

Hotspot detection and a nonstationary process variance function estimation

by

Eunice Jungeun Kim

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:

Zhengyuan Zhu, Major Professor

Petrutza Caragea

Philip M. Dixon

Mark S. Kaiser

Daniel J. Nordman

Iowa State University

Ames, Iowa

2013

Copyright © Eunice Jungeun Kim, 2013. All rights reserved.

TABLE OF CONTENTS

LIST OF TABLES	v
LIST OF FIGURES	vi
ACKNOWLEDGEMENTS	viii
CHAPTER 1. OVERVIEW	1
1.1 Motivation	1
1.2 Structure	2
1.3 Abstracts	3
CHAPTER 2. A TWO-STAGE SPATIAL SAMPLING DESIGN FOR HOTSPOT DETECTION	5
2.1 Introduction	5
2.2 Objectives, assumptions, and notations	7
2.3 Sampling designs maximizing detection probability	13
2.3.1 One-stage designs	14
2.3.2 Detection probabilities of one-stage designs	15
2.3.3 Two-stage sampling to maximize detection probability	20
2.3.4 Theoretical properties of two-stage sampling	23
2.4 Numerical results	25
2.5 Beryllium clean-up study at Ames Laboratory	28
2.6 Conclusion	33

CHAPTER 3. DIFFERENCE-BASED VARIANCE FUNCTION ESTIMATION OF A ONE-DIMENSIONAL NONSTATIONARY PROCESS	34
3.1 Introduction	34
3.1.1 Motivation	34
3.1.2 Literature Review	35
3.2 Model and Definition	37
3.3 Theoretical Results	41
3.3.1 Local variogram estimator	41
3.3.2 Bias of the estimator	45
3.3.3 Variance of the estimator	47
3.3.4 Risk of Local Variogram Estimator	50
3.4 Algorithm and Bandwidth Selection	53
3.5 Simulation Study	55
3.5.1 Set-up	55
3.5.2 Discussion of Results	56
3.6 Discussion	66
CHAPTER 4. VARIANCE FUNCTION ESTIMATION OF TWO-DIMENSIONAL NONSTATIONARY PROCESS	68
4.1 Introduction	68
4.2 Data Model and Method	70
4.2.1 Notations and Definitions	71
4.2.2 Method	74
4.3 Properties of Difference Filter	79
4.3.1 Configuration of Difference Filter	80
4.3.2 Determining Weights	81
4.3.3 L -filter variogram	83

4.4	Simulation Study	85
4.4.1	Data Model and Measures of Estimation	86
4.4.2	Results	89
4.5	Discussion	98
CHAPTER 5. SUMMARY AND DISCUSSION		100
APPENDIX A. ADDITIONAL MATERIAL		102
A.1	Derivations	102
A.1.1	Proof of Equation (2.3) in Chapter 2	102
A.1.2	Derivations for Chapter 3	103
A.2	Filter Weights	104
A.2.1	Simple Differencing: Symmetric Weight	104
A.2.2	Variance Minimization under Independent and Identically Distributed Errors: Hall-Kay-Titterington Weight	106
BIBLIOGRAPHY		108

LIST OF TABLES

Table 2.1	<i>Floor-by-Floor</i> Detection Probabilities from a Case Study	32
Table 2.2	Overall <i>Floor</i> Detection Probabilities from avCase Study	32
Table 3.1	A Sine Standard Deviation Function Estimation Summary . . .	63
Table 3.2	A Quadratic Standard Deviation Function Estimation Summary	64
Table 3.3	Bandwidth Selection Results	65
Table 4.1	Number of Fourth Order Terms in $(D_i^2)^2$	77
Table 4.2	Number of Fourth Order Terms in $D_i^2 D_j^2$ with One Node Overlap	78
Table 4.3	Number of Fourth Order Terms in $D_i^2 D_j^2$ with Two Nodes Overlap	79
Table 4.4	L -Filter Variogram Values for Symmetric and HKT Weight Filters	84
Table 4.5	Exponential Variogram Functional Values	84
Table 4.6	Comparing Five Symmetric Weight Filters via Discretely Inte- grated MSE , Median Absolute Deviation, and Maximum Abso- lute Deviation	93
Table 4.7	Line versus Y Configuration Estimation Results	94
Table 4.8	Nearest Neighbor Simple Differencing versus Second-Nearest Neigh- bor Simple Differencing	95

