# Geometry of high dimensional Gaussian data 

Department of Mathematics, Linköping University<br>Olof Samuel Mossberg

LiTH-MAT-EX-2023/13-SE

Credits: 16 hp
Level: G2
Supervisor: Martin Singull,
Department of Mathematics, Linköping University
Examiner: Xiangfeng Yang,
Department of Mathematics, Linköping University
Linköping: March 2024

## Abstract

Collected data may simultaneously be of low sample size and high dimension. Such data exhibit some geometric regularities consisting of a single observation being a rotation on a sphere, and a pair of observations being orthogonal.

This thesis investigates these geometric properties in some detail. Background is provided and various approaches to the result are discussed. An approach based on the mean value theorem is eventually chosen, being the only candidate investigated that gives explicit convergence bounds. The bounds are tested employing Monte Carlo simulation and found to be adequate.

## Keywords:

HDLSS, high dimensional data, stochastic boundedness, asymptotic orthogonality, geometry, multivariate normal distribution.

## URL for electronic version:

http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-201265

## Sammanfattning

Data som insamlas kan samtidigt ha en liten stickprovsstorlek men vara högdimensionell. Sådan data uppvisar vissa geometriska mönster som består av att en enskild observation är en rotation på en sfär, och att ett par av observationer är rätvinkliga.

Den här uppsatsen undersöker dessa geometriska egenskaper mer detaljerat. En bakgrund ges och olika typer av angreppssätt diskuteras. Till slut väljs en metod som baseras på medelvärdessatsen eftersom detta är den enda av de undersökta metoderna som ger explicita konvergensgränser. Gränserna testas sedermera med Monte Carlo-simulering och visar sig stämma.

## Nyckelord:

HDLSS, högdimensionell data, stokastisk begränsning, asymptotisk ortogonalitet, geometri, multivariat normalfördelning.

## URL för elektronisk version:

http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-201265

## Acknowledgements

I would like to thank Martn Singull for suggesting a thesis topic that turned out to be interesting, and for his comments on how to improve typography and text. I would also like to thank Xiangfang Yang for his initial help with the thesis and for his role as the examiner.

I would also like to thank and commend Felix Faber for being a stimulating conversationalist, in both private and mathematical matters.

## Contents

1 Preliminaries ..... 1
1.1 Introduction ..... 1
1.2 Aims and outline ..... 2
2 Theoretical background ..... 5
2.1 Definitions ..... 5
2.2 Some fundamental results ..... 7
2.2.1 Taylor's theorem ..... 7
2.2.2 Mean value theorem ..... 8
2.2.3 A note on the mean value theorem ..... 8
2.2.4 Markov's inequality ..... 10
2.2.5 Law of large numbers ..... 11
2.2.6 Continuous mapping theorem ..... 11
2.2.7 Slutsky's theorem ..... 12
2.2.8 Delta method ..... 12
3 Main results ..... 13
3.1 Examples ..... 13
3.2 Euclidean norm of high dimensional data ..... 15
3.3 Pairwise Euclidean distance of high dimensional data ..... 17
3.4 Pairwise orthogonality of high dimensional data ..... 18
4 Simulations ..... 23
4.1 Initial simulations ..... 23
4.2 Simulations testing the bounds for the Eulidean norm ..... 25
4.3 Simulations testing the bounds for pairwise orthogonality ..... 27
5 Discussion ..... 29
Mossberg, 2024. ..... ix
A Simulation code ..... 33
A. 1 Initial simulations code ..... 33
A. 2 Euclidean norm code ..... 34
A. 3 Pairwise orthogonality code ..... 35

## Chapter 1

## Preliminaries

This chapter contains a brief description of the concept of High Dimensional Low Sample Size Statistics (HDLSS) and its applications. The chapter also includes the aims and outline of this thesis.

### 1.1 Introduction

HDLSS is the study of a particular type of data set. These data sets are high dimensional in the sense that a data point is a vector in $\mathbb{R}^{d}$, where $d$ is large in comparison to the sample size $n$ of the data set [6.

Multidimensional data can be represented as matrix [9, where the row vectors correspond to independent observations and each column vector corresponds to a variable,

$$
\mathbf{X}_{n \times d}=\left[\begin{array}{ccc}
x_{11} & \ldots & x_{1 d} \\
\vdots & \ddots & \\
x_{n 1} & & x_{n d}
\end{array}\right]
$$

Many traditional statistical applications involve a scenario of many measurements of one variable. Many measurements of one variable are known as a "large sample size". Often deployed statistical methods involve the asymptotic behaviour of a statistic as the sample size grows without limit, i.e., is "large" [3]. An example of statistical analysis utilizing "large" sample sizes is the central limit theorem, where the asymptotic properties enable statistical inference. For the matrix $\mathbf{X}$ above, this type of application reduce the matrix to a column vector $\mathbf{X}_{n \times 1}$.

By contrast, the asymptotic behaviour in HDLSS derives from the number of variables being large. In the limiting case of one high-dimensional observation, the matrix reduces to a row vector $\mathbf{X}_{1 \times d}$. As it happens, these row vectors display some geometrical regularities.

Roughly speaking, independent $d$-dimensional normal vectors $Z_{i}$ will be distributed on a sphere of radius $\sqrt{d}$ and the observations will be orthogonal to each other, i.e., $\left\|Z_{1}\right\|=\left\|Z_{2}\right\|=\sqrt{d}$ and $Z_{1} \perp Z_{2}$ for independent highdimensional $Z_{1}, Z_{2}$.

While the HDLSS configuration may appear exotic, they are widely used in statistics and machine learning. One example is gene expression. So-called Genome Wide Association Studies may involve measuring tens of thousands of genes. However, until recently, a problem was that collecting and sequencing the genome was expensive, thus limiting sample sizes to perhaps a few dozen people [1].

Another example is medical imaging which involves a large number of measurements. The data gathered is a picture that can be represented as a high dimensional vector, but the sample size may be limited due to costs. In other cases, cost is not a limiting factor; rather, what is measured is only available in a limited quantity. Spectral measurements in chemometrics represent one such example, where one measurement yields information about many spectrum channels [6].

### 1.2 Aims and outline

This thesis concerns the two geometrical properties of HDLSS data, that for $Z_{1}, Z_{2}$ d-dimensional multivariate normal random vectors we have that
i) $\frac{\left\|Z_{1}\right\|}{\sqrt{d}}=1$ in some sense as $d \longrightarrow \infty$,
ii) $Z_{1}$ and $Z_{2}$ are asymptotically orthogonal in some sense as $d \longrightarrow \infty$.

The precise statement of claims i) and ii) will be made later, but claim i) will be of the form

$$
\left\|Z_{i}\right\|=\sqrt{d}+\text { "bounded random variation" }
$$

and in a similar fashion claim ii) will be of the form

$$
\operatorname{ang}\left(Z_{1}, Z_{2}\right)=\frac{\pi}{2}+\text { "bounded random variation", }
$$

where "bounded random variation" will be assigned an exact meaning. This thesis aims to prove i) and ii). The outline of this thesis is as follows.

Chapter 2 contains definitions and some fundamental results from analysis and probability theory. Here, the exact meaning of "some bounded variation" - commonly referred to as "stochastic boundedness" - will be defined. The chapter also includes a detailed discussion of the mean value theorem.

Chapter 3 contains a preliminary excursion that consists of a few examples using some of the tools from Chapter 2 that make i) and ii) plausible. The main result of this thesis is contained in the latter part, where the claims i) and ii) are precisely formulated, demonstrated and explicit bounds for the stochastic boundedness are derived.

Chapter 4 contains some simulations to demonstrate the validity of claims i) and ii) and also contains simulations to corroborate the explicit bounds derived in Chapter 3 .

Chapter 5 contains a brief discussion regarding the results.

