Facultad de Ciencias

Study of the variance of grid-based estimators

Estudio de la varianza de estimadores basados en rejillas

Trabajo de fin de grado
para acceder al

GRADO EN MATEMÁTICAS

Autor: Sara Ruiz Daza

Director: Domingo Gómez Pérez

Codirector: Ana I. Gómez Pérez

Septiembre de 2020

Agradecimientos

En primer lugar, me gustaría dar las gracias a Domingo y Ana, por proponerme un proyecto tan interesante, y sobre todo, por su paciencia, motivación, entrega y recibirme siempre con una sonrisa en su despacho.

También me gustaría agradecer estos seis años a mis 11 compañeros del Doble Grado de Física y Matemáticas. Muchas gracias Pablo Andrés por ver siempre el lado positivo de las cosas y ser ejemplo; David, por tu enorme amistad; Nayara, por estar siempre ahí; Toraya, por ser compañera en grandes momentos de la carrera; Fer, por tu gran apoyo este último curso, juntos sobrevivimos; Celia, por ser tan buena persona y contagiarlo; Pablo, por todas tus enseñanzas; Poles, por transmitir tu poca cordura; Pedro, por los cafés y momentos filosóficos; María, por tener siempre una sonrisa en la cara; y Senovilla, por ser matemático desde antes de empezar la carrera. No estaría ahora presentando este TFG si no fuera por vosotros. ¡Seguiremos celebrando éxitos juntos!

Finalmente, quiero dar las gracias a mi familia y amigos por todo el apoyo en estos años. Comienza un nuevo camino y quiero que sigáis formando parte de él.

Study of the variance of grid-based estimators

Abstract

keywords: Covariogram model, cycloid grid, planar curve, rhomboid grid, unbiased stereology, variance estimator

The main purpose of this project is to estimate the length of a planar curve and to find an expression for its variance. A common method is the unbiased Buffon-Steinhaus estimator which is based on intersection counting with a square grid. Whereas the procedure seems easy to understand, the error variance prediction formulae is a nontrivial problem. An effective approach has been made by Gual-Arnau and Cruz-Orive [2000] and it will be followed in this project. It is based in the study of the covariogram for periodic measurement functions by exploiting symmetric properties. These theoretical results have been applied by Gomez et al. [2016] to study the variance components due to orientations and to intersection counting for different DNA molecules.

In particular, in this thesis, a focus has been set on the study of the $\gamma_{3,6,7}$ curve which was described by Pausinger and Vartziotis [2018] and is considered highly isotropic. Modifications of its isotropic character may help to understand the orientation component of the variance estimator. Furthermore several models have been suggested and tested in order to improve previous estimators. Finally, preceding studies have been compared with non-square grids such as rhomboid and cycloid grids.

Estudio de la varianza de estimadores basados en rejillas

Resumen

palabras clave: Modelo de covariograma, rejilla cicloide, curva plana, rejilla romboide, estereología insesgada, estimador de la varianza

El principal objetivo de este trabajo es estimar la longitud de una curva plana y encontrar una expresión que prediga su varianza. Un método común e insesgado es el de Buffon-Steinhaus, que se basa en contar intersecciones con una rejilla cuadrada. Aunque el procedimiento es fácil de entender, encontrar un estimador de la varianza no es un problema trivial. Gual-Arnau and Cruz-Orive [2000] trata este problema y será seguido en este proyecto. Está basado en el estudio de covariogramas para funciones periódicas, aprovechando sus propiedades simétricas. Estos resultados teóricos se aplican en Gomez et al. [2016] para estudiar las componentes de la varianza debido a la orientación y a la superposición con la rejilla para distintas moléculas de ADN.

En particular, en este proyecto hemos estudiado la curva $\gamma_{3,6,7}$ que es descrita en Pausinger and Vartziotis [2018] y es considerada bastante isotrópica. Modificaciones en su carácter isotrópico mediante deformaciones pueden ayudar a entender el componente de la varianza debido a la orientación. Además, distintos modelos han sido sugeridos y probados con el fin de reducir el error de previos estimadores. Finalmente, se ha extendido su estudio para el caso de rejillas no cuadradas como la romboide o la cicloide.

Contents

1 Introduction

1.1. Motivation

While Algebra or Calculus are 'popular' areas in Mathematics, Stereology is not widely known despite it has a large range of applications. Its main purpose is to estimate geometric quantities of 3D objects in n-dimensional Euclidean spaces (such us length, area surface or volume) when direct measurements are not possible. It is based on sections and intersections with test probes (e.g., lines, planes, grids...).

There are two main branches of Stereology: 'design stereology' and 'model stereology' as described by Cruz-Orive [2017]. The first one concerns objects assumed to be fixed and bounded (e.g., a rock) and sampling must to be performed with isotropic uniform random test systems. In model stereology the target object is considered to be random (e.g., chemical compounds) and the sampling does not have so many requirements. In this project, we are going to focus only on design-based stereology.

One of its main practical purposes is to extract quantitative information from images by using simple and fast algorithms. Other common application of this field is to extract information about three-dimensional objects from two-dimensional measurements made on sampled sections. This can be applied to areas such us microscopy, topography, radiology and, more recently, neuroscience. And as it will be shown in this thesis, these methods combine different areas of mathematics such as Probability, Geometry or Computational Science. Cruz-Orive et al. [2014] reports satisfactory results obtained from the use of stereological methods applied to the human brain to accurately measure the subcortical surface area from digitized images. Another example can be found in the article by Reed and Howard [2003], which measures the eye volumes in Dover sole (*Solea Solea)*. Finally, Batra et al. [1995] estimate capillary length in a left ventricle of a rat heart. In a recent publication, Gomez et al. [2016] estimate the length of DNA molecules from images of their flattened projections. In particular, this project is inspired in this paper and more details will be given below.

Generally speaking, in this thesis, we are interested in obtaining the length of a planar curve and in studying the accuracy of different models suggested. An effective approach is the unbiased Buffon-Steinhaus estimator. And one of the most interesting facts of this method is that the error variance can be predicted. Therefore, we will start the chapter with an introduction to the Buffon-Steinhaus estimator focusing on derivation of the variance estimation.

1.2. Curve length estimator by intersection counting

Let $Y \subset \mathbb{R}^2$ a bounded, planar, piecewise smooth curve of finite length B. A grid can be used to estimate its length. However, as it was introduced above, it has to be isotropic uniform random (IUR). That means, the test lines are both isotropic in direction and uniform random in position. The orientation of the test lines will be isotropic if the angle between a fixed axis and the test lines is $UR[0, 2\pi)$ and the test lines will be uniform random in position if the distance between a fixed origin and the nearest test line is $UR[0, T)$ where T is the gap length between the test lines. As a result, the method can be proven to be unbiased. Then, an unbiased estimator of B is:

$$B_{est} = \frac{\pi}{2} \cdot \frac{a}{l} \cdot I \tag{1}$$

where I is the total number of intersections between the curve and the test lines, and a/l is the grid constant. This grid constant depends on the arrange of the test lines and is given by the area enclosed per unit length as explained by Howard and Reed [1998]. For a parallel-line grid, the grid spacing is the inter-line spacing of the grid T, and the B_{est} expression matches the one given by the *"Buffon's needle"* relationship reported by Schroeder [1974]. Table 1 illustrates the value of the different grid constants handled in this thesis.

Type of grid	Grid constant value
Parallel lines	T
Square grid	$\frac{1}{2}T$
Cycloid grid	πr
Rhomboid grid	$\frac{T}{2} sin(\frac{\pi}{\sqrt{3}})$

Table 1: Grid constants for the most commonly used test probes. T is the gap length between straight lines in the case of parallel lines and square grids. For a rhomboid grid, T represents the side length of one of its rhombus. Finally, r is the radius of the generating circle of a cycloid grid.

Moreover, all the grids used are periodic in space. Thus, we can define the *fundamental tile* of a grid as the primitive cell which can build up the whole grid using only translations. In the case of square and rhomboid grids, the fundamental tile will be squares and rhombus of length size T. In Fig. 3 a sketch of the fundamental tile of a cycloid grid is shown and $T = 4\pi r$ is fixed in one of the axis.

1.2.1. Length estimation with a square grid - Buffon-Steinhaus method

Taking into account the B_{est} formulae introduced above in Eq. (1) and the grid constants, the unbiased Buffon- Steinhaus estimator which is based on intersection counting with a square grid, can be expressed as,

$$\hat{B} = \frac{\pi}{4} \cdot T \cdot I$$

where T is the gap length or grid spacing of a square grid and I the total number of intersections between the curve and the test lines, as illustrated in Fig. 1.

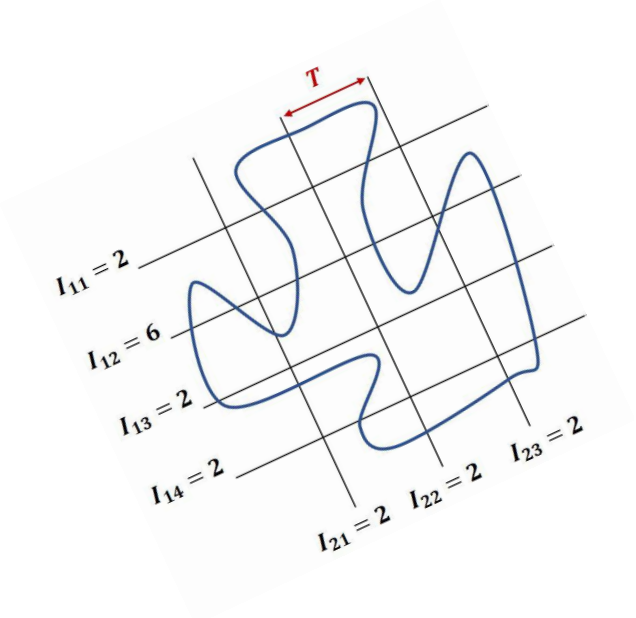


Figure 1: Square grid superimposition on a planar curve. T is the gap between test lines and $I = 18$ is the total number of the intersections.

Firstly, the purpose is to estimate the finite length B of the curve Y. An efficient approach starts considering the total orthogonal projection $l(w)$ of Y into a fixed axis making an angle w, as illustrated in Fig. 2. However, before to attack the root of this problem, it is convenient to introduce the Cauchy's projection formula with a simple example as is done by Kendall and Morgan [1963].

Consider a straight segment of length b. Then, the projection of this segment onto a fixed axis making an angle w is

$$proj_w(b) = b\,|cos(w)|.$$

Assuming that the angle w is uniform random in the interval $[0, 2\pi)$, i.e., $w \sim$ UR[0, 2π), the mean value of $proj_w(b)$ can be expressed as

$$E[proj_w(b)] = \frac{1}{2\pi}\int_0^{2\pi} b\,|cos(w)|\,dw = \frac{2b}{\pi}$$

and therefore an unbiased estimator of b will be

$$\tilde{b} = \frac{\pi}{2}\,proj_w(b)$$

On the other hand, the curve Y can be represented as

$$Y = \bigcup_{i=1}^{N} y_i$$

where y_i denotes the ith straight line segment of Y with length b_i, and N is the finite total number of segments. As a consequence, the total length of Y can be expressed as

$$B = \sum_{i=1}^{N} b_i. \tag{2}$$

Coming back to the original problem, the total orthogonal projection $l(w)$ of Y will be

$$l(w) = \sum_{i=1}^{N} b_i \, |cos(w)|$$

and therefore, the total length of Y may be estimated by

$$\tilde{B}(w) = \frac{\pi}{2} \, l(w). \tag{3}$$

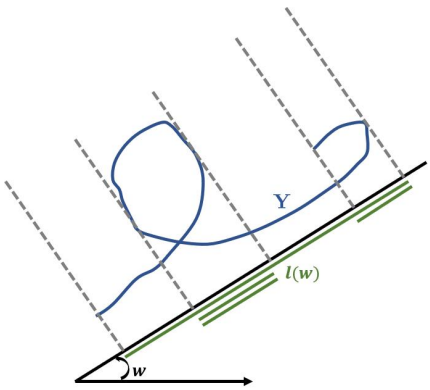


Figure 2: Sketch of the Cauchy method applied to a finite plane curve Y. The estimated length B is $\pi/2$ times the mean of its total projected length $l(w)$ when w is UR[0,2π).

This is an unbiased estimator of B. Furthermore, in the case of a square grid, two projections $l(w) = l(w + \pi)$ are estimated at the same time and, as a consequence, it is enough to take $w \sim$UR$[0, \pi)$. Measuring also the projected length $l(w + \frac{\pi}{2})$, another unbiased estimator for B is

$$\hat{B}(w) = \frac{\pi}{4} \left[l(w) + l\left(w + \frac{\pi}{2}\right) \right]. \tag{4}$$

Both methods can be compared in terms of their estimated variance. And as it is expected, $\hat{B}$ is more precise that $\tilde{B}$ since more observations are being taken. In addition, the more isotropic (in geometry) the curve, the closer the two projections are because they do not change significantly with the angle. Therefore, the less isotropic the curve is, the more $\hat{B}$ and $\tilde{B}$ differ.

In summary, it will imply that the error variance will be smaller in the case of isotropic planar curves. These assumptions will be tested later for different curves.

Furthermore, the total orthogonal projected length $l(w)$ and $l(w + \pi/2)$ can be estimated from the intersection counting between the grid and the planar curve. An unbiased estimator is:

$$\hat{l}(w) = T \sum_{j=1}^{n_1} I_{1j}(w) \qquad \hat{l}\left(w + \frac{\pi}{2}\right) = T \sum_{j=1}^{n_2} I_{2j}\left(w + \frac{\pi}{2}\right). \tag{5}$$

It is helpful to introduce the notation $I_1(w) = \sum_{j=1}^{n_1} I_{1j}(w)$ and $I_2(w) = \sum_{j=1}^{n_1} I_{2j}(w + \pi/2)$ representing the total number of intersection between the curve and the n_1 vertical and n_2 horizontal test lines of the grid respectively. For the example illustrated in Fig. 2, I_1=6 and I_2=12.

1.2.2. Length estimation with a cycloid grid

The Buffon-Steinhaus method is based on intersection counting with a square grid. Despite of the fact that is an unbiased method, some orientations seems to be favored. For instance, if the line segments of the curve are parallel to the test lines, some intersections are missing and as consequence, the obtained estimated length will be lower than the true value.

This fact leads to the study of an alternative grid. Here, a cycloid grid is suggested and will allow us to solve the problem outlined above. Its graph is illustrated in Fig. 3. Furthermore, as a reminder, the test lines of any system must be IUR, as was explained in section 1.2.1. Since this particular test system will contains all possible orientations of the line segments, there is not need of making isotropic orientations of the test system. Therefore we can use one given orientation of the test probe and we just need to be sure that they are uniform random in position.

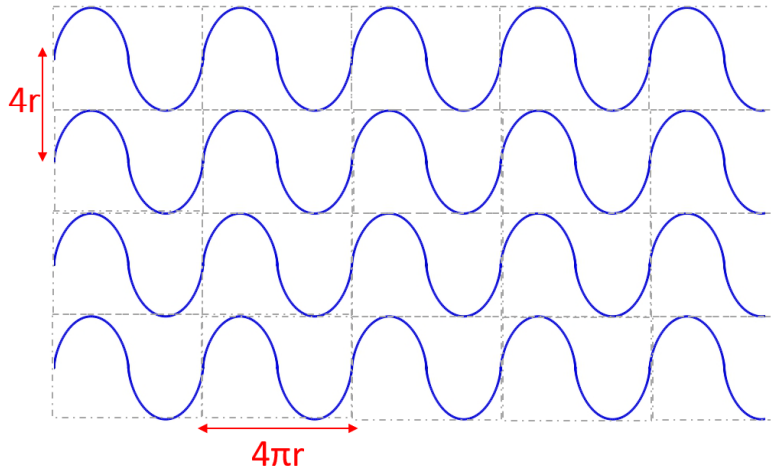


Figure 3: Cycloid test system with radius $r = 0.5$ and grid spacing of 2 units. Each of the cells encloses a fundamental tile.

The cycloid lines are generated by a circle rolling along the horizontal plane, with

$$x = r(t - sint)$$
$$y = r(1 - cost)$$

where t is a real parameter that corresponds to the angle through which the rolling circle has rotated and r is the radius of the circle which generates the cycloid arc. The grid spacing is set

equal to four times the radius, as mentioned by Cruz-Orive et al. [2014]. Therefore, the grid constant is $a/l = 4 \cdot r \cdot \pi/4 = \pi r$ (see Table 1). This test system was studied by Cruz-Orive et al. [2014] where they obtained satisfactory estimations of the surface area by overlaying this test system over digitized brain sections. Previous work by Gokhale et al. [2004] propose to use virtual cycloids to estimate the surface area in thick tissue sections.

