THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Spatial Analyses of Nerve Patterns and Global Testing Approaches.

Konstantinos Konstantinou

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology
Göteborg, Sweden 2024

Spatial Analyses of Nerve Patterns and Global Testing Approaches.
Konstantinos Konstantinou
Göteborg 2024

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Typeset with LaTeX
Printed in Göteborg, Sweden 2024

Spatial Analyses of Nerve Patterns and Global Testing Approaches.

Konstantinos Konstantinou

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

This thesis consists of two parts: statistical analyses of and models for epidermal nerve fibers (ENFs), and extensions of global envelope tests. In the first part, the main goal was to improve our understanding of the ENF structure changes that occur as a result of nerve damage due to neurological disorders, such as diabetic neuropathy. For this purpose, different stochastic models were proposed. The ENF data were treated as point patterns in three-dimensional boxes, and samples from subjects suffering from diabetic neuropathy and healthy volunteers were considered. In Paper I, we introduced a new summary that measured the volume of the epidermis covered by the nerves and examined second-order properties of the underlying processes. Further, a three-dimensional point process model for the nerve structure was developed. The two-dimensional version of the model captured the planar spatial structure. However, the complete model could not capture the attraction between the nerve fiber endings in the data. Therefore, in Paper II a pairwise interaction Markov model allowing neighboring nerve endings to interact was proposed. In Paper III, we considered the two-dimensional projections of the ENF patterns and developed spatial thinning models to study the nerve death process. Insights from our analyses indicated that nerve mortality is guided by a biological process that favors the removal of isolated nerve trees. The goodness-of-fit of the models in Papers I-III was evaluated using global envelope tests. In the second part, we extended the global envelope tests for quantile regression and for comparison of distributions of n samples. In Paper IV, we proposed non-parametric, permutation based global tests, that allowed for simultaneous inference of the quantile regression process. In Paper V, we proposed graphical n-sample tests for correspondence of distributions based on the global envelope testing framework. Further, we presented a detailed discussion regarding the graphical interpretation of the test results for each suggested test statistic.

Keywords: Anisotropy, death process, diabetic neuropathy, epidermal nerve fibers, global quantile regression, permutation test, point processes.

List of publications

This thesis is partly based on the author's licentiate thesis and is based on the work represented by the following papers:

I. **Konstantinou, K.** and Särkkä, A.(2021). Spatial modeling of epidermal nerve fiber patterns. *Statistics in Medicine*, doi: https://doi.org/10.1002/sim.9194.

II. **Konstantinou, K.** and Särkkä, A.(2022). Pairwise interaction Markov model for the 3D epidermal nerve fiber endings. *Journal of Microscopy*,doi: https://doi.org/10.1111/jmi.13142.

III. **Konstantinou, K.**, Ghorbanpour, F., Picchini, U., Loavenbruck, A. and Särkkä, A.(2023). Statistical modeling of diabetic neuropathy: Exploring the dynamics of nerve mortality. *Statistics in Medicine*, doi: https://doi.org/10.1002/sim.9851

IV. Mrkvička, T., **Konstantinou, K.**, Kuronen, M and Myllymäki, M (2024). Global quantile regression. Submitted to *Journal of Computational and Graphical Statistics*.

V. **Konstantinou, K.**, Mrkvička, T. and Myllymäki, M (2024). The power of visualizing distributional differences: Graphical n-sample tests. Submitted to *Computational Statistics*.

Author contributions

I. I developed the two-step NOC-like model and extended the concept of reactive territories in three dimensions. I implemented the methods and conducted all statistical analyses. I did most of the writing for the publication.

II. I implemented the estimation methods of the pairwise interaction Markov field model and carried out the simulation study. I conducted all statistical analyses and did most of the writing for the publication.

III. I developed the spatial thinning models, conducted the simulation study, and performed the inference and evaluation of the models. I did most of the writing for the publication.

IV. I co-developed the global quantile regression test and implemented some of the permutation strategies in the GET package in R. I conducted the simulation study and did most of the writing for the publication.

V. I developed and implemented the global permutation tests in the GET package in R. I conducted the simulation study and did most of the writing for the publication.

List of Figures

Acknowledgements

First and foremost, I want to express my sincere gratitude to my supervisor Aila Särkkä for introducing me to the wonderful subject of spatial point processes, for always being available to meet and advise me, and for always carefully reading and providing constructive feedback on my manuscripts. Thank you Aila for your continuous support and encouragement. Further, I want to thank Tomáš Mrkvička, Mari Myllymäki, and Mikko Kuronen, co-authors of Paper IV, it was a pleasure working together. I especially want to thank Mari for the information and hospitality during my visits to Helsinki. I also want to thank my co-supervisors, Umberto Picchini and Ottmar Cronie, as well as my examiner Petter Mostad, for the fruitful discussions and guidance. I'd like to thank William R. Kennedy, Gwen Wendelschafer-Crabb, and the other members of the Kennedy Lab for sharing the epidermal nerve fiber data, as well as Adam Loavenbruck for addressing my neurology and data-related questions.

I want to express my sincere gratitude to the Swedish Research Council for financially supporting my research. Further, I want to thank the Chalmers Research Foundation, the Royal Society of Science and Knowledge in Gothenburg, the Wilhelm and Martina Lundgren's Science Foundation, and the Linda Peetre Memorial Fund for covering my conference and research visit related expenses. This work would not have been possible without their contribution. I am also grateful to all my colleagues in the mathematical department for making this such a lovely workplace. I would also like to thank all the department's administrative staff for always being willing and able to help me with any non-mathematical issues that come up.

Finally, I would like to express my gratitude to my friends and family for their love, continuous encouragement, and unwavering support. This thesis would not have been feasible without them.

Contents

1 Introduction

In the first part of this thesis, statistical analyses and point process models for the epidermal nerve fibers (ENFs) are presented. ENFs are dendroidal, unmyelinated thin sensory nerve fibers found in the epidermis, the outermost living layer of the skin. They pass across the dermis, the skin layer beneath the epidermis, enter, and grow within the epidermis with or without branching until they terminate (see Figure 2.1). The ENF endings are responsible for transferring signals such as pain and heat to the central nervous system. Peripheral neuropathies, such as diabetic neuropathy, a neuropathic disorder caused by diabetes, damage the nerves. This damage translates into symptoms such as neuropathic pain and loss of sensation. As there is no current treatment able to restore the nerve fibers functionality, diagnosis of the neuropathy at an early stage is important.

The ENF data were treated as realisations of multitype point processes, consisting of the locations where the nerve fibers enter the epidermis, branch, and terminate observed in three-dimensional boxes. Throughout this thesis, those points will be referred to as base, branching, and end points, respectively. One of the main goals of this thesis was to further enhance our understanding of the biological process that leads to the morphological alterations in nerve fibers caused by neuropathy, and therefore data from healthy volunteers and patients suffering from diabetic neuropathy were considered. A better understanding of the underlying process can contribute to the development of more efficient techniques for early identification of the disease. To achieve our objectives, in Papers I and II we investigated the three-dimensional structure of the nerve trees, and in Paper III we investigated the death process of the nerve trees. Unlike the previous models that solely examine the base and end point locations, we included the first branching points in the analyses.

In Paper I, the three-dimensional structure of the ENF endings was investigated, to the best of our knowledge, for the first time. The nerve tree structure within the individual nerve fibers and in each disease group was investigated.

Our findings indicated that the segments connecting the end and branching points in the two groups have significant distributional differences. Then, we examined possible competitive behaviors between the nerve trees. For this purpose, the concept of epidermal active territory (EAT), a tool that approximates the volume of the epidermis covered by the nerve trees was introduced. The EAT values for each nerve tree were then attached as marks to the base points. No evidence was found to suggest that the nerve trees compete with each other in terms of mark correlation, a summary statistic often used in the analysis of marked point processes. Finally, we constructed a three-dimensional two-step spatial point process model for the end points that included some ingredients from the earlier introduced models. In the first step, the branching point locations were constructed towards open space, while in the second step, the end point clusters were constructed around the branching points. The two-dimensional version of the model fitted the data quite well while the 3D version failed to capture the structure of the data at intermediate distances.

The model proposed in Paper I was further developed in Paper II. This extension allowed interaction between nerve fiber end points. In this model, the planar point patterns were simulated using the 2D version of the model proposed in Paper I. Then, the z-coordinates of the points were constructed using a pairwise interaction Markov random field model (Christoffersen et al., 2021). For evaluating the model we considered groupwise pooled second-order summary statistics from the healthy and mild diabetic samples. To assess the goodness of fit we used directional summary functions from spatial statistics due to the anisotropic nature of the data. We found some indication of variations in the degree of interaction and the extent of the interaction zone between the groups.

The dynamics of the nerve mortality process were studied in Paper III. For this purpose, we considered ENF patterns projected onto the plane, and developed spatial thinning models for the healthy samples. Initially, we developed an independent p-thinning model, i.e. a model where each nerve tree is retained with probability p independently of the other nerves, to test the hypothesis of random nerve tree mortality, that is there is no particular biological process guiding the nerve removal. However, our results indicated that such a model was not enough to capture the changes in the spatial structure of the base and end points. Therefore, we developed a parametric dependent thinning model in which the retention probability was given as a function of the distance between base points. In particular, the retention probability was high for points with close neighbors and low for isolated points. To infer the distribution of the model parameter we used an approximate Bayesian computation method(Sisson et al., 2018) and the model was evaluated with respect to spatial and non-spatial summary statistics. According to our results, the dependent thinning model could generate patterns that look like early-stage diabetic point pattern replicates. As a result, this study indicated that as the neuropathy advances, the isolated nerve trees die first. Therefore, this was the first model that could explain the increased clustering observed in the diabetic patterns.

The goodness-of-fit of the models developed for the ENF data were evaluated using the global envelope tests (Myllymäki et al., 2017). These tests were initially developed for solving the multiple testing problem in spatial statistics, and extended to other various areas since then. In the second part of this thesis, we further extended the global testing framework. In Paper IV, global quantile regression (GQR), an extension of the global envelope tests to the quantile regression model (Koenker and Bassett Jr, 1978) was developed. The GQR test is a Monte Carlo permutation test, that for any set of quantiles, simultaneously tests if a covariate of interest has a significant effect on the conditional quantile distribution of a response variable, even in the presence of some nuisance covariates affecting the response distribution. We refer to this procedure as simultaneous inference for the quantile regression process. The test also allows for graphical interpretation, that is, it illustrates the set of quantiles responsible for the rejection of the null hypothesis. As global envelope tests are based on ranks, there are no parametric assumptions for the test to be valid. The only assumption is the exchangeability of the test vector under the permutation strategy. As this assumption is violated for some permutation schemes when nuisance parameters are present, we conducted a simulation study to investigate the empirical significance levels as well as the power of the tests in different settings motivated by some applications. We showed that the Freedman-Lane permutation (Freedman and Lane, 1983) was liberal when extreme quantiles and significant nuisance effects were considered. For this purpose, we proposed four permutation strategies that were close to exact. One of the permutation strategies satisfied exchangeability but was valid only for categorical nuisance. The other three permutation schemes assumed specific nuisance effects. We showed that these schemes achieved correct nominal significance levels, as long as the nuisance and interesting covariates were not correlated. In the case of correlated covariates, the test was liberal under permutation model misspecification, i.e., when the permutation assuming a location shift nuisance effect was used when the real nuisance effect was not a location shift, and conservative when the permutation model was correctly specified. Finally, when the interesting covariate is categorical the GQR test provides an n-sample graphical test of correspondence of distributions, even in the presence of nuisance variables.