## Chapter 2

## Theoretical background

This chapter will revise some fundamental definitions and results from analysis and probability theory, with particular emphasis on a different perspective on the mean value theorem.

These definitions and results will then be utilized in Chapter 3 to corroborate, precisely formulate, and ultimately demonstrate claims i) and ii) from Section 1.2 .

### 2.1 Definitions

In this section, some basic notions from vector geometry and probability theory are defined. First, we define the concept of the transpose matrix, which allows convenient and compact notation and definition for future operations.

Definition 1 (Transpose). The transpose $\mathbf{M}^{\prime}$ of a matrix $\mathbf{M} \in \mathbb{R}^{n \times d}$ is defined as the new matrix

$$
\begin{equation*}
\left(\mathbf{M}^{\prime}\right)_{i j}=\mathbf{M}_{j i} \tag{2.1}
\end{equation*}
$$

With the transpose defined, the Euclidean norm of a vector has a compact definition.

Definition 2 (Euclidean norm). The Euclidean norm, or Euclidean distance from the origin, of a column vector $v \in \mathbb{R}^{d}$, is defined as

$$
\begin{equation*}
\|v\|=\sqrt{v^{\prime} v} \tag{2.2}
\end{equation*}
$$

Henceforth all vectors will be understood to be column vectors throughout this thesis.

Another central concept in vector geometry is the dot product.
Definition 3 (Dot product). The dot product of two vectors $u, v \in \mathbb{R}^{d}$ is defined as

$$
\begin{equation*}
u \cdot v=u^{\prime} v \tag{2.3}
\end{equation*}
$$

Remark. Definition 3 is identical to $\|u\|\|v\| \cos \theta$ in the standard basis in $\mathbb{R}^{d}$, where $\theta$ is the angle between $u$ and $v$. A consequence of this is that two non-zero vectors are perpendicular if and only if their dot product is zero [7.

Next, let us recall some basic concepts from probability. A random variable is a real-valued function defined on some sample space $\Omega$.

A continuous random variable has a probability density function. The probability density function $f_{X}$ of a random variable is the function that satisfies

$$
\mathrm{P}(X \in A)=\int_{A} f_{X}(x) d x
$$

where $A \subseteq \mathcal{X}$, and $\mathcal{X}=\operatorname{Im}(X)$ is the image of $X$.
Finally, the so-called expectation of a random variable will be central to our reasoning. The expectation of a random variable $X$ is denoted by $\mathrm{E}[X]$ and for a continuous random variable $X$ is calculated from the relationship

$$
\mathrm{E}[X]=\int_{\mathcal{X}} x f_{X}(x) d x
$$

The precise distribution of the HDLSS random variables we will consider, is the multivariate normal.

Definition 4 (Multivariate normal density). The random vector $X=\left(\xi_{1}, \ldots, \xi_{d}\right)^{\prime}$ has a $d$-dimensional multivariate normal distribution, denoted $X \sim \mathcal{N}_{d}(\mu, \boldsymbol{\Sigma})$, if its density is

$$
\begin{equation*}
f_{X}(x)=\frac{1}{(2 \pi)^{d / 2}|\boldsymbol{\Sigma}|^{1 / 2}} e^{-\frac{1}{2}(x-\mu)^{\prime} \boldsymbol{\Sigma}^{-1}(x-\mu)} \tag{2.4}
\end{equation*}
$$

where $|\cdot|$ denotes the determinant of a matrix and the matrix $\boldsymbol{\Sigma}$ is defined by

$$
(\boldsymbol{\Sigma})_{i, j}=\mathrm{E}\left[\xi_{i}-\mu_{i}\right]\left[\xi_{j}-\mu_{j}\right]=\operatorname{Cov}\left(\xi_{i}, \xi_{j}\right) .
$$

The matrix $\boldsymbol{\Sigma}$ has to be positive definite, i.e., the quadratic form $x^{\prime} \boldsymbol{\Sigma} x$ has to be positive for any nonzero $x \in \mathbb{R}^{d}$.
Remark. Later we will utilize the standard $d$-dimensional multivariate normal density. This is not a restriction, since $X \sim \mathcal{N}_{d}(\mu, \boldsymbol{\Sigma})$ can be transformed to the coordinates

$$
Z=\boldsymbol{\Sigma}^{-1 / 2}(X-\mu),
$$

which has the Jacobian determinant $\left|\mathbf{J}_{Z}\right|=|\boldsymbol{\Sigma}|^{1 / 2}$. It then follows from equation (2.4) and the transformation theorem that

$$
f_{Z}(z)=\frac{1}{(2 \pi)^{d / 2}} e^{-\frac{1}{2} z^{\prime} z}
$$

or equivalently, $Z \sim \mathcal{N}_{d}(0, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix [5].
We will also require a sense of convergence of random variables, while there are several such notions available, we will use what is called "convergence in probability".

Definition 5 (Convergence in probability). A sequence of random variables $\left\{X_{n}\right\}_{n=1}^{\infty}$ is said to converge in probability to another random variable $X$, denoted $X_{n} \xrightarrow{\mathrm{p}} X$, if for every $\delta>0$

$$
\begin{equation*}
\lim _{n \longrightarrow \infty} \mathrm{P}\left(\left|X_{n}-X\right|>\delta\right)=0 . \tag{2.5}
\end{equation*}
$$

Critically we will also use the concept of "stochastically bounded", alternatively called bounded in probability [8].

Definition 6 (Bounded in probability). A sequence of random variables $\left\{X_{n}\right\}_{n=1}^{\infty}$ is said to bounded in probability, written $X_{n}=O_{\mathrm{p}}\left(a_{n}\right)$, if for every $\varepsilon>0$ there exists $C_{\varepsilon}$ and $n_{\varepsilon}$ such that for every $n>n_{\varepsilon}$

$$
\begin{equation*}
\mathrm{P}\left(\left|X_{n} / a_{n}\right|>C_{\varepsilon}\right)<\varepsilon . \tag{2.6}
\end{equation*}
$$

Definition 6 would provide the exact meaning of "bounded random variation" mentioned in Section 1.2 , which will be discussed later.

### 2.2 Some fundamental results

In this section, we revisit some useful results from analysis and probability theory.

### 2.2.1 Taylor's theorem

We first formulate Taylor's theorem, a result pertaining to the exact relationship between a function $f$ and its approximation in the standard basis of the space of polynomials of order $r, \mathbb{P}_{r}$.

Theorem 1 (Taylor's theorem). If a function $f$ has derivatives of order $r+1$, then

$$
\begin{equation*}
f(x)-\sum_{i=0}^{r} \frac{f^{(i)}(a)(x-a)^{i}}{i!}=\int_{a}^{x} \frac{f^{(r+1)}(t)}{r!}(x-t)^{r} d t \tag{2.7}
\end{equation*}
$$

Proof. The theorem follows from the fundamental theorem of calculus and partial integration. See [3] for details.

### 2.2.2 Mean value theorem

We next consider the mean value theorem, which can be considered a consequence of Taylor's theorem but can also be independently formulated.

Theorem 2 (Mean value theorem). If $f$ is continuously differentiable on ( $a, x$ ) and continuous on $[a, x]$, then

$$
\begin{equation*}
f^{\prime}(c)=\frac{f(x)-f(a)}{x-a} \tag{2.8}
\end{equation*}
$$

for some $c \in(a, x)$.
Proof. One way to prove this is to set $r=0$ in equation 2.7 and use the mean value theorem for integrals. See [4] for other ways to prove this.

### 2.2.3 A note on the mean value theorem

Theorem 2 states that a differentiable function $f$ on $(a, x)$ and continuous on [ $a, x]$ may be written as

$$
\begin{equation*}
f(x)=f(a)+f^{\prime}(c)(x-a), \tag{2.9}
\end{equation*}
$$

where $c \in(a, x)$ if $x>a$. If instead $x<a$, equation 2.9) still applies, but $c \in(x, a)$.