1.2.3. Length estimation with a rhomboid grid

The square grid can be easily programmed and its symmetry allows for a fast method. However, in this thesis, grids with different angles (rhomboid grid) will be studied since it seems to have less orientations favored than a square grid. Using the Legendre-Gauss quadrature method:

$$\int_a^b f(x)dx \;=\; \frac{b-a}{2}\int_{-1}^{1} f(x(t))\,dt \;\approx\; \frac{b-a}{2}\left[f\left(x\left(-\frac{1}{\sqrt{3}}\right)\right)+f\left(x\left(\frac{1}{\sqrt{3}}\right)\right)\right]$$

where $x = \frac{a+b}{2} + \frac{b-a}{2}t$, $t \in [-1, 1]$, the following integral may be computed:

$$\int_0^{\pi} f(x)dx \;\approx\; \frac{\pi}{2}\left[f\left(\frac{\pi}{2}\left(1-\frac{1}{\sqrt{3}}\right)\right)+f\left(\frac{\pi}{2}\left(1+\frac{1}{\sqrt{3}}\right)\right)\right]$$

Therefore, if we change the grid angle, instead of using the nodes 0 and $\frac{\pi}{2}$ in the interval [-1,1], we will use $\frac{\pi}{2}\left(\frac{-1}{\sqrt{3}}\right)$ and $\frac{\pi}{2}\left(\frac{1}{\sqrt{3}}\right)$ in the interval $[0,\pi]$.

Finally, the length is given by Eq. (1). The grid constant is the area of the test system per unit length. For a rhombus of side length T and angle $\pi/\sqrt{3}$, the area is base times height, that is $T^2 \sin(\pi/\sqrt{3})$. Hence, the grid constant is $a/l = (T^2 \sin(\pi/\sqrt{3}))/2T$, and the total length can be unbiased estimated as

$$B_{est} \;=\; \frac{\pi}{4} \cdot T \cdot \sin\left(\frac{\pi}{\sqrt{3}}\right) \cdot I \tag{6}$$

where I is the number of intersections between the curve and the test lines.

1.3. Estimation of the variance using covariograms

In this section a general estimation of the variance using covariograms will be presented. In particular, in section 1.3.1 this developed method will be applied to get the variance predictor of the Buffon-Steinhaus estimator. Moreover, it will be used to get new models for the variance estimator in section 2.

Let f be a function with bounded support. Then, the covariogram is defined as

$$g(h) = \int f(x)\; f(x+h)\; dx. \tag{7}$$

Now, consider V as a target parameter which want to be estimated, namely the area under the curve f. The parameter V can be expressed by the integral

$$V = \int f(x)\, dx \tag{8}$$

where f is called the measurement function and it is a non-random periodic function of period r, of bounded support, square integrable and piecewise continuous. In addition, f is perfectly determined.

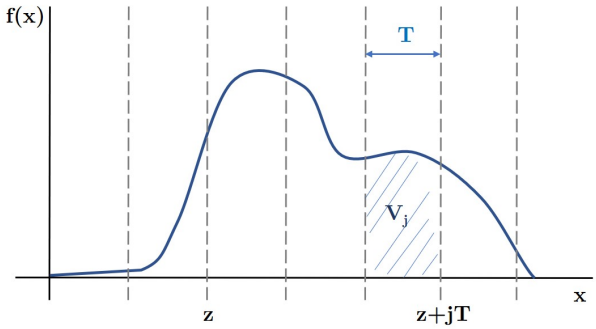


Figure 4: Measurement function f. Samples are taken at space T.

By taking observations (using a regular sampling test-lines of spacing T) at $\{z + jT\ ,\ j = ... - 2, -1, 0, 1, 2, ...\}$, as illustrated in Fig. 4, an unbiased estimator of V is

$$\hat{V} = T \sum_{j=-\infty}^{\infty} f(z + jT) \equiv \vartheta(z). \tag{9}$$

First, note that the infinity sum is well defined since f is of bounded support. Second, $\vartheta(z)$ is a function of z of period T. Thus, it is sufficient to study the range [0,T).

Assuming $z \sim$ UR(0,T], the variance of the random variable $\vartheta(z)$ is calculated

$$var[\vartheta(z)] = E[\vartheta^2(z)] - (E[\vartheta(z)])^2 = \int_0^T \frac{1}{T}\vartheta^2(z)dz - \left(\int_0^T \frac{1}{T}\vartheta(z)dz\right)^2. \tag{10}$$

The first term in the right-hand can be expressed

$$\frac{1}{T}\int_0^T \vartheta^2(z)dz = \frac{1}{T}\int_0^T T^2 \sum_p \sum_q f(z + pT)f(z + qT)dz$$

using q = p + k, the preceding equation becomes

$$T\int_0^T \sum_k \sum_p f(z + pT)f(z + pT + kz)dz = T\sum_k \int_0^T \sum_p f(z + pT)f(z + pT + kT)dz =$$
$$= T\sum_k \int_{-\infty}^{\infty} f(x)f(x + kT)dx)$$

which is the covariogram definition (see Eq. (7)). Therefore,

$$\int_0^T \frac{1}{T}\vartheta^2(z)dz = T\sum_k g(kT). \tag{11}$$

The second term in the right-hand side of Eq. (10) can be expressed as

$$\left(\int_0^T \frac{1}{T}\vartheta(z)dz\right)^2 = \left(\int_0^T \frac{1}{T}T\sum_j f(z+jT)\right)^2 = \left(\sum_j \int_0^T f(z+jT)\right)^2 = \left(\sum_j Vj\right)^2 = V^2.$$

Thus, $\hat{V}$ is an unbiased estimator of V. Moreover, $V^2 = (\int f(x)dx)^2 = \int g(h)dh$ and, as a consequence,

$$\left(\int_0^T \frac{1}{T}\vartheta(z)dz\right)^2 = \int_{-\infty}^{\infty} g(h)dh \tag{12}$$

Using Eq. (11) and (12) in the right-hand side of Eq. (10), the variance may be expressed as

$$var[\vartheta(z)] = T\sum_{k=-\infty}^{\infty} g(kT) - \int_{-\infty}^{\infty} g(h)dh. \tag{13}$$

Therefore, the variance has two components. The first one is the approximate value of the integral $\int g(h)dh$, whereas the second component is the integral itself. Hence, it can be interpreted as the difference between the histogram area and the curve area. When the grid spacing T is large, this formula has only a few terms and it can be computed quickly. However, for smaller grid spacing, it may be impossible to solve directly.

Now, it is desired to find an alternative variance formula based on the Fourier transform of the covariogram g. Since $\vartheta(z)$ is periodic function (of period T), the estimator can be expressed as

$$\vartheta(z) = \sum_{n=0}^{\infty}\left(a_n cos\left(\frac{2\pi nz}{T}\right) + b_n sin\left(\frac{2\pi nz}{T}\right)\right)$$

where

$$a_n = \frac{2}{T}\int_0^T \vartheta(z)cos\left(\frac{2\pi nz}{T}\right)dz \qquad b_n = \frac{2}{T}\int_0^T \vartheta(z)sin\left(\frac{2\pi nz}{T}\right)dz.$$

Alternatively,

$$\vartheta(z) = \sum_{n=-\infty}^{\infty} c_n e^{\frac{2\pi inz}{T}}$$

where

$$c_n = \begin{cases} \frac{1}{2}(a_n - ib_n) = \frac{1}{T}\int_0^T \vartheta(z)e^{-\frac{2\pi inz}{T}} & if \quad n > 0 \\ \frac{1}{2}(a_n + ib_n) = \frac{1}{T}\int_0^T \vartheta(z)e^{\frac{2\pi inz}{T}} & if \quad n < 0 \end{cases}$$

Using Eq. (10), for $n > 0$ it can be written

$$c_n = \frac{1}{T}\int_0^T \vartheta(z)e^{-\frac{2\pi i n z}{T}} = \frac{1}{T}\int_0^T T\sum_{j=-\infty}^{\infty} f(z+jT)e^{-\frac{2\pi i n z}{T}} = \int_{-\infty}^{\infty} f(z)e^{-\frac{2\pi i n z}{T}}dz$$

and the conjugate $\bar{c_n}$ for $n < 0$. Hence, $|c_n|^2 = c_n\bar{c_n} = \frac{1}{4}(a_n^2 + b_n^2)$

Furthermore, Parseval theorem

$$\frac{1}{T}\int_0^T \vartheta^2(z)dz = \frac{1}{2}\sum_{n=0}^{\infty}(a_n^2 + b_n^2)$$

can be applied and the first term in the right-hand side of Eq. (10) will become

$$\frac{1}{T}\int_0^T \vartheta^2(z)dz = \sum_{n=-\infty}^{\infty} c_n\bar{c_n}.$$

A new expression of the variance is obtained

$$var[\vartheta(z)] = \sum_{-\infty}^{\infty} c_n\bar{c_n} - \int_{\infty}^{\infty} g(h)dh.$$

Now, consider the Fourier transform

$$F(u) = \int_{-\infty}^{\infty} f(z)e^{-2\pi i u z}dz$$

Then,

$$var[\vartheta(z)] = \sum_{n=-\infty}^{\infty} F\left(\frac{n}{T}\right)\bar{F}\left(\frac{n}{T}\right) - \int_{-\infty}^{\infty} g(h)dh \tag{14}$$

where $\bar{F}$ is the conjugate of F. The Fourier transform of the covariogram g defined in Eq. (7) is

$$G(u) = \int_{-\infty}^{\infty} g(h)e^{-2\pi i u h}dh = \int_{\infty}^{\infty} f(x)dx\int_{-\infty}^{\infty} f(x+h)e^{-2\pi i u h}dh =$$
$$= \int_{-\infty}^{\infty} f(x)dx\int_{-\infty}^{\infty} f(r)e^{-2\pi i(r-x)}dh = F(u)\bar{F}(u)$$

Replacing this result in Eq. (14), taking into account that $G(0) = \int_{-\infty}^{\infty} g(h)dh$ and that the covariogram is symmetric $g(-h) = g(h)$ about $h = 0$, a new version of the variance is given

$$var[\vartheta(z)] = \sum_{n=-\infty}^{\infty} G\left(\frac{n}{T}\right) - G(0) = 2\sum_{n=1}^{\infty} G\left(\frac{n}{T}\right) \tag{15}$$

This lastest version of the variance shows that it only depends linearly on the covariogram G and on the test-lines spacing T. Therefore, if the covariogram were known, the estimation variance would be a trivial exercise. Unfortunately, this is not the case. As mentioned above, only samples of f at $\{z + jT\ ,\ j = ... - 2, -1, 0, 1, 2, ...\}$ are given.

Consider suppose that the covariogram is a polynomial

$$g(h) = \sum_{j=0}^{r} a_j |h|^j$$

where a_j are unkown coefficients. The Fourier transform is

$$G(u) = 2\int_{-\infty}^{\infty} \left(\sum_{j=0}^{r} a_j |h|^j \right) e^{-2\pi i u h}\, dh \;=\; 2\sum_{j=0}^{r} a_j \int_{-\infty}^{\infty} h^j e^{-2\pi i u h}\, dh$$

An efficient way to calculate this integral is using the Gamma Function

$$\Gamma(m) = \int_0^{\infty} y^{m-1} e^{-y} dy$$

and applying the variable change $y = m - 1$ and $y = 2\pi u i h$, the required integral becomes

$$\frac{1}{(2\pi i u)^{j+1}} \int_0^{\infty} y^j e^{-y} dy = \frac{\Gamma(j+1)}{(2\pi i u)^{j+1}}$$

As the variance is a real function, it is only needed the real part of $G(u)$. Then, only the terms of odd order contribute, i.e. of the form $j = 2k - 1$ with $k = 1, 2, ...$ Using Eq. (15),

$$var[\vartheta(z)] = 2\sum_{n=1}^{\infty} \Re\left[G\left(\frac{n}{T}\right)\right] = 2\sum_{k=1}^{s} a_{2k-1} T^{2k} \left[2\, \frac{\Gamma(2k)}{(-1)^k (2\pi)^{2k}} \sum_{n=1}^{\infty} \frac{1}{n^{2k}} \right]. \tag{16}$$

On the other hand, from Abramowitz and Stegun [1999], the Bernoulli number $B_{2k} \equiv B_{2k}(0)$, i.e. the evaluation of the Bernoulli polynomial in 0, can be expressed as

$$B_{2k}(0) = \frac{(-1)^{k-1}\, 2\, \Gamma(2k+1)}{(2\pi)^{2k}} \sum_{n=1}^{\infty} \frac{1}{n^{2k}} \tag{17}$$

Thus, the estimation variance from Eq. (16) can be written as

$$var[\vartheta(z)] = -\sum_{k=1}^{\frac{r+1}{2}} T^{2k}\, a_{2k-1} \frac{B_{2k}}{k} = -a_1 B_2 T^2 - \frac{1}{2} a_3 B_4 T^4 - \dots \tag{18}$$

In particular, $a_1 = g'(0)$ under the polynomial model and $B_2 = 1/6$. Hence,

$$var[\vartheta(z)] \;\simeq\; -\frac{1}{6}\, g'(0)\, T^2 \tag{19}$$

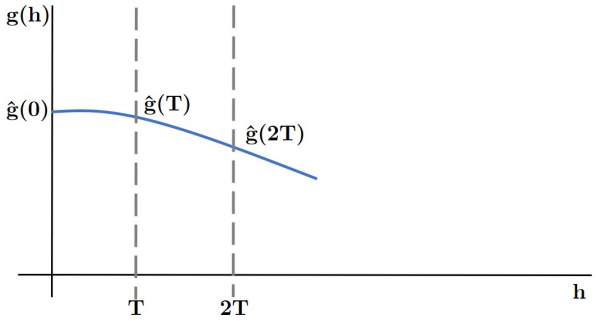


Figure 5: Quadratic curve fitting of the covariogram.

and the next term with $B_4 = -1/30$ will be $\frac{1}{60}\, a_3\, T^4$. Therefore, Eq. (19) provides a good estimation of the variance. Indeed, the variance increase quadratically with the test-lines spacing T.

As illustrated in Fig. 5, the next step is to fit the covariogram with a parabola

$$g(h) = a_0 + a_1 h + a_2 h^2 \tag{20}$$

though the three sample points $\{(jT,\ \hat{g}(jT))\ ; j = 0, 1, 2\}$. The covariogram can be expressed as

$$\begin{aligned} \hat{g}(jT) &= T \sum_{r=1}^{S-j} f_r f_{r+j} \\ f_j &\equiv f(z + jT) \qquad \hat{g}(jT) \equiv T g_j \end{aligned} \tag{21}$$

where S is the total number of test lines and f_i denotes the ith section area observed at the ith abscissa. It is possible to write

$$g_0 \equiv \sum_{i=0}^{S} {f_i}^2 \qquad g_1 \equiv \sum_{r=0}^{S-1} f_i f_{i+1} \qquad g_2 \equiv \sum_{r=0}^{S-2} f_i f_{i+2} \tag{22}$$

Furthermore, replacing the (known) sample points in the Eq. (20)

$$\begin{aligned} T g_0 &= a_0 \\ T g_1 &= a_0 + a_1 T + a_2 T^2 \\ T g_2 &= a_0 + a_1 (2T) + a_2 (2T)^2 \end{aligned}$$

which is a system of equations with three variables: a_0, a_1 and a_2. Finally, this coefficients can be obtained

$$a_0 = T g_0 \qquad a_1 = \frac{-3g_0 + 4g_1 - g_2}{2} \qquad a_2 = \frac{g_2 + g_0 - 2g_1}{2T} \tag{23}$$

Particularly, the value obtained for a_1 can be replaced in Eq. (19), since $g'(0) = a_1$, and the new estimation of the variance can be written as

$$var[\vartheta(z)] \ \simeq\ \frac{1}{6}\ \frac{3g_0 - 4g_1 + g_2}{2}\ T^2. \tag{24}$$

The definition of the coefficient error is

$$ce[\vartheta(z)] = \frac{var[\vartheta(z)]^{1/2}}{\vartheta(z)}. \tag{25}$$

In this way, the variance predictor is normalised and Eq. (24) can be expressed as

$$ce[\vartheta(z)] \simeq \frac{1}{2\sum f_i}\sqrt{\frac{3\sum {f_i}^2 - 4\sum f_i f_{i+1} + \sum f_i f_{i+2}}{3}}$$

because $\vartheta(z) = T\sum f_i$. Here $ce[\vartheta(z)]$ is the estimated coefficient error whereas $CE[\vartheta(z)]$ will denote the empirical one in the following sections.