In Paper V, the global envelope tests were extended for comparing the distributions of n-samples. For this end, simple permutations, i.e., data permutations between the n samples, were used, and functional test statistics capturing different distributional contrasts were considered. A simulation study was conducted to evaluate the performance of the proposed tests with regard to statistical power in different simulated scenarios. The performance of the tests was further compared with the performance of the classical two-sample Kolmogorov-Smirnov test. According to our results, the proposed global tests were more powerful than the Kolmogorov-Smirnov test in all studied settings. Finally, we provided detailed guidelines on how the graphical illustration of the test result should be interpreted for each test statistic.

The thesis is structured as follows. In Section 2, we describe the epidermal nerve fiber dataset and in Section 3, we briefly present some theoretical aspects of the point process theory used in Papers I-III. In Section 4, we present some additional statistical tools for simulation, estimation and hypothesis testing used throughout the thesis. In Section 5, we present the permutation strategies considered for the global tests proposed in Papers IV and V. A summary of the appended papers is given in Section 6. Discussion about the main contributions and future work are presented in Section 7.

2 Data

For the statistical analyses of Papers I-III, the epidermal nerve fiber data (Kennedy and Wendelschafer-Crabb, 1993; Kennedy et al., 1996, 1999) were considered. The epidermal nerve fiber data were collected through suction-induced skin biopsies, a procedure where a portion of the epidermis was removed, mounted on a slide, and stained for imaging. Then, confocal microscopy was used to manually trace the locations where the nerve fibers enter the epidermis, branch, and terminate (Kennedy and Wendelschafer-Crabb, 1993). In this thesis, an epidermal nerve tree is represented by its base, first branching, and end points. The structure of the ENF data, including the complete nerve trees is presented in Figure 2.1. The three types of points used in the analyses are marked with different colors.

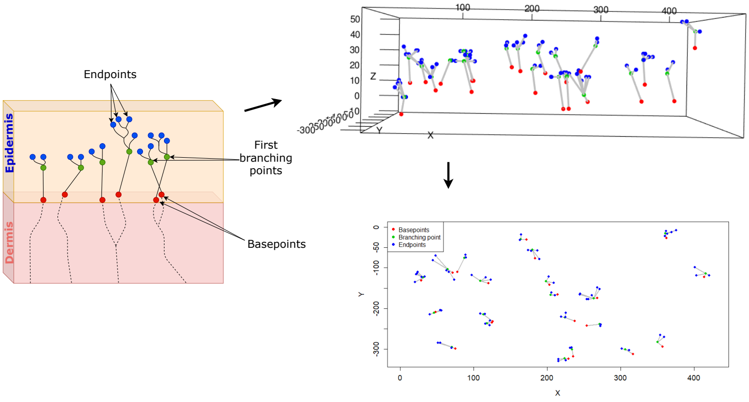

Figure 2.1: An illustration of the ENF data structure (left). An example of the three-dimensional multitype point patterns of the ENF base, branching, and end points (top right) and the corresponding two-dimensional point pattern of the projections (bottom right). The connections between the branching points and the base and end points are also illustrated.

The data were regarded as realisations of point processes in three-dimensional boxes and the projections of the data as point processes in the corresponding rectangles. Each point pattern contained three different types of points: base, branching, and end points. Planar point process models for the base point patterns were suggested in Andersson et al. (2019) and Andersson and Mrkvička (2020). The base points were found to be clustered indicating that the nerve fibers may branch prior to entering the epidermis. The second type of points were the first branching points, which will be referred to as branching points throughout the thesis. Those were the points where the fibers begin to branch in the epidermis. The branching points were included in the analysis for the first time in this thesis. The epidermal nerve fiber endings were the third type of points. Those are the points responsible for feeling heat and pain and hence their spatial structure is important. Earliest spatial analysis of the ENFs, focused on studying the spatial structure of the ENF projections onto the plane. The main findings suggested that there is a negative correlation between the spatial intensity of the base and end point patterns with the degree of the neuropathy (Kennedy et al., 1996, 1999; Myllymäki et al., 2012; Andersson et al., 2016; Olsbo et al., 2013). Furthermore, several studies proposed clustered point process models for the planar spatial structure of the ENFs end points (Olsbo et al., 2013; Andersson et al., 2016; Ghorbanpour et al., 2021; Garcia et al., 2020). In particular, in this thesis we used modelling ideas from the Non-Orphan Cluster model developed in Olsbo et al. (2013) and the Uniform Cluster Centre model developed in Andersson et al. (2016).

The ENF dataset is a hierarchically structured point pattern collection comprising data from healthy volunteers and patients suffering from diabetic neuropathy. The main hierarchies are the degree of diabetic neuropathy, i.e. healthy, mild, moderate, or severe, the different subjects, and samples within the subjects. In this thesis, we concentrated on data collected from 8 mild diabetic subjects and 32 healthy controls. Furthermore, we focused on samples obtained from the foot since research has shown that changes in the physiology of the ENFs occur at an early stage in the distant body regions (Kennedy et al., 1999). For each subject, three to six skin samples were available.

3 Spatial point processes

This chapter aims to provide a brief introduction to the theoretical concepts used in the first three papers. The basic concepts, definitions, and a brief overview of spatial summary functions for spatial point processes are recalled and discussed. For a more mathematically rigorous treatment of the topic, the reader is referred to the literature (Illian et al., 2008; Møller and Waagepetersen, 2004; Chiu et al., 2013). The definitions and notations given here mainly follow the book by Illian et al. (2008). Throughout this work, $\mathbb{R}^d$ and $\mathcal{B}(\mathbb{R}^d)$, denote the d-dimensional Euclidean space and its corresponding Borel sets, respectively. Further, all the subsets under consideration are assumed to be Borel sets. The indicator function is denoted by $\mathbb{1}\{\cdot\}$ and the $\neq$ above summation sign denotes the summation over all distinct pairs of points. The probability of an event A is denoted by $\mathbb{P}(A)$ and the Lebesgue measure by $|\cdot|$.

3.1 Basic definitions

Spatial point processes are mathematical models suitable for characterizing a random set of points. The process is usually defined in the entire space $\mathbb{R}^d$ but is only observed in a bounded observation window $W \subset \mathbb{R}^d$. Realizations of point processes are called point patterns or point configurations and the points of the process are often referred to as events. Point processes are central in many applications and serve as models for a wide range of physical phenomena. Such applications include astronomy, which involves modeling the spatial locations of galaxies and stars, forestry, which involves modeling the spatial distribution and interactions between e.g. different tree species, and medical applications, for instance modeling the changes in the spatial structure of the termination locations of epidermal nerve fibers as diabetic neuropathy progresses. The latter application was investigated in this thesis.

Point processes are assumed to be *simple*, meaning that at every distinct location the process places at most one point, and *locally finite*, indicating that for every bounded set B, the random variable $N_X(B)$ associated with the number of points of the process $X = \{X_i\}$ in B, is finite. The notation X_i denotes the locations of random points in D. In mathematical notation, this is written as follows.

(i) X is simple, i.e. $\mathbb{P}(X_i \neq X_j) = 1, \quad \forall\, i \neq j$

(ii) X is locally finite, i.e. for any bounded set B, we have that $N_X(B) < \infty$

The *intensity measure* of a point process, $\Lambda(B) : \mathcal{B}(\mathbb{R}^d) \to [0, \infty)$, is defined as the expected number of points of the process X in B, i.e $\Lambda(B) = E[N_X(B)]$. The intensity measure $\Lambda(B)$ can typically be expressed as

$$\Lambda(B) = \int_B \lambda(x)dx, \tag{3.1}$$

where the function $\lambda : \mathbb{R}^d \to [0, \infty)$, is called the *intensity function*.

A point process X is *stationary* if its distribution is invariant under translations. For stationary point processes, the intensity measure is $\Lambda(B) = \lambda \mid B \mid$. This implies that the intensity function $\lambda(x) \equiv \lambda$ is constant, and hence it is called the intensity of the process, which is interpreted as the mean number of points per unit volume. A point process with constant intensity function is called homogeneous and a point process with non-constant intensity function is called inhomogeneous. Therefore, all stationary point processes are automatically homogeneous. On the other hand, testing for the stationarity assumption of the underlying process from just one realization is statistically impossible, and justifying stationarity is mainly based on application related arguments. Lastly, a point process X is *isotropic* if its distribution is invariant under rotations around the origin. In the appended Papers, the point patterns containing the projections of the ENF base, branching, and end points onto the plane, are assumed to be subsets of realizations from stationary planar point processes. The three-dimensional point patterns are assumed to be subsets of realizations from stationary three-dimensional point processes.

There are three main types of point processes, namely clustered, regular, and completely spatially random (CSR) processes. In a clustered process, the points are arranged in clusters, in a regular process, there are repulsive dependencies between the points, and in the CSR process, there is no structure, i.e. the points are uniformly and independently distributed in space. An illustration of the three main types of point processes is given in Figure 3.1. More complicated types of point processes may be created by incorporating different types of dependencies at different scales, for instance, a clustered process with repulsion between the clusters. Therefore, characterizing the structure of the point pattern data at hand is one of the main topics in point process theory.

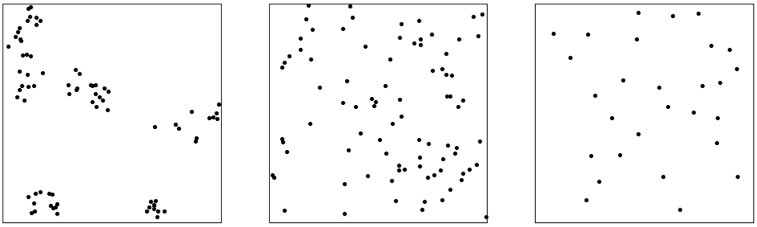

Figure 3.1: The three main types of point patterns: are clustered (left), completely spatially random (middle), and regular (right).

It is important to note that the spatial structure of a three-dimensional pattern and the pattern consisting of the events projected onto the plane might have different clustering behaviors. An example of such a case is illustrated in Figure 3.4. The pattern on the left (purple) is more clustered than the right pattern (red) in $\mathbb{R}^2$ but less clustered in $\mathbb{R}^3$.

3.1.1 Poisson point process

A Poisson point process is a mathematical model describing the complete spatial randomness case. A point process is a homogeneous Poisson point process with intensity $\lambda \geq 0$ if

(i) the random number of points of the process in a set B follows a Poisson distribution with the expectation $\lambda|B|$, i.e $N_X(B) \sim Pois(\lambda|B|)$, and

(ii) given $N_X(B) = n$, the n points are uniformly and independently allocated in B.

Despite its simplicity, the Poisson process plays a fundamental role in the characterisation of the spatial structure of spatial point patterns. As many theoretical properties and spatial summary functions can explicitly be derived for the Poisson point process, it is used as a reference model. That is, the summary functions estimated from empirical point patterns are compared to the theoretical values under CSR.

Furthermore, it serves as a basic building block for constructing more complex clustered and regular point processes. For instance, it serves as a model for the parent process in Neyman-Scott point processes, a family of models for clustered point patterns (see Section 3.4.1), as well as an initial model in the hardcore type of processes, a family of models for regular point patterns.