In equation 2.9), $c$ is in fact a function of $x$, emphasizing the composite nature of this equation and rewriting yields

$$
\begin{equation*}
\left(f^{\prime} \circ c\right)(x)=\frac{f(x)-f(a)}{x-a} \tag{2.10}
\end{equation*}
$$

Thus, the linear coefficient in equation 2.9 can be expressed as a function of $x$. The function $f^{\prime} \circ c$ will play a role in later calculations on norms and angles, where the first argument of the composition will be the derivative of the square root and the arccosine, respectively. More specifically, we will utilize the maximum modulus of the function defined by equation 2.10 .

Obtaining this maximum is straightforward in the case of $g(x)=\sqrt{x}$. Let $x \neq a$ and equation 2.10 then yields

$$
\begin{align*}
\max _{x \geq 0}\left|\left(g^{\prime} \circ c\right)(x)\right|=\max _{x \geq 0} \frac{1}{2 \sqrt{c(x)}} & =\max _{x \geq 0} \frac{\sqrt{x}-\sqrt{a}}{x-a} \\
& =\max _{x \geq 0} \frac{\sqrt{x}-\sqrt{a}}{(\sqrt{x}-\sqrt{a})(\sqrt{x}+\sqrt{a})}=\frac{1}{\sqrt{a}} . \tag{2.11}
\end{align*}
$$

The calculation is illustrated in Figure 2.1 for $a=1$.


Figure 2.1: Remainder for the mean value theorem square root function expansion. Note that the maximum modulus is bounded.

We shall also use a similar result for $h(x)=\arccos (x)$. In this case, let $x \neq 0$, $a=0$ and equation 2.10 yields the maximum modulus

$$
\begin{align*}
\max _{x \in[-1,1]}\left|\left(h^{\prime} \circ c\right)(x)\right| & =\max _{x \in[-1,1]}\left|\frac{-1}{\sqrt{1-c(x)^{2}}}\right|=\max _{x \in[-1,1]}\left|\frac{\arccos x-\frac{\pi}{2}}{x}\right|  \tag{2.12}\\
& =\frac{\pi}{2}
\end{align*}
$$

This can be seen because if $x>0$, then $h^{\prime} \circ c$ is decreasing

$$
\frac{d}{d x}\left(\frac{\arccos (x)-\frac{\pi}{2}}{x}\right)=\frac{d}{d x}\left(-\frac{\arcsin (x)}{x}\right)=-\frac{\frac{x}{\sqrt{1-x^{2}}}-\arcsin (x)}{x^{2}}<0
$$

where we have used the fact that $x=\sin (y)$ and $\tan (y)>y$ for $y \in\left(0, \frac{\pi}{2}\right)$. A similar argument will show that $h^{\prime} \circ c$ is increasing for $x<0$. For $x=0$ note that

$$
\lim _{x \rightarrow 0}\left(\frac{\arccos (x)-\frac{\pi}{2}}{x}\right)=\lim _{x \longrightarrow 0}\left(-\frac{\arcsin (x)}{x}\right)=\lim _{y \longrightarrow 0}\left(-\frac{1}{\frac{\sin (y)}{y}}\right)=-1 .
$$

Finally, evaluating $h^{\prime} \circ c$ at the boundary and confirming the claim of its maximum modulus

$$
\left.\frac{\arccos (x)-\frac{\pi}{2}}{x}\right|_{x=1}=\left.\frac{\arccos (x)-\frac{\pi}{2}}{x}\right|_{x=-1}=-\frac{\pi}{2}
$$

The calculation is illustrated in Figure 2.2 .


Figure 2.2: Remainder for the mean value theorem arccosine function expansion. Note that the maximum modulus is bounded.

### 2.2.4 Markov's inequality

Markov's inequality relates a non-negative random variable to its expectation.
Theorem 3 (Markov's inequality). For $a>0$ and $\xi \geq 0$ the following holds

$$
\mathrm{P}(\xi \geq a) \leq \frac{\mathrm{E}[\xi]}{a}
$$

Proof. See [5] for details.

### 2.2.5 Law of large numbers

We will heavily utilize Bernoulli's celebrated theorem, the law of large numbers, that rigorously establishes the intuitive notion that averaging measurements will eventually increase precision.
Theorem 4 (Law of large numbers). Let $\xi_{1}, \xi_{2}, \ldots, \xi_{d}$ be i.i.d. random variables with finite variance, $\mathrm{E}[\xi]=\mu$ and let

$$
\bar{\xi}_{d}=\frac{1}{d} \sum_{k=1}^{d} \xi_{k}
$$

it then follows that $\bar{\xi}_{d} \xrightarrow{\mathrm{p}} \mu$ as $d \rightarrow \infty$.
Proof. To show that $\bar{\xi}_{d} \xrightarrow{\mathrm{p}} \mu$, it is sufficient according to Definition 5 to demonstrate that for every $\delta>0$,

$$
\lim _{d \longrightarrow} \mathrm{P}\left(\left|\bar{\xi}_{d}-\mu\right|>\delta\right)=0
$$

Let $\operatorname{Var}(\xi)<\infty$ and noting that independence implies zero covariance, then for every $\delta>0$,

$$
\begin{aligned}
\mathrm{P}\left(\left|\bar{\xi}_{d}-\mu\right|>\delta\right) & =\mathrm{P}\left(\left|\frac{1}{d} \sum_{k=1}^{d} \xi_{k}-\mu\right|>\delta\right)=\mathrm{P}\left(\left|\frac{1}{d} \sum_{k=1}^{d}\left(\xi_{k}-\mu\right)\right|^{2}>\delta^{2}\right) \\
& \leq \frac{\mathrm{E}\left[\sum_{k=1}^{d}\left(\xi_{k}-\mu\right)\right]^{2}}{n^{2} \delta^{2}}=\frac{\sum_{k=1}^{d} \mathrm{E}\left[\xi_{k}-\mu\right]^{2}+\sum_{k \neq l} \operatorname{Cov}\left(\xi_{k}, \xi_{l}\right)}{n^{2} \delta^{2}} \\
& =\frac{d \operatorname{Var}(\xi)+d(d-1) \cdot 0}{d^{2} \delta^{2}}=\frac{\operatorname{Var}(\xi)}{d \delta^{2}} \longrightarrow 0, d \longrightarrow \infty,
\end{aligned}
$$

where the inequality follows from an application of Theorem 3] See [2] for more details.

### 2.2.6 Continuous mapping theorem

The following result will be used and establishes that the limit of a continuous mapping is the mapping of the limit.
Theorem 5 (Continuous mapping theorem). Let $\left\{X_{n}\right\}_{n=1}^{\infty}, X_{i} \in \mathbb{R}^{k}$, be a sequence of random variables and $g: \mathbb{R}^{k} \rightarrow \mathbb{R}^{m}$ a continuous function. It then follows that if $X_{n} \xrightarrow{\mathrm{p}} X$ then $g\left(X_{n}\right) \xrightarrow{\mathrm{p}} g(X)$.
Proof. See [10] for details.

### 2.2.7 Slutsky's theorem

Slutsky's theorem is central to both probability theory and statistical inference. We mainly use it as an aid to establish the final result, the delta method.