1.3.1. Estimation of the variance by intersection counting with a square grid

As it was presented in 1.2.1, the Buffon-Steinhaus method is based on intersection counting with a squared grid. In this new section, the estimation variance described above will be adapted to the particular case of a square grid intersecting with any curve. Hence, the covariogram can be defined as

$$g(h) = \int_0^{\pi} l(w)\, l(w+h)\, dw \tag{26}$$

where $l(w)$ is the orthogonal projected length (here the measurement function of period π) and $w \sim \mathrm{UR}[0,\pi)$, as it was introduced in Section 1.2.1.

The sample points will be $\{(jT, \hat{g}(jT)); j = 0, \frac{\pi}{2}, \pi\}$. Later it will be assumed that is enough to sample in $\{(jT, \hat{g}(jT); j = 0, \frac{\pi}{2}\}$ because of the symmetry of the grid. Furthermore, Eq. (21) can be written as

$$\hat{g}(jT) = T\sum_{r=1}^{S-j} l_r l_{r+j}$$
$$l_j \equiv l(w+jT) \qquad \hat{g}(jT) \equiv T g_j$$

and taking into account the symmetry of the square grid, namely $l(w) = l(w+\pi)$, Eq. (22) will become

$$\begin{aligned} g_0 &= l^2(w) + l^2\left(w+\frac{\pi}{2}\right) \\ g_{\frac{\pi}{2}} &= l(w)\, l\left(w+\frac{\pi}{2}\right) + l\left(w+\frac{\pi}{2}\right) l(w+\pi) = 2\, l(w)\, l\left(w+\frac{\pi}{2}\right) \\ g_{\pi} &= l(w)\, l(w+\pi) + l\left(w+\frac{\pi}{2}\right) l\left(w+\frac{3\pi}{2}\right) = l^2(w) + l^2\left(w+\frac{\pi}{2}\right) = g_0 \end{aligned} \tag{27}$$

The coefficient of interest a_1 of Eq. (23) is

$$a_1 = -2\,(g_0 - g_{\frac{\pi}{2}}) \tag{28}$$

Then, to estimate the variance, it may be applied Eq. (19) with $g'(0) = a_1$ and $T = \pi/2$,

$$var[\vartheta(z)] \simeq \frac{1}{3}\left(g_0 - g_{\frac{\pi}{2}}\right)\frac{\pi^2}{4} = \frac{\pi^2}{12}\left[l(w) - l\left(w + \frac{\pi}{2}\right)\right]^2. \tag{29}$$

From this equation, it can be observed that the variance exhibits quadratic growth with the projections difference. Therefore, for closer projections, this term becomes very small and the variance tends to zero. And, in particular, isotropic curves, will have similar projections in both directions and the variance is expected to be negligible.

Finally, for the latter discussion convenience, the square coefficient of error (variance predictor divided by the true length), the empirical mean and the error variance of the estimator are defined respectively as follows,

$$ce^2[\hat{B}(w)] = \frac{var[\tilde{B}(w)]}{B^2} \tag{30a}$$

$$\mathbb{E}_e[\hat{B}(w)] = \frac{1}{M}\sum_{k=1}^{M}\hat{B}(w_k) \tag{30b}$$

$$Var_e[\hat{B}(w)] = \frac{1}{M}\sum_{k=1}^{M}\left[\hat{B}(w_k) - \mathbb{E}_e[\hat{B}(w)\right]^2 \tag{30c}$$

where M is the corresponding replications of $\hat{B}(w)$ computed in the results section and w_k are the values in the set:

$$w_k = (U + k - 1)\frac{\pi}{2M}\ ,\ k = 1, 2, ..., M \qquad U \sim UR[0, 1). \tag{31}$$

In fact, it is possible to obtain a more accurate estimator of the variance than Eq. (29) by exploiting the periodicity of the measurement function and the properties of the Bernoulli polynomials.

Suppose that the covariogram is defined between 0 and π and consider the second-degree polynomial $g(x) = a_0 + a_1x + a_2x^2$. The Bernoulli polynomials of degree $\alpha \geq 0$ have several properties:

- The Bernoulli polynomial is symmetric, i.e. $(-1)^\alpha B_\alpha(x) = B_\alpha(1 - x)$
- Its integral between zero and one is zero.

By assumption, the covariogram is modeled as a polynomial. Gual-Arnau and Cruz-Orive [2000] proofs that if a polynomial of second degree satisfies $g(x) = g(\pi - x)$, then, following the steps described in Appendix 5, there exist β_0, β_1 such that

$$g(x) = \beta_0 + \beta_1 B_2\left(\frac{x}{\pi}\right).$$

Gual-Arnau and Cruz-Orive [2000] considered higher degree polymomials and related the degree with the differentiability of the measurement function. Following their notation, and taking into account that we have only two samples at angles $0, \frac{\pi}{2}$, Eq. 13 gives:

$$\frac{\pi}{2}\left(g(0)+g\left(\frac{\pi}{2}\right)\right)-\int_0^{\pi} g(x)dx \;=\; \frac{\pi\beta_1}{2}\left(B_{2m+2}(0)+B_{2m+2}\left(\frac{1}{2}\right)\right) \tag{32}$$

where m denotes the (differentiability) class C^m of the measurement function.

On the one hand, for $m = 0$, and considering the Bernoulli polynomials of degree 2, the right-hand side of Eq. (32) becomes

$$-\frac{\pi\beta_1}{2}\left(-\frac{1}{12}\right)$$

On the other hand, for $m = 1$, and considering the Bernoulli polynomials of degree 4, the right-hand side of Eq. 32 becomes

$$-\frac{\pi\beta_1}{2}\left(-\frac{1}{240}\right)$$

Finally, recalling the values of β_1, β_0, the substitution done in equation Eq. (29) gives the following expression for the variance estimator:

$$var[\vartheta(z)] \;=\; \frac{\pi^2}{240}\left[l(w)\;-\;l\left(w\;+\frac{\pi}{2}\right)\right]^2 \tag{33}$$

1.4. Moore–Aronszajn theorem

As it may be noticed, it has been assumed that the length of the curve can be represented as a sum of observations of the measurement function (see Eq. (9)). However, this is not a trivial assumption and it resides in the Moore–Aronszajn theorem.

Moore–Aronszajn theorem (Aronszajn [1950]) **:** Suppose $K(x, y)$ is a symmetric, positive definite kernel on a set X. Then, there is a unique Hilbert space of functions on X for which K is a reproducing kernel. Even more, the functions f in this Hilbert space can be represented as

$$f(x) = \sum_{i=1}^{\infty} \alpha_i K(x, z_i)$$

A proof and a nice introduction to Reproducing kernels can be found in the article by Aronszajn [1950].

We are not going deeper in this theorem. Nevertheless, in our case, it is necessary to check that the polynomial $g(h)$ from Eq. (20) satisfies the hypothesis of the theorem, i.e, that the function is positive definite:

$$a_0^2\, g_2\,(x_0 - x_0) \;+\; a_0\, a_1\, g_2(x_1 - x_0) \;+\; a_1\, a_0\,(x_0 - x_1) \;+\; a_1^2\, g_2\,(x_1 + x_1) \;\geq\; 0$$

Doing the substitutions for a_0, a_1, g_0 and g_1 described above and using $x_0 = 0$ and $x_1 = \pi/2$, this expression was quickly checked (e..g. using *Sage*). Therefore, this theorem ensures the existence of such decomposition of the covariogram presented in this thesis and it will be applied later to get new expressions of the variance predictor by exploiting the covariogram model.

1.5. Curve $\gamma_{3,6,7}$

In Gomez et al. [2016] two DNA molecules were studied: one considered "fairly isotropic" and one considered "fairly anisotropic" in space. They were compared in terms on the variance estimator due to the use of test lines (also called local error or Cavalieri component) and due to the orientation of the test system (also known as the Cauchy component). Motivated by this paper, we decided to study in more detail this orientation component. To carry out this research, we suggested to test the same estimators used in the paper cited with a highly isotropic curve: $\gamma_{3,6,7}$.

The $\gamma_{3,6,7}$ is defined via the sum of exponentials as described by Pausinger and Vartziotis [2018]. Let $a = (a_0, \ldots, a_m)$ denote a vector of positive integers $a_0, \ldots, a_m$ with $m \geq 1$. The closed curve $\gamma_a : \mathbb{R} \to \mathbb{C}$ is defined as

$$\gamma_a(t) = \sum_{j=0}^{m} e^{2\pi i a_j t}$$

where i denotes the unit imaginary number. In particular, the curve $\gamma_{3,6,7}$, is defined as following

$$\gamma_{3,6,7}(t) = e^{2\pi i 3t} + e^{2\pi i 6t} + e^{2\pi i 7t}$$

The $\gamma_{3,6,7}$ selection has been based on:

- The γ_a graphs can be highly symmetric. The $\gamma_{3,6,7}$ curve is symmetric with respect to the x-axis, as illustrated in Fig. 6. Furthermore, it seems to be fairly isotropic and therefore it will allow us to propose and study length and variance estimators for this kind of curves.

- The self intersections of γ_a have been studied by Pausinger and Vartziotis [2018]. However, the $\gamma_{3,6,7}$ is mentioned as a curve intersecting in "a very well structured way; which turn out to be very hard to describe in an explicit fashion". These self intersections provide a new scenario to see how the length and variance estimators behave.

For the $\gamma_{3,6,7}$ the following aspects will be studied:

1. Behaviour of the $\tilde{B}(w)$ estimator as a function of the angle $w \in [0, \pi]$.

2. Behaviour of the $\hat{B}(w)$ estimator using two orthogonal projected lengths.
3. Test the length and error variance estimators based on intersection counting with square grids.
4. Behaviour of the length and error variance estimators under deflections.
5. Comparison with DNA molecules studied by Gomez et al. [2016].
6. Test of the new variance predictors proposed in Section 2.

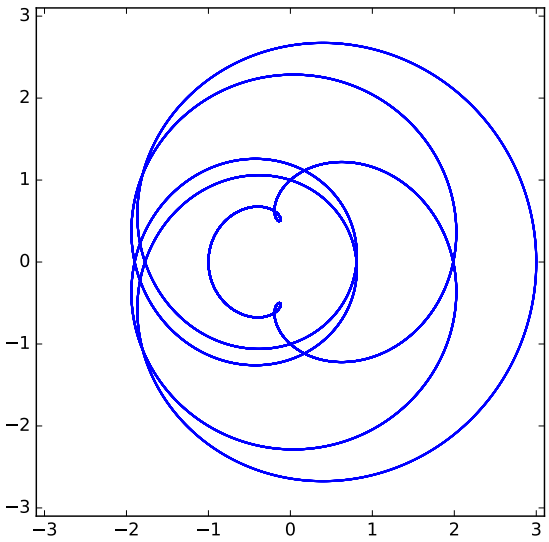


Figure 6: Plot of the curve $\gamma_{3,6,7}$.

2 Models

In the previous section, the variance estimation has been obtained by fitting the covariogram with a second-degree polynomial (see Eq. (20)). An alternative polynomial suggested by Cruz-Orive and Gual-Arnau [2002] was:

$$M(x) \;=\; a\,\pi\,B_4\!\left(\frac{x}{\pi}\right) \;+\; b$$

where a and b are unknown numerical coefficients and $B_4(x)$ is the Bernoulli polynomial of degree 4.

Now, we are going to propose several new models which attempt to model the behaviour of the Cauchy component in the variance estimator. Taking into account that the covariogram is assumed to be periodic and symmetric $g(h) = g(\pi - h)$, we can consider the Bernoulli polynomials of even degree $n \geq 0$ since they satisfy $B_n(1-x) = B_n(x)$, as it was mentioned in Section 1. In particular, we will use the second, fourth and sixth degrees which are illustrated in Fig. 7 . Thus, if for instance we are sampling on the semicircle the covariogram has maximum values at 0 and π and a minimum at $\pi/2$.

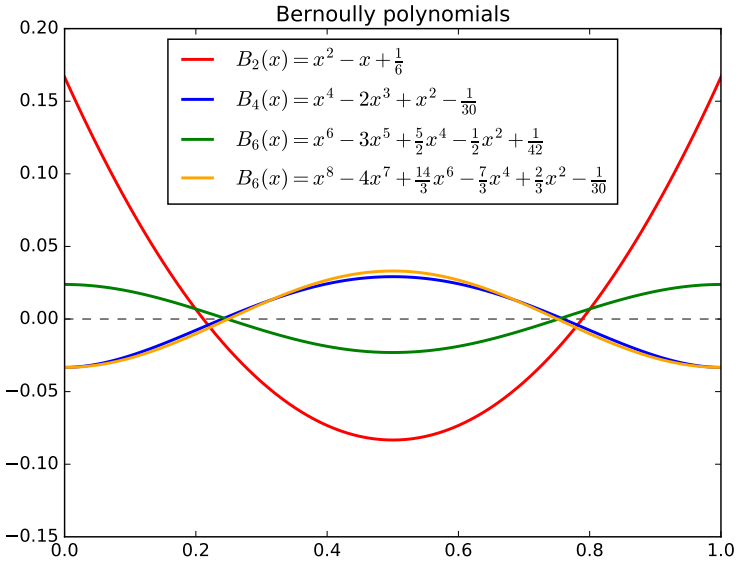


Figure 7: Bernoulli polynomials of degree 2, 4, 6 and 8.

2.1. Proposed models with a square grid

1^{st} model

It is trivial that the periodic functions in $[0, 2\pi)$ are in correspondence with continuos functions in the circle $\mathbb{S}$. The first suggested model is:

$$M_1\ (x, y)\ =\ a((x-1)^2 + y^2) + b$$

where a and b are real parameters defined as follows.

In particular, and as we are sampling on the semicircle, it may result useful define the following change of variables:

$$\begin{aligned} \psi\colon [0, 2\pi) &\to \mathbb{S}^1 \\ x &\mapsto (cos(x), sin(x)) \end{aligned}$$

and $M_1 = M \circ \psi^{-1}$. Therefore, on the one hand

$$\begin{aligned} M_1(1,0) &= M \circ \psi^{-1}(1,0) = M(0) \\ M_1(0,1) &= M \circ \psi^{-1}(0,1) = M\left(\frac{\pi}{2}\right) \end{aligned} \tag{34}$$

and on the other hand,

$$\begin{aligned} M_1(1,0) &= b \\ M_1(0,1) &= 2a + b \end{aligned} \tag{35}$$

Then, from Eq. (34) and (35), the coefficients a and b can be obtained solving a simple linear system of equations. Namely,

$$a = \frac{1}{2}\left[M\left(\frac{\pi}{2}\right) - M(0)\right] \qquad\qquad b = M(0)$$

Moreover, using polar coordinates the following integral can be computed,

$$\int_0^{\pi} M_1(x, y)\ d\mathbb{S} = 2\pi a + b\pi = \pi\ M\left(\frac{\pi}{2}\right)$$

Finally, to estimate $var[\vartheta(z)]$ we may apply Eq. (13),

$$var[\hat{\vartheta}(z)] = \frac{\pi}{2}\left(M(0) - M\left(\frac{\pi}{2}\right)\right)$$

and from Eq. (27),

$$var[\hat{\vartheta}(z)] = \frac{\pi}{2}\left[l(w) - l\left(w + \frac{\pi}{2}\right)\right]^2 \tag{36}$$

2^{nd} **model:**

Here we will follow the reproducing kernel proposed by [Brauchart and Dick, 2013, Eq. 2.5]. Similarly to model $M_1(x, y)$, we define:

$$M_2(x, y) = a\left(1 - \frac{1}{\pi}\sqrt{(x-1)^2 + y^2}\right) + b$$

where a and b are real parameters defined as follows.