LIST OF FIGURES

Figure 2.1	Two-Stage Systematic Sampling Diagram	9
Figure 2.2	Examples of Three One Dimensional Spatial Sampling Designs .	15
Figure 2.3	Comparing Detection Probability of Systematic, Markov Chain Design, One-per-Stratum Design, and Simple Random Sampling	19
Figure 2.4	Optimal First-Stage Sampling Floor Proportion τ versus Sample Proportion α	24
Figure 2.5	Expected <i>Floor</i> Detection Count Profiles and the Optimal Sample Splitting Proportion α^* for Different Proportions of <i>Floor</i> Con- tamination c 's	27
Figure 2.6	Spedding Hall Door-Top Beryllium Census in Log Scale.	29
Figure 2.7	Wilhelm Hall Door-Top Beryllium Census in Log Scale.	30
Figure 3.1	Step-wise Standard Deviation Function Estimation	58
Figure 3.2	Sinusoidal Standard Deviation Function Estimation	59
Figure 3.3	Difference-Based versus Likelihood-Based Estimation Summary Using Discretely Integrated <i>MSE</i> and Maximum Absolute Devi- ation	60
Figure 4.1	Bias in Estimation Using a Symmetric Weight Filter	76
Figure 4.2	Filter Configurations	80
Figure 4.3	Three Model Functions of $\sigma(s_x, s_y)$	87
Figure 4.4	Examples of Nonstationary Data	87

Figure 4.5	Estimation Results of Symmetric Weight Filters versus Hall-Kay-Titterington Weight Filters	90
Figure 4.6	Side-by-Side Boxplots for the Comparison of Weighting Schemes	91
Figure 4.7	A Comparison of Symmetric Weight Filter Configurations	95
Figure 4.8	A Study of Line Filter Scale Effect on Estimation Depended on Data Dependency	96
Figure 4.9	A Study of Line Filter Scale Effect on Estimation Depended on Data Dependency and Grid Size	96

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Dr. Zhengyuan Zhu for guiding me in various aspects of conducting research and the writing of this dissertation. One of the most valuable lessons I have learned is that often writing down what I have and organizing it once in a while helps to make a progress. For those who are procrastinating in writing, do start now. I would also like to thank my committee members and co-workers for their efforts and contributions to this work: Dr. Dixon, Dr. Kaiser, Dr. Caragea, Dr. Nordman, Tom Wessels, Jim Withers, and the rest of Ames Laboratory TASF G-40 team. I would additionally like to thank Dr. Fienberg for his guidance throughout the initial stages of my graduate career and Dr. Kass for his inspirational teaching of how to be a good teacher by his walk. I am also grateful for the intellectually challenging and warmly inviting times I have shared with the teachers (e.g. Dr. Steele) and colleagues who have walked the paths alongside of me. Thank you, all!

Aileen, Kevin, Judy, Daniel M., Don B., Minsun, Youngmin, Dr. Tokdar, and Frances, your insights and words of encouragement have often inspired me and renewed my hopes for completing my graduate education. My parents have asked me a few times to quit the PhD program so that I would not agonize over any unsolved problems. I am very glad they spoke such words to me, so that I could once again go against their words and complete my mission. I bet those words helped me to muster up courage and prove myself to you, mom, dad, and E. I am indebted to your silent cheer and the last minute words of encouragement. Lastly, my husband, Ildoo, has also shown immeasurable patience and support while I was in residence in Ames. Thank you, God, for the opportunity to stay in school extra long and bringing him into my life.

CHAPTER 1. OVERVIEW

The topics discussed in this dissertation are related to correlated data analysis. Throughout the dissertation, I assume that the data are smoothly varying with negligible measurement errors. I consider the cases of transect data for simplicity and extend the idea to lattice data, which may be a post-processed form of geospatial data that is often encountered. The focus is on defining the properties of our statistical approaches and comparing against alternative approaches.

1.1 Motivation

Chapter 2 of this dissertation is concerned with proposing a cost-effective method for spatial sampling for hotspot detection. Due to the correlated nature of the spatial data, spatial sampling would benefit from a stratification of the sampling domain. A sequential sampling design over a stratified domain is proposed which not only gives a higher detection probability than a one-stage design but also provides an economical strategy to implement a sampling design.

Chapter 3 and 4 is concerned with estimating variance function from a nonstationary spatial process. Nonstationarity is a frequently encountered feature of spatial data. When one needs to predict or estimate a possible range of values for a particular location, variance estimation at the location is necessary to provide a prediction interval or a confidence interval. Differencing nonstationary random field for variance function estimation reduces the estimation bias by bypassing the mean function estimation, and

this nonparametric method provides a flexible application and simple implementation.

1.2 Structure

Here is an overview of how the discussion progresses in the next four chapters. In Chapter 2 we compare the detection probabilities of four one-stage sampling plans. Then a two-stage systematic sampling design is proposed, which consistently detects problematic areas with higher probability than any one-stage design of equivalent sample size. For a two-stage design, a sample splitting proportion parameter is dependent on a hotspot dispersal scenario. We determine the optimal value of the parameter via a simulation study and apply the proposed design to a case study.