Theorem 6 (Slutsky's theorem). Let $X_{n}$ and $Y_{n}$ be sequences of random variables. Suppose that

$$
X_{n} \xrightarrow{\mathrm{~d}} X \text { and } Y_{n} \xrightarrow{\mathrm{p}} a,
$$

where $a$ is a constant. Then

$$
\begin{aligned}
& X_{n}+Y_{n} \xrightarrow{\mathrm{~d}} X+a \\
& X_{n}-Y_{n} \xrightarrow{\mathrm{~d}} X-a \\
& X_{n} \cdot Y_{n} \xrightarrow{\mathrm{~d}} X \cdot a \\
& \frac{X_{n}}{Y_{n}} \xrightarrow{\mathrm{~d}} \frac{X}{a}, \text { for } a \neq 0 .
\end{aligned}
$$

Proof. The reasoning relies on the distance $\left|Y_{n}-a\right|$ being under control. See (5] for details.

### 2.2.8 Delta method

The next theorem was the inspiration for the method that we will ultimately use to establish the geometric claims in detail. An example in Chapter 3 will illustrate why we will not use the powerful delta method directly.

Theorem 7 (First order delta method). Let $X_{n}$ be a sequence of random variables that satisfies $\sqrt{n}\left(X_{n}-\theta\right) \longrightarrow \mathcal{N}\left(0, \sigma^{2}\right)$ in distribution. For a given function $g$ and a specific value of $\theta$, suppose that $g^{\prime}(\theta)$ exists and is not 0 . Then

$$
\sqrt{n}\left(g\left(X_{n}\right)-g(\theta)\right) \longrightarrow \mathcal{N}\left(0,\left(\sigma g^{\prime}(\theta)\right)^{2}\right) \text { in distribution. }
$$

Proof. The theorem follows from Taylor's Theorem 1 and Slutsky's theorem 6. see [3].

## Chapter 3

## Main results

This chapter will utilize the results from Chapter 2 to make credible, precisely formulate and ultimately demonstrate claims i) and ii) from Section 1.2 .

Section 3.1 contains several examples that contain analysis of claims i) and ii) utilizing the law of large numbers and the delta method from Chapter 2 While these powerful methods allow for simple derivation of claims i) and ii), they have drawbacks that are briefly considered.

Section 3.2 and 3.4 contain the precise statement of i) and ii) and a proof that remedies the mentioned drawbacks illustrated in Section 3.1. The method employed here yields explicit stochastic bounds. These bounds are corroborated by simulations in Chapter 4. Finally, Section 3.3 contains a simple yet interesting consequence of claim i) that is formulated and proved as a corollary.

### 3.1 Examples

Theorem 4 and Theorem 5 from Chapter 2 can make plausible the claim that for high $d$-dimensional multivariate $Z$ it is roughly the case that $\|Z\|=\sqrt{d}$. This is illustrated in Example 1

Example 1 (Euclidean norm from the law of large numbers). Let $Z=\left(\xi_{1}, \ldots, \xi_{d}\right)^{\prime}$ $\sim \mathcal{N}_{d}(0, \mathbf{I})$. Then according to Definition 2

$$
\|Z\|=\sqrt{Z^{\prime} Z}=\sqrt{\sum_{k=1}^{d} \xi_{k}^{2}}=\sqrt{d} \sqrt{\bar{\xi}^{2}}
$$

where $\overline{\xi^{2}}=\frac{1}{d} \sum_{k=1}^{d} \xi_{k}^{2}$. From Theorem 4 it now follows that $\overline{\xi^{2}} \xrightarrow{\mathrm{p}} 1$ and then
it follows from Theorem 5 that $\sqrt{\xi^{2}} \xrightarrow{\mathrm{p}} 1$. Thus the claim that $\|Z\| \approx \sqrt{d}$ has been made more plausible.

Theorem 4 and Theorem 5 from Chapter 2 can also make plausible the claim that for high $d$-dimensional multivariate $Z_{1}, Z_{2}$ it is roughly the case that $Z_{1} \perp Z_{2}$. This is illustrated in Example 2.

Example 2 (Pairwise orthogonality from the law of large numbers). Let $Z_{1}=$ $\left(\xi_{1}, \ldots, \xi_{d}\right)^{\prime}, Z_{2}=\left(\eta_{1}, \ldots, \eta_{d}\right)^{\prime} \stackrel{\mathrm{iid}}{\sim} \mathcal{N}_{d}(0, \mathbf{I})$. According to the remark subsequent to Definition 3 the angle between $Z_{1}$ and $Z_{2}$ is

$$
\begin{aligned}
\operatorname{ang}\left(Z_{1}, Z_{2}\right)=\arccos \left(\frac{Z_{1} \cdot Z_{2}}{\left\|Z_{1}\right\|\left\|Z_{2}\right\|}\right) & =\arccos \left(\frac{\sum_{i=1}^{d} \xi_{i} \eta_{i}}{\sqrt{\sum_{i=1}^{d} \xi_{i}^{2}} \sqrt{\sum_{i=1}^{d} \eta_{i}^{2}}}\right) \\
& =\arccos \left(\frac{\overline{\xi \eta}}{\sqrt{\xi^{2}} \sqrt{\eta^{2}}}\right)
\end{aligned}
$$

and from Theorem 4 and 5 and the independence of $\xi$ and $\eta$ it follows that

$$
\frac{\bar{\xi} \eta}{\sqrt{\overline{\xi^{2}}} \sqrt{\overline{\eta^{2}}}}=\frac{\frac{1}{d} \sum_{i=1}^{d} \xi_{i} \eta_{i}}{\sqrt{\overline{\xi^{2}}} \sqrt{\overline{\eta^{2}}}} \xrightarrow{\mathrm{p}} \frac{\mathrm{E}[\xi \eta]}{\sqrt{\mathrm{E}\left[\xi^{2}\right]} \sqrt{\mathrm{E}\left[\eta^{2}\right]}}=\frac{0 \cdot 0}{1 \cdot 1}=0 .
$$

Another application of Theorem 5 yields

$$
\operatorname{ang}\left(Z_{1}, Z_{2}\right) \xrightarrow{\mathrm{p}} \frac{\pi}{2},
$$

which is the desired result $Z_{1} \perp Z_{2}$.
Remark. Example 1 and 2 shows that claims i) and ii) are both closely connected to the law of large numbers.

Claim i) can also be approached by means of Theorem 7. This is illustrated in Example 3 .

Example 3 (Delta method approach). Let $Z=\left(\xi_{1}, \ldots, \xi_{d}\right)^{\prime} \sim \mathcal{N}_{d}(0, \mathbf{I})$. An approach to showing that $\|Z\| \approx \sqrt{d}$ is to choose $g(x)=\sqrt{x}, X_{d}=\frac{1}{d} \sum_{k=1}^{d} \xi_{k}^{2}$, $\theta=\mathrm{E}\left[X_{d}\right]=1$ in Theorem 7. Then

$$
\|Z\|=\sqrt{d}+2^{1 / 2} \underbrace{\sqrt{d} \frac{\sqrt{X_{d}}-1}{\sqrt{2}}}_{\xrightarrow[\mathrm{d}]{\longrightarrow} \mathcal{N}(0,1)} \xrightarrow{\mathrm{d}} \sqrt{d}+2^{1 / 2} \mathcal{N}(0,1)=\sqrt{d}+O_{\mathrm{p}}(1)
$$

according to Theorem 7 .

Let us briefly consider examples 113. Example 1 and 2 have made the claims i) and ii) of the geometric properties of HDLSS data more plausible, however, no information about the limiting process is acquired. This is partially remedied by the approach with the delta method in Example 3 , where the behavior of the stochastic boundedness can be deduced as $O_{\mathrm{p}}(1)$. Full knowledge of the stochastic boundedness, however, evades the delta method approach. The details of the stochastic boundedness are lost in the limiting process when Theorem 7 is derived, a consequence of the central limit theorem, and the remainder vanishes in probability in the proof.

The subsequent sections of this chapter will deal with claims i) and ii) in more detail, establishing explicit bounds for the asymptotic behavior of the Euclidean norm and angle in the spirit of Definition 6. The method will essentially be a "zeroth order" delta method, dealing explicitly with the remainder term.