Similarly to model $M_1(x, y)$, we define the following change of variables:

$$\begin{aligned} \psi\colon [0, \pi] &\to \mathbb{S}^1 \\ x &\mapsto (cos(2x), sin(2x)) \end{aligned}$$

and $M_2 = M \circ \psi^{-1}$. Therefore, on the one hand

$$\begin{aligned} M_2(1, 0) &= M \circ \psi^{-1}(1, 0) = M(0) \\ M_2(-1, 0) &= M \circ \psi^{-1}(-1, 0) = M\left(\frac{\pi}{2}\right) \end{aligned} \tag{37}$$

Sampling on the semicircle, we can compute the following integral

$$\int_0^{\pi} M_2(x, y) d\mathbb{S} \;=\; \pi(a + b) \;-\; \frac{4a}{\pi}$$

Moreover, since $M_2(1, 0) = a + b = g_0$ and $M_2(-1, 0) = a\left(1 - \frac{2}{\pi}\right) + b = g_{\frac{\pi}{2}}$, we get

$$a = \frac{\pi}{2}\left(g_0 - g_{\frac{\pi}{2}}\right) \qquad\qquad b = \left(1 - \frac{\pi}{2}\right)g_0 + \frac{\pi}{2}g_{\frac{\pi}{2}}.$$

Finally, similarly to Eq. (32), the variance can be expressed as

$$var[\hat{\vartheta}(z)] = \left(2 - \frac{\pi}{2}\right)\left[g(0) - g\left(\frac{\pi}{2}\right)\right] \tag{38}$$

and from Eq. (27),

$$var[\hat{\vartheta}(z)] = \left(2 - \frac{\pi}{2}\right)\left[l(w) - l\left(w + \frac{\pi}{2}\right)\right]^2. \tag{39}$$

3^{rd} **model**

In the same way as M_2, it can be proposed:

$$M_3(x, y) \;=\; a\left(1 - \frac{1}{\pi}\sqrt{(x-1)^2 + y^2}\right) \;+\; b\left(1 - \frac{1}{\pi}\sqrt{x^2 + (y-1)^2)}\right)$$

where a and b are real parameters defined as follows. We will require that $a > b$ in order to ensure that the covariogram has a maximum value at 0 (as it was explained above, Bernouilli

polynomials of even degree have maximum values at 0) and therefore it will fit better with the measurement function.

On one hand,

$$\int_0^{\pi} M_3(x,y)\, d\mathbb{S} = \frac{a}{\pi}(\pi^2-4) + \frac{b}{\pi}(4\sqrt{2}-8+\pi^2).$$

On the the other hand, using that $M_3(1,0) = g_0$ and $M_3(-1,0) = g_{\pi/2}$, we can write

$$g_0 = M_3(1,0) = a + b(1-\sqrt{2}/\pi)$$

$$g_{\pi/2} = M_3(-1,0) = a(1-2/\pi) + b(1-\sqrt{2}/\pi).$$

Then, $a = \pi/2(g_0 - g_{\pi/2})$ and $b = \frac{\frac{\pi}{2}g_{\pi/2}+(1-\pi/2)g_0}{1-\sqrt{2}/\pi}$

Therefore, the variance may be expressed as:

$$var[\hat{\vartheta}(z)] = \frac{\pi}{2}\left(g(0) + g\left(\frac{\pi}{2}\right)\right) - \frac{a}{\pi}(\pi^2-4) - \frac{b}{\pi}(4\sqrt{2}-2+\pi^2).$$

$$= \frac{1}{2(\pi-\sqrt{2})}\,(K_1\, g_0 \;-\; K_2\, g_{\pi/2})$$

where

$$K_1 = \pi^2(\sqrt{2}-1) + \pi(3\sqrt{2}-4) - 12\sqrt{2} + 16 \;\simeq\; 3.8798$$

$$K_2 = \pi^2(\sqrt{2}-1) + \pi(5\sqrt{2}-4) - 4\sqrt{2} \;\simeq\; 8.079313$$

Note: The condition $a > b$ implies that this formulae should be only used in that case. In other case, we can use M_2.

4^{th} model

Now, we want to increase the degree of the Bernoulli polynomial. As it was illustrated in Fig. 7, Bernoulli polynomials of degree 4 have a minimum value at 0 and maximum values at 1/2 and 1. Moreover, the measurement function has a maximum value at 0. Therefore, a good idea is to shift $B_4(x)$ a value of 1/2 along the x-axis. This leads us to suggest a similar model to the one proposed by Cruz-Orive and Gual-Arnau [2002]:

$$M_4(x) \;=\; a\,\pi\, B_4\left(\frac{x}{\pi} - \frac{1}{2}\right) \;+\; b.$$

On the one hand, $g_0 = M_4(0) = a\pi B_4(1/2) + b$ and $g_{\pi/2} = M_4(\pi/2) = a\pi B_4(1) + b$. Then,

$$a = \frac{g_{\pi/2} - g_0}{\pi(B_4(1) - B_4(1/2))} \;=\; \frac{16}{\pi}\,(g_0 - g_{\pi/2})$$

$$b = g_0 - \frac{B_4(1/2)}{B_4(1) - B_4(1/2)}(g_{\pi/2} - g_0) \;=\; \frac{8}{15}g_0 + \frac{7}{15}g_{\pi/2}.$$

On the other hand,

$$\int_0^{\pi} M_4(x,y)\, d\mathbb{S} \;=\; \frac{1}{16}\pi^2 a + \pi b.$$

Finally,

$$var[\hat{\vartheta}(z)] = \frac{31\pi}{30}(-g_0 + g_{\pi/2})$$

and from Eq. (27),

$$var[\hat{\vartheta}(z)] = \frac{89\pi}{270}\left[l(w) - l\left(w + \frac{\pi}{2}\right)\right]^2. \tag{40}$$

5^{th} **model**

Another model can be proposed, just by increasing the degree of the Bernoulli polynomial:

$$M_5(x) \;=\; a\,\pi\, B_6\left(\frac{x}{\pi}\right) \;+\; b$$

where a and b are real parameters defined as follows. Note that $B_6(x)$ has a maximum value at 0 (see Fig. 7), thus a shift is no need.

On the one hand, using that $g_0 = M_5(0) = a\pi B_6(0)+b$ and $g_{\pi/2} = M_5(\pi/2) = a\pi B_6(1/2)+b$, we get,

$$a \;=\; \frac{g_0 - g_{\pi/2}}{\pi(B_6(0) - B_6(1/2))} \;=\; \frac{64}{3\pi}(g_0 - g_{\pi/2})$$

$$b \;= g_0 - \frac{g_0 - g_{\pi/2}}{B_6(0) - B_6(1/2)} \;=\; \frac{31}{63}g_0 \;+\; \frac{32}{63}g_{\pi/2}$$

On the other hand, taking into account that $\int_0^{\pi} B_6(x)dx = 0$, we have,

$$\int_0^{\pi} M_5(x,y)\, d\mathbb{S} \;=\; \pi b$$

Finally, the variance can be expressed as

$$var[\hat{\vartheta}(z)] \;=\; \frac{\pi}{126}\left(g_0 - g_{\pi/2}\right)$$

and from Eq. 27,

$$var[\hat{\vartheta}(z)] = \frac{\pi}{126}\left[l(w) - l\left(w + \frac{\pi}{2}\right)\right]^2 \tag{41}$$

As we will see in the Results of section 4, this Cauchy component will have a good performance, so we will follow with the local error component. Thus the total variance predictor can be expressed as:

$$var[\hat{B}_2] \;=\; var_2(\hat{B}_1) \;+\; var[\hat{B}_2|w] \tag{42}$$

where the first term is the variance component due to orientations and the latter term is the local error component due to intersection counting with the test lines. The notation employed is explained by Gomez et al. [2016]. The first component receives the name of Cauchy component, whereas the second one is also called Cavalieri component).

Following the indications given by Gomez et al. [2016], both terms will have the following expressions

$$var_2(\hat{B}_1) \;=\; \frac{\pi}{126}\left[\left(l(w) - l\left(w+\frac{\pi}{2}\right)\right)^2 - (\hat{\sigma}_1^2 + \hat{\sigma}_2^2)\right] \tag{43}$$

$$var[\hat{B}_2|w] = \left(\frac{\pi}{4}\right)^2 (\hat{\sigma}_1^2 + \hat{\sigma}_2^2) \tag{44}$$

and, $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ can be computed from Eq. 24, namely

$$\hat{\sigma}_i^2 \;=\; \frac{T^2}{12}(3C_{0i} - 4C_{1i} + C_{2i}) \;,\; i = 1,2 \;,\; n_i \geq 3$$

$$\hat{\sigma}_i^2 \;=\; \frac{T^2}{6}(C_{0i} - C_{1i}) \;,\; i = 1,2 \;,\; n_i = 2$$

$$c_{ki} = \sum_{j=1}^{n_i-k} I_{ij} I_{i,j+k} \;,\; k = 0,1,2 \;,\; i = 1,2$$

where I_{1j} and I_{2j} are the total number of intersections between the curve and the jth vertical and the jth horizontal test lines of the grid respectively.

6^{th} model

A final model can be proposed, using the 8^{th} Bernoulli polynomial:

$$M_6(x) \;=\; a\,\pi\,B_8\left(\frac{x}{\pi}\right) \;+\; b$$

where a and b are real parameters defined as follows. Note that $B_8(x)$ has a maximum value at 0 (see Fig. 7), thus a shift is not considered.

On the one hand, using that $g_0 = M_6(0) = a\pi B_8(0) + b$ and $g_{\pi/2} = M_6(\pi/2) = a\pi B_8(1/2) + b$, we get,

$$a \;=\; \frac{g_0 - g_{\pi/2}}{\pi(B_8(0) - B_8(1/2))} \;=\; \frac{-256}{17\pi}(g_0 - g_{\pi/2})$$

$$b \;= g_0 - \frac{g_0 - g_{\pi/2}}{B_8(0) - B_8(1/2)} \;=\; \frac{127}{255}g_0 \;+\; \frac{128}{255}g_{\pi/2}$$

On the other hand, taking into account that $\int_0^\pi B_8(x)dx = 0$, we have,

$$\int_0^\pi M_6(x,y)\, d\mathbb{S} \;=\; \pi b.$$

Finally, the variance can be expressed as

$$var[\hat{\vartheta}(z)] \;=\; \frac{\pi}{510}\Big(g_0 - g_{\pi/2}\Big)$$

and from Eq. (27),

$$var[\hat{\vartheta}(z)] = \frac{\pi}{510}\left[l(w) - l\left(w+\frac{\pi}{2}\right)\right]^2 \tag{45}$$

Note that while using higher degrees of the Bernouilli polynomial, the variance estimator becomes smaller. Therefore we can be wondering whether or not it is worth increasing the degree. Gual-Arnau and Cruz-Orive [2000] claims that for models expressed with B_{2m+2}, m may be equal to the smoothness constant of the measurement function. Consequently, we will not go further with higher degrees.

2.2. Proposed models with a cycloid grid

A variance predictor for the cycloid grid was proposed by Cruz-Orive et al. [2014]. It is called the "fakir predictor" and it is based on intersection counting with a cycloid grid. In particular, it is necessary to distinguish between the fundamental tiles of the cycloid grid as illustrated in Fig. 3 with 20 fundamental tiles. Thus, the matrix I_{kl} will represent the number of intersections between the curve and the klth fundamental tile of the cycloid grid. Let, $k = 1, 2, \ldots, K$ and $l = 1, 2, \ldots, L$. In the Fig. 3, $K = 5$ and $L = 4$.

We denote the sum of all the matrix elements, i.e. the total number of intersections, by I defined as:

$$\sum_{k=1}^{K}\sum_{l=1}^{L} I_{kl} \;=\; I.$$

Furthermore, the fakir predictor is described as:

$$var[\hat{\vartheta}(z)] \;=\; 0.00569\sum_{r=1}^{2}(3c_{0r} - 4c_{1r} + c_{2r}) \tag{46}$$

where,

$$c_{m1} = \sum_{k=1}^{K-m} \sum_{l=1}^{L} I_{k,l}\, I_{k+m,l}\ ,$$
$$c_{m1} = 0\ , \quad K \leq m\ ,$$
$$c_{m2} = \sum_{k=1}^{K} \sum_{l=1}^{L-m} I_{k,l}\, I_{k,l+m}\ ,$$
$$c_{m2} = 0\ , \quad L \leq m\ , \quad m = 0, 1, 2$$

Finally, this project will be end up with a new model proposal for the cycloid grid. Taking into account the dimensions of the fundamental tile of the cycloid grid which is in the rectangle $R = [0, 4\pi r) \times [0, 4r)$ (see Fig. 3), the covariogram model can be expressed as

$$g(h_1, h_2) = \int_0^{4r} \int_0^{4\pi r} f(x, y) f(x + h_1, y + h_2)\, dx\, dy$$

Using the Moore–Aronszajn theorem, the kernel of the covariogram model can be approximated with two Bernouilli polynomials as:

$$M_c(h_1, h_2) = aB_n\left(\frac{h_1}{4\pi r}\right) B_m\left(\frac{h_2}{4r}\right) + b \tag{47}$$

where B_n and B_m are Bernouilli polynomials of degree n and m respectively, and a and b are real parameters defined as follows.

- **For m=n=2 :**

As it was done for the previous models, using that $g_0 = M_c(0, 1)$ and $g_{\pi/2} = M_c(1, 0)$, we get,

$$a = \frac{96(\pi^2 g_0 - \pi^2 g_{\pi/2})r^2}{4\pi^2 + 4(\pi - 2\pi^2)r - 1}, \qquad b = \frac{1}{3}\frac{12\pi^2 g_{\pi/2} - 8(\pi^2 g_0 - \pi^2 g_{\pi/2})r^2 - 12(2\pi^2 g_{\pi/2} - \pi g_0)r - 3g_0)}{4\pi^2 + 4(\pi - 2\pi^2)r - 1}.$$

Furthermore, the following integral can be computed

$$\int_0^{4r} \int_0^{4\pi r} M_c(h_1, h_2)\, dh_1\, dh_2 = 16\pi b r^2$$

and by Eq. (13) the variance can be expressed as

$$var[\hat{\vartheta}(z)] = \frac{1}{6}\, \frac{256K_1 r^4 + 384K_2 r^3 - 93K_3 r^2 - 3K_4 g_0 - 3K_5 g_{\pi/2} - 12K_6 r}{4\pi^2 + 4(\pi - 2\pi^2)r - 1} \tag{48}$$

where

$$K_1 = \pi^3 g_0 - \pi^3 g_{\pi/2}\ , \quad K_2 = 2\pi^3 g_{\pi/2} - \pi^2 g_0\ , \quad K_3 = 4\pi^3 g_{\pi/2} - \pi g_0$$
$$K_4 = \pi - 4\pi^3\ , \quad K_5 = \pi - 4\pi^3\ , \quad K_6 = (2\pi^3 - pi^2)g_0 + (2\pi^3 - \pi^2)g_{\pi/2}$$

- **For m=n=4 :**

We need to introduce a shift along the x-axis in order to have a maximum value at 0, as it was done for M_4. Namely,

$$M_c(h_1, h_2) \;=\; aB_n\left(\frac{h_1}{4\pi r} - \frac{1}{2}\right) B_m\left(\frac{h_2}{2r} - \frac{1}{2}\right) + b$$

Repeating the same process, we obtain the following variance predictor:

$$var[\hat{\vartheta}(z)] \;=\; \frac{1}{6}\,\frac{220K_1r^4 + 384K_2r^3 - 96K_3r^2 - 3K_4g_0 - 3K_5g_{\pi/2} - 12K_6r}{4\pi^2 + 4(\pi - 2\pi^2)r - 1} \tag{49}$$

where $K_1, K_2, ..., K_6$ were defined above.

As a comment, future work will include programming and testing this model. In this project, we will only study the performance of the model defined in Eq. (46).

3 Simulations

This brief chapter is dedicated to introduce the techniques used in section 4. First of all, most of the code have been written in Python and Blender has been used for the simulation. The variance predictors have been studied by means of Monte Carlo superimpositions of a curve onto the different grids. In the case of the cycloid grid, the intersection counting with the curve has a high computation cost. Therefore, the *sweep-line* method has been proposed. In the following sections we will go through the details of the implementation in Blender, the Monte Carlo simulations and the sweep-line method.

3.1. Simulation in Blender

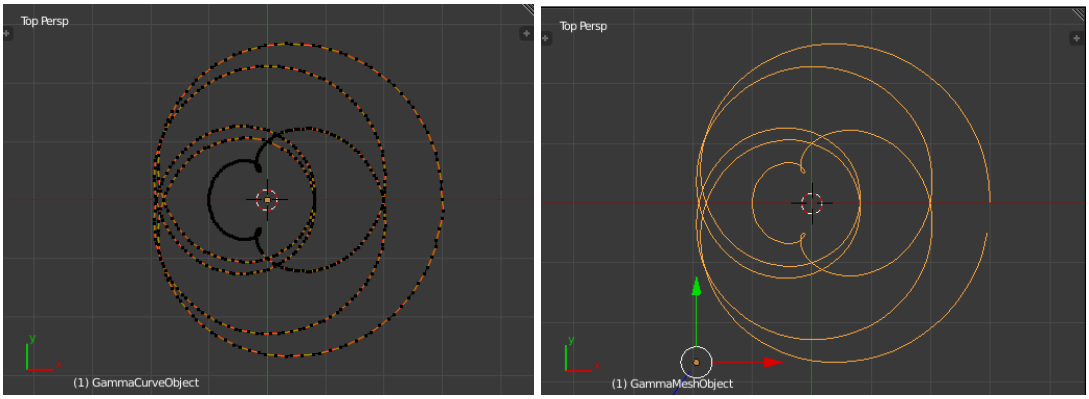
Blender is a free an open software which allows for the curves conversion into polygonal of small linear segments, intersection counting and running of Python code.