In Chapter 3 a difference-based nonparametric variance function estimator is proposed. First ‘local variogram’ is defined for a nonstationary process by assuming local stationarity. The local variogram possesses the same idea as a variogram in a stationary process but with the variance scale factor multiplied locally. We derive the basic properties of the local variogram estimator and its asymptotic rate of risk. We contrast the difference-based nonparametric estimation to Anderes and Stein (2011) local-likelihood-based estimation through a simulation study.

The estimator is extended from being applied to one-dimension in Chapter 3 to a two-dimensional setting in Chapter 4. As the number of dimensions increases, the number of directions grows, and there are many more choices of directions, scales, and weight options for the differencing. Because we consider square lattice data, the directions for the differencing filter from one to two dimensions would increase twofold. Adding to the complexities is the configuration of the differencing filters in the high-dimensional support of data. In one dimension a line is the only configuration. On a two-dimensional plane, the configuration starts to bear a significant meaning and provides a wide variety of choices. Consequently weights assigned to each point of a filter configuration are another

attribute of the filter that needs to be determined. After exploring the components of the filter for the spatially correlated data variance function estimation in two dimensions, in Chapter 4, we detail the simulation study and suggest specific difference filters for a nonparametric variance function estimation. The statistically efficient averaging idea also applies to this local smoothing approach, as the larger the number of data points to consider, the more precision we have of the estimator. However, the extent of gathering multiplicity should be balanced with the size of the neighborhood.

In Chapter 5, I briefly review the materials in the three main chapters and conclude.

1.3 Abstracts

Chapter 2 Abstract

A two-stage spatial sampling design for detecting contaminated areas is proposed for effective decontamination planning. A two-stage design has a higher or equal hotspot detection probability than a one-stage design under fixed budget constraints. The proposed design uses the expected relative size of the contaminated area and the overall sampling rate as the two control variables in determining an optimal sample splitting proportion for a two-stage design. Results are shown through simulation studies and theoretical derivation.

Chapter 3 Abstract

Many spatial processes exhibit nonstationary features. We estimate a variance function from a single process observation where the errors are nonstationary and correlated. We assume that the mean process is smooth and that the error process is a product of a smooth variance function and a second-order stationary process. A difference-based approach for a one-dimensional nonstationary process is developed along with a bandwidth selection method which takes into account the error dependence structure. The

asymptotic properties of the estimator are investigated, and the estimation results are compared to that of a local-likelihood approach proposed by Anderes and Stein (2011). Simulation study shows that our method has a smaller integrated MSE, fixes the boundary bias problem, and requires far less computing time as the evaluation of likelihood with matrix inversion is not necessary.

Chapter 4 Abstract

A difference-based variance function estimation is developed for a two-dimensional nonstationary process with correlated errors. There are a few practical guides for selecting a difference filter of its shape, scale, and weight depending on the degree of correlation in the data. When the data is strongly correlated, a symmetric weighting scheme is preferred; and when the data is weakly correlated or independent, the Hall-Kay-Titterington weight is preferred. A few practical guides for a two-dimensional linear filters in this chapter should be easily adopted in practice.

CHAPTER 2. A TWO-STAGE SPATIAL SAMPLING DESIGN FOR HOTSPOT DETECTION

2.1 Introduction

We are interested in designing a sampling plan to detect remedial units containing contaminant hotspots. We assume that contaminated hotspots are spatially clustered and that remediation is performed over a neighborhood of hotspots to reduce the risk of even low levels of exposure. In a building, for example, a contamination remediation unit may be a room, floor, or a section of a floor, and often a sampling unit is smaller than a remediation unit. Our goal is to maximize the detection probability of a remediation unit under the constraint of a fixed budget. Hence, arranging sampling units into an equal-sized remediation unit, which is equivalent to a contamination classification unit, is the first step in implementing our proposed sampling plan.