### 3.2 Euclidean norm of high dimensional data

Claim i) is the geometric property of HDLSS data, that the data will roughly be distributed on a sphere with radius $\sqrt{d}$. In the following theorem, this statement is made precise and proven.

Theorem 8 (Euclidean norm of a HDLSS random variable). Let $Z=\left(\xi_{1}, \ldots, \xi_{d}\right)^{\prime}$ $\sim \mathcal{N}_{d}(0, \mathbf{I})$. With the distance defined as in Definition 2 and the stochastically bounded term defined as in Definition 6 it follows that

$$
\begin{equation*}
\|Z\|=d^{1 / 2}+O_{\mathrm{p}}(1) \tag{3.1}
\end{equation*}
$$

Proof. First note that according to Definition 2 we have

$$
\|Z\|=\sqrt{Z^{\prime} Z}=\sqrt{\sum_{k=1}^{d} \xi_{k}^{2}}=\sqrt{\sum_{k=1}^{d} Y_{k}}
$$

where $Y_{k} \sim \chi^{2}(1)$. Utilizing equation 2.9 and expanding around $a=\mathrm{E}\left[Y_{k}\right]=1$ yields

$$
\begin{equation*}
\sqrt{\sum_{k=1}^{d} Y_{k}}=\sqrt{d} \sqrt{\bar{Y}}=\sqrt{d}\left(1+\frac{1}{2 \sqrt{c}}(\bar{Y}-1)\right)=\sqrt{d}+\frac{\sqrt{d}(\bar{Y}-1)}{2 \sqrt{c}} \tag{3.2}
\end{equation*}
$$

for some $c$ that is a function of $\bar{Y}$, and range as discussed in Section 2.2.3. It remains to be shown that the last term in equation 3.2 is $O_{\mathrm{p}}(1)$.

According to Definition 6 we need to show that for every $\varepsilon>0$ there exists $C_{\varepsilon}$ such that whenever $d>N_{\varepsilon}$,

$$
\begin{equation*}
\mathrm{P}\left(\left|\frac{\sqrt{d}(\bar{Y}-1)}{2 \sqrt{c}}\right|>C_{\varepsilon}\right)<\varepsilon \tag{3.3}
\end{equation*}
$$

First note that

$$
\begin{aligned}
\mathrm{P}\left(\left|\frac{\sqrt{d}(\bar{Y}-1)}{2 \sqrt{c}}\right|>C_{\varepsilon}\right) & \leq \mathrm{P}\left(\left|\max _{\omega \in \Omega}\left(\frac{1}{2 \sqrt{c \circ \bar{Y}(\omega)}}\right) \sqrt{d}(\bar{Y}-1)\right|>C_{\varepsilon}\right) \\
& =\mathrm{P}\left(|\sqrt{d}(\bar{Y}-1)|>C_{\varepsilon}\right)
\end{aligned}
$$

since

$$
\begin{equation*}
\left\{\omega \in \Omega:\left|\frac{\sqrt{d}(\bar{Y}(\omega)-1)}{2 \sqrt{c}}\right|>C_{\varepsilon}\right\} \subseteq\left\{\omega \in \Omega:|\sqrt{d}(\bar{Y}(\omega)-1)|>C_{\varepsilon}\right\} \tag{3.4}
\end{equation*}
$$

To see this, let

$$
s \in\left\{\omega \in \Omega:\left|\frac{\sqrt{d}(\bar{Y}(\omega)-1)}{2 \sqrt{c}}\right|>C_{\varepsilon}\right\} .
$$

It then follows that

$$
\left|\max _{\omega \in \Omega}\left(\frac{\sqrt{d}}{2 \sqrt{c \circ \bar{Y}(\omega)}}\right)(\bar{Y}(s)-1)\right| \geq\left|\frac{\sqrt{d}(\bar{Y}(s)-1)}{2 \sqrt{c}}\right|>C_{\varepsilon},
$$

hence

$$
\begin{aligned}
s & \in\left\{\omega \in \Omega:\left|\max _{\omega \in \Omega}\left(\frac{d}{2 \sqrt{c \circ \bar{Y}(\omega)}}\right)(\bar{Y}(\omega)-1)\right|>C_{\varepsilon}\right\} \\
& =\left\{\omega \in \Omega:|\sqrt{d}(\bar{Y}(\omega)-1)|>C_{\varepsilon}\right\}
\end{aligned}
$$

where the equality is a consequence of equation 2.11) and that $\operatorname{Im}(\bar{Y})=[0, \infty)$. Now, utilizing Markov's inequality from Theorem 3 and the fact that $\operatorname{Var}(\bar{Y})=$ $\frac{2}{d}$ it finally follows that

$$
\begin{aligned}
\mathrm{P}\left(\left|\frac{\sqrt{d}(\bar{Y}-1)}{2 \sqrt{c}}\right|>C_{\varepsilon}\right) & \leq \mathrm{P}\left(|\sqrt{d}(\bar{Y}-1)|>C_{\varepsilon}\right) \\
& =\mathrm{P}\left(|\sqrt{d}(\bar{Y}-1)|^{2}>C_{\varepsilon}^{2}\right) \leq \frac{d \mathrm{E}[\bar{Y}-1]^{2}}{C_{\varepsilon}^{2}}=\frac{2}{C_{\varepsilon}^{2}}<\varepsilon,
\end{aligned}
$$

if we choose $C_{\varepsilon}>\sqrt{\frac{2}{\varepsilon}}$ and this choice will always satisfy equation (3.3).

Remark. In arriving at $C_{\varepsilon}$ in Theorem 8, no consideration of the dimension $d$ was necessary, i.e., for every $\varepsilon>0$ there exists a $C_{\varepsilon}$ such that equation (3.3) is satisfied for all $d \in \mathbb{N}$. Simulations in the next chapter will corroborate this result.

In addition, with minimal adjustments in the proof an analogous result of equation 3.1 holds for other $L_{p}$ norms than $p=2$, i.e., under similar assumptions on $Z$, it holds that

$$
\|Z\|_{p}=d^{1 / p}+O_{\mathrm{p}}(1)
$$

### 3.3 Pairwise Euclidean distance of high dimensional data

We formulate a related property of HDLSS data, the fact that the pairwise Euclidean distance is approximately a deterministic number, as a corollary to Theorem 8 .

Corollary 8.1. Let $Z_{1}=\left(\xi_{1}, \ldots, \xi_{d}\right)^{\prime}, Z_{2}=\left(\eta_{1}, \ldots, \eta_{d}\right)^{\prime} \stackrel{\text { iid }}{\sim} \mathcal{N}_{d}(0, \mathbf{I})$ and independent. Then with the definitions of Theorem 8, it follows that

$$
\begin{equation*}
\left\|Z_{1}-Z_{2}\right\|=(2 d)^{1 / 2}+O_{\mathrm{p}}(1) . \tag{3.5}
\end{equation*}
$$

Proof. For univariate standard, independent normal distributions $\xi_{i}$ and $\eta_{i}$ it holds that $\xi_{i}-\eta_{i} \stackrel{\mathrm{~d}}{=} \sqrt{2} \zeta_{i}$ where $\zeta_{i} \sim \mathcal{N}(0,1)$. Hence,

$$
\begin{aligned}
\left\|Z_{1}-Z_{2}\right\|=\sqrt{\sum_{k=1}^{d}\left(\xi_{k}-\eta_{k}\right)^{2}} & \stackrel{\mathrm{~d}}{=} \sqrt{\sum_{k=1}^{d}\left(\sqrt{2} \zeta_{k}\right)^{2}}=\sqrt{2} \sqrt{\sum_{k=1}^{d} \zeta_{k}^{2}} \\
& =\sqrt{2}\left(d^{1 / 2}+O_{\mathrm{p}}(1)\right)=(2 d)^{1 / 2}+O_{\mathrm{p}}(1)
\end{aligned}
$$

where the penultimate equality is an application of Theorem 8 and the ultimate equality is an immediate property of stochastic boundedness.