In order to estimate the length of a curve we can use Eqs. (3) and (4). The orthogonal projected length may be estimated via Eq. (5) and therefore, we need to calculate the intersections number between the curve and the test system which needs to cover the curve entirely. This is done by segmenting the curve, that is, the curve will be converted into a finite set of line segments as illustrated in Fig. 8(a) and 8(b). For each of the segments (with known endpoint coordinates), it will be check whether or not it intersect the test system. This avoids double counting of a segment intersecting two of the test system lines.

In the case of the square grid, we need to obtain the total orthogonal projected length $l(w)$ and the projected length $l(w + \pi/2)$ onto fixed test lines. Finally, we can repeat this process rotating the test lines clockwise by a fixed number of degrees. However, in practice we can get the same results by rotating the curve anti-clockwise around the origin, and this is much easier to program.

Note that the accuracy of the method can change as function of the constant of the grid. Therefore, it is convenient to choose an appropriate offset for the test system. This first check will be done for all the grids handled in this thesis.

Finally, another fact to take into account are the units used. One unit of Blender is equal 37.54 nm for the DNA molecule.



(a) (b)

Figure 8: (a) The $\gamma_{3,6,7}$ curve is converted into a connected polygonal of small line segments with *Blender*. (b) Curve $\gamma_{3,6,7}$ represented in *Blender*.

3.2. Sweep-line method

As it was mentioned above, we calculate the intersections between the curve and the grid by converting the planar curve into a set of line segments and then checking for crossing between any of the segments of the grid and curve. For linear test systems, such us the square or rhomboid grids, the count of the number of intersections between the curve (with unknown equation) and the grid is relatively easy and fast - the test lines are already straight line segments and there is no need to divide them. However, for a cycloid grid, it is much more slow and tricky. Cruz-Orive et al. [2014] proposes the use of "bounding boxes" which allow for a non-exhaustive search of intersections. An alternative method is the sweep-line (or Bentley–Ottmann) algorithm described by Smid [2003].

The purpose of the sweep-line method is to find all the intersection points of a given set of n line segments. Its main idea is to think in a vertical line that sweeps from right to left through all the line segments looking for intersections. Therefore, all the intersections that lies on the left side of the vertical line have been already detected and the algorithm will be $O((n+k)log(n))$ where n is the number of segments and k is the number of segment intersections.

Lets consider the example displayed in Fig. 9. We define the *event points* as the end point of a segment. In our case, we will have twelve event points and we sort them based on x-coordinate. The interesting x-coordinates are the ones that are events or an intersection point between segments. In addition, we can declare a *Segment list* (SL) which keeps the "activate" segments and changes when the sweep-line moves to the right. Only two things can happen. The first one, a segment can be added to the list if the sweep-line lies on the left end-point of the segment. And the second one, a segment can be removed from the list if the sweep-line finds the right end-point of the segment. This list is sorted by the y-coordinate.

An example is given following Fig. 9, we start sweeping from left to right. The first event point we find is 1, so we add segment a to SL. Then, we find event 2 and we add b to SQ. Now, we need to figure out if there is an intersection between segment a and b. In this case there is not and we proceed. We find then event 3 and the segment c is added to SQ. We have three active

segments and have to check if segment c intersects a or b. It is not the case and we continue. We find event 4 and segment d is activated. We may check the intersections between d and a, c and b. However, we can notice that a is between d and c. So d can not cross c until there is an intersection between d and a or and between a and c - we only need to compare the neighbours. The same happens between d and b, there is an intersection, but we will not detect it right now. The next event point is 5 which is the right end-point of a, hence the segment a is removed from SQ. Now, we have to compare segments d and c because vector a is not lying anymore between them. Since there is not intersection, we proceed.

The next event point is 6, so e is activated and we check if it intersects with d. The answer is negative and we continue with event 7 which removes c from SQ. At this point, we find the intersection between d and b and we add it as an additional event point, because once we move there, the order of the segments sorted by the y-coordinate changes. We continue with the algorithm until we process all the segments as illustrated at the top of Fig. 9.

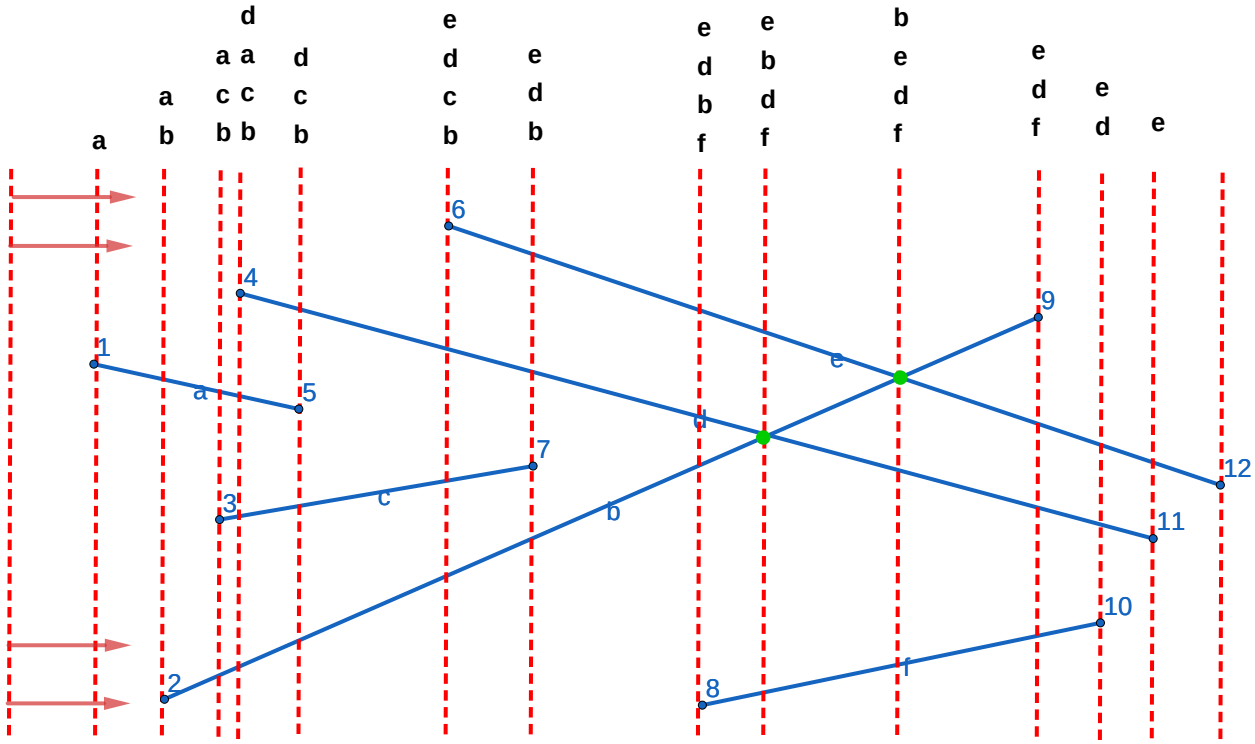


Figure 9: Sketch of the sweep-line algorithm with 6 line segments.

3.3. Monte Carlo simulation

The performance of the variance predictors have been studied by means of Monte Carlo simulations. For each of the curves analysed, the behaviour of the Cauchy estimators $\tilde{B}(w)$ and $\hat{B}(w)$ as function of the angle w were computed according Eqs. (3) and (4). $M = 256$ values were taken in the interval $[0, 2\pi)$ following Eq. (31). Afterwards, the empirical mean and the

error variance were calculated via Eqs. (30a) and (30c).

Furthermore, all the variance predictors were computed for the same w_k values, $\{var[\vartheta(z)], k = 1, 2, \ldots, M = 1024\}$. In addition, recalling that the curve must be UIR in the fundamental tile of the grid, the curve coordinates of the M associate points are $\{x_1, x_2, ..., x_M\} \subset \mathbb{R}^2$ and they are generated by rotation angles w_k in the interval $[0, 2\pi)$. Therefore, curve location and rotation of the M superimpositions of the curve onto the grid can be expressed with the following expression: $\{(x_k, w_k), k = 1, 2, ..., M\}$.

With new notation, the Eqs. (30b) and (30c) can be reformulate as:

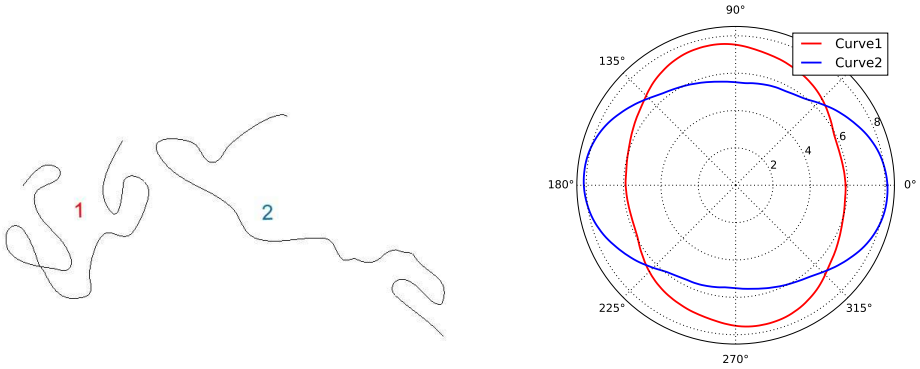
$$E_e[\hat{B}(w, x)] = \frac{1}{M} \sum_{k=1}^{M} \hat{B}(w_k, x_k) \tag{50}$$

$$Var_e[\hat{B}(w, x)] = \frac{1}{M} \sum_{k=1}^{M} \left[\hat{B}(w_k, x_k) - E_e[\hat{B}(w, x)]\right]^2. \tag{51}$$

4 Results

This chapter will be dedicated to present all the results obtained. First of all, the results obtained by Gomez et al. [2016] have been replicated in order to check the good performance of the algorithms. Furthermore, the graphs will be shown in such a way that they can be easily compared with the new graphs obtained.

4.1. DNA Molecules with a square grid



(a) (b)

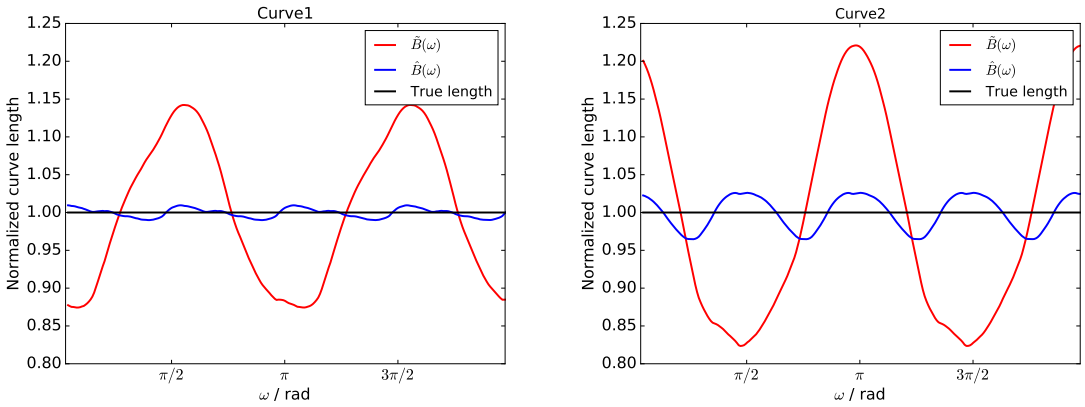
Figure 10: (a) *Curve1* considered fairly isotropic on the left side and *Curve2* considered fairly anisotropic on the right side. (b) Projection length $l(w)$ as a function of the angle w, in red for *Curve1* and in blue for *Curve2*.

In order to check the good performance of the developed code, this project starts replicating the results obtained by Gomez et al. [2016]. Furthermore, to compare the figures with the results obtained in the next sections, it may be helpful to normalize the scale of the plots and show them with the same notation that is being used in this thesis. Therefore, this sub-section will be focus on the present those results needed from the mentioned paper, summarize their consequences and motivate further studies carried out in this project.

Briefly explained, there are two curves (ADN molecules): one considered "fairly isotropic" (here called "*Curve1*") and one considered "fairly anisotropic" (here called "*Curve2*"). Both curves are illustrated in Fig 10(a) and have almost the same length (393 and 392 nm respectively). A good way to compare them is by studying the total orthogonal projected length $l(w)$

as a function of the angle $w \sim UR[0, 2\pi)$. Fig. 10(b) shows their corresponding roses of projections, i.e., the projection length $l(w)$ as a function of the angle w. As it can be observed, the more anisotropic is the curve, the more flattened is the projection.

Furthermore, for each of the curve it has been studied the behaviour of $\hat{B}(w)$ and $\tilde{B}(w)$ estimators. As it was introduced above, $\tilde{B}(w)$ takes into account the total orthogonal projected lengths of the curve, whereas $\hat{B}(w)$ uses two orthogonal projections. Fig. 11(a) and 11(b) show both estimators as a function of the angle w. In order to be able to compare both curves simultaneously, the y-axis is normalized by the "*true length*". This *true length* is obtained as the sum of the lengths of the small line segments in which is divided the curve. In addition, all the plots were performed with M = 256 values of w in the interval $[0, 2\pi)$ and the grid spacing was T=25 nm, which corresponds to $E[I] \approx 20$ intersections between the grid and the curve.

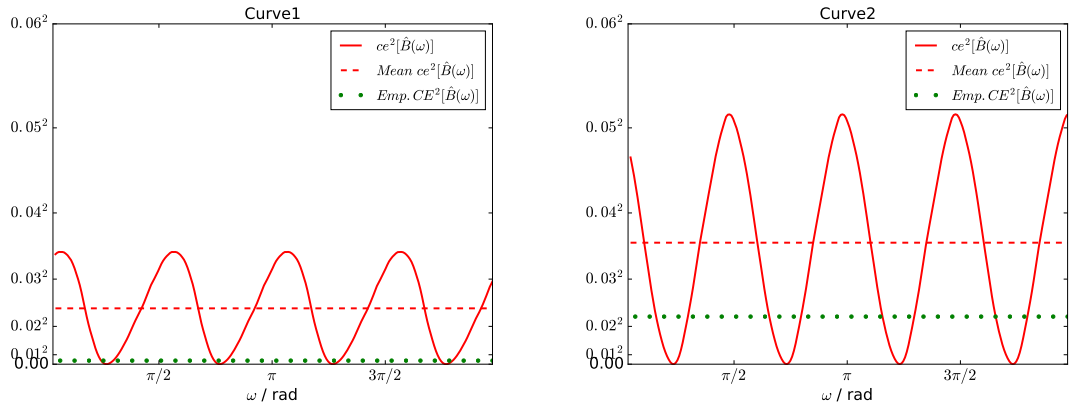


(a) (b)

Figure 11: Normalized estimator $\tilde{B}(w)$ in red and $\hat{B}(w)$ in blue as a function of the angle w, for *Curve1* (a) and *Curve2* (b). The black line represents computed *true length* for both curves.

In addition, in Fig. 11 can be observed how both estimators fluctuate around the *true length* computed via Eq. (2). Moreover, the use of the two orthogonal projections reduce this shift. And in particular, in the case of the anisotropic curve, this fluctuation is larger, as it was expected from the rose projection graph (see Fig. 10(b)).

Besides, the predictor error variance can be estimated via Eq. (33). It is shown with a solid line in Fig. 12(a) and 12(b) for the two tested curves. Their means are shown in a dashed red line. In the case of the *Curve1*, $\text{ce}^2[\hat{B}(w)]$ has a mean value of 0.0244^2 and for *Curve2*, it is 0.0360^2. Finally, the empirical $\text{CE}^2[\hat{B}(w)]$ is computed by Eq. (30c) and shown with a dotted green line. For *Curve1*, $\text{CE}^2[\hat{B}(w)]$ is equal 0.0062^2 and for *Curve2* is 0.0225^2.



(a) (b)

Figure 12: Model based prediction of the square coefficient of error of the two projection Cauchy estimator $\text{ce}^2[\hat{B}(w)]$ in solid line. The dashed line is the mean value of $\text{ce}^2[\hat{B}(w)]$ and the dotted green line represents the empirical $\text{CE}^2[\hat{B}(w)]$.

4.2. Curve $\gamma_{3,6,7}$

So far, the results obtained by Gomez et al. [2016] have been introduced. The comparison made for *Curve1* and *Curve2* raise the following questions:
what would happen if we test a much more isotropic curve? would we be able to improve the variance estimation? and for such very isotropic curve what would happen with the predictor if we deform the curve?

That leads us to the study of the $\gamma_{3,6,7}$ and afterwords to the analysis of the variance control.