In the field of industrial hygiene ‘sampling’ refers to collecting contaminants for analysis, while in the field of statistics ‘sampling’ refers to selecting a subset of a population to make statistical inference on the extent of contaminant dispersion. In both communities, formulating an economical and efficient sampling strategy is important especially in determining the extent of contamination in a given area. Singer (1972, 1975) provides a Fortran program for computing detection probability of elliptically shaped hotspots using square, rectangular, and hexagonal grid sampling. Parkhurst (1984) uses Singer (1972) and demonstrates that sampling on a triangular grid gives better coverage than on a square grid and results in 23% fewer sampling sites when fixing the maximum distance

between the sampling sites and the location of a potential hotspot on both grids. He notes that when the hotspots are regularly dispersed, one-per-stratum random sampling has a more consistent hotspot detection probability than regular sampling, and in such a case random sampling on a square grid should be easier to implement than on a triangular grid. Gilbert (1982) and Zirschky and Gilbert (1984) examine the grid spacing issue, the detection probability of an elliptically shaped hotspot, the Type II error of a grid sampling plan, and the detection probability of multiple hotspots. When there is insufficient information on the shape of a hotspot, the spacing between grid points should be finer. Otherwise, the detection risk can be calculated, *a priori* compounded with a statistical distribution, based on the sample size, the sampling grid, and the ratio of a hotspot major semi-axis to grid spacing. Gilbert summarizes his previous work on grid sampling for hotspot detection in Chapter 10 of his book *Statistical Methods for Environmental Pollution* by Gilbert (1987).

In spatial sampling literature, it is known that simple random sampling is not very efficient for spatial sampling. Breidt (1995) has introduced the Markov chain design as a general spatial sampling framework that contains a few design parameters to make sampling locations systematically dispersed with added randomness. As this design is a compromise between a systematic sampling plan and a one-per-stratum spatial sampling design, its detection probability of spatial clusters is slightly less than a systematic sampling plan but greater than a one-per-stratum design. Thompson (1990) has introduced adaptive cluster sampling designs to estimate the total population of rare and clustered spatial phenomena. In the first stage, one takes a simple random sample of sampling sites. In the second stage, the first-stage measurements are used to identify areas of interest to sample further. Christman (2003) combines the work of Thompson (1990) and Breidt (1995) and proposes an adaptive two-stage one-per-stratum sampling of rare, dispersed populations. She proves that there is an increased efficiency of estimation in two-stage sampling over one-stage systematic sampling when the same sample size is

used.

We propose a two-stage systematic sampling plan that maximizes the detection probability of a remediation unit with hotspots. We refer to a remediation unit as a *floor*. In first-stage sampling, we sample a fixed proportion of all *floors*. In second-stage, we sample the remaining part of the *floors* that do not have any hotspots detected in the first-stage sampling. This strategy requires dividing every *floor* and the corresponding sampling resource into two parts to perform an adaptive sampling design.

In Section 2.2, I state the sampling objective and assumption, describe the sampling plan, and expound on the assumptions for a data model. In Section 2.3, I consider an optimal design that maximizes detection probability. In Section 2.3.1 several one-stage sampling designs are reviewed, and in Section 2.3.2 I compare their detection probabilities and show that a systematic design gives the highest detection probability. In Section 2.3.3 I describe the procedure of a two-stage systematic sampling plan. In Section 2.3.4 I prove the effectiveness of a two-stage systematic design over any one-stage design. In Section 2.4 a simulation study gives an optimal set of two-stage design parameters. In Section 2.5 I verify the effectiveness of a two-stage design using a beryllium decontamination case study, which was conducted at the Ames Laboratory in 2010-2011 under the supervision of Tom E. Wessels and James H. Withers. In Section 2.6 we conclude with some remarks.

2.2 Objectives, assumptions, and notations

Industrial hygienists and statisticians alike are interested in developing a sampling plan that identifies the locations of hotspots (exposure sites) with high sensitivity given a fixed budget. To minimize the exposure risk, we keep the unit of remediation larger than a sampling unit, for example, as an integer multiple of a sampling unit. Let *room* represent a sampling unit and *floor* represent a unit of remediation. In this new language, the objective is to detect as many contaminated *floors* as possible so as to decontaminate

and reduce harmful element exposure. Define the contamination of a *floor* as having at least one sampling unit whose measurement exceeds a threshold, and the detection of a *floor* as detecting at least one of those sampling units. To detect and declare a *floor* contaminated, we need to detect one contaminated *room* in a given *floor*.

The following assumptions are made to simplify the presentation of our method.

Assumption 1. Contamination exposure sites are of equal and fixed size clusters.

Assumption 2. A sampling domain is a transect.

Assumption 3. Contamination exposure sites are independent across *floors* and the probability of *floor* contamination is fixed.

Assumption 4. There is no uncertainty in declaring a site or *room* contaminated.

Assumption 5. There is at most one hotspot per *floor*.

Assumption 6. There are the same number of *rooms* per *floor*.

