Grama, I. G. and Haeusler, E., 2000. Large deviations for martingales via Cramér’s method. *Stoch. Process. Appl.* **85**, 279–293.
- [20] Grama, I. G. and Haeusler, E., 2006. An asymptotic expansion for probabilities of moderate deviations for multivariate martingales. *J. Theoret. Probab* **19**, 1–44.
- [21] Granger, C. and Joyeux, R. 1980. An introduction to long memory time series models and fractional differencing. *J. Time Series Anal.* **1** 15–29.
- [22] Hall, P. 1987. On Kullback-Leibler Loss and Density Estimation. *Ann. Statist.* 15, no. 4, 1491-1519.
- [23] Hall, P. and Morton, S. 1993. On the estimation of entropy. *Ann. Inst. Statist. Math.* **45**, 69-88.
- [24] Hallin, M., Lu, Z. and Tran, L. T., 2004. Local linear spatial regression. *Ann. Statist.* **32**, 2469–2500.
- [25] Heinrich, L., 1990. Some bounds of cumulants of m -dependent random fields. *Math. Nachr.* **149**, 303–317.
- [26] Hosking, J. R. M. 1981. Fractional differencing. *Biometrika* **68** 165–176.
- [27] Joe, H. 1989. On the estimation of entropy and other functionals of a multivariate density. *Ann. Inst. Statist. Math.* **41**, 683-697.
- [28] Johnstone, I. M., 1999. Wavelet shrinkage for correlated data and inverse problems: adaptivity results. *Statist. Sinica* **9**, 51–83.
- [29] Joutard, C., 2006. Sharp large deviations in nonparametric estimation. *J. Nonparametr. Stat.* **18**, 293–306.
- [30] Joutard, C., 2013. Strong large deviations for arbitrary sequences of random variables. *Ann. Inst. Statist. Math.* **65**, 49–67.
- [31] Jurecková, J., Kallenberg, W. C. M., Veraverbeke, N., 1988. Moderate and Cramér-type large deviation theorems for M -estimators. *Statistics & probability letters* **6**, 191-199.

- [32] Khinchin, A. I., 1929. Über einen neuen Grenzwertsatz der Wahrscheinlichkeitsrechnung. (in German), *Math. Ann.* **101**, 745–752.
- [33] Koul, H. L., Mimoto, N. and Surgailis, D., 2016. A goodness-of-fit test for marginal distribution of linear random fields with long memory. *Metrika* **79**, 165–193.
- [34] Kozachenko, L. F. and Leonenko, N. N. 1987. Sample estimate of entropy of a random vector. *Problems of Information Transmission*, **23**, 95-101.
- [35] Lahiri, S. N. and Robinson, P. M., 2016. Central limit theorems for long range dependent spatial linear processes. *Bernoulli* **22**, 345–375.
- [36] Laurent, B. 1996. Efficient estimation of integral functionals of a density. *Ann. Statist.* **24**, 659-681.
- [37] Laurent, B. 1997. Estimation of integral functionals of a density and its derivatives. *Bernoulli* **3**, 181-211.
- [38] Lee, S.-H., Tan, V. Y. F. and Khisti, A., 2016. Streaming data transmission in the moderate deviations and central limit regimes. *IEEE Trans. Inform. Theory* **62**, 6816–6830.
- [39] Lee, S.-H., Tan, V. Y. F. and Khisti, A., 2017. Exact moderate deviation asymptotics in streaming data transmission. *IEEE Trans. Inform. Theory* **63**, 2726–2736.
- [40] Leonenko, N., Pronzato, L. and Savani, V. 2008. A class of Rényi information estimators for multidimensional densities. *Ann. Statist.* **36**, 2153–2182. Corrections, *Ann. Statist.* **38** (2010), 3837-3838.
- [41] Malevich, T.L., Abdalimov, B., 1979. Probabilities of large deviations for U -statistics. *Theory of Prob. and Appl.* **24**, 215-219.
- [42] Michel, R., 1976. Nonuniform central limit bounds with applications to probabilities of deviations. *Ann. Probab.* **4**, 102–106.
- [43] Mörters, P. 2008. Large deviation theory and applications. Available online at: <https://people.bath.ac.uk/maspm/LDP.pdf>
- [44] Nagaev, S. V., 1965. Some limit theorems for large deviations. *Teor. Veroyatn. Primen.* **10**, 231–254.
- [45] Nagaev, S. V., 1979. Large deviations of sums of independent random variables. *Ann. Probab.* **7**, 745–789.
- [46] Nair, C., Prabhakar, B. and Shah, D. On entropy for mixtures of discrete and continuous variables. arXiv:cs/0607075
- [47] Peligrad, M., Sang, H., Zhong, Y. and Wu, W. B., 2014a. Exact moderate and large deviations for linear processes. *Statist. Sinica* **24**, 957–969.