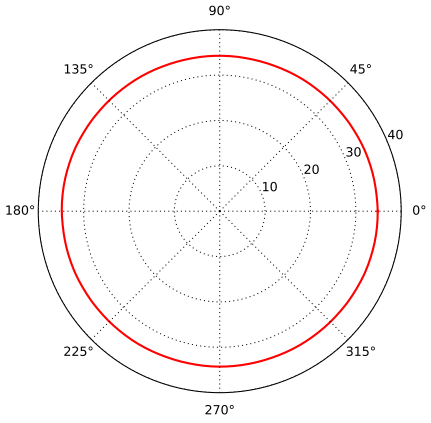
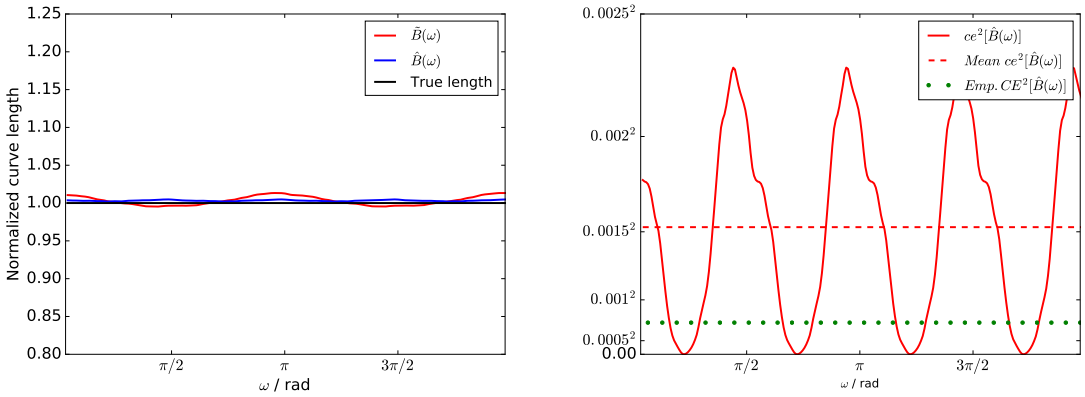


Figure 13: Total orthogonal projected length (rose of projections) $l(w)$ as a function of the angle w of the curve $\gamma_{3,6,7}$.

The curve $\gamma_{3,6,7}$ has been generated in Python using the parametric equations (see Fig. 6). A list of point coordinates has been introduced in *Blender* which allows for polygonal division

of the curve into small linear segments, as illustrated in Fig. 8(a). The endpoint coordinates of these segments are fixed with respect to the origin of a coordinate system represented in Fig. 8(b). In this graph it can be observed that the $\gamma_{3,6,7}$ does not appear as a close curve in $(3, 0)$ and this is a problem arising from the use of *Blender*. By closing the curve at the end, *Blender* segments around the $(3, 0)$ with two different vector orientations and, as a result, the number of intersections with the x-axis is counted twice. Therefore, it has been decided not to close the curve. Note that it will not affect the results because all the graphs are normalized by the computed length.

The $\gamma_{3,6,7}$ curve with, *true length* of 2027 nm, gives the visual impression of being highly isotropic in space. As it was done for the *Curve1* and the *Curve2*, this can be checked by studying the projection length $l(w)$. Fig. 13 shows the corresponding projection. Its fairly circular shape assures its isotropic character in space.



(a) (b)

Figure 14: (a) Cauchy estimators $\tilde{B}(w)$ and $\hat{B}(w)$ of the curve $\gamma_{3,6,7}$. (b) The square coefficient of the error of the two projection Cauchy estimator $ce^2[\hat{B}(w)]$ in solid line. The dashed line is the mean value of $ce^2[\hat{B}(w)]$. And the dotted green line represents the empirical error variance.

On the one hand, the Cauchy estimators $\tilde{B}(w)$ and $\hat{B}(w)$ are shown in Fig. 14(a). Note that the graph preserve the y-axis scale of Fig. 11. Therefore, it can be directly compared and conclude that both estimators have a better performance with the $\gamma_{3,6,7}$ curve than with the curves studied by Gomez et al. [2016] due to its isotropic nature.

On the other hand, the variance predictor has been studied in Fig. 14(b). The mean $\text{ce}^2[\hat{B}(w)]$ is 0.015^2 and, as expected, it is significantly smaller than for the *Curve1* and *Curve2* because of its isotropic character. The empirical $\text{CE}^2[\hat{B}(w)]$ value is 0.0008^2 which may be considered negligible indeed. The numerical results obtained are summarized in Table 2.

	Curve 1	Curve 2	$\gamma_{3,6,7}$
True length (nm)	393	392	2027
$\tilde{B}$ (nm)	394	351	2027
$\hat{B}$ (nm)	392	391	2027
Mean ce$^2[\hat{B}(w)]$	0.0244^2	0.0360^2	0.0015^2
Emp. CE$^2[\hat{B}(w)]$	0.0062^2	0.0225^2	0.0008^2

Table 2: The *True length* length is calculated as the sum of the lengths of the straight line segments in which are divided the curves. $\tilde{B}$ and $\hat{B}$ are computed via Eq. (3) and (4) respectively. The Mean ce$^2[\hat{B}(w)]$ and Emp. CE$^2[\hat{B}(w)]$ are obtained from Fig. 12(a), 12(b) and 14(b).

4.2.1. Variance control

In the previous sub-section we have concluded that the more isotropic is the curve, the smaller is the variance predictor. In order to quantify how much this predictor is affected by the isotropy character of the curve, it has been suggested subject the $\gamma_{3,6,7}$ to deformations. These deformations have been performed in *Blender* by scaling along the x-axis and the y-axis independently.

Fig. 15 shows how the square coefficient of the error and the empirical error of the two projection Cauchy estimation increase drastically when the curve is deformed. The coefficients of error for $\gamma_{3,6,7}$ (in red) are almost negligible compared to the same curve when a resize factor of 1.5 (in green) and 2 (in blue) have been applied along the x-axis. The lenght of the curve has been modified from 2027 nnm to 2588 and 3153 nm in each case.

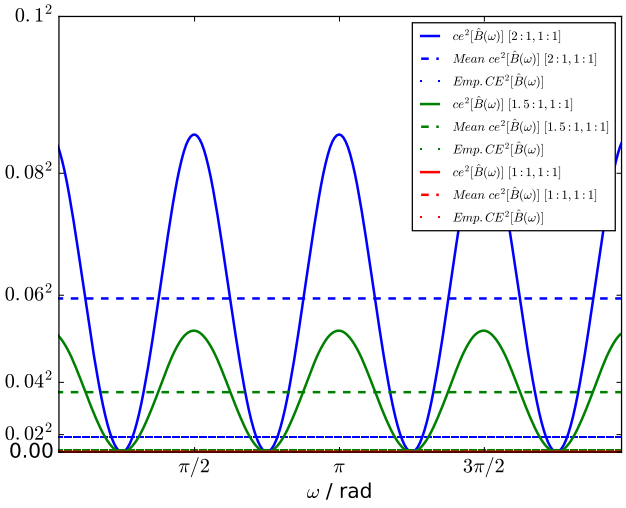


Figure 15: In red, the model based predictor of the square coefficient of the error, the mean ce$^2[\hat{B}(w)]$ and the empirical CE$^2[\hat{B}(w)]$ for the $\gamma_{3,6,7}$. Idem in blue for a resize with a factor of 1.5 along the x-axis and in green for a factor of 2.

By increasing the length along the x-axis, the curve lose its isotropic nature and the Cauchy

estimator become less effective. The $ce^2[\hat{B}(w)]$ is 0.0371^2 for the smallest transformation and 0.0593^2 for the biggest one. The empirical $CE^2[\hat{B}(w)]$ has a value of 0.0068^2 when a resize factor of 1.5 is applied and a value of 0.0189^2 when a resize of 2 is implemented. Similar results were obtained when the y-axis was deformed.

4.3. Models tested

For the models introduced in Chapter 2, we want to study how the Cauchy component of the variance estimators behave for two different curves: the *Curve2* (fairly anisotropic) and the $\gamma_{3,6,7}$ curve (highly isotropic). The same square grid than above will be used and the results will be compared with the previous model.

Fig. 15 represents the square coefficient of the error of the two projection Cauchy estimator for the six models presented. In the case of M_1, the variance estimator is given by Eq. 36. From this graph, we can conclude that the Mean $ce^2[\hat{B}(w)]$ is 0.2215^2 for *Curve2* and 0.0094^2 for the $\gamma_{3,6,7}$, which is about $\frac{120}{\pi}$ times larger than the previous model studied.

For M_2, we got that the Mean $ce^2[\hat{B}(w)]$ is 0.1158^2 for *Curve2* and 0.005^2 for the $\gamma_{3,6,7}$ and we can conclude that this model works better than M_1, but still it is worse than the starting model.

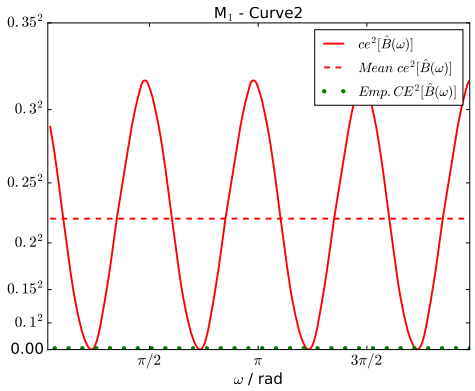
In the case if M_3, the variance estimator is negative. As a reminder, we defined this model when $a > b$, and doing the calculations we observed that this never happens. As a consequence, the variance estimator obtained is negative and the model is rejected.

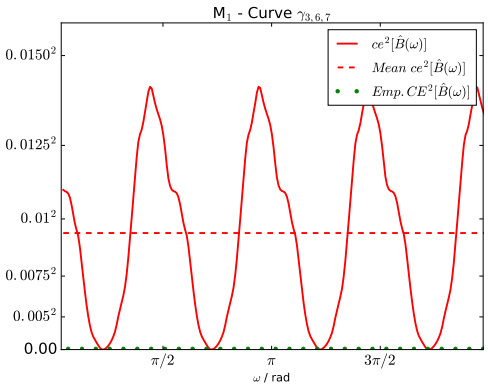
M_4 introduces a shift along the x-axis in order to have the minimum of the $B_2(x)$ at zero. However, this shift does not improve the results and the Mean $ce^2[\hat{B}(w)]$ is 0.1799^2 for *Curve2* and 0.0077^2 for the $\gamma_{3,6,7}$.

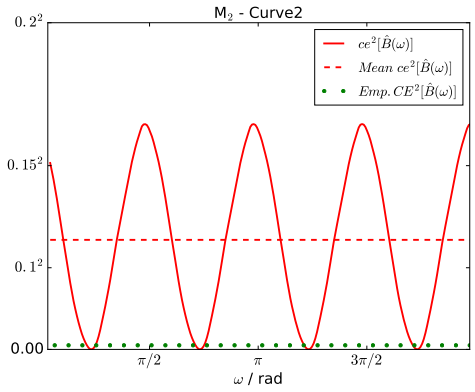
M_5 incorporates the Bernouilli polynomial of 6^{th} degree. In this case, the proposal model works better than the one introduced by Cruz-Orive and Gual-Arnau [2002]. The Mean $ce^2[\hat{B}(w)]$ is 0.0279^2 and 0.0012^2 for the *Curve2* and the $\gamma_{3,6,7}$ curve respectively.

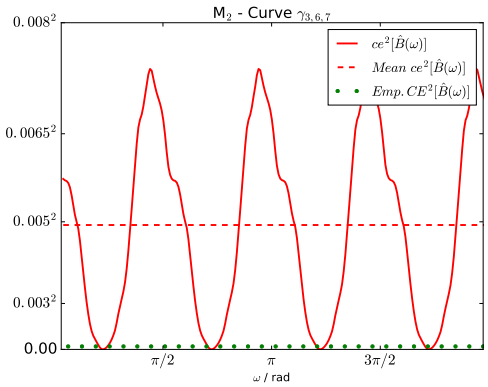
The last model, M_6, uses the Bernouilli polynomial of 8^{th} degree. The square coefficient of the error is reduced as well, being the Mean $ce^2[\hat{B}(w)]$ equals to 0.0129^2 for the *Curve2* and equals to 0.0006^2 for the $\gamma_{3,6,7}$ curve. However the empirical error is larger than the square coefficient of the error and, therefore, the Cauchy variance predictor is likely underestimated.

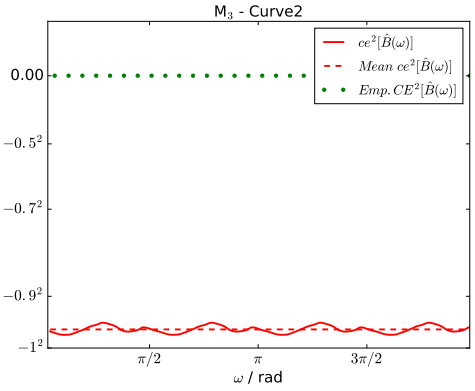
Finally, Table 3 summarized the square coefficient error for all the models studied, as well as the empirical error.

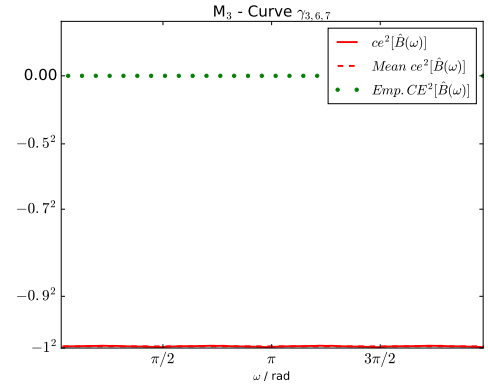

M1 - Curve2
$ce^2[\hat{B}(\omega)]$
$Mean\ ce^2[\hat{B}(\omega)]$
$Emp.\ CE^2[\hat{B}(\omega)]$
0.35^2
0.3^2
0.25^2
0.2^2
0.15^2
0.1^2
0.00
$\pi/2$
π
$3\pi/2$
ω / rad


M1 - Curve $\gamma_{3,6,7}$
$ce^2[\hat{B}(\omega)]$
$Mean\ ce^2[\hat{B}(\omega)]$
$Emp.\ CE^2[\hat{B}(\omega)]$
0.0150^2
0.0125^2
0.01^2
0.0075^2
0.005^2
0.00
$\pi/2$
π
$3\pi/2$
ω / rad


M2 - Curve2
$ce^2[\hat{B}(\omega)]$
$Mean\ ce^2[\hat{B}(\omega)]$
$Emp.\ CE^2[\hat{B}(\omega)]$
0.2^2
0.15^2
0.1^2
0.00
$\pi/2$
π
$3\pi/2$
ω / rad


M2 - Curve $\gamma_{3,6,7}$
$ce^2[\hat{B}(\omega)]$
$Mean\ ce^2[\hat{B}(\omega)]$
$Emp.\ CE^2[\hat{B}(\omega)]$
0.008^2
0.0065^2
0.005^2
0.003^2
0.00
$\pi/2$
π
$3\pi/2$
ω / rad


M3 - Curve2
$ce^2[\hat{B}(\omega)]$
$Mean\ ce^2[\hat{B}(\omega)]$
$Emp.\ CE^2[\hat{B}(\omega)]$
0.00
-0.5^2
-0.7^2
-0.9^2
-1^2
$\pi/2$
π
$3\pi/2$
ω / rad


M3 - Curve $\gamma_{3,6,7}$
$ce^2[\hat{B}(\omega)]$
$Mean\ ce^2[\hat{B}(\omega)]$
$Emp.\ CE^2[\hat{B}(\omega)]$
0.00
-0.5^2
-0.7^2
-0.9^2
-1^2
$\pi/2$
π
$3\pi/2$
ω / rad