### 3.4 Pairwise orthogonality of high dimensional data

Claim ii) is that one of the geometric properties of HDLSS data is that two observations will be roughly pairwise orthogonal. In the following theorem, this statement is made precise and proven.

Theorem 9 (Pairwise orthogonality of HDLSS random variables). Let $Z_{1}=$ $\left(\xi_{1}, \ldots, \xi_{d}\right)^{\prime}, Z_{2}=\left(\eta_{1}, \ldots, \eta_{d}\right)^{\prime} \stackrel{\text { iid }}{\sim} \mathcal{N}_{d}(0, \mathbf{I})$ and independent. With the angle defined according to the remark subsequent to Definition3 and the stochastically bounded term defined according to Definition 6, it follows that

$$
\begin{equation*}
\operatorname{ang}\left(Z_{1}, Z_{2}\right)=\frac{\pi}{2}+O_{\mathrm{p}}\left(d^{-1 / 2}\right) \tag{3.6}
\end{equation*}
$$

Proof. First note that according to Definition 3we have

$$
\operatorname{ang}\left(Z_{1}, Z_{2}\right)=\arccos \left(\frac{Z_{1} \cdot Z_{2}}{\left\|Z_{1}\right\|\left\|Z_{2}\right\|}\right)
$$

which is well-defined because of the Cauchy-Schwarz inequality. Utilizing equation 2.9 and expanding around $a=0$ yields

$$
\begin{align*}
\operatorname{ang}\left(Z_{1}, Z_{2}\right)=\arccos \left(\frac{Z_{1} \cdot Z_{2}}{\left\|Z_{1}\right\|\left\|Z_{2}\right\|}\right) & =\frac{\pi}{2}+\frac{-1}{\sqrt{1-c^{2}}}\left(\frac{Z_{1} \cdot Z_{2}}{\left\|Z_{1}\right\|\left\|Z_{2}\right\|}\right)  \tag{3.7}\\
& =\frac{\pi}{2}+\frac{-1}{\sqrt{1-c^{2}}} \frac{\sum_{i=1}^{d} \xi_{i} \eta_{i}}{\sqrt{\sum_{i=1}^{d} \xi_{i}^{2}} \sqrt{\sum_{i=1}^{d} \eta_{i}^{2}}}
\end{align*}
$$

for some $c$ that is a function of both $Z_{1}$ and $Z_{2}$, and range as discussed in Section 2.2.3. It remains to be shown that the last term in equation (3.7) is $O_{\mathrm{p}}\left(d^{-1 / 2}\right)$.

According to Definition 6 we need to show that for every $\varepsilon>0$ there exists $C_{\varepsilon}$ such that whenever $d>N_{\varepsilon}$

$$
\begin{equation*}
\mathrm{P}\left(\left|\sqrt{d} \frac{-1}{\sqrt{1-c^{2}}} \frac{\sum_{i=1}^{d} \xi_{i} \eta_{i}}{\sqrt{\sum_{i=1}^{d} \xi_{i}^{2}} \sqrt{\sum_{i=1}^{d} \eta_{i}^{2}}}\right|>C_{\varepsilon}\right)<\varepsilon . \tag{3.8}
\end{equation*}
$$

Note that according to Theorem 4 and Theorem 5 it follows that for $\zeta \sim \mathcal{N}(0,1)$

$$
\sqrt{\frac{1}{n} \sum_{i=1}^{d} \zeta_{i}^{2}} \xrightarrow{\mathrm{p}} \sqrt{\mathrm{E}\left[\zeta^{2}\right]}=1 .
$$

This together with Definition 5 implies that for every $\varepsilon>0$ and every $\delta>0$ there exists a $N_{\delta, \varepsilon}$ such that if $d>N_{\delta, \varepsilon}$ then

$$
\begin{equation*}
\mathrm{P}\left(\left|\sqrt{\frac{1}{d} \sum_{i=1}^{d} \zeta_{i}^{2}}-1\right|>\delta\right)<\frac{\varepsilon}{3} \tag{3.9}
\end{equation*}
$$

Further note that since $\xi_{i}$ and $\eta_{j}$ are independent,

$$
\mathrm{E}\left[\overline{\xi_{\eta}}\right]^{2}=\mathrm{E}\left[\frac{1}{d} \sum_{i=1}^{d} \xi_{i} \eta_{i}\right]^{2}=\frac{1}{d^{2}} \mathrm{E}\left[\sum_{i=1}^{d}\left(\xi_{i} \eta_{i}\right)^{2}+\sum_{i \neq j} \xi_{i} \eta_{i} \xi_{j} \eta_{j}\right]=\frac{1}{d} .
$$

Using an argument analogous to the one in equation (3.4, the result from equation 2.12 and

$$
\operatorname{Im}(\mathcal{Z})=\operatorname{Im}\left(\frac{Z_{1} \cdot Z_{2}}{\left\|Z_{1}\right\|\left\|Z_{1}\right\|}\right)=[-1,1]
$$

where $\mathcal{Z}$ is defined as the argument of the arccosine in equation (3.7), it now
follows that

$$
\left.\left.\begin{array}{rl} 
& \mathrm{P}\left(\left|\sqrt{d} \frac{-1}{\sqrt{1-c^{2}}} \frac{\sum_{k=1}^{d} \xi_{k} \eta_{k}}{\sqrt{\sum_{k=1}^{d} \xi_{k}^{2}} \sqrt{\sum_{k=1}^{d} \eta_{k}^{2}}}\right|>C_{\varepsilon}\right) \\
\leq & \mathrm{P}\left(\left|\sqrt{d} \max _{\omega \in \Omega}\left(\left|\frac{-1}{\sqrt{1-(c \circ \mathcal{Z}(\omega))^{2}}}\right|\right) \frac{\sum_{k=1}^{d} \xi_{k} \eta_{k}}{\sqrt{\sum_{k=1}^{d} \xi_{k}^{2}} \sqrt{\sum_{k=1}^{d} \eta_{k}^{2}}}\right|>C_{\varepsilon}\right) \\
= & \mathrm{P}\left(\left|\sqrt{d}\left(\frac{\pi}{2}\right) \frac{\frac{1}{d} \sum_{k=1}^{d} \xi_{k} \eta_{k}}{\sqrt{\frac{1}{d} \sum_{k=1}^{d} \xi_{k}^{2}} \sqrt{\frac{1}{d} \sum_{k=1}^{d} \eta_{k}^{2}}}\right|>C_{\varepsilon}\right) \\
= & \mathrm{P}\left(\left|\frac{\sqrt{d} \pi \frac{1}{d} \sum_{k=1}^{d} \xi_{k} \eta_{k}}{2 \sqrt{\frac{1}{d} \sum_{k=1}^{d} \xi_{k}^{2}} \sqrt{\frac{1}{d} \sum_{k=1}^{d} \eta_{k}^{2}}}\right|>C_{\varepsilon}\right.
\end{array}\left|\sqrt{\frac{1}{d} \sum_{k=1}^{d} \xi_{k}^{2}}-1\right| \leq \delta \cap\left|\sqrt{\frac{1}{d} \sum_{k=1}^{d} \eta_{k}^{2}}-1\right| \leq \delta\right)\right)
$$

if we choose $C_{\varepsilon}>\frac{\pi}{2(1-\delta)^{2}} \sqrt{\frac{3}{\varepsilon}}$ and let $d>N_{\delta, \varepsilon}$.
Remark. The integer $N_{\delta, \varepsilon}$ may be found as the smallest integer satisfying equation (3.9) which can alternatively be expressed as

$$
\begin{equation*}
N_{\delta, \varepsilon}=\left\lceil\inf \left(\underset{d}{\arg \max }\left[\int_{d(1-\delta)^{2}}^{d(1+\delta)^{2}} \frac{1}{2^{d / 2} \Gamma(d / 2)} s^{d / 2-1} e^{-s / 2} d s>1-\frac{\varepsilon}{3}\right]\right)\right\rceil \tag{3.10}
\end{equation*}
$$

where $\lceil\bullet\rceil$ is the ceiling function and $[\bullet]$ is the Iverson bracket defined by

$$
[Q]= \begin{cases}1, & \text { if } Q \text { is true } \\ 0 & \text { otherwise }\end{cases}
$$