3.2 Functional summary statistics

We recall that the ENF base, branching, and end points were treated as realisations of stationary point processes in a three-dimensional box and their projections as stationary point processes in a rectagular window. In this section, functional summary statistics for stationary point processes are briefly recalled.

3.2.1 Summary functions

The *empty space distribution function* $F(r) : [0, \infty) \to [0, 1]$ gives the probability that the open ball around an arbitrary point $x \in \mathbb{R}^d$ with radius r, $b(x, r)$, contains at least one event of the point process X. For stationary point processes, since the probability $\mathbb{P}(N_X(b(x, r)) = 0)$ does not depend on x, it is sufficient to consider x to be the origin o. In mathematical notation, $F(r)$ is expressed as

$$F(r) = 1 - \mathbb{P}(N_X(b(o, r)) = 0). \tag{3.2}$$

Similarly, the *nearest neighbor distance distribution function* $G(r) : [0, \infty) \to [0, 1]$ gives the probability that the ball around an arbitrary point x of the process with radius r, $b(x, r)$, contains its nearest neighboring point x of the process X. Assuming that the process is stationary we can assume that x is the origin. The nearest neighbour function $G(r)$ is given by

$$G(r) = 1 - \mathbb{P}_o(N_X(b(o, r) \setminus \{o\}) = 0) \tag{3.3}$$

where $\mathbb{P}_o$ is the Palm distribution of X, a conditional probability distribution given that there is an event in the origin.

The J function can be constructed from the nearest neighbor function $G(r)$ and empty space function $F(r)$. The J function is given by

$$J(r) = \frac{1 - G(r)}{1 - F(r)}, \quad \text{when} \quad F(r) < 1.$$

The values for the summary functions for the three main types of point patterns are interpreted as follows

(i) For CSR point patterns the following is true for $r \geq 0$

$$F(r) = G(r) = 1 - e^{-\lambda \pi r^2} \quad \text{and} \quad J(r) \equiv 1. \tag{3.4}$$

(ii) For regular point patterns, we have that for $r \geq 0$

$$G(r) < 1 - e^{-\lambda \pi r^2} < F(r) \quad \text{and} \quad J(r) > 1. \tag{3.5}$$

(iii) For clustered patterns, we have that for $r \geq 0$

$$F(r) < 1 - e^{-\lambda \pi r^2} < G(r) \quad \text{and} \quad J(r) < 1. \tag{3.6}$$

It is important to note that the G, F, and J functions are appropriate for describing the spatial structure at small scales, since they consider the nearest events, but cannot provide any information about the structure at larger scales. Further, we should be careful when interpreting values of the J function as $J \equiv 1$ does not imply that X is the homogeneous Poisson process (Bedford and Van den Berg, 1997). Moreover, since $\lim_{r \to \infty} 1 - F(r) = 0$ the variance of the estimate $\widehat{J}(r)$ increases with increasing r.

3.2.2 Second-order characteristics

Ripley's K function proposed by Ripley (1977) is a second-order summary function that can characterize the structure of a point process at different scales. For stationary and isotropic point processes with intensity λ, Ripley's K function has a straightforward interpretation. In particular, $\lambda K(r)$ gives the expected number of further points of the process X within distance r from an arbitrary point x of the process. As X is stationary, we can assume that x is the origin o. In mathematical terms, Ripley's K function is defined as

$$\lambda K(r) = E_o[N_X(b(o, r) \setminus \{o\})] \tag{3.7}$$

where E_o is the expectation with respect to the Palm distribution, which is interpreted as the conditional expectation given that there is an event in the origin. The values of Ripley's K function for point patterns in $\mathbb{R}^d$ can be interpreted as follows

(i) For CSR patterns, $K(r) = |b(o,r)|, \quad r \geq 0$

(ii) For regular patterns, $K(r) < |b(o,r)|, \quad r \geq 0$

(iii) For clustered patterns, $K(r) > |b(o,r)|, \quad r \geq 0$

where $b(o,r)$ is the d-dimensional ball centered at the origin o with radius r. A more interpretative version of the K function is the so-called centered L function. The centered L function is given by

$$L(r) - r = \sqrt[d]{\frac{K(r)}{|b(o,1)|}} - r, \quad r > 0. \tag{3.8}$$

The theoretical value for the CSR case is $L(r) - r \equiv 0$, and hence we can determine if a pattern is clustered or regular by comparing the value of the summary function directly with zero.

3.2.3 Edge corrections

Naive estimators of the summary functions ignore neighboring points that might not have been observed, i.e. points outside the observation window W, which makes the naive estimators biased. In point process literature, this issue is referred to as edge effects. An illustration of this issue is displayed in Figure 3.2.

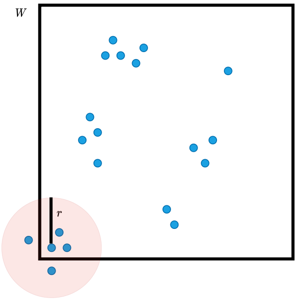

Figure 3.2: Example of edge effects. The points outside the observation window W are not observed and therefore, if they are ignored the estimators become biased.

Therefore, to construct unbiased estimators, some edge correction weights $C(x_i, x_j)$ are included in the estimators. Even though all the summary functions introduced earlier need to be edge corrected, this Section presents edge correction schemes only for the K function. The most common corrections for Ripley's K function are the *translation, isotropic* and *minus sampling* corrections. For a stationary point process, the translation correction weights $C(x_i, x_j)$ are defined as follows.

$$C(x_i, x_j) = \frac{1}{|W \cap W_{x_i - x_j}|} \tag{3.9}$$

where $W_{x_i-x_j}$ denotes the translated window $W_{x_i-x_j}$ and $|\cdot|$ the Lebesgue measure. If the process is also isotropic, the isotropic correction is defined as

$$C(x_i, x_j) = \frac{\nu_1(\partial b(x_i, || \; x_i - x_j \; ||) \cap W)}{2\pi \; || \; x_i - x_j \; ||} \tag{3.10}$$

where ν_1 denotes the length of a curve, $|| \cdot ||$ denotes the Euclidean metric, and ∂ denotes the boundary of a set. The above expression can be interpreted as the proportion of the perimeter of the ball that lies within the window W. In the minus sampling correction, only the points that have a distance larger than r from the boundary of the window are used as reference points in the estimation of the summary function. A visual interpretation of the different edge correction schemes is given in Figure 3.3

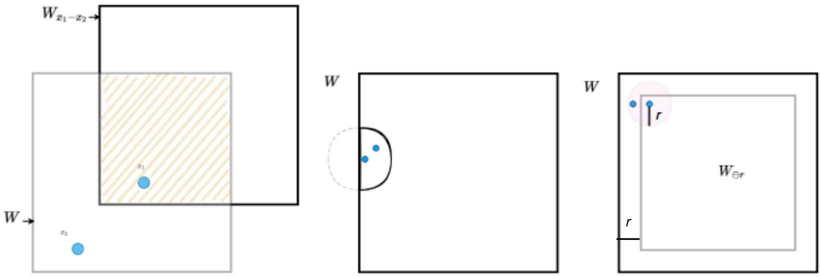

Figure 3.3: Illustration of the translation (left), isotropic (middle) and minus sampling (right) edge corrections

Therefore, an estimator for Ripley's K function corrected for edge effects is given by the following formula

$$\widehat{K}(r) = \frac{1}{\widehat{\lambda} n} \sum_{x_1, x_2 \in X \cap W}^{\neq} C(x_1, x_2) \mathbb{1}\{|| \; x_1 - x_2 \; || \leq r\}, \quad r \geq 0, \tag{3.11}$$

where $C(x_i, x_j)$ is one of the edge correction weights, $\widehat{\lambda}$ an estimator of the process intensity and n the total number of events.

3.2.4 Extensions of the K function

Bivariate K function

The K function can be extended for multitype point processes. Let X_a and X_b be two stationary, possibly dependent, point processes observed in W, and let λ_a, λ_b be the intensities of X_a and X_b, respectively. Then, $\lambda_b K_{a,b}(r)$ gives the expected number of further points of the process X_b in the d-dimensional ball with radius r centered at an arbitrary point x of the process X_a. If the process is stationary, x can be assumed to be the origin o. In mathematical notation, this is expressed as

$$\lambda_b K_{a,b}(r) = E_a[N_{X_b}(b(o,r) \setminus \{o\})], \tag{3.12}$$

where E_a is a conditional expectation given there is an event of X_a in the origin o. Important to note that the *bivariate $K_{a,b}(r)$ function* coincides with the original Ripley's K function when X_a and X_b are the same processes. An estimator for $K_{a,b}(r)$ corrected for edge effects can be obtained by

$$\widehat{K}_{a,b}(r) = \frac{1}{\widehat{\lambda}_b \widehat{\lambda}_a |W|} \sum_{i=1}^{n_a} \sum_{j=1}^{n_b} C(x_i^a, x_j^b) \mathbb{1}\{0 < \| x_i^a - x_j^b \| \leq r\}, \quad r \geq 0, \tag{3.13}$$

where $C(x_i^a, x_j^b)$ is an edge correction term, n_a and n_b are the numbers of points and $\widehat{\lambda}_a$ and $\widehat{\lambda}_b$ the intensity estimates of X_a and X_b, respectively.

Cylindrical K function

The isotropic K function is not an appropriate summary statistic for non-isotropic point patterns due to its symmetric structuring element, i.e. a d-dimensional ball. Directional K functions with non-symmetric structuring elements have been suggested as extensions of Ripley's K function for anisotropic point processes (see e.g Stoyan and Stoyan (1994); Rajala et al. (2018)). The three-dimensional ENF end point clusters are anisotropic, as the behavior in the xy plane differs from the behavior in the z direction, and therefore, we considered the cylidrical K function. The *cylindrical K function $K_{cyl}^u(r)$* is a directional K function that uses a cylindrical structuring element (Møller et al., 2016). Similar to Ripley's K function the values of the cylindrical variant can be interpreted as the expected number of further events within distance r from an arbitrary event x, that are within $B^u(r, w)$, i.e. the shape created by the intersection of a cylinder with fixed half-width w and direction u with spheres of radius $r > 0$, divided by the process intensity. Choosing appropriate directions u mainly depends on the nature of the application. For instance, the cylindrical $K^u_{cyl}(r)$ function can be estimated towards directions of the coordinate axes to test the assumption of isotropy. If $K^u_{cyl}(r)$ is not the same in each direction u, it is an indication of anisotropy. An estimator for $K^u_{cyl}(r)$ corrected for edge effects is given by

$$\widehat{K}^u_{cyl}(r) = \frac{1}{\hat{\lambda^2}} \sum_{x_1, x_2 \in X \cap W}^{\neq} C(x_1, x_2) \mathbb{1}[x_1 - x_2 \in B^u(r, w)], \qquad r > 0, \tag{3.14}$$

where $C(x_1, x_2)$ is an edge correction term, $B^u(r, w)$ denotes the shape created by the intersection of a cylinder with fixed half-width w and direction u with spheres of radius $r > 0$ and $\hat{\lambda^2}$ is an estimator for the process squared intensity. As the half width w is fixed, $K^u_{cyl}(r)$ is defined as a function of distance r. In a similar fashion as for the usual K function, more interpretative variants can be defined for both the bivariate and cylindrical K functions. The structuring elements of the isotropic K function and the cylindrical K function oriented towards the z-axis are illustrated in Figure 3.4.