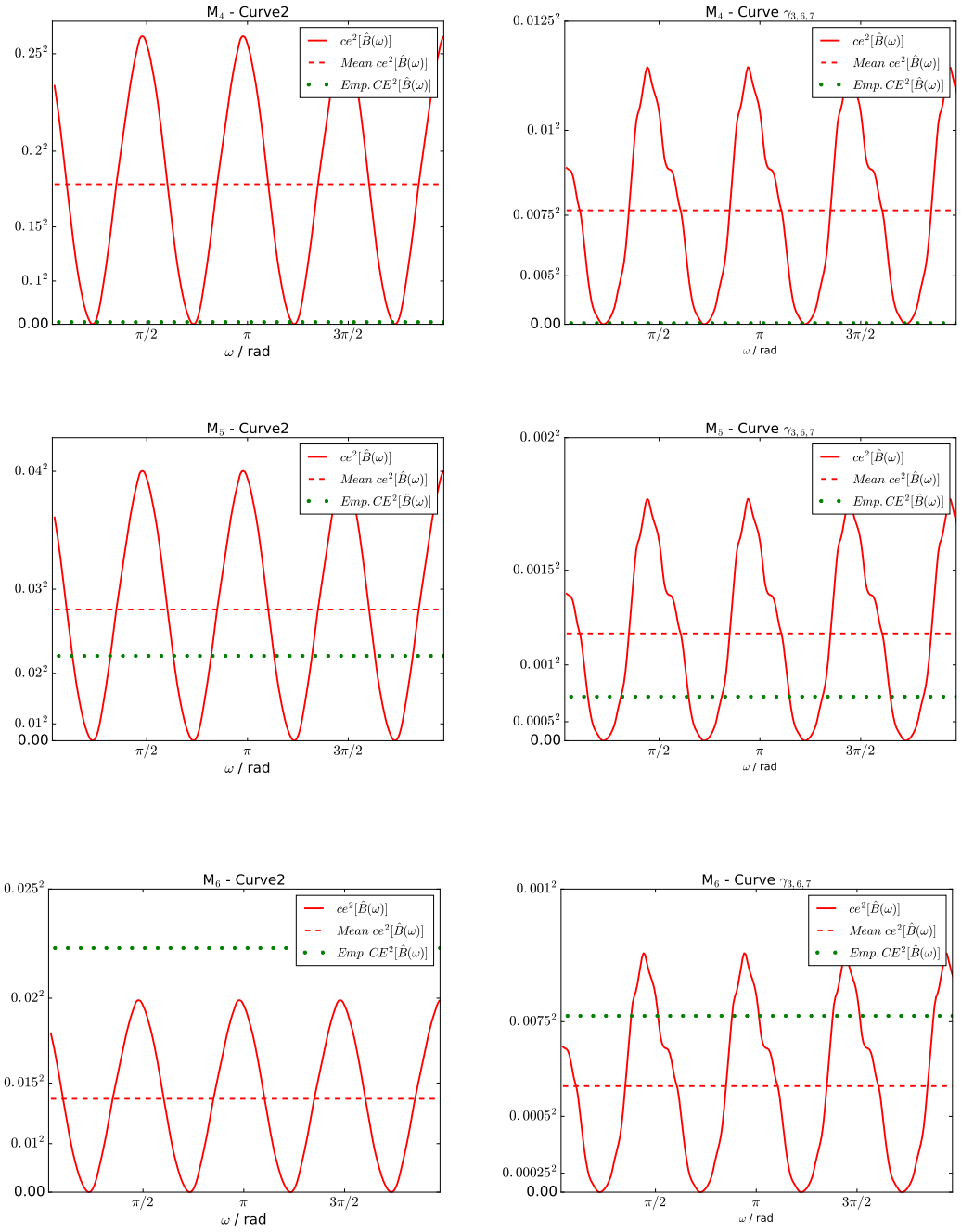
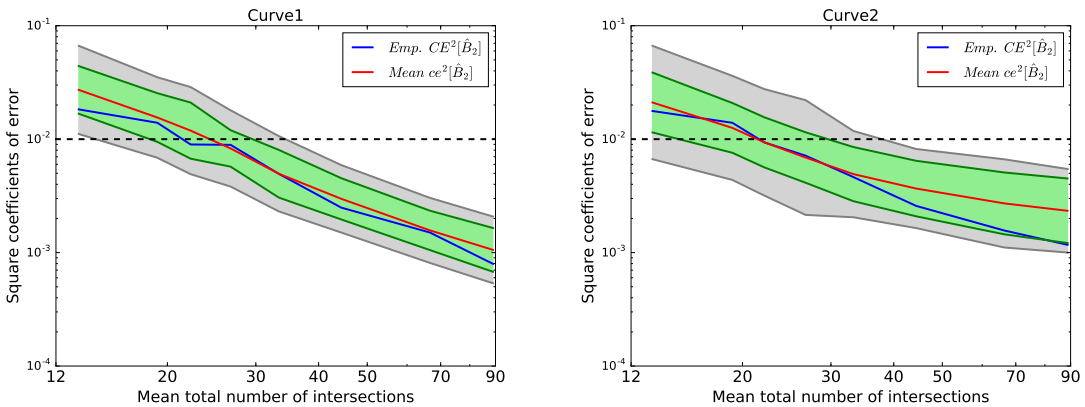


Figure 15: Model based predictions of the square coefficient of error of two projection Cauchy estimator for *Curve2* on the left side and for $\gamma_{3,6,7}$ curve on the right. In solid line is represented the square coefficient error of the two projection Cauchy estimator $\text{ce}^2[\hat{B}(w)]$. The dashed line is the mean value of $\text{ce}^2[\hat{B}(w)]$ and the dotted green line is the empirical error variance.

Model	*Curve2*	$\gamma_{3,6,7}$
Previous model	0.0358^2	0.0015^2
Model 1	0.2216^2	0.0094^2
Model 2	0.1158^2	0.048^2
Model 3	< 0	< 0
Model 4	0.1799^2	0.0077^2
Model 5	0.0279	0.0012^2
Model 6	0.0139	0.0006^2
Empirical error	0.0224^2	0.0007^2

Table 3: Predictions of the square coefficient error of two projection Cauchy estimator for the six models proposed. Here "previous model" refers to the model proposed by Cruz-Orive and Gual-Arnau [2002] and the empirical error variance is determined by Eq. (30c).

For all the models analysed above, M_5 has the better Cauchy estimator behaviour. Therefore, we proceed studying the Cavaleri component of the variance predictor. Figs. 16(a) and 16(b) show in red how the $M = 1024$ Monte Carlo replicates the $ce^2[\hat{B}_2]$ given by Eq. 18 and in blue the empirical $CE^2[\hat{B}_2]$ obtained from Eq. 51 as function of the mean total number of intersections E[I] for *Curve1* and *Curve2* respectively. The graph corresponds to eight different values of grid spacing, from T = 5.63 to 37.54 nm. The 95 % and 100 % confidence bands of $ce^2[\hat{B}_2]$ are displayed in green and grey respectively. That means the 95 % and all the data is enclosed by those bands. The empirical curve is always within the 95 % confidence band.



(a) (b)

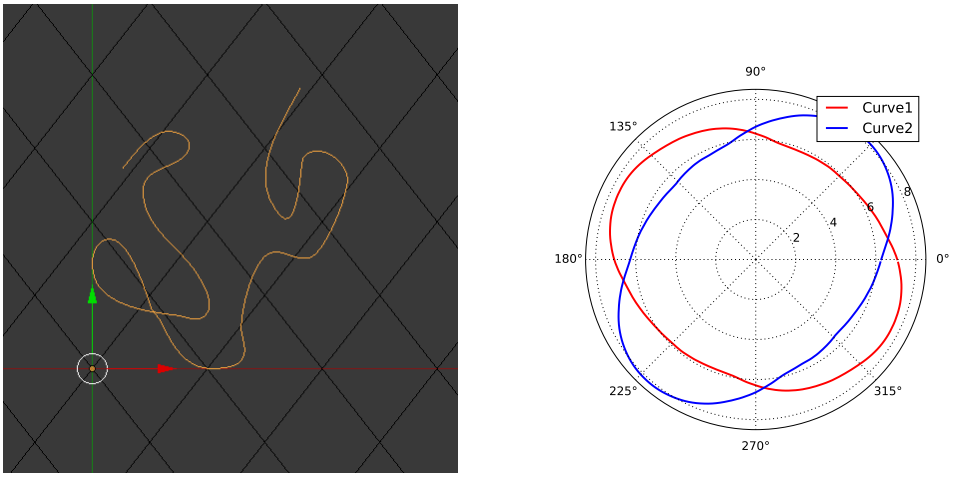
Figure 16: Monte Carlo simulation of Emp $CE^2[\hat{B}_2]$ in blue and $ce^2[\hat{B}_2]$ in red. The green and grey regions are the 95 % and 100 % confidence bands for the 1024 replications of the variance predictor. The black dashed line has been drawn as reference for coefficients of error of 10^{-2}.

Fig.16 can be compared with [Gomez et al., 2016, Fig. 4(c,d)]. In that paper, $ce^2[\hat{B}_2]$ was

below 10^{-2} for $E[I] \approx 27$ intersections, whereas in our case, this limit is reached for $E[I] \approx 20$ intersections which correspond to a grid spacing of 27.28 nm.

4.4. DNA Molecules with a rhomboid grid

All the previous models have been tested with a square grid. In this last part of the project, we want to study the behaviour of new grids. In particular, this section is fully dedicated to rhomboid grids, wheres the next one describes cycloid grids.



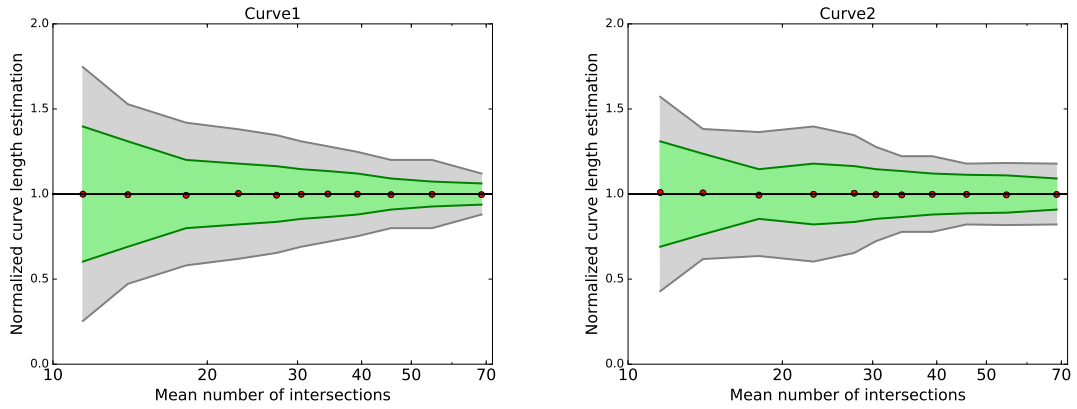
(a) (b)

Figure 17: (a) Screen capture of a rhomboid grid overlaid on the *Curve1*. (b) Total projected length (rose of projections) $l(w)$ as a function of the angle w of the DNA curves.

For a rhomboid grid, displayed in Fig. 17(a), we need to use the Gauss quadrature introduced in Section 1.2.3. The sampling nodes were: $\frac{\pi}{2}\frac{1}{\sqrt{3}}$ and $-\frac{\pi}{2}\frac{1}{\sqrt{3}}$. We can quickly check that total projected length of the *Curve1* and *Curve2* has the same shape that Fig. 10(b). The only difference between both graphs should be a slight rotation as illustrated in Fig. 17(b).

In order to set the appropriate grid dimensions, a study of the number of intersections and length estimation as function of the length size has been carried out. Fig 18(a) and 18(b) show the Monte Carlo simulation for different dimensions of the grid for *Curve1* and *Curve2*. The shaded regions are 95 % and 100 % confidence bands. The red dots represent the estimated length computed using the Gaussian quadrature. Since the method is unbiased, the red dots are over the *true length*. From both plots, we can conclude that an appropriate grid length may be T = 23 nm, which corresponds to $E[I] \approx 23$ intersections. Therefore, all the following results have this configuration setup.

Similarly to the square grid, for a rhomboid grid the Cauchy's formula can be applied easily and the $\tilde{B}(w)$ and $\hat{B}(w)$ estimators have been obtained and shown in Fig. 19(a) and 19(b) for *Curve1* and *Curve2* respectively. These graphs can be directly compared with Fig. 11(a) and 11(b). In particular, for a non-square grid, the Cauchy estimator with two projections have larger fluctuations throughout the angles.



(a) (b)

Figure 18: Simulation of the intersection counting and the normalized curve length estimator (B_{est}) as function of the grid dimensions *Curve1* (a) and *Curve2* (b). The red dots are the empirical estimation lengths for different intersections which corresponds to grid sizes from T=7.5 to 37.5 nm. The green area represents the 95 percentile of the replications and the gray one encloses all the data. 1024 Monte Carlo simulations were run for each grid spacing.

In Fig. 19(c) and 19(d), the square coefficient error and the empirical error are represented. The corresponding values for $\text{ce}^2[\hat{B}(w)]$ are 0.0237^2 and 0.0351^2 for the *Curve1* and *Curve2* respectively. Although the Cauchy's estimators showed a better behaviour for a rhomboid grid, the variance estimator as function of the angle and its average remain almost the same (when it is compared with Fig. 12(a) and 12(b)). This might be explained by the fact that in both cases the maximums and minimums of $\hat{B}$ are closed to $\pi/2 + 2k\pi, k \in \mathbb{Z}$. In addition, the $\text{ce}^2[\hat{B}(w)]$ value is closer to the empirical $\text{CE}^2[\hat{B}(w)]$ than in the case of a square grid.

Furthermore, we can compare the error due to the length size of the grid. For that, we can use the B_{est} which takes into account the number of intersections. Fig. 20(a) and 20(b) illustrate the Emp. CE^2 for different grid dimensions. The 95 % and 100 % confidence bands are represented as well. As a reference, we can compare the results with the same study carried out with a square grid by Gomez et al. [2016]. In both cases, the value of $\text{ce}^2[\hat{B}(w)]$ goes below 10^{-2} when we reach about E(I) $\approx$ 27 intersections.

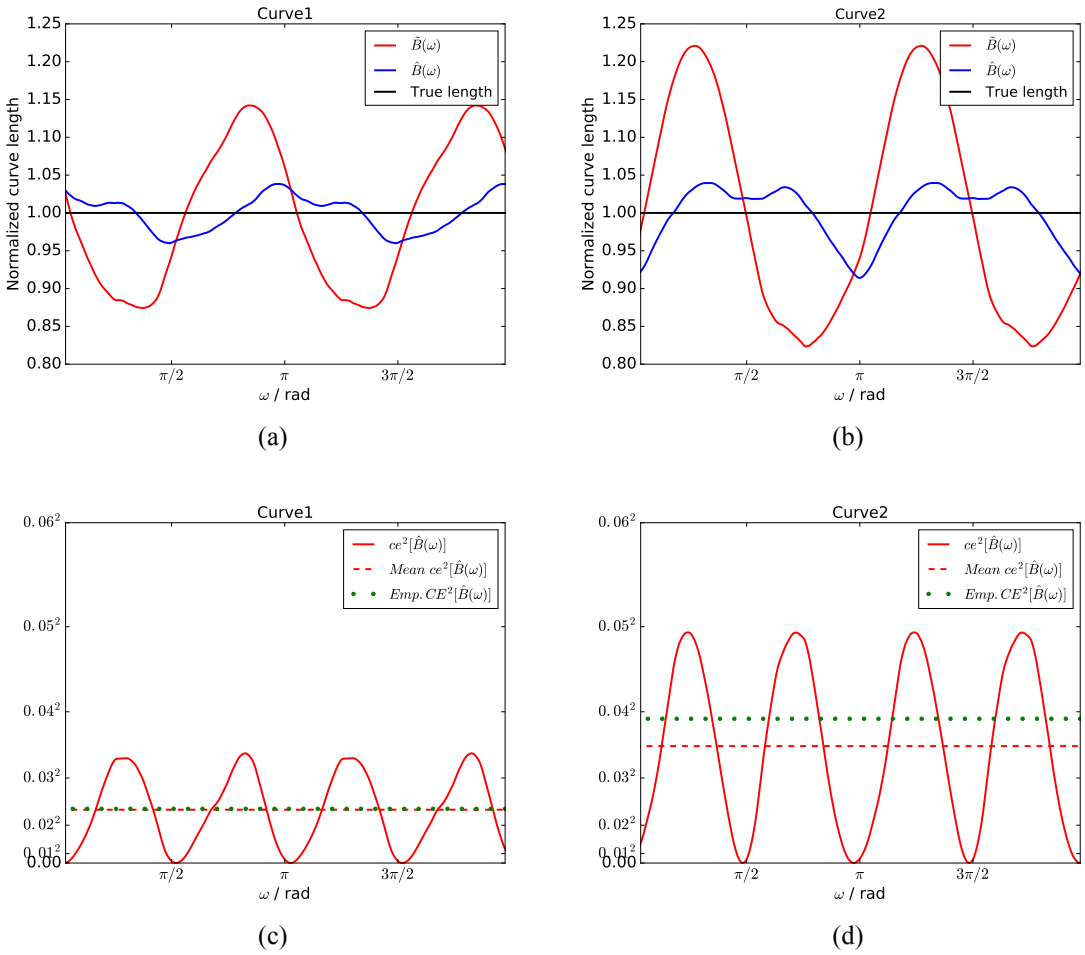
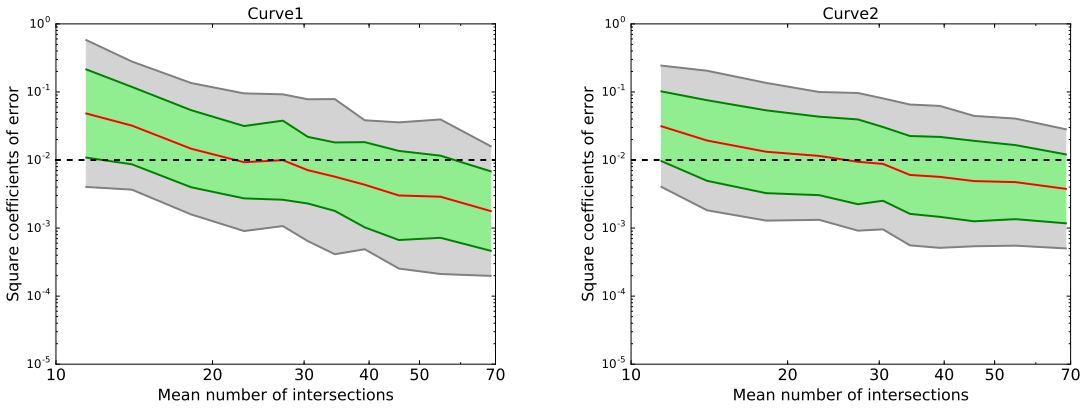


Figure 19: Top: Normalized Cauchy estimator $\tilde{B}(w)$ in red and $\hat{B}(w)$ in blue as a function of the angle w, for *Curve1* (a) and *Curve2* (b) obtained with a rhomboid grid. In black is represented the *true length* for both curves. Bottom: The square coefficient of the error of the two projection Cauchy estimator $\text{ce}^2[\hat{B}(w)]$ in solid line, for *Curve1* (c) and *Curve2* (d). The dashed line is the mean value of $\text{ce}^2[\hat{B}(w)]$ and the dotted line represents the empirical $\text{CE}^2[\hat{B}(w)]$.