## Chapter 4

## Simulations

In this chapter, the method and result of simulations will be discussed, with the primary goal being to corroborate the derived $C_{\varepsilon}$ relationship for the two cases of stochastic boundedness discussed in Section 3.2 and 3.4 .

$$
\begin{equation*}
C_{\varepsilon}>\sqrt{\frac{2}{\varepsilon}} \tag{4.1}
\end{equation*}
$$

and

$$
\begin{equation*}
C_{\varepsilon}>\frac{\pi}{2(1-\delta)^{2}} \sqrt{\frac{3}{\varepsilon}} \tag{4.2}
\end{equation*}
$$

respectively.
The immediate question concerning equation (4.1) and 4.2 is if the regions are correct, i.e., does choosing $C_{\varepsilon}$ consistent with the respective regions satisfy equation (3.3) and (3.8) - something simulations can make more credible. Simulations also indirectly test the reasoning in Section 2.2 .3 and equation (3.4).

### 4.1 Initial simulations

Some initial simulations may be carried out to illustrate the geometric properties mentioned in Chapter 1 .

Make a draw $D_{d}$ from a $d$-dimensional multivariate random variable, calculate its Euclidean norm $\left\|D_{d}\right\|$ and create the ordered pair $\left(d,\left\|D_{d}\right\|\right)$. Connect the pairs by a line graph. The result can be inspected in Figure 4.1 and for reference $\sqrt{d}$ is plotted in the same figure.


Figure 4.1: Draws of the Euclidean norm of a $d$-dimensional random vector with $\sqrt{d}$ plotted for reference.

Several observations may be made. First, the simulation is consistent with equation (3.1), i.e. the deviation from $\sqrt{d}$ looks to be bounded. Second, the absolute difference between $\sqrt{d}$ and the draw $\left\|D_{d}\right\|$ appears to be fixed, while the relative difference appears to tend to zero.

Next, we can illustrate the orthogonality property. Make a draw of a pair $D_{d}^{\prime}$ of independent $d$-dimensional multivariate random variables, calculate the angle $\operatorname{ang}\left(D_{d}^{\prime}\right)$ between them, and create the ordered pair $\left(d, \operatorname{ang}\left(D_{d}^{\prime}\right)\right)$. Connect the pairs by a line graph. The result can be inspected in Figure 4.2 and for reference $\frac{\pi}{2}$ is plotted in the same figure.


Figure 4.2: Draws of an angle between two $d$-dimensional random vectors with $\frac{\pi}{2}$ plotted for reference.

We note that Figure 4.2 is consistent with equation (3.6), and here both the absolute and relative error appears to tend to zero.

### 4.2 Simulations testing the bounds for the Eulidean norm

To test the bounds in equation (4.1), we need to verify that if we choose $C_{\varepsilon}$ in a way consistent with this equation, then

$$
\mathrm{P}\left(|\|Z\|-\sqrt{d}|>C_{\varepsilon}\right)<\varepsilon
$$

To test this, we will generate $N$ draws from a $d$-dimensional multivariate normal distribution. Having acquired a Monte Carlo sample of the distribution, we can calculate an ordered pair $\left(\varepsilon, C_{\varepsilon}\right)$ from the equality

$$
\mathrm{P}\left(|\|Z\|-\sqrt{d}|>C_{\varepsilon}\right)=\varepsilon
$$

The result can then be inspected visually in the form of a line graph connecting the pairs and be compared to the region in the $\left(\varepsilon, C_{\varepsilon}\right)$-plane prescribed by equation 4.1). If the bound is successful, the region indicated by equation (4.1) should be above the line graph.

The result for $d=100$ dimensions and $N=100000$ draws is illustrated in Figure 4.3 .


Figure 4.3: Plot of the region specified by equation (4.1) and the $\left(\varepsilon, C_{\varepsilon}\right)$ pairs acquired from a Monte Carlo simulation connected by a line graph.

### 4.3 Simulations testing the bounds for pairwise orthogonality

In a similar fashion to Section 4.2, to test the bounds given by equation 4.2) we need to verify that if we choose $C_{\varepsilon}$ in a way consistent with this equation, then

$$
\mathrm{P}\left(\sqrt{d}\left|\arccos \left(\frac{Z_{1} \cdot Z_{2}}{\left\|Z_{1}\right\|\left\|Z_{2}\right\|}\right)-\frac{\pi}{2}\right|>C_{\varepsilon}\right)<\varepsilon
$$

To test this, we will generate $N$ independent pairs from a $d$-dimensional multivariate normal distribution. Having acquired a Monte Carlo sample of the distribution, we can calculate an ordered pair $\left(\varepsilon, C_{\varepsilon}\right)$ from the equality

$$
\mathrm{P}\left(\sqrt{d}\left|\arccos \left(\frac{Z_{1} \cdot Z_{2}}{\left\|Z_{1}\right\|\left\|Z_{2}\right\|}\right)-\frac{\pi}{2}\right|>C_{\varepsilon}\right)=\varepsilon .
$$

The result can then be inspected visually in the form of a line graph connecting the pairs and be compared to the region in the $\left(\varepsilon, C_{\varepsilon}\right)$-plane prescribed by equation (4.2). If the bound is successful, the region indicated by equation (4.2) should be above the line graph.

The result for $d=100$ dimensions and $N=100000$ draws is illustrated in Figure 4.4.


Figure 4.4: Plot of the region specified by equation 4.2 and the $\left(\varepsilon, C_{\varepsilon}\right)$ pairs acquired from a Monte Carlo simulation connected by a line graph.

## Chapter 5

## Discussion

This thesis aimed to formulate and prove in a precise manner the key geometric properties of HDLSS random variables, the fact that they are pairwise orthogonal and distributed on a sphere. This has been accomplished.

The method employed, utilizing the mean value theorem, allowed for testable stochastic bounds $C_{\epsilon}$. The $C_{\epsilon}$ bounds were tested and found successful in Chapter 4 . One observation that can be made from Chapter 4 is that the bounds are generous. One possible reason for the large gap is the utilization of the mean value theorem in the way employed. Another possible reason is the use of Markov's inequality, which is a coarse inequality. It would therefore be interesting to explore alternatives to the use of the mean value theorem and Markov's inequality. One replacement for Markov's inequality could be the Chernoff bound

$$
\mathrm{P}(X \geq a) \leq \inf _{t \geq 0}\left(\mathrm{M}_{X}(t) e^{-t a}\right)
$$

where $\mathrm{M}_{X}(t)$ is the generating function for $X$. The Chernoff bound is less coarse so it is possible that the bounds would be less generous. The use of the mean value theorem could be replaced by conditioning and larger utilization of convergence in probability.

Another interesting aspect to explore further would be to attempt a generalization of the result. As noted in Section 3.2 the result generalizes to the $p$-norm. Inspired by the $p$-norm result, one could for example explore what conditions on the norm are necessary to have an analogous result for the spherical distribution or what the necessary conditions on the inner product are to have an analogous result for the orthogonality property.