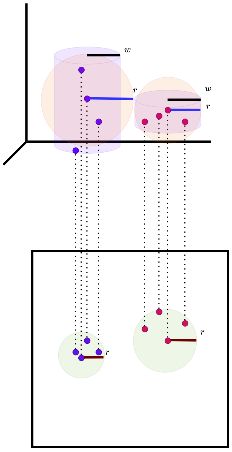

Figure 3.4: Illustration of potential differences in clustering in different dimensions. The projections of the "red" pattern are less clustered in $\mathbb{R}^2$ but more clustered than the "purple" pattern in $\mathbb{R}^3$. The structuring elements for the isotropic Ripley's K in $\mathbb{R}^2$ (bottom) and the cylindrical K function in $\mathbb{R}^3$ (top) are also illustrated.

Pooled K function

The aforementioned spatial summary functions are appropriate for describing the spatial structure of a single point pattern. When point pattern replicates are available, the summary functions can be pooled to characterize the average spatial structure (Diggle et al., 1991; Baddeley et al., 1993; Diggle et al., 2000). For instance, the ENF data set is hierarchically structured into groups depending on the neuropathy severity, subjects within those groups, and several samples within each subject, and we want to compare the average spatial structure of the ENF patterns between the groups. The following methodology can be applied to extend the samplewise summary functions to subjectwise and groupwise functions. Firstly, samplewise summary functions $K_{ij}(r)$ for sample $j \in \{1, ..., m_i\}$ of subject i can be estimated for each point pattern, as introduced in Section 3.2.2. Then, subject specific $\bar{K}_i(r)$ functions can be obtained as a weighted mean of the $K_{ij}(r)$ functions for all subjects $i \in \{1, ..., m\}$ as

$$\bar{K}_i(r) = \sum_{j=1}^{m_i} w_{ij} K_{ij}(r). \tag{3.15}$$

Finally, the subjectwise $\bar{K}_i(r)$ functions are combined to obtain the groupwise $\bar{K}_g(r)$ function for the group $g \in \{\text{healthy, diabetic}\}$

$$\bar{K}_g(r) = \sum_{i=1}^{m} w_i \bar{K}_i(r). \tag{3.16}$$

Several weighting schemes can be applied to calculate the weights. As our samples cannot be assumed to have the same intensity, we used a square number weight scheme (Diggle et al., 2000). An explicit description of this weighting scheme is as follows. Let n_{ij} be the number of points in sample j of subject i, and let $n_i = \sum_{k=1}^{m_i} n_{ik}$ be the total number of points in the samples from subject $i \in \{1, ..., m\}$. Then the square point number weights for the subjectwise $\bar{K}_i(r)$and groupwise $\bar{K}_g(r)$ are given by

$$w_{ij} = \frac{n_{ij}^2}{\sum_{k=1}^{m_i} n_{ik}^2}, \quad w_i = \frac{n_i^2}{\sum_{k=1}^{m} n_k^2}. \tag{3.17}$$

3.2.5 Mark correlation function

Often in applications additional information, often referred to as marks, are attached to the points. Generally, the marks are either integers, i.e. different types of points in multitype point patterns, or real-valued numbers, i.e. size-related features such as height, diameter, or width, even though the mark space can be any Polish space, i.e., a complete and separable metric space (Eckardt and Moradi, 2024; Cronie et al., 2024). The mark correlation function is a second-order characteristic for marked point processes that is used to detect spatial dependencies between the marks. The classical mark correlation function $K_{mm}(r)$ for quantitative marks is defined as follows.

$$K_{mm}(r) = \frac{c_{mm}(r)}{\mu^2} \quad r \geq 0 \tag{3.18}$$

where $c_{mm}(r) = E_{o,r}[m_o \cdot m_r]$ is the two-point (Palm) conditional expectation given that there is an event in the origin and an event distance r away and $\mu^2 = c_{mm}(\infty)$ is the squared mean mark. The mark correlation function can be interpreted as follows

(i) If there is no correlation between the marks (given two associated spatial locations distance r apart), then $K_{mm}(r) \equiv 1$.

(ii) If there is a negative correlation between the marks, for instance, there is competition between the points, we expect smaller than average marks for close points and hence $K_{mm}(r) < 1$.

(iii) If there is a positive correlation between the marks, for instance, the points benefit from being close together, we expect larger than average marks and hence $K_{mm}(r) > 1$.

In general, the mark correlation function can be defined naturally through a test function $t(m_o, m_r)$, through the summary statistic $c_t(r) = E_{o,r}[t(m_o, m_r)]$. For instance, the classical mark correlation function (Stoyan, 1984) above is obtained with the test function $t(m_o, m_r) = m_o \cdot m_r$. An estimator of the mark correlation $K_t(r)$ based on test function t, for $r \geq 0$ and corrected for edge effects is given by

$$\widehat{K}_t(r) = \frac{1}{\widehat{\lambda}^2 \widehat{\mu}^2} \sum_{x_1, x_2 \in X \cap W}^{\neq} C(x_1, x_2) t(m(x_1), m(x_2)) \mathbb{1}\{\| x_1 - x_2 \| \leq r\}, \tag{3.19}$$

where $\widehat{\lambda}$ is an estimator for the process intensity, $\widehat{\mu}$ is an estimator for the mean of the mark distribution, and $C(x_1, x_2)$ is an edge correction term.

Often the null hypothesis of random labeling, i.e., the marks have a common mark distribution and are independent of the point locations and each other, is of interest. Simulation envelopes under random labeling, can be constructed using a Monte Carlo method where at each iteration the marks are randomly permuted between the points, the locations of the points are kept fixed, and the mark correlation function for the permuted marked patterns is computed (D'Angelo et al., 2023). Simulations using this method construct marked point patterns with randomly labeled marks which are then used to create simulation based envelopes, i.e. an acceptance region under the null hypothesis of random labeling of the marks.

3.3 Spatially thinned processes

Spatial thinning of a point process X is one of the fundamental operations for point processes. A thinned process $X_{th} \subset X$ is a point process obtained after a thinning operation was applied to X, i.e., a probabilistic rule determining which points $x \in X$ should be retained in X_{th}. For each point $x \in X$, a probabilistic model defines the retention probability $\pi(x)$. Independent thinning models are the thinning models in which $\pi(x)$ is independent of other points in X. A special case is the p-thinning model in which for every point $x \in X$ the retention probability p is constant. When X is stationary, second-order properties such as Ripley's K and L functions are invariant under independent thinning operations. On the contrary, this invariance property does not hold under a dependent thinning operation, a thinning operation in which the retention probability depends on other points in X, that is $\pi(x) = \pi(x \mid X)$. In paper III, a dependent thinning model for the ENF data was proposed.

3.4 Point process models for ENF patterns

This Section focuses on point process models for the ENF patterns, i.e. clustered point process models. The Neyman-Scott family of cluster point process models is described first, followed by cluster models developed specifically for the ENFs.

3.4.1 Neyman-Scott point processes

Neyman-Scott point processes are cluster processes originally introduced by Neyman and Scott (1952) to model the locations of galaxies in space. The construction of a Neyman-Scott point process is rather simple. Initially, the process for the cluster centers, i.e. the so-called parent process, should be specified. Typically, a Poisson point process P with intensity λ_p is used as the parent process. Then, a distribution for the number of daughter points N_c per cluster center $c \in P$ is chosen. The daughter points $x \in X_c$ are distributed in space according to a scattering distribution δ independently of each other. The final process X is then $X = \cup_c X_c$, hence the parent points are not included.

A Neyman-Scott process X is stationary and if the scattering distribution is isotropic then X is also isotropic. Moreover, the first and second-order properties of X can be derived explicitly. Let $\alpha = E[N_c]$ be the expectation of the distribution of the number of offsprings N_c, then the intensity of X is

$$\lambda = \lambda_p \alpha \tag{3.20}$$

Now let $p_k = \mathcal{P}(N_c = k)$, for $k \in \mathbb{N} \cup \{0\}$ and $F_d(r)$ be the distribution function of the random distance between two independent points in the same cluster X_c. Then, Ripley's K function for X is given by

$$K(r) = \pi r^2 + \frac{1}{\lambda\alpha} \sum_{i=2}^{\infty} p_i i(i-1) F_d(r), \quad r \geq 0. \tag{3.21}$$

The most notable examples of Neyman-Scott point processes are the Matérn and Thomas cluster point processes. In both point process models, the distribution of the random number of offsprings follows a Poisson distribution with expectation α, that is $N_c \sim Poisson(\alpha)$. In the Matérn cluster process, the scattering distribution δ is a uniform distribution in the ball $b(c, R)$, for $c \in P$ and some radius R. In the Thomas cluster process, δ is a Gaussian distribution with variance σ^2. In both cases, δ is isotropic, and hence X is a stationary and isotropic process. When the distribution of N_c is considered to be a discrete distribution other than Poisson, generalizations of the Matérn and Thomas processes can be obtained (see e.g Andersson and Mrkvička (2020)).

3.4.2 Cluster models for ENFs

In earlier studies on the nerve fibers, different types of cluster models for the planar spatial structure of the endpoints were suggested. The models presented here are models for the end points and are constructed conditioned on the observed base point patterns. Each model consists of three main components, namely the length of the segments L, the angle Φ between the segments connecting base and end points and the x axis, and the tree size S, i.e. number of end points per base point. For every component, distributions are suggested. The main assumption that is common in both models is the independence between the different components, which simplifies the parameter estimation

procedure significantly.
The Non-Orphan Cluster (NOC) model (Olsbo et al., 2013) is given by

$$\begin{aligned} &L{\sim}Gamma(\alpha,\beta)\\ &\Phi \mid \mu{\sim}VonMises(\mu,\kappa)\\ &S{\sim}Jonquiere(\delta,\gamma) \end{aligned} \tag{3.22}$$

where μ is the so-called open space direction, defined for each base point as the direction opposite to the closest other base point.

The Uniform Cluster Centre (UCC)(Andersson et al., 2016) is given by

$$\begin{aligned} &L{\sim}Gamma(\alpha,\beta)\\ &\Phi \mid \mu{\sim}VonMises(\mu,\kappa)\\ &S-1{\sim}NegativeBinomial(k,p) \end{aligned} \tag{3.23}$$

where there is no preference for μ, i.e. $\mu \sim \text{Unif}(0, 2\pi)$.

Both models use two parameter distributions for the tree size as the one parameter Poisson distribution was not flexible enough. On the other hand, the main difference between the models is the direction in which the end point clusters are sent towards. In the NOC model, this direction depends on the base point pattern as the clusters are sent towards open space, while in the UCC model the clusters are sent towards a random direction.

4 General statistical tools

In this Section, some statistical tools used for simulation and estimation throughout the thesis are briefly described.

4.1 Metropolis-Hastings algorithm

Markov Chain Monte Carlo (MCMC) methods are statistical methods for inference and simulation from a target density $\pi(x)$ (see e.g Brooks et al. (2011)). If certain conditions are satisfied, a Markov chain $Y_0, Y_1, ...$ having the target distribution as its limiting distribution can be constructed. The Metropolis-Hastings algorithm is an MCMC algorithm, that requires the target distribution to have probability density (or probability mass) function to be known up to a constant (Metropolis et al., 1953; Hastings, 1970). Hence, the Metropolis-Hastings algorithm is also useful for simulating spatial point processes with a density function that has an intractable normalizing constant. The algorithm described in this section conditions on the number of points in the point pattern, i.e $N_X(B) = n$. Therefore, we are interested in simulating from the conditional unnormalized density h_n such as

$$\pi(x_1, x_2, ..., x_n) \propto h_n(x_1, x_2, ..., x_n). \tag{4.1}$$

If certain conditions are satisfied, the algorithm creates a Markov chain of point processes $Y_0, Y_1, ...$ whose stationary distribution converges to the target distribution π. In a *systematic updating scheme* (Møller and Waagepetersen, 2004) we cycle through each point in every iteration.