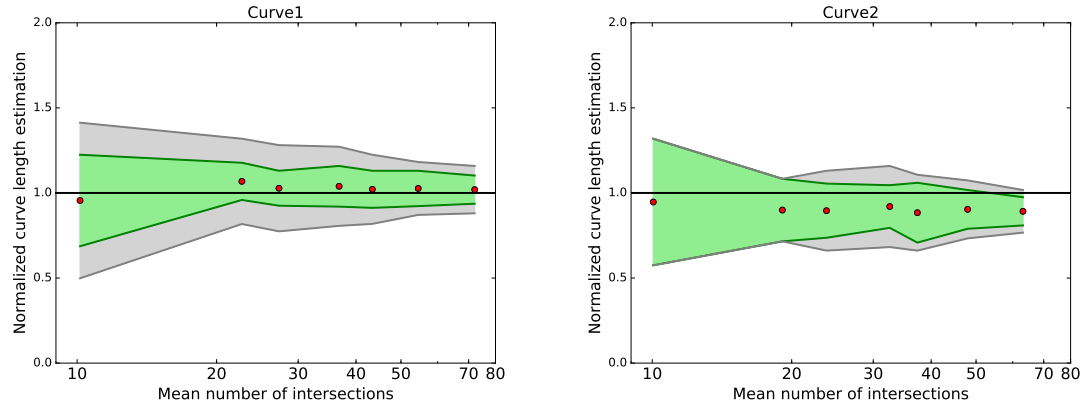
To sum up, both grids studied behave similarly in terms of decreasing the square coefficient of the errors when the number of intersections raise. The main difference can be appreciated in the $\hat{B}$ and $\tilde{B}$ estimators. A further study can consist in deciding whether it is worth reducing the square coefficients of error while increasing the complexity of the algorithm with the rhomboid grid.



(a) (b)

Figure 20: Simulation of the empirical error function of the number of intersections between the curve and the rhomboid grid in logarithmic scale. The red line represents the Emp. CE^2, the green region is the 95 % confidence band of such error and gray area encloses all the empirical error coefficients (100 % confidence band). 1024 simulations were run for each grid spacing. The black dashed line has been drawn as reference for coefficients of error of 10^{-2}.

4.5. Cycloid grid



(a) (b)

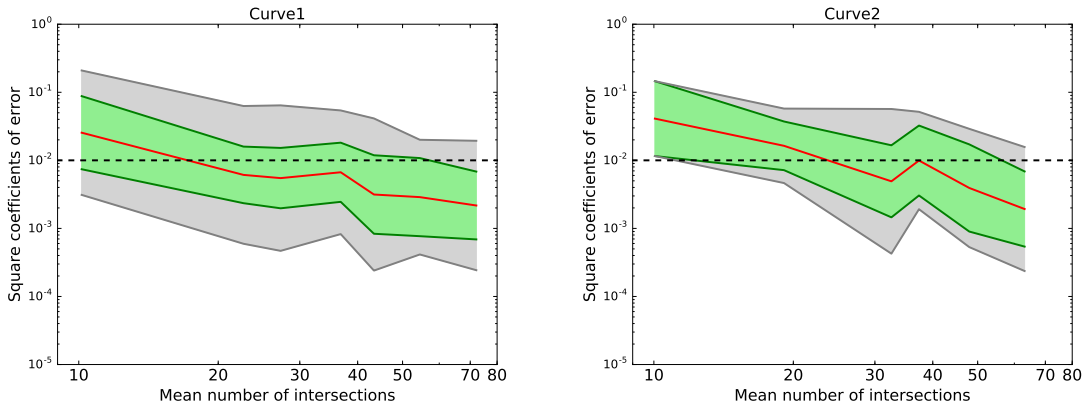
Figure 21: Simulation of the intersection counting and the normalized curve length estimator (B_{est}) as function of the grid dimensions *Curve1* (a) and *Curve2* (b). The red dots are the empirical estimation lengths for different intersections which corresponds to radius from 1.12 to 7.51 nm. The green area represents the 95 percentile of the replications and the gray one encloses all the data. 1024 simulations were run for each grid spacing.

To generate the cycloid grid, two parameters have to be specified: the radious of the circle that creates the cycloid and the gap or distance between the test lines. As it was suggested by Cruz-Orive et al. [2014], this gap is four times the value of the radius r. Furthermore, the cycloid grid involves all the possible orientations, and therefore there is no need to make the orientation of the grid IR. In this case, the curve is overlaid using the same orientations and the 1024

simulations are obtained with shifts of the curve of $4\pi r/\sqrt{1024}$ along the x-axis and $4r/\sqrt{1024}$ along the y-axis.

To set an appropriate value of the radius, it has been studied the normalized curve length estimated as function of the radius via Eq. 1. Fig 21 shows the estimated curve length as function of the mean number of intersections in logarithmic scale for *Curve1* and *Curve2*. The red dots should approximately coincide with the true length (black horizontal line). This is not achieved and a tentative explanation might be that we do not have the cycloid with all the possible orientations. That is, the cycloid grid is segmented to get the number of intersections, and when we approximate the curve by linear segments, we loss all the possible orientations. Therefore, we tried to take smaller segments but the computational time increased. Similar issues arise in the simulations given by Cruz-Orive et al. [2014].

Finally, we wanted to compare how the empirical error is reduced while decreasing the length of the fundamental tile (i.e. the radius of the rolling circumference which generates the cycloid arcs). As it can be observed, the slope seems smother than for square and rhomboid grids. However, we have to say that this is a very preliminary graph because as it was said in the paragraph above, the method does not look to be unbiased. Further details are given in the discussion Chapter 5.



(a) (b)

Figure 22: Simulation of the empirical error for different as functions of the number of intersections between the curve and the cycloid grid in logarithmic scale. The read line represent the Emp. CE^2, the green region is the 95 % confidence band of such error and gray area encloses all the empirical error coefficients (100 % confidence band). 1024 simulations were run for each grid spacing.

5 Summary and Outlook

In this project, we wanted to study the variance of grid-based estimators. That is, by superimpositions of a curve onto a test system. Chapter 1 was fully dedicated to give an overview of the length estimators derived from Buffon-Steinhaus unbiased method. In addition, it was observed that certain orientations seemed to be favored when a square grid was superposed. For instance, if the small linear segments, in which are divided the curve, are parallel to the test lines, they will not cut the grid, not intersections will be count and the length will be smaller than one expected. As a consequence, a cycloid grid was proposed since it has all the possible orientations and none will biased the method. An alternative test system was suggested as well: the rhomboid grid.

To end up with this section, we gave the theoretical tools to estimate the variance components (Cauchy and Cavalieri) from different models by exploiting the covariogram features.

Moreover, Chapter 2 was divided in two main parts. In the first one (Subsection 2.1), we proposed six different models to reduce the error variance predictor for square grids using different polynomials. For all of them, the Cauchy component of the variance predictor were derived. In particular, for M_5, the Cavalieri component was obtained as well. The second part (Subsection 2.2) was dedicated to the study of two models for a cycloid grid: the first one was introduced by Cruz-Orive et al. [2014] and the second one was suggested using the covariogram model.

For the simulation, explained into Chapter 3, we used Blender and all the code generated was in Python. In particular, to reduce the computational time for a cycloid grid, the sweep-line algorithm was used in the simulations. Finally, a description of how we checked the performance of our estimators by means of Monte Carlo simulations was given.

In Chapter 4, we started replicating the study done by Gomez et al. [2016] for two different curves: *Curve1* (fairly isotropic) and *Curve2* (fairly anisotropic). And this analysis motivated the study of the $\gamma_{3,6,7}$, curve which seemed highly isotropic in space. Using the unbiased estimators of B, $\tilde{B}$ and $\hat{B}$, its length was predicted. Fig. 14(b) illustrated how the Cauchy component was reduced by a factor of 264 in comparison with the *Curve1* and by a factor of 576 with respect to the *Curve2*.

Another study carried out with the $\gamma_{3,6,7}$ curve was the analysis of the variance error prediction under deflections along the x and y axis. By deformations of its shape, the curve loses its isotropic character and the Cauchy component increases significantly as shown in Fig 15. This results agrees with the differences observed between *Curve1* and *Curve2* by Gomez et al. [2016]. Thus, we can conclude that the more isotropic is the curve, the smaller is the error va-

riance predictor given by Eq. (33).

Furthermore, we tested the six variance predictors models introduced in Subsection 2.1 for the $\gamma_{3,6,7}$ (highly isotropic) and *Curve2* (fairly anisotropic). The results obtained were summarized in Table 3. In particular, it was observed that the M_5 model had the best Cauchy error variance predictor, therefore the Cavalieri component was studied as well. Considering both components, the total variance predictor had a slighly better performance than the previous one studied.

Considering the cycloid grid and bearing in mind that we do not want only accuracy methods, but also we want them as fast as possible, two main troubles had to be solved. On the one hand, we had to calculate the intersections between the curve and the cycloid lines. Searching directly through all the segments intersections was the trivial solution. However, it is $O(n^2)$ algorithm and hence non-viable for us. On the other hand, if we get the intersection coordinates, we still have to check in which fundamental tile are they in order to apply the first model introduced in Subsection 2.2. The difficulties counting the intersections were cited by [Cruz-Orive et al., 2014, Appendix C].

By computing directly the number of intersections we got the results explained in Subsection 4.5. This grid is aimed to be used to count manually thus it is hard to simulate the performance. However, we believe that a better implementation of the sweep-line algorithm (going deeper in the filed of Computational geometry) plus a K-D Tree data structure (to check whether or not an intersection point is within the fundamental tile), might solve the problem.

Finally, another non-square grid wanted to be tested and we proposed the rhomboid grid. Using the Gaussian quadrature we obtained the nodes where the observations should be taken. The results were similar to the square grid, however, the computational time and complexity of the code were a bit problematic than for a square grid. Further studies can compare cycloid, rhomboid, square and parallel test systems in terms of ease of use, efficiency and variance estimator error reduction.

References

M. S. Abramowitz and I. Stegun. Ia.(1964), handbook of mathematical functions. *Washington: National Bureau of Standards*, page 923, 1999.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

S. Batra, M. König, and L. Cruz-Orive. Unbiased estimation of capillary length from vertical slices. 1995.

Blender. Blender description. URL https://www.blender.org/.

J. Brauchart and J. Dick. A simple proof of stolarsky's invariance principle. *Proceedings of the American Mathematical Society*, 141(6):2085–2096, 2013.

L. Cruz-Orive. Stereology: A historical survey. *Image Analysis Stereology*, 36(3):153–177, 2017. ISSN 1854-5165. doi: 10.5566/ias.1767. URL https://www.ias-iss.org/ojs/IAS/article/view/1767.

L. Cruz-Orive and X. Gual-Arnau. Precision of circular systematic sampling. *Journal of microscopy*, 207(3):225–242, 2002.

L. Cruz-Orive, J. Gelŝvartas, and N. Roberts. Sampling theory and automated simulations for vertical sections, applied to human brain. *Journal of Microscopy*, 253(2):119–150, 2014. doi: 10.1111/jmi.12103.

A. M. Gokhale, R. A. Evans, J. L. Mackes, and P. R. Mouton. Design-based estimation of surface area in thick tissue sections of arbitrary orientation using virtual cycloids. *Journal of Microscopy*, 216(1):25–31, 2004. doi: 10.1111/j.0022-2720.2004.01392.x.

A. I. Gomez, M. Cruz, and L. M. Cruz-Orive. On the precision of curve length estimation in the plane. *Image Analysis & Stereology*, 35(1):1–14, 2016.

X. Gual-Arnau and L. M. Cruz-Orive. Systematic sampling on the circle and on the sphere. *Advances in Applied Probability*, 32(3):628–647, 2000.

V. Howard and M. Reed. *Unbiased stereology: three-dimensional measurement in microscopy*. Oxford, 1998.

M. Kendall and P. Morgan. Geometrical probability charles griffin. *Ltd., London*, pages 58–59, 1963.

F. Pausinger and D. Vartziotis. On the symmetry of finite sums of exponentials. *Elemente der Mathematik*, 9 2018. ISSN 0013-6018.

M. G. Reed and C. Howard. The sampling error of rotator estimates of eye volume. *Image Analysis & Stereology*, 22(1):57–61, 2003.

L. L. Schroeder. Buffon's needle problem: An exciting application of many mathematical concepts. *The Mathematics Teacher*, 67(2):183–186, 1974.

M. Smid. Computing intersections in a set of line segments : the Bentley-Ottmann algorithm. 2003.

Hints on the derivation of the variance prediction formulae when the covariogram is a second-degree polynomial

When the covariogram is a second-degree polynomial:

$$g(x) = a_0 + a_1 x + a_2 x^2$$

the symmetry property, namely $g(x) = g(\pi - x)$, can be applied

$$a_0 + a_1 x + a_2 x^2 = a_0 + a_1(\pi - x) + a_2(\pi - x)^2$$

then,

$$2(a_1 + a_2\pi)x - (a_1\pi + a_2\pi^2) = 0 \qquad \forall x$$

Hence, the following relation between the coefficients is obtained:

$$a_2 = \frac{-a_1}{\pi}$$

Moreover, replacing this expression in $g(x)$:

$$g(x) = a_0 + a_1\left(x - \frac{x^2}{\pi}\right)$$

and using the definition of the Bernoulli polynomial of second degree $B_2(x) = x^2 - x + 1/6$,

$$g(x) = a_0 + a_1\pi\left[B_2(0) - B_2\left(\frac{x}{\pi}\right)\right] = \beta_0 + \beta_1 B_2\left(\frac{x}{\pi}\right)$$

where $\beta_0 = a_0 + a_1\pi B_2(0)$ and $\beta_1 = -a_1\pi$.

To compute the variance, the first term in the right-hand of Eq. 13 is:

$$\frac{\pi}{2}(g_0 + g_1) = \frac{\pi}{2}\left(g(0) + g\left(\frac{\pi}{2}\right)\right) = \beta_0\pi + \beta_1\pi(B_2(0) + B_2(1/2))$$

and the second term:

$$\int_0^\pi g(x)dx = \int_0^\pi \beta_0 + \beta_1 B_2\left(\frac{x}{\pi}\right) = \int_0^\pi \beta_0 + \beta_1\left(\frac{x^2}{\pi^2} - \frac{x}{\pi} + \frac{1}{6}\right)dx = \beta_0\pi$$

Therefore,

$$var[\vartheta(z)] = \frac{1}{24}\beta_1\pi$$

Finally, using $\beta_1 = -a_1\pi$ and Eq. 28,

$$var[\vartheta(z)] = \frac{\pi^2}{12}\left(l(w) - l\left(w + \frac{\pi}{2}\right)\right)^2$$