In Assumption 1, a clustered arrangement is a realistic description of the contamination process. The assumption for equal and fixed size clusters is not but simplifies the theoretical derivation. Assumption 2 can be justified for cases where the observations in one direction has a strong spatial correlation while in the orthogonal direction they are weakly correlated. In such a scenario, we shall display the data as a transect. When the observations exhibit strong spatial dependency in all directions, one should use a two-dimensional grid. Here we present a scenario of one-dimensional sampling plan. In a two-dimensional spatial design, a more complicated calculation of a hotspot detection probability is required based on the assumptions of the shape of a hotspot. Assumption 3 describes remediation units, *floors*, as containing physically independent characteristics for *floor* contamination probability. Assumption 4 describes a case where observations have strong signals and small measurement errors, i.e. a small coefficient of variation.

Since the observations are strongly correlated and the hotspots are clustered, the sampling sensitivity is not affected by a small measurement error. Assumption 5 and 6 are, again, for the simplification of detection probability calculation. These assumptions allow us to explicitly derive theoretical results without complex details.

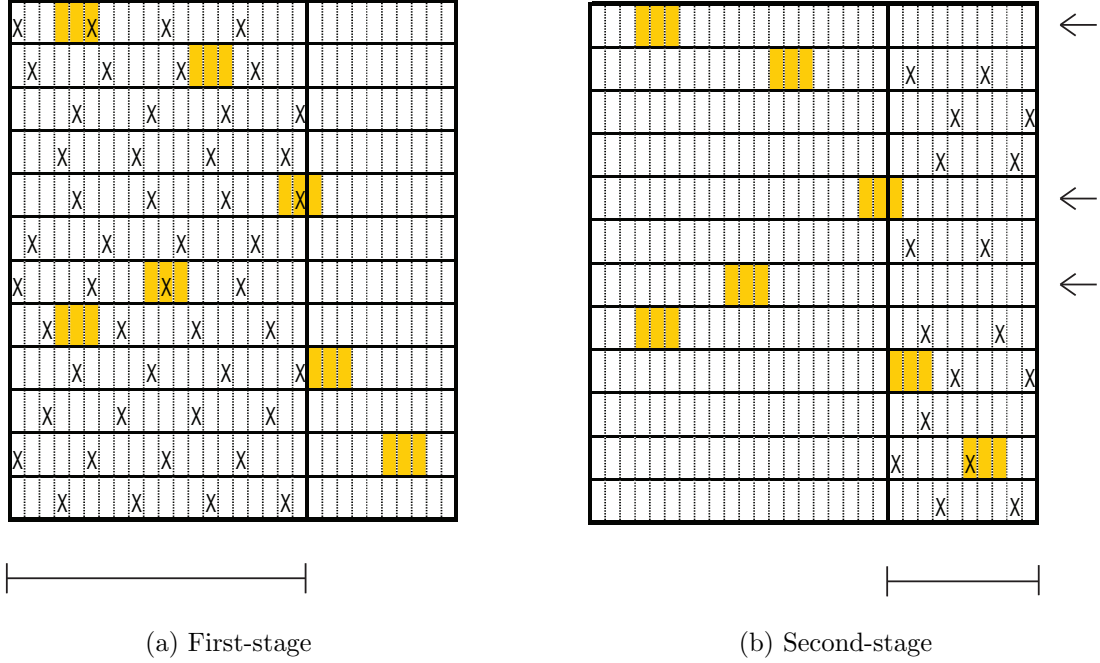


Figure 2.1: A two-stage systematic sampling diagram with the first and second stages. Each horizontal section represents a *floor*, each cell represents a *room*, ‘X’ marks sampling sites, i.e. *rooms*, and the shaded area signifies hotspots. In (b), the *floors* marked with arrows, which are the first, fifth and seventh *floors* from the top, did not require sampling.

We define the data model parameters for contamination distribution as follows:

- n : sample size
- T : number of *floors* (remediation sections) in the sampling venue
- R : number of *rooms* (sampling locations) per *floor*
- N : total number of *rooms*, $N = TR$

- r : sampling rate defined as the number of samples divided by the number of total sample sites, i.e. n/N .
- c : probability that a *floor* is contaminated
- p : relative size of contamination given a *floor* contamination
- b : size of a contamination as an integer multiple of a sampling unit
- τ : proportion of a *floor* sampled in first-stage
- α : proportion of samples used in first-stage

Remark When τN is not an integer, we round it up as for the number of *rooms* in first-stage sampling.

Remark We define a proportion τ for the number of sampling sites N and a proportion α for the number of sample n . The two proportions should be the same when a two-stage sampling requires sampling over the total area. However, probabilistically some *floors* do not require full sampling over two stages by the proposed design. The efficiency gain in our two-stage design comes from saving a portion of the sampling sites from sampling as shown by the example in Figure 2.1. The relationship between α and τ is detailed in Figure 2.4 where $\alpha \geq \tau$.