Given the state at iteration k, $Y_k = \mathbf{x}_k = (x_1, ..., x_n)$, assume that for the i-th coordinate x_i of $\mathbf{x}_k$ we propose a new point $\xi \sim q_i(\mathbf{x}_k, \cdot)$ from a proposal density q_i. The Hastings ratio $r_i(\mathbf{x}_k, \xi)$ is given by

$$r_i(\mathbf{x}_k, \xi) = \frac{h_n((\mathbf{x}_k \setminus x_i) \cup \xi) q_i((x_1, ..., x_{i-1}, \xi, x_{i+1}, ..., x_n), x_i)}{h_n(\mathbf{x}_k) q_i(\mathbf{x}_k, \xi)}. \tag{4.2}$$

Choosing a symmetric proposal, i.e. a proposal density such that $q_i(x, y) = q_i(y, x)$, simplifies the Hastings ratio $r_i(\mathbf{x}_k, \xi)$ to

$$r_i(\mathbf{x}_k, \xi) = \frac{h_n((\mathbf{x}_k \setminus x_i) \cup \xi)}{h_n(\mathbf{x}_k)}. \tag{4.3}$$

A proposed state ξ is accepted with acceptance probability $a_i(\mathbf{x}_k, \xi)$ given by

$$a_i(\mathbf{x}_k, \xi) = \min(1, r_i(\mathbf{x}_k, \xi)) \tag{4.4}$$

otherwise x_i is left unchanged. Moreover, properties of the Markov chain created by the specific algorithm, such as irreducibility and reversibility can be proved. For a more mathematically rigorous treatment of the topic you are referred to Chapter 7 in Møller and Waagepetersen (2004). In Paper II, a Metropolis-Hastings algorithm with a uniform proposal was used to simulate from the model. A pseudo-algorithm for a generic Metropolis-Hastings for point processes with conditional density h_n and a fixed number of points is given in Algorithm 1 below.

Algorithm 1

```
Input: A point pattern $Y_0$ = $(x_1, ..., x_n)$ and $M$ the number of iterations
Output: A realisation from $X$ given $N_X(B) = n$
for $m \leftarrow 0, ..., M$ do
    Given that $Y_m = \mathbf{x}_m$
    for $j \leftarrow 1, ..., n$ do
        Draw $\xi \sim q_j(\mathbf{x}_m, \cdot)$
        Calculate $r_j(\mathbf{x}_m, \xi)$ using (4.2)
        Calculate $a_j(\mathbf{x}_m, \xi)$ using (4.4)
        Draw $U \sim$ Uniform$(0, 1)$
        if $U < a_j(\mathbf{x}_m, \xi)$ then
            Set $Y_m = (x_1, ..., x_{j-1}, \xi, x_{j+1}, ..., x_n)$
        else
            Set $Y_m = \mathbf{x}_m$
        end if
    end for
end for
```

4.2 Approximate Bayesian computation

Approximate Bayesian computation (ABC) (Sisson et al., 2018; Marin et al., 2012) refers to a family of methods for Bayesian parameter inference of a statistical model $\mathcal{M}(\theta)$ when the likelihood of the model is unknown or when the likelihood is computationally expensive to approximate, but given a vector of parameters θ, data can efficiently be simulated from $\mathcal{M}(\theta)$. ABC methods require the specification of a prior distribution $P(\theta)$ for the parameter vector θ. Furthermore, a set of informative summary statistics $S(\cdot)$ and a tolerance parameter ϵ need to be specified. Unfortunately, choosing $S(\cdot)$ is often not a straightforward task. Recently, machine learning methods for learning the summary statistics for ABC have been developed (Jiang et al., 2017; Wiqvist et al., 2019). A pseudocode for a simple ABC acceptance-rejection sampling (Pritchard et al., 1999), is given in Algorithm 2. This method samples parameters from the approximate pseudo-posterior

$$P_\epsilon(\theta \mid S(y)) \propto P(\theta) \int \mathbb{1}_{||s^*-s||<\epsilon} P(s|\theta) ds^*$$

where $s^* = S(y^*)$ and $s = S(y)$ are obtained from the simulated data y^* and empirical data y, respectively.

Algorithm 2 ABC acceptance-rejection sampler

```
Input: prior $P(\theta)$, model $\mathcal{M}(\theta)$, summary statistic $S(\cdot)$, threshold $\varepsilon > 0$,
positive integer $N$.
Output: posterior draws $(\theta_1, ..., \theta_N)$.
for $i \leftarrow 1, ..., N$ do
    repeat
        Draw from prior $\theta^* \sim P(\theta)$
        Simulate $\mathcal{M}(\theta^*) \rightarrow y^*$
        Compute $S(y^*)$
    until $\|S(y^*) - S(y)\| < \epsilon$
$\theta_i \leftarrow \theta^*$
end for
```

On the other hand, when the posterior is not similar to the prior, this method has a very small acceptance rate, and hence it is computationally inefficient. Therefore, the tolerance parameter ϵ should be chosen to balance both the precision of the approximate posterior and acceptance rates. To this end, methods for tuning ϵ have been proposed (Simola et al., 2021; Drovandi and Pettitt, 2011). To address this issue, in Paper II we used a more sophisticated

ABC rejection sampler in which we (i) simulated a large number of $\theta^* \sim P(\theta)$; (ii) conditionally of θ^*, simulated data y^* as $\mathcal{M}(\theta^*) \to y^*$; (iii) reduced both y^* and y to a low-dimensional set of summary statistics $S(y^*)$ and $S(y)$, respectively, and evaluated their proximity using the Euclidean distance, i.e. $|| \ S(y^*) - S(y) \ ||$; and finally, (iv) retained θ^* if $|| \ S(y^*) - S(y) \ || < \epsilon$, for some small $\epsilon > 0$, and rejected otherwise.

4.3 Shift plots and qq plots

The shift function introduced in Doksum and Sievers (1976) is a statistical tool for comparing two distributions, graphically. Letting $X \sim F$ and $Y \sim G$, the shift function is defined as the function $\Delta(x)$ such as $F(x) = G(x + \Delta(x))$. Solving for $\Delta(x)$ we get that $\Delta(x) = G^{-1}(F(x)) - x$. Hence, $\Delta(\cdot)$ expresses the amount of 'shift' required so that X and Y coincide. Further, $\Delta(x) \equiv 0$ implies that X and Y have the same distribution. The qq (quantile quantile) plot is closely related to the shift function. When comparing two random samples with a qq-plot, the quantiles of one sample are plotted against the quantiles of the other. If they have the same distribution, the points should fall on the line $y = x$. The shift function $\Delta(x)$ is the shortest distance between the qq-plot points and the line $y = x$.

Simultaneous 95% confidence bands for $\Delta(x)$ can be created using the Kolmogorov-Smirnov statistic. Hence, if the line $y = 0$ lies within the confidence bands of the estimated shift function $\hat{\Delta}(x)$ then F and G are statistically indistinguishable. One advantage of this statistical tool is that if F and G differ, visual inspection of the shift plot can provide information on how the distributions differ. Examples of how shift plots can be used to compare the distribution of two random random variables are shown in Figure 4.1.

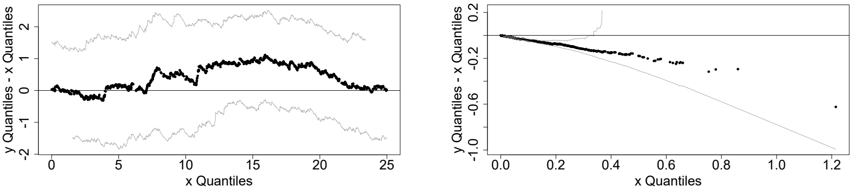

Figure 4.1: Illustration of the shift plots for two random variables that are statistically indistinguishable (left) and from different distributions (right), with 95% confidence bands based on the Kolmogorov-Smirvov statistic.

4.4 Global envelope tests

Global envelope tests are graphical, Monte-Carlo based tests for multivariate or functional data introduced in Myllymäki et al. (2017) originally developed to solve the multiple comparison problem in spatial statistics. Let $T(r)$ denote the statistic of interest, e.g. Ripley's K function, and $\mathcal{R} = (r_1, \ldots, r_d)$ be a d-dimensional vector containing the values where $T(\cdot)$ is evaluated. Further, let the d-dimensional discretization of the empirical statistic be denoted by $\mathbf{T}_0 = (T_0(r_1), \ldots, T_0(r_d))$ and the corresponding statistics for s simulated data sets under the "null model" by $\mathbf{T}_1$, ..., $\mathbf{T}_s$.

Global envelopes are non-parametric test as they are constructed by ranking the extremeness of the test vectors $\mathbf{T}_0, \ldots, \mathbf{T}_s$. The rankings are obtained through a ranking measure E. Some common ranking measures are the unscaled MAD measure (Ripley, 1981), the pointwise rank measure (Myllymäki et al., 2017), the extreme rank length (ERL) measure (Narisetty and Nair, 2016), the continuous rank measure (Hahn et al., 2015) and the area measure (Mrkvička et al., 2022). Figure 4.2 illustrates a toy example with five test statistics $\mathbf{T}_1, \ldots, \mathbf{T}_5$ where large values are considered extreme. In this example, the variance of the test statistic increases from the left to the right.

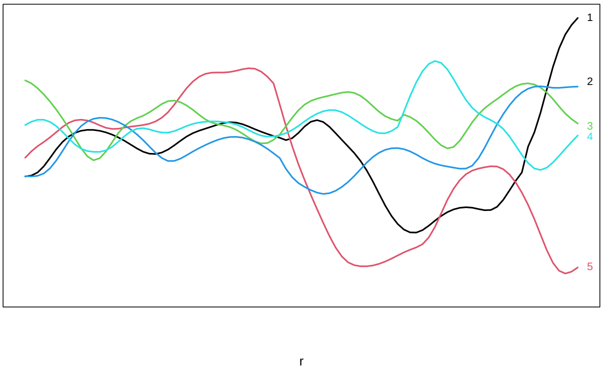

Figure 4.2: A toy example for describing the different measures. According to the unscaled MAD measure the black curve (1) is the most extreme, according to the ERL measure the green curve (3) is the most extreme and according to the continuous and area rank measures the red curve (5) is the most extreme.

According to the unscaled MAD measure, $\mathbf{T_1}$ is the most extreme as it obtains the overall highest value. According to the rank measure all test statistics are the most extreme as they obtain the largest value at some range. The ERL, area and continuous measures were proposed as solutions to breaking the ties in the rank measure. According to the ERL measure, $\mathbf{T_3}$ is the most extreme as it is above the other statistics over a longer domain of ranges. According to the continuous rank and area measures $\mathbf{T}_5$ is the most extreme. The continuous and area measures consider the deviation and the area between the test statistics, respectively. Then both of them adjust for the variance of the test statistic, i.e., smaller variance gives more extreme measure. For a more rigorous treatment to the topic see Myllymäki and Mrkvička (2019) and references therein.