## Bibliography

[1] Makoto Aoshima, Dan Shen, Haipeng Shen, Kazuyoshi Yata, Yi-Hui Zhou, and J. S. Marron. A survey of high dimension low sample size asymptotics. Australian $\mathcal{E}$ New Zealand Journal of Statistics, 60(1):4-19, 2018.
[2] Gunnar Blom, Jan Enger, Gunnar Englund, Jan Grandell, and Lars Holst. Sannolikhetsteori och statistikteori med tillämpningar. Studentlitteratur AB, 2017.
[3] George Casella and Roger L. Berger. Statistical inference. Duxbury advanced series. Brooks/Cole, Cengage Learning, 2008.
[4] Göran Forsling and Mats Neymark. Matematisk analys : en variabel. Liber, 2011.
[5] Allan Gut. An intermediate course in probability. Springer texts in statistics. Springer, 2009.
[6] Peter Hall, J. S. Marron, and Amnon Neeman. Geometric representation of high dimension, low sample size data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(3):427-444, 2005.
[7] Ulf Janfalk. Linjär algebra. [Linköpings universitet], 2021.
[8] Svante Janson. Probability asymptotics: notes on notation, 2011.
[9] Richard A. Johnson and Dean W. Wichern. Applied Multivariate Statistical Analysis: Pearson New International Edition., volume 6th edition, Pearson new international edition of Pearson Custom Library. Pearson, 2014.
[10] Aad W. van der Vaart. Asymptotic statistics. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, 1998.

## Appendix A

## Simulation code

This appendix contains the code used to create the graphs in chapter 4

## A. 1 Initial simulations code

\%The draw of the angle of a single pair \%of d-dimensional random vectors
\%d dimensions
d=10000;
for $i=1$ :d
$r=\operatorname{randn}(i, 1)$;
$X(i)=\operatorname{sqrt}\left(\left(r .{ }^{\prime}\right) * r\right)$;
end
$\mathrm{x}=\operatorname{linspace}(0, \mathrm{~d}, \mathrm{~d})$;
z = sqrt(x);
plot(x,X, 'color', 'b')
hold on
plot(x,z, 'color', 'r', 'LineWidth',1.5)
xlabel('d')
ylabel('Radius')
ylim([0 1.3*sqrt(d)])
legend('Simulated draws','\surdd')
legend('Location','eastoutside')
legend boxoff
set(gca,'box','off')

```
%The draw of the angle of a single pair
%of d-dimensional random vectors
%d dimensions
d=10000;
for i=1:d
    r = randn(i,1);
    s = randn(i,1);
    X(i) = acos((r'*s)/((sqrt(s.'*s))*(sqrt(r.'*r))));
end
x = linspace(0,d,d);
plot(x,X, 'color', 'b')
z = ((pi/2)*ones(1, d));
hold on
plot(x,z, 'color', 'r', 'LineWidth',1.5)
xlabel('d')
ylabel('Angle')
ylim([0 3])
legend('Simulated draws','\pi/2')
legend('Location','eastoutside')
legend boxoff
set(gca,'box','off')
```


## A. 2 Euclidean norm code

\%Computed and simulated quantiles for the Euclidean norm \%d dimensions
\%N simulations
d = 100;
$\mathrm{N}=100000$;
$\mathrm{X}=\operatorname{zeros}(1, \mathrm{~N})$;
for $i=1: N$
r = randn(d,1);
X(i) $=\operatorname{sqrt}\left((r .)^{\prime}\right)$ r) - sqrt(d);
end

C = 120;
[counts, centers] = hist(X, (2*C));
P = counts/ N ;

```
%m}\mathrm{ is the vector of quantiles,
%p the vector of corresponding probabilities.
m = zeros(1,C);
p = zeros(1,C);
for k=1:C
    PE = 0;
    for i=1:(2*C)
            if abs(centers(i)) > (k/50)
                PE = PE + P(i);
            end
        end
    m(k) = k/50;
    p(k) = PE;
end
figure
x = linspace(0.055,1,100);
z = sqrt(x);
y = sqrt(2)./z;
ylim([0 6])
patch([x fliplr(x)], [y max(ylim)*ones(size(y))], 'b','FaceAlpha',.3)
xlabel('\epsilon')
ylabel('C(\epsilon)')
hold on
plot(p,m, 'color', 'r')
legend('Computed region','Monte carlo')
legend('Location','eastoutside')
legend boxoff
```


## A. 3 Pairwise orthogonality code

\%Computed and simulated quantiles for orthogonality property \%d dimensions
\%N simulations
d = 100;
$\mathrm{N}=100000$;
$\mathrm{X}=\operatorname{zeros}(1, \mathrm{~N})$;
for $\mathrm{i}=1: \mathrm{N}$

```
    %r = randn(d,1);
    a = randn(d,1);
    b = randn(d,1);
    %l = sqrt((r.')*r);
    %rv = l - sqrt(d);
    X(i) = sqrt(d)*(acos((a'*b)/(sqrt(a'*a)*sqrt(b'*b)))-pi/2);
end
```

C = 200;
[counts, centers] = hist(X, (2*C));
P = counts/ N ;
p = zeros (1, C);
m = zeros (1,C);
PE = 0;
for $k=1$ : $C$
PE = 0;
for $i=1:(2 * C)$
if abs(centers(i)) > (k/50)
PE = PE + P(i);
end
end
$m(k)=k / 50$;
$p(k)=P E ;$
end
figure
$\mathrm{x}=$ linspace (0.077,1,100);
delta = 0.01;
z = ((1-delta)~2)*sqrt(x);
y = pi*sqrt(3/4)./z;
ylim([0 10])
patch([x fliplr(x)], [y max(ylim)*ones(size(y))], 'b','FaceAlpha',.3)
xlabel('\epsilon')
ylabel('C(\epsilon)')
hold on
plot(p,m, 'color', 'r')
legend('Computed region', 'Monte carlo')
legend('Location','eastoutside')
legend boxoff

## Copyright

The publishers will keep this document online on the Internet - or its possible replacement - from the date of publication barring exceptional circumstances.

The online availability of the document implies permanent permission for anyone to read, to download, or to print out single copies for his/her own use and to use it unchanged for non-commercial research and educational purpose. Subsequent transfers of copyright cannot revoke this permission. All other uses of the document are conditional upon the consent of the copyright owner. The publisher has taken technical and administrative measures to assure authenticity, security and accessibility.

According to intellectual property law the author has the right to be mentioned when his/her work is accessed as described above and to be protected against infringement.

For additional information about the Linköping University Electronic Press and its procedures for publication and for assurance of document integrity, please refer to its www home page: http://www.ep.liu.se/.

## Upphovsrätt

Detta dokument hålls tillgängligt på Internet - eller dess framtida ersättare från publiceringsdatum under förutsättning att inga extraordinära omständigheter uppstår.

Tillgång till dokumentet innebär tillstånd för var och en att läsa, ladda ner, skriva ut enstaka kopior för enskilt bruk och att använda det oförändrat för ickekommersiell forskning och för undervisning. Överföring av upphovsrätten vid en senare tidpunkt kan inte upphäva detta tillstånd. All annan användning av dokumentet kräver upphovsmannens medgivande. För att garantera äktheten, säkerheten och tillgängligheten finns lösningar av teknisk och administrativ art.

Upphovsmannens ideella rätt innefattar rätt att bli nämnd som upphovsman i den omfattning som god sed kräver vid användning av dokumentet på ovan beskrivna sätt samt skydd mot att dokumentet ändras eller presenteras i sådan form eller i sådant sammanhang som är kränkande för upphovsmannens litterära eller konstnärliga anseende eller egenart.

För ytterligare information om Linköping University Electronic Press se förlagets hemsida http://www.ep.liu.se/.
(C) 2024, Olof Samuel Mossberg