In Figure 2.1 the two-stage sampling idea is illustrated via a realization of a systematic sampling plan with the X's representing sampling sites. The left diagram shows a realization of first-stage sampling, and the right diagram a realization of second-stage sampling. We take the diagram as a cross-section of a building where each horizontal line represents a *floor* with $T = 12$ total number of *floors*. Every *floor* does not need to have the same side sampled, but the sampling sites in each *floor* are required to be contiguous. For simplicity, we keep the first-stage sampling area to be all in the same side, on the panel to the left of the double vertical lines. To the right of the doubled lines

is the second-stage sampling area. On each *floor*, each cell represents a *room*, and there are $R = 30$ *rooms*. There is at most one hotspot cluster per *floor*, and it is represented by the shaded boxes of size $b = 3$ *rooms*. Say that the sample size is $n = 66$. Then the overall sampling rate is $r = n/N = 66/(12 \times 30) = 11/60$. Here we set the first-stage *floor* splitting proportion $\tau = 0.6$, and therefore the number of *rooms* in first-stage is $0.6 \times 360 = 216$ *rooms*. Let the number of first-stage sample $\alpha n = 48$, and then the fixed space between sample points should be 5. We have $\alpha \geq \tau$ because in the first-stage τ proportion of all *floors* should be sampled with αn sampling points but in the second-stage not all *floors* need $1 - \tau$ proportion sampled with $(1 - \alpha)n$ sampling points. In this illustrative example, we see that the first, the fifth, and the seventh *floors* from the top are exempt from second-stage sampling because the hotspots on the respective *floors* are detected. For the maximum detection probability given a fixed budget we carry out systematic sampling over two-stage sampling where not every *floor* is fully sampled, yet the probabilistic design should render a more or less efficient second-stage sampling depending on the first-stage sampling result.

We can still show that a two-stage sampling design should be more effective than a one-stage design even without Assumptions 1, 3, and 5. When the size of hotspots varies, counter to Assumption 1, the detection probability of each *floor* fluctuates accordingly. The *floors* with large hotspots have larger *floor* detection probability, and this higher detection probability is shared among the remaining floors by reducing the number of *floors* to be sampled and increasing the *floor* detection probability in second-stage sampling. Therefore, a two-stage design *floor* detection probability should be greater than a one-stage design, hence the greater efficiency of using the sampling resource. When the floor contamination probability varies among floors, counter to Assumption 3, the floor with a higher contamination probability has a higher detection probability. As in the case of varying the size of hotspots the sampling resource in first-stage sampling is to be used more efficiently in second-stage, and the sampling rate should be greater than that

of a one-stage design. Lastly, when there is more than one hotspot (contamination cluster) per *floor*, counter to Assumption 5, the detection probability should increase from assuming only one hotspot per *floor*. By the same reasoning in the previous arguments countering Assumptions 1 and 3, a two-stage design benefits from the increased *floor* detection probability and becomes more efficient than a one-stage design.

Based on Assumptions 1-6 we use the above notations to define random variables as follows:

- C : total number of *floors* contaminated $C \sim \text{Bin}(T, c)$.
- C_1 : number of *floors* that contain contamination in the first-stage sampling area.
 $C_1|C \sim \text{Bin}(C, \tau)$
- C_2 : number of *floors* that contain contamination in the area not sampled in the first-stage. $C_2 = C - C_1$.
- D_0 : number of contaminated *floors* detected in one-stage sampling. $D_0|C \sim \text{Bin}(C, d_0)$ where d_0 is the detection probability.
- D_1 : number of contaminated *floors* detected in first-stage for a two-stage sampling plan. $D_1|C_1 \sim \text{Bin}(C_1, d_1)$ where d_1 is the *floor* detection probability of first-stage sampling.
- D_2 : number of contaminated floors detected in second-stage for the two-stage sampling plan. $D_2|C_2 \sim \text{Bin}(C_2, d_2)$ where d_2 is the *floor* detection probability of the second-stage sampling.

Breidt (1995) has proposed a variant of a systematic sampling plan called a Markov chain design in which the danger of administering a systematic sampling plan is avoided in the case of a periodic dispersion of hotspots.

We define a Markov chain sampling transition probability matrix for a one-dimensional design as follows:

Definition 2.2.1 A Markov chain sampling transition probability matrix $\mathbf{P}_{N_l \times N_l}$ for a one-dimensional design is an $N_l \times N_l$ matrix with the (i, j) element given by $\mathbf{P}(i, j)$ where N_l is the number of sampling locations per stratum. $\mathbf{P}(i, j)$ is the probability of sampling from location j in one stratum conditioned on its neighboring sampling location i in its stratum. For each row i , we need $\mathbf{P}(i, j) \geq \mathbf{P}(i, k)$ if $|i - j| < |i - k|$ in one-dimensional sampling. Every row should sum to 1. That is, $\sum_j \mathbf{P}(i, j) = 1$ for all i .