Let α be the significance level and let $E_i < E_j$ be interpreted as $\mathbf{T}_i$ is more extreme than $\mathbf{T}_j$. Now, let $E_{(\alpha)} \in \mathbb{R}$ be the largest E_i such that

$$\sum_{i=0}^{s} \mathbb{1}(E_i < E_{(\alpha)}) \leq \alpha(s+1)$$

and let $I_{(\alpha)}$ denote the set of vectors less than or as extreme as $E_{(\alpha)}$. Then, a $100(1-\alpha)\%$ global envelope is the band $(\mathbf{T}_{low}^{(\alpha)}, \mathbf{T}_{upp}^{(\alpha)})$ given by

$$T_{low}^{(\alpha)}(r_k) = \min_{i \in I_{(\alpha)}} T_i(r_k) \quad \text{and} \quad T_{upp}^{(\alpha)}(r_k) = \max_{i \in I_{(\alpha)}} T_i(r_k) \quad \text{for } k = 1, ..., d.$$

where $\mathbf{T}_{low}^{\alpha} = (T_{low}^{\alpha}(r_1), \ldots, T_{low}^{\alpha}(r_d))$ and $\mathbf{T}_{upp}^{\alpha} = \big(T_{upp}^{\alpha}(r_1), \ldots, T_{upp}^{\alpha}(r_d)\big)$.

The global envelope can be interpreted as the acceptance region of the test, that is the test rejects the null hypothesis if and only if the empirical statistics $\mathbf{T}_0$ goes outside the global envelope at any point $r \in \mathcal{R}$. Moreover, it allows for graphical interpretation as the range of values responsible for the rejection of the null hypothesis are graphically illustrated. An example of a 95% global envelope (shaded area) constructed under the null hypothesis of CSR for the centered L function, i.e. $L(r) - r$, is shown in Figure 4.3. The null hypothesis is rejected as the empirical test statistic (solid line) does not lie completely within the envelope (shaded area), indicated by red points. Moreover, the reason for rejecting the null hypothesis is the clustering observed at small ranges.

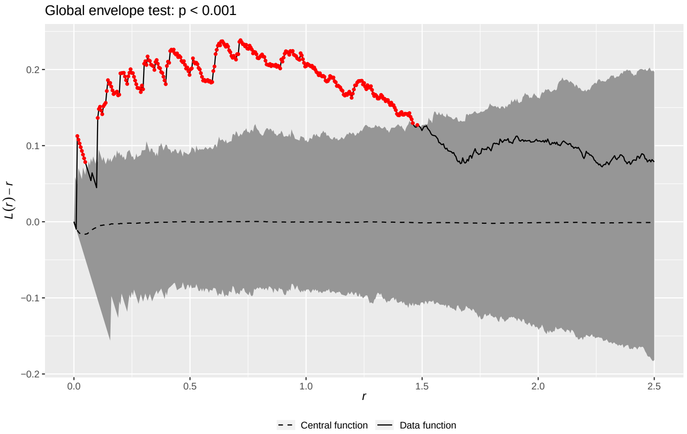

Figure 4.3: The $L(r) - r$ function of a point pattern (solid line) with 95% global envelope (shaded region) constructed under the null hypothesis of CSR.

The validity of global envelope tests does not depend on the distribution of the test statistic. Therefore, any test statistic may be used for testing. However, in order for the global envelopes to achieve desired type I errors, the test statistics $\mathbf{T}_0, ..., \mathbf{T}_s$ must be exchangeable, that is for any permutation σ and any measurable set A the following property must hold

$$\mathbb{P}\left((T_0, T_1, \ldots, T_s) \in A\right) = \mathbb{P}\left((T_{\sigma(0)}, T_{\sigma(1)}, \ldots, T_{\sigma(s)}) \in A\right).$$

Note, that exchangeability depends on the permutation strategy or statistical model used to obtain the replications of the test statistic under the null hypothesis.

In Papers I-III, global envelope tests were applied to summary functions, e.g. the K function, for goodness-of-fit evaluation of the proposed models. In Papers IV and V, global tests for quantile regression and for comparing the distibutions of n populations were proposed.

4.5 Quantile regression

Quantile regression, a statistical model developed in Koenker and Bassett Jr (1978), models the quantiles of the conditional response variable $Y \in \mathbb{R}^{n\times 1}$

given a set of covariates $\mathbf{X} \in \mathbb{R}^{n \times p}$ linearly on $\mathbf{X}$ as detailed in Equation (4.5)

$$Q_{Y|\mathbf{X}}(\tau) = \inf\{y : F_{Y|\mathbf{X}}(y) \geq \tau\} = \mathbf{X}^T \boldsymbol{\beta}(\tau) \qquad \text{for } \tau \in [0,1]. \tag{4.5}$$

For instance, for the $\tau = 0.5$ quantile this model corresponds to median regression, a model for the conditional median of the response Y given a set of covariates $\mathbf{X}$. Estimating the regression coefficients $\boldsymbol{\beta}(\tau) = (\beta_1(\tau), \dots, \beta_p(\tau))$ involves solving the optimization problem

$$\hat{\boldsymbol{\beta}}(\tau) = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{X}_i^T \boldsymbol{\beta})$$

where ρ_τ: $\mathbb{R} \to \mathbb{R}^+$ such that $\rho_\tau(u) = u\tau$ if $u \geq 0$ and $\rho_\tau(u) = -u(1-\tau)$ if $u < 0$. This optimization problem can be efficiently solved using linear programming methods (Dantzig, 2016; Portnoy and Koenker, 1997).

For a fixed quantile τ, there exist three main approaches for constructing confidence intervals for $\beta_j(\tau)$, $j \in \{1, \dots, p\}$, in the literature. The first approach constructs confidence intervals using asymptotic results (Koenker and Machado, 1999), the second approach, by inverting rank-scores (Gutenbrunner et al., 1993), and the third approach uses resampling methods(Efron, 1992). For a more rigorous description of the methods see Koenker (2005) and references therein.

The aforementioned inference methods are suitable only for local inference, i.e., testing the null hypothesis $H_0 : \beta_j(\tau) = 0$, for fixed τ. On the other hand, simultaneous inference of the quantile regression process, $\boldsymbol{\beta}(\tau)$ for $\tau \in [0,1]$, might often be of interest. In this problem, the following null hypothesis is of interest

$$H_0 : \boldsymbol{\beta}(\tau) = \mathbf{0}, \text{ for all } \tau \in [0,1]. \tag{4.6}$$

In paper IV, global quantile regression, a statistical framework allowing for global inference of the quantile regression process was proposed. The tests presented are global envelope tests (Myllymäki et al., 2017).

5 Permutation strategies

This Section describes the permutation strategies developed to simulate "null data", i.e., data under the null hypothesis, for global testing. For instance, in Paper IV testing the null hypothesis (4.6) was of interest. Throughout this section $\mathbf{Y} \in \mathbb{R}^{n\times 1}$ denotes the response variable, $\mathbf{X} \in \mathbb{R}^{n\times p}$ denotes the interesting covariates, $\mathbf{Z} \in \mathbb{R}^{n\times q}$ denotes the nuisance covariates and $\mathcal{T} = \{\tau_1, \ldots, \tau_d\}$ a discrete set of quantiles.

A simple permutation scheme permutes the rows of a data vector $\mathbf{Y}$ while keeping the covariates fixed. Assuming there exist no nuisance covariates affecting the response distribution, this permutation generates new data so that $\mathbf{X}$ does not affect the distribution of the response variable $\mathbf{Y}$, i.e., data under the null hypothesis (4.6). On the other hand, the statistical power of this permutation test is low when the response distribution is affected by nuisance covariates. For this purpose, more sophisticated permutation strategies were developed in Paper IV.

In particular, we are interested in performing inference for $\beta(\tau)$ in the quantile regression model

$$Q_{\mathbf{Y}|\mathbf{X},\mathbf{Z}}(\tau) = \mathbf{X}\beta(\tau) + \mathbf{Z}\gamma(\tau) \quad \text{for all } \tau \in \mathcal{T}. \tag{5.1}$$

The general framework for constructing the global tests is detailed in Algorithm (3). In this Section, the permutation strategies developed for the global tests (step 2 of Algorithm (3)) are presented.

Algorithm 3 Global inference for quantile regression (5.1) using permutation schemes

Input: Response $\mathbf{Y}$, interesting covariates $\mathbf{X}$, nuisance covariates $\mathbf{Z}$.
Output: A global envelope test.

1. For observed data, compute the test vector

$$\mathbf{T}_0 = (\beta_1(\tau_1), \ldots, \beta_1(\tau_d), \ldots \beta_p(\tau_1), \ldots, \beta_p(\tau_d)) \tag{5.2}$$

   containing all the coefficients of the vectors $\boldsymbol{\beta}(\tau_1), \ldots, \boldsymbol{\beta}(\tau_d)$, rearranged for better visualization.

2. Simulate s replicates of data under the null hypothesis (4.6).

3. Compute the test vectors for the s simulated data, and obtain $\mathbf{T}_1, \ldots, \mathbf{T}_s$.

4. Apply a global envelope test to $\mathbf{T}_0, \mathbf{T}_1, \ldots, \mathbf{T}_s$.

5.1 Within categorical Nuisance

This permutation strategy can be used when $\mathbf{Z}$ contains only one categorical nuisance covariate with K levels. In this setting, new data $\mathbf{Y}^*$ under the null model (4.6) are generated as follows.

1. Split the data into subsets based on the K levels of $\mathbf{Z}$. Let $(\mathbf{Y}^{(k)}, \mathbf{X}^{(k)}, \mathbf{Z}^{(k)})$ be the K subsets, with $k = 1, \ldots, K$.

2. Within each subset k, permute $\mathbf{Y}^{(k)}$ using a simple permutation, while keeping $\mathbf{X}^{(k)}$ and $\mathbf{Z}^{(k)}$ fixed. Then new data $\mathbf{Y}^*$ are obtained by appending the permuted subsets $\mathbf{Y}^{*(k)}$.

5.2 Freedman-Lane based permutations

The main idea of the Freedman-Lane permutation proposed in (Freedman and Lane, 1983) for general linear models, is to generate new data $\mathbf{Y}^*$ by applying a simple permutation on the residuals of the reduced model, i.e., the regression model with only the nuisance covariates. In Paper IV, we adapted the Freedman-Lane permutation for the quantile regression using the following steps.

1. Fit the reduced model $Q_{\mathbf{Y}|\mathbf{Z}}(\tau) = \mathbf{Z}\gamma(\tau)$ for all $\tau \in \mathcal{T}$. Let $\widehat{\gamma}(\tau)$ be the estimated coefficients.
2. Compute the residuals of the reduced model $\epsilon_{\mathbf{Z}}(\tau) = \mathbf{Y} - \mathbf{Z}\widehat{\gamma}(\tau)$ for all $\tau \in \mathcal{T}$.
3. Apply a simple permutation to the residual matrix $\epsilon_{\mathbf{Z}}$. Let $\epsilon_{\mathbf{Z}}{}^*$ be the permuted residual matrix.
4. The new data $\mathbf{Y}^*$ are obtained by $\mathbf{Y}^*(\tau) = \mathbf{Z}\widehat{\gamma}(\tau) + \epsilon_{\mathbf{Z}}{}^*(\tau)$.