- [48] Peligrad, M., Sang, H., Zhong, Y. and Wu, W. B., 2014b. Supplementary material for the paper “Exact moderate and large deviations for linear processes”. *Statist. Sinica*, 15 pp, available online at: <http://www3.stat.sinica.edu.tw/statistica/>
- [49] Petrov V. V., 1954. A generalization of the Cramér limit theorem. (in Russian), *Uspehi Matem. Nauk* **9**, 195–202.
- [50] Petrov V. V., 1965. On the probabilities of large deviations for sums of independent random variables. *Theory Probab. Appl.* **10(2)**, 287–298.
- [51] Petrov V. V., 1975. *Sums of Independent Random Variables*. Springer-Verlag.
- [52] Petrov V. V., 1995. *Limit Theorems of Probability Theory*. Oxford University Press, Oxford.
- [53] Petrov V. V., Robinson J., 2006. On large deviations for sums of independent random variables. Available online at: <http://www.maths.usyd.edu.au/u/pubs/publist/preprints/2007/petrov-2.pdf>
- [54] Picard, D. and Tribouley, K., 2000. Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* **28**, 298–335.
- [55] Privalov I. I., 1984. *Introduction to the Theory of Functions of a Complex Variable*. (in Russian), 13th Ed., Nauka, Moscow.
- [56] Rubin, H. and Sethuraman, J., 1965. Probabilities of moderate deviations. *Sankhyā Ser. A* **27**, 325–346.
- [57] Sang, H. and Xiao, Y., 2018. Exact moderate and large deviations for linear random fields. *J. Appl. Probab.* **55(2)**, 431–449.
- [58] Seneta, E., 1976. *Regularly Varying Functions*. Lecture Notes in Mathematics **508**, Springer, Berlin.
- [59] Seoh, M., Ralescu, S. S., Puri, M. L., 1985. Cramér Type Large Deviations for Generalized Rank Statistics. *Ann. Probab.* **13**, 115–125.
- [60] Slastnikov, A. D., 1978. Limit theorems for probabilities of moderate deviations. *Teor. Veroyatn. Primen.* **24**, 340–357.
- [61] Statulevičius V.A., 1966. On large deviations. *Z. Wahrsch. verw. Gebiete* **6**, 133–144.
- [62] Surgailis, D., 1982. Zones of attraction of self-similar multiple integrals. *Lith Math J* **22** 185–201.
- [63] Tsybakov, A. B. and van der Meulen, E. C. 1994. Root-n consistent estimators of entropy for densities with unbounded support. *Scand. J. Statist.*, **23**, 75–83.
- [64] van der Vaart, A. W., 1998. *Asymptotic Statistics*. Cambridge University Press, Cambridge.

- [65] Vandemaële, M., Veraverbeke, N., 1982. Cramér Type Large Deviations for Linear Combinations of Order Statistics. *Ann. Probab.* **10**, 423-434.
- [66] Vu, V. Q., Yu, B. and Kass, R. E. 2007. Coverage-adjusted entropy estimation. *Statist. Med.*, **26**, 4039-4060.
- [67] Wu, W. B. and Zhao, Z., 2008. Moderate deviations for stationary processes. *Statistica Sinica* **18**, 769–782.
- [68] Zhang, S. and Wong, M., 2003. Wavelet threshold estimation for additive regression models. *Ann. Statist.* **31**, 152–173.

VITA

Aleksandr Beknazaryan was born in Hrazdan, Armenia, on October 16, 1989. He received Bachelor's degree in Mathematics in 2010 from Yerevan State University, a Master's degree in Mathematics from Yerevan State University in 2012, a Master of Advanced Study degree in Mathematics from the University of Cambridge in 2013 and a PhD in Mathematics from Kazan State Power Engineering University in 2015. He is currently a PhD student in Mathematics with concentration in Statistics at the University of Mississippi.