An example of $\mathbf{P}(i, j)$ for one-dimensional five sampling unit stratum is

$$\mathbf{P}_{5 \times 5} = \frac{1}{15} \begin{pmatrix} 4 & 3 & 3 & 3 & 2 \\ 4 & 4 & 3 & 2 & 2 \\ 3 & 3 & 3 & 3 & 3 \\ 2 & 2 & 3 & 4 & 4 \\ 2 & 3 & 3 & 3 & 4 \end{pmatrix}.$$

If $\mathbf{P}(i, i) = 1$ for all i 's and $\mathbf{P}(i, j) = 0$ for all $i \neq j$, then \mathbf{P} should revert the sampling design to a systematic sampling plan. If $\mathbf{P}(i, j) = 1/N_l$ for all i and j from $1, \dots, N_l$, then \mathbf{P} generates a one-per-stratum sampling design.

2.3 Sampling designs maximizing detection probability

In this section we compare *floor* detection probabilities of several designs under the assumption that hotspots are spatially correlated. Define *floor* contamination as having at least one sampling unit whose measurement exceeds a threshold, and a *floor* detection as detecting at least one of those sampling units. In Section 2.3.1 we briefly review several sampling plans, and in Section 2.3.2 we compare detection probabilities of one-stage spatial sampling plans. In Section 2.3.3 I describe the proposed two-stage sampling plan that maximizes the contaminated *floor* detection probability. In Section 2.3.4 I demonstrate theoretically that the detection probability of a two-stage design is greater than that of a one-stage design.

2.3.1 One-stage designs

We review four spatial sampling designs and one of their variants. A simple random sampling is an equal probability selection of independent rN sampling sites from N possible sites. Since a hotspot dispersal scenario is likely to be clustered in space, a simple random sampling plan should be the least efficient one in a cluster detection. A systematic spatial sampling is a selection of a fixed location in every stratum with a random starting point. Since nearby sample measurements are likely to be more similar and redundant than sample measurements that are further apart, a one-per-stratum design is more economical in detecting contaminated floors than a simple random design. There is a potential disadvantage when contamination is periodically dispersed, and the distance between sampling locations is the same as the periodic dispersal distance. However, that is a very unlikely scenario for contamination dispersal. A one-per-stratum design is a compromise between a simple random sampling and a systematic sampling where the sampling location in each stratum is random. A general form of a systematic sampling design and a one-per-stratum design is proposed in Breidt (1995) and called a Markov chain design. A transition probability matrix, defined in Definition 2.2.1, helps to run a Markov chain from one end of a sampling transect to the other in order to select the sampling location within a stratum dependent on the neighboring sampling locations. A Markov chain design provides a sensible spatial sampling approach by imposing a minimum and a maximum distance between neighboring sample sites. It is less flexible than a one-per-stratum design yet less rigid than a systematic design.

Figure 2.2 shows a realization of three of the four one-dimensional spatial sampling designs described above. Each design has a sample size $n = 14$ where the number of total sampling sites $N = 70$. At the top, we see that a simple random sampling plan has a large chance to miss a clustered contamination due to a large portion of the floor being uncovered. A Markov chain design in the middle has some variability in sampling locations within a stratum, yet it does not look drastically different from a systematic

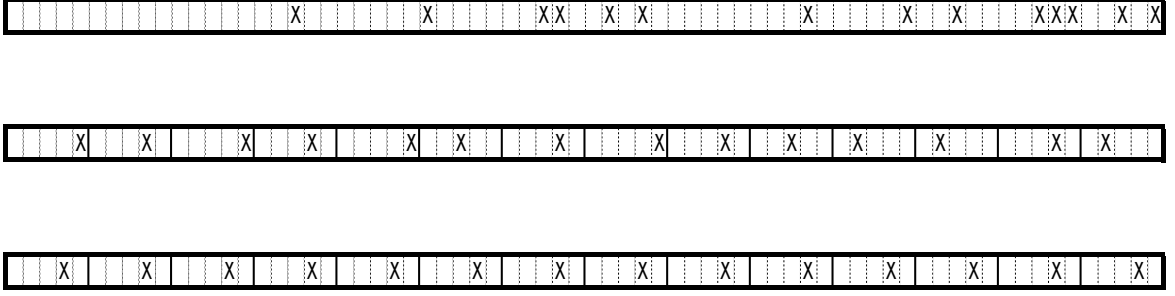


Figure 2.2: Examples of three sampling designs in one dimensional display. The top is a realization of simple random sampling (SRS), the middle is a Markov-chain design, and the bottom is a systematic sampling (SYS) design.

sampling design in the bottom. A Markov chain design could be designed more like a systematic design when the transition probability matrix has a positive probability close to or only on the diagonal, or it could be designed like a one-per-stratum design when the transition probabilities are uniform across each row.