Then, the following d quantile regression models for the data $\mathbf{Y}^*(\tau_1), \ldots, \mathbf{Y}^*(\tau_d)$ are considered.

$$Q_{\mathbf{Y}^*(\tau_1)|\mathbf{X},\mathbf{Z}}(\tau_1) = \mathbf{X}\beta(\tau_1) + \mathbf{Z}\gamma(\tau_1), \ldots, Q_{\mathbf{Y}^*(\tau_d)|\mathbf{X},\mathbf{Z}}(\tau_d) = \mathbf{X}\beta(\tau_d) + \mathbf{Z}\gamma(\tau_d).$$

5.3 Removal of nuisance effects

Three of the permutation strategies in Paper IV, construct global tests for $\beta(\tau)$ using Algorithm (3) but considering the model

$$Q_{\epsilon_{\mathbf{Z}}|\mathbf{X}}(\tau) = \mathbf{X}\beta(\tau) \quad \text{for all } \tau \in \mathcal{T} \tag{5.3}$$

instead of the full model (5.1), where $\epsilon_{\mathbf{Z}}$ denotes the data after the estimation and removal of the nuisance effect. Then new data $\epsilon_{\mathbf{Z}}{}^*$ are generated by simple permutations of $\epsilon_{\mathbf{Z}}$. An illustration of this procedure is shown in Figure 5.1.

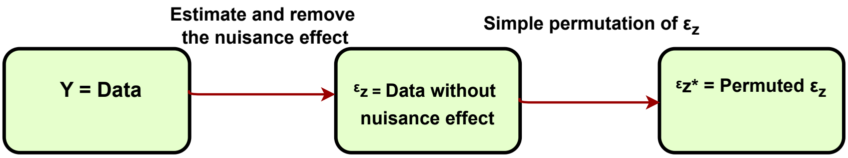
Estimate and remove the nuisance effect
Simple permutation of εz
Y = Data
εz = Data without nuisance effect
εz* = Permuted εz

Figure 5.1: The general framework of the permutations with removal of the nuisance effect.

The first permutation is obtained if a location shift nuisance effect is assumed, i.e. the nuisance covariates affect only the location of the response distribution. The second permutation is obtained if a location and scale shift nuisance effect is assumed, i.e. the nuisance covariates affect both the location and scale of the response distribution, and the third permutation is obtained if a general quantile nuisance effect is assumed.

6 Summary of papers

6.1 Paper I

In this paper, we studied the 3D spatial structure of epidermal nerve fibers by representing each nerve tree by the base point, first branching point and end points. For the analysis, we considered skin samples obtained from the right feet of 32 healthy volunteers and 8 mild diabetic neuropathy subjects. Unlike earlier models that included only the base and end points, we included the first branching points into the analysis. Our analyses indicated that the branching points are naturally better choice for endpoint cluster centers than the base points, and unlike the base points, that are clustered, they are indistinguishable from completely spatially random processes. Moreover, we compared the three-dimensional structure of the end point patterns between the two disease groups using summary functions for point patterns. Even though the planar point patterns of the mild group are more clustered than the point patterns in the healthy group, no significant difference in clustering was found between the three-dimensional point patterns of the two groups in terms of second-order summary statistics.

To study the tree structure within the individual nerve trees, we considered the branch lengths and angles of the first segments, the tree segment connecting the base to the branching points, and the later segments, the tree segments connecting the branching points to the end points. We used shift plots to compare the branch lengths and angular distributions of the two tree segments. Our statistical analysis suggested that the first segment length is significantly larger and the first segments grow more vertically than the later segments. Further, we compared the tree structure between the groups. Our results indicated that there are significant differences between the angular distributions of the later segments of the two groups.

Moreover, we extended the concept of reactive territories, in our paper called epidermal active territories (EAT), introduced in Andersson et al. (2016) for 2D point patterns, to 3D point patterns. The epidermal active territory is defined as the volume of the area in the epidermis covered by individual nerve fibers. Our results showed that the total volume of the epidermis covered by the nerve trees was larger in the healthy group than in the mild group. In addition, possible competitive behavior between individual nerves was examined by using the mark correlation function of the base point process with the epidermal active territories as marks. No mark correlation was detected between the marks.

Finally, we proposed a two-step point process model for the end points conditioned on the base point patterns. In the first step, we conditioned on the base points and sent the first branching points towards open space, as in the NOC model in Olsbo et al. (2013). In the second step, the endpoint clusters are constructed around the simulated branching points. The two-dimensional version of the model fitted the data quite well, while the three-dimensional version revealed that there are interactions between the endpoints that were not captured by the model. After the paper had published, we noticed that the shift plots presented in the Paper are misinterpreted, as we ignored the cumulative nature of the plots.

6.2 Paper II

Inspired by Christoffersen et al. (2021) we developed a 3D point process model, that allowed the end points to interact with each other. The model consisted of two steps. In the first step, the planar point patterns were obtained using the two-dimensional version of the model introduced in Paper I. In the second step, the process in the z-direction X_z given the planar process X_p was constructed using a pairwise interaction Markov random field model. In the model, two points are considered neighbors if they lie within a cylindrical interaction region, i.e. a cylinder with halfwidth w and height $2t$ centered in one of the points. The conditional density consisted of two parts, one modelling the cylindrical interaction (γ) between the endpoints and a hardcore ball of radius h not allowing points to be closer than the points in the data. The parameters of the model, $\theta = (h, w, t, \gamma)$, were estimated by maximizing the pseudolikelihood over a grid of values for the cylinder parameters (w, t) using the minimum distance between the endpoints in the data multiplied by $\frac{n-1}{n}$ as an estimate for the hardcore distance h. To reduce the bias due to edge effects, minus sampling was used. The parameter estimates suggest that in both groups, after a hardcore radius h the end points attract each other. The attraction is larger in the healthy group than in the mild diabetic group.

Furthermore, a Markov chain Monte Carlo algorithm, where the number of points in the planar process X_p are fixed, was used to simulate from the model. We used a systematic updating scheme cycling over the point indexes 1 to n and using a uniform proposal for a new point in W_z. Due to the anisotropic nature of the data, the goodness-of-fit of the model was evaluated using the cylindrical K function. Simulations from the model were able to capture both the complete spatial structure of the endpoints and the structure of the endpoints with respect to their branching points. After the publication of the paper, we noticed that the reviewer's suggestion of changing Figure 3 from a boxplot (see Figure 6.1) to a dotplot made the interpretation of the results harder.

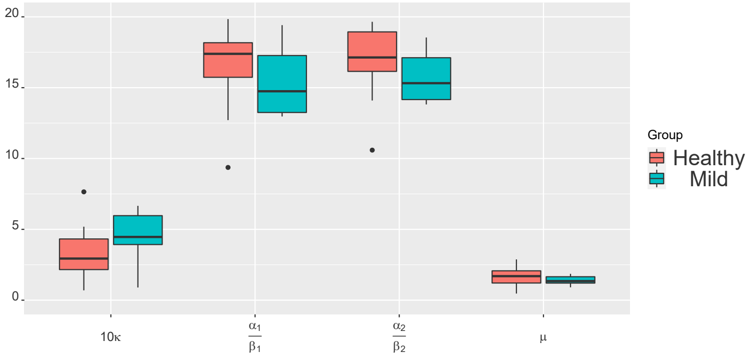

Figure 6.1: Original boxplot with the parameter estimates in Paper II.

6.3 Paper III

In this paper, we investigated the dynamics of the nerve mortality process caused by the progression of the severity of the neuropathy. For the statistical analyses, we considered bivariate planar point patterns consisting of the locations of base and end points of ENFs from diabetic patients diagnosed with mild and moderate diabetic neuropathy as well as healthy controls. To study the nerve death process, we developed spatial thinning models for the base point patterns, where whole nerve trees, i.e. base points and the end points connected to them, are removed according to a probabilistic model.

Initially, we tested the hypothesis of random nerve mortality, i.e., there exists no particular biological process guiding the nerve removal and hence the nerves are removed completely at random and independently of the removal of other points. In this case, the null model corresponds to an independent p-thinning model, a model in which each nerve tree is retained with constant probability p. To obtain simulated mild diabetic ENF data, this model was applied to the healthy ENF data with retention probability estimated as the ratio between the corresponding mean mild group and mean healthy group base point intensities. Goodness-of-fit assessment of the model was performed using global envelope tests with the centered Ripley's L function as the test statistic. Our findings indicated that the simulated mild diabetic patterns, i.e., the thinned healthy patterns, did not capture the spatial structure in the mild diabetic ENF patterns.

To further investigate the behavior of the nerve mortality process, we developed a parametric dependent thinning model in which the retention probability for each nerve tree depended on the distance to the closest other nerve tree. In particular, the retention probability was larger for nerve trees with close neighboring trees and lower for isolated trees. The parameter of the model controlled the behavior of those retention probabilities. More specifically, it favored the removal of isolated trees for small values of the parameter, and for large parameter values, the model converged to an independent p-thinning. The model parameter was estimated by an approximate Bayesian computation method. For the ABC method, we used an exponential prior and chose as the summary statistic the scale for which the estimated empty space function for the base point configurations evaluates to 30%. The quality of the proposed inference procedure was assessed in a simulation study. According to our results, the 95% predictive envelopes for several spatial and non spatial statistics obtained from simulations from the posterior predictive distribution of the model captured the structure in the mild ENF data well. Therefore, the increased ENF clustering in the diabetic patterns can be explained by the model, as the isolated nerve trees tend to die first. Finally, we investigated thinning models applied to mild ENF data. We observed that an independent p-thinning of the base points, and hence dependent thinning of endpoints, was sufficient to capture the structure in the ENF data from the moderate diabetic neuropathy group. On the other hand, the sample size for this analysis was small and therefore further studies are required.

6.4 Paper IV

In this paper, we extended the global envelope tests used for goodness-of-fit purposes, e.g. in Papers I-III, to perform simultaneous inference for quantile regression (Koenker and Bassett Jr, 1978) . To this end, we developed global quantile regression, a statistical framework suitable for testing whether a covariate has an effect on any set of quantiles of the response distribution. The proposed method is based on the quantile regression model and the global envelope tests (Myllymäki et al., 2017) with permutations to simulate data under the "null model", that is the quantile regression process $\boldsymbol{\beta}(\tau) \equiv 0$ for $\tau \in [0, 1]$. To the best of our knowledge, this is the first method in the literature suitable for simultaneous inference of the quantile regression process that requires no regularity assumptions to be valid.

Initially, we considered the case in which there are no nuisance covariates affecting the response distribution. In this case, a simple permutation strategy permuting the response variable is sufficient to construct an exact global test. On the other hand, when nuisance covariates are present simple permutation breaks the dependence between the interesting covariates $\mathbf{X}$ and the nuisance covariates $\mathbf{Z}$, and hence simple permutation is no longer valid. For this purpose, we proposed five different permutation strategies and investigated their performance under different settings in a simulation study. When the nuisance covariate is categorical, permuting the response variable within the levels of the nuisance provides a permutation scheme that satisfies exchangeability and hence is exact.

When both continuous and categorical nuisance covariates are present, we investigated the performance of Freedman-Lane (FL) (Freedman and Lane, 1983) based permutations as this permutation is considered the best for generalized linear models. Our findings indicated that FL-based permutations are liberal in the presence of significant nuisance effects when extreme quantiles are considered in the global test. Therefore, we proposed three permutation strategies that are close to exact even when extreme quantiles are considered. Two of the permutations assume a specific nuisance effect, i.e. location or location-scale shift of the response distribution, while the third permutation estimates the nuisance quantile effect from the data. Even though the latter is always valid, it has lower power for small datasets as the quantile effect is badly estimated.

Furthermore, we investigated the validity of the methods when the nuisance and interesting covariates are correlated. Our findings suggested that the global test is liberal for a permutation under model misspecification and conservative otherwise. Finally, one should point out that the choice of the permutation strategy should follow the proposed guidelines, and therefore an initial exploratory analysis is suggested. In the paper, we illustrate this procedure through two data examples from forestry and economics. Last but not least, in the special case where the interesting covariate is categorical the global test corresponds to a graphical n-sample correspondence of the distribution test even when nuisance covariates are present. This problem was further investigated in Paper V.

6.5 Paper V

In this paper, we extended the global envelopes to test the equality hypothesis of distributions of n samples. The proposed tests are permutation tests and assume that the underlying distributions are not affected by any nuisance covariates. Therefore, new data under the null hypothesis of equality of the n distributions are simulated using data permutations between the n samples.

For testing, we used five test statistics capturing different distributional contrasts. Firstly, we considered test statistics in the form of empirical cumulative distribution functions and kernel-estimated density functions of the distributions of the samples. Secondly, we proposed test statistics expressing pairwise deviations between the empirical cumulative distribution functions and between the empirical quantiles of two samples. Lastly, we applied the global quantile regression framework developed in Paper IV. We considered a quantile regression model with a categorical interesting covariate and the quantile regression process as our test statistic. Furthermore, we proposed global tests based on combinations of the previously mentioned statistics.

The proposed tests are graphical as they are based on the global envelope testing framework. Therefore, in addition to a p-value of the test results, they also provide a graphical illustration of the reason for rejecting the null hypothesis. As the graphical interpretation is dependent on the test statistic used, we provided detailed guidelines on how to interpret the test results correctly.

The performance of the proposed tests with regard to statistical power was evaluated through a simulation study. The suggested tests were further compared with the performance of the classical two-sample Kolmogorov-Smirnov test, a graphical test for comparing the distributions of two samples. The performance of the tests was investigated in five simulated experiments. In each experiment, we considered a different distributional difference between the two samples. According to our results, the proposed tests outperformed the classical test in all studied settings. Further, the combined tests,i.e. test combining different test statistics, performed quite well in all settings, even though they were outperformed by some other test in each case. However, since distributional differences are usually unknown apriori, we recommended using combined tests as they provide balanced performance and rich graphical interpretation. Finally, we applied the recommended test to two data examples, one from ecology and one from auxology.

7 Conclusion and future work

This thesis contributes to a better understanding of the biological mechanisms guiding changes in the nerve structure due to the progression of diabetic neuropathy. In contrast to the planar point process models for the base and end points in the literature, in this thesis the first three-dimensional point process models for the complete nerve structure were developed. Moreover, a possible description of how the healthy ENF trees die due to neuropathy progression, as well as an explanation for the increased clustering observed in the diabetic patterns was given for the first time. Finally, even though the methods and models in this thesis had been developed having the ENF data in mind, they can be used for point patterns in general.

This thesis further contributes to the general statistical literature as extensions of the global envelope tests for inference in quantile regression and for comparison of n distributions were developed. To the best of our knowledge, the former extension was the first test that could perform global inference of the quantile regression process without making any regularity assumptions. The framework is generic and can be used to study the quantile effects of a covariate on the response distribution at different sets of quantiles or for comparing the distributions of n populations. The latter test was further investigated and graphical n sample tests were developed.

There are several potential paths for future research. For instance, extending the proposed three-dimensional models by considering interactions between the different types of points or by considering approximating the open space direction using multiple neighbouring base points. Another future direction is training machine learning models to diagnose the severity of the neuropathy. For instance, training a convolutional neural network model for the original microscopy images, or using a random forest with nerve information such as counts, lengths, or cluster radius of the ENFs. On the other hand, such models require a larger ENF dataset. Therefore, if data collection is expensive, one may simulate artificial data from the models described in this thesis. Another future path is further developing the permutation strategies for the global quantile regression test to study interactions or random effects. For instance, the latter will allow the test to be used for the ENF data. In this setting, potential nuisance variables to be considered are age, gender, and BMI. The dependencies within the ENF samples, may also be reduced by either considering the average nerve lengths of each ENF sample, or by using a sophisticated sampling strategy.

Bibliography

Andersson, C., Guttorp, P., and Särkkä, A. (2016). Discovering early diabetic neuropathy from epidermal nerve fiber patterns. *Statistics in medicine*, 35(24):4427–4442.

Andersson, C. and Mrkvička, T. (2020). Inference for cluster point processes with over-or under-dispersed cluster sizes. *Statistics and Computing*, 30(6):1573–1590.

Andersson, C., Rajala, T., and Särkkä, A. (2019). A bayesian hierarchical point process model for epidermal nerve fiber patterns. *Mathematical biosciences*, 313:48–60.

Baddeley, A., Moyeed, R., Howard, C., and Boyde, A. (1993). Analysis of a three-dimensional point pattern with replication. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 42(4):641–668.

Bedford, T. and Van den Berg, J. (1997). A remark on the van lieshout and baddeley j-function for point processes. *Advances in Applied Probability*, 29(1):19–25.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC press.

Chiu, S. N., Stoyan, D., Kendall, W. S., and Mecke, J. (2013). *Stochastic geometry and its applications*. John Wiley & Sons.

Christoffersen, A. D., Møller, J., and Christensen, H. S. (2021). Modelling columnarity of pyramidal cells in the human cerebral cortex. *Australian & New Zealand Journal of Statistics*.

Cronie, O., Julia, J., and Konstantinos, K. (2024). Discussion of the Paper "Marked Spatial Point Processes: Current state and Extensions to Point Processes on Linear Networks". *Journal of Agricultural, Biological and Environmental Statistics*.

Dantzig, G. (2016). *Linear Programming and Extensions*, volume 48. Princeton University Press.

Diggle, P. J., Lange, N., and Beneš, F. M. (1991). Analysis of variance for replicated spatial point patterns in clinical neuroanatomy. *Journal of the American Statistical Association*, 86(415):618–625.

Diggle, P. J., Mateu, J., and Clough, H. E. (2000). A comparison between parametric and non-parametric approaches to the analysis of replicated spatial point patterns. *Advances in Applied Probability*, 32(2):331–343.

Doksum, K. A. and Sievers, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika*, 63(3):421–434.

Drovandi, C. C. and Pettitt, A. N. (2011). Estimation of parameters for macroparasite population evolution using approximate bayesian computation. *Biometrics*, 67(1):225–233.

D'Angelo, N., Adelfio, G., Mateu, J., and Cronie, O. (2023). Local inhomogeneous weighted summary statistics for marked point processes. *Journal of Computational and Graphical Statistics*, pages 1–15.

Eckardt, M. and Moradi, M. (2024). Marked Spatial Point Processes: Current state and Extensions to Point Processes on Linear Networks. *Journal of Agricultural, Biological and Environmental Statistics*, pages 1–33.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer.

Freedman, D. and Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298.

Garcia, N. L., Guttorp, P., and Ludwig, G. (2020). Interacting cluster point process model for epidermal nerve fibers. *Spatial Statistics*, 35:100414.

Ghorbanpour, F., Särkkä, A., and Pourtaheri, R. (2021). Marked point process analysis of epidermal nerve fibres. *Journal of Microscopy*.

Gutenbrunner, C., Jurečková, J., Koenker, R., and Portnoy, S. (1993). Tests of linear hypotheses based on regression rank scores. *Journaltitle of Nonparametric Statistics*, 2(4):307–331.

Hahn, U. et al. (2015). A note on simultaneous monte carlo tests. *CSGB Research Reports, Department of Mathematics, Aarhus University*.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. John Wiley & Sons.

Jiang, B., Wu, T.-y., Zheng, C., and Wong, W. H. (2017). Learning summary statistic for approximate bayesian computation via deep neural network. *Statistica Sinica*, pages 1595–1618.

Kennedy, W. R., Nolano, M., Wendelschafer-Crabb, G., Johnson, T. L., and Tamura, E. (1999). A skin blister method to study epidermal nerves in peripheral nerve disease. *Muscle & Nerve: Official Journal of the American Association of Electrodiagnostic Medicine*, 22(3):360–371.

Kennedy, W. R. and Wendelschafer-Crabb, G. (1993). The innervation of human epidermis. *Journal of the neurological sciences*, 115(2):184–190.

Kennedy, W. R., Wendelschafer-Crabb, G., and Johnson, T. (1996). Quantitation of epidermal nerves in diabetic neuropathy. *Neurology*, 47(4):1042–1048.

Koenker, R. (2005). *Quantile Regression*. Cambridge U. Press.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Koenker, R. and Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association*, 94(448):1296–1310.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Møller, J., Safavimanesh, F., and Rasmussen, J. G. (2016). The cylindrical-function and poisson line cluster point processes. *Biometrika*, 103(4):937–954.

Møller, J. and Waagepetersen, R. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Boca Raton: CRC Press.

Mrkvička, T., Myllymäki, M., Kuronen, M., and Narisetty, N. N. (2022). New methods for multiple testing in permutation inference for the general linear model. *Statistics in Medicine*, 41(2):276–297.

Myllymäki, M. and Mrkvička, T. (2019). Get: Global envelopes in r. *arXiv preprint arXiv:1911.06583*.

Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H., and Hahn, U. (2017). Global envelope tests for spatial processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):381–404.

Myllymäki, M., Panoutsopoulou, I., and Särkkä, A. (2012). Analysis of spatial structure of epidermal nerve entry point patterns based on replicated data. *Journal of microscopy*, 247(3):228–239.

Narisetty, N. N. and Nair, V. N. (2016). Extremal depth for functional data and applications. *Journal of the American Statistical Association*, 111(516):1705–1714.

Neyman, J. and Scott, E. (1952). A theory of the spatial distribution of galaxies. *The Astrophysical Journal*, 116:144.

Olsbo, V., Myllymäki, M., Waller, L. A., and Särkkä, A. (2013). Development and evaluation of spatial point process models for epidermal nerve fibers. *Mathematical biosciences*, 243(2):178–189.

Portnoy, S. and Koenker, R. (1997). The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 12(4):279–300.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798.

Rajala, T., Redenbach, C., Särkkä, A., and Sormani, M. (2018). A review on anisotropy analysis of spatial point patterns. *Spatial Statistics*, 28:141–168.

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):172–192.

Ripley, B. D. (1981). *Spatial statistics*. Wiley, New York.

Simola, U., Cisewski-Kehe, J., Gutmann, M. U., and Corander, J. (2021). Adaptive approximate bayesian computation tolerance selection. *Bayesian analysis*, 16(2):397–423.

Sisson, S. A., Fan, Y., and Beaumont, M. (2018). *Handbook of approximate Bayesian computation*. CRC Press.

Stoyan, D. (1984). On correlations of marked point processes. *Mathematische Nachrichten*, 116(1):197–207.

Stoyan, D. and Stoyan, H. (1994). *Fractals, random shapes and point fields: methods of geometrical statistics*, volume 302. Wiley-Blackwell.

Wiqvist, S., Mattei, P.-A., Picchini, U., and Frellsen, J. (2019). Partially exchangeable networks and architectures for learning summary statistics in approximate bayesian computation. In *International Conference on Machine Learning*, pages 6798–6807. PMLR.