2.3.2 Detection probabilities of one-stage designs

We compute the *floor* detection probabilities for three spatial sampling designs shown in Figure 2.2. Let D denote the event that one or more of the sample n is in a hotspot of size b sampling units located in N sampling sites. Then, the detection probability of a size n simple random sample, representing the event of detection as D_{SRS} , is

$$Pr(D_{SRS}) = 1 - \frac{\binom{N-b}{n}}{\binom{N}{n}}. \quad (2.1)$$

For a systematic sampling plan with a sampling rate of r , the *floor* detection represented by D_{SYS} , is in the following lemma.

Lemma 2.3.1 *Let the hotspot cluster size be b sampling units long on a transect. Let N denote the total number of sampling sites and $r = n/N$ a sampling rate. A systematic*

sampling plan detection probability of a b -sampling-unit large contamination is $\min(br, 1)$, that is,

$$Pr(D_{SYS}) = \min(br, 1). \quad (2.2)$$

Proof of Lemma 2.3.1. Consider a case where N is a multiple of a stratum size N_l . In other words, the sample size n is an integer defined as N/N_l . The selection probability of each sampling site within a stratum is $1/N_l = n/N = r$. For sample of size n to be in a hotspot of size b sampling units, we find b different sampling possibilities where $b \leq N_l$ and N_l different cases where $b > N_l$. Therefore the probability of *floor* detection is br when $b \leq \frac{1}{r}$ or else the *floor* detection probability is 1.

Consider the other case where the total sampling sites N is not an integer multiple of a stratum size N_l and the sample size depends on the randomly chosen starting point. Let $N_{l'}$ be the size of the remainder sampling sites when N is divided by N_l , i.e. $N = N_l n + N_{l'}$, and $N_{l'}$ is possibly $1, 2, \dots, N_l - 1$. Again, the selection probability for every sampling site within a full-length stratum is $1/N_l$ because there are N_l possible sites in these n number of strata. For the $n+1^{st}$ stratum at either end of the transect the selection probability is not $1/N_{l'}$ but rather $1/N_l$ because we place $N_{l'}/N_l$ probability of using the $n+1^{th}$ sample in this last incomplete stratum and each sampling site within this incomplete stratum should have equal probability of selection. Therefore, the expected sample size at a fixed sampling rate $1/N_l$ is $E(\text{sample size}) = n + \frac{N_{l'}}{N_l}$. Therefore, the probability of D_{SYS} is $br = b/N_l$ when the size of contamination $b < \frac{1}{r}$ and 1 when $b \geq \frac{1}{r}$. \square

We calculate a Markov chain design detection probability via simulation because as the chain gets longer or as the size of a stratum becomes relatively larger, the calculation becomes complicated. Where $N = 150$ we test three sampling rates $r = 1/6, 1/10$, and $1/15$ and use the following corresponding transition probability matrices:

for $r = \frac{1}{6}$

$$\mathbf{P}_{6 \times 6} = \frac{1}{12} \begin{pmatrix} 5 & 3 & 2 & 2 & 0 & 0 \\ 3 & 3 & 2 & 2 & 2 & 0 \\ 2 & 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 & 2 \\ 0 & 2 & 2 & 2 & 3 & 3 \\ 0 & 0 & 2 & 2 & 3 & 5 \end{pmatrix};$$

for $r = \frac{1}{10}$

$$\mathbf{P}_{10 \times 10} = \frac{1}{20} \begin{pmatrix} 4 & 3 & 3 & 3 & 3 & 2 & 2 & 0 & 0 & 0 \\ 3 & 4 & 3 & 2 & 2 & 2 & 2 & 2 & 0 & 0 \\ 2 & 3 & 4 & 3 & 3 & 2 & 1 & 1 & 1 & 0 \\ 1 & 2 & 3 & 4 & 3 & 2 & 1 & 1 & 1 & 0 \\ 1 & 2 & 2 & 3 & 3 & 3 & 2 & 2 & 1 & 1 \\ 1 & 1 & 2 & 2 & 3 & 3 & 3 & 2 & 2 & 1 \\ 0 & 1 & 1 & 1 & 2 & 3 & 4 & 3 & 2 & 1 \\ 0 & 1 & 1 & 1 & 2 & 3 & 3 & 4 & 3 & 2 \\ 0 & 0 & 2 & 2 & 2 & 2 & 2 & 3 & 4 & 3 \\ 0 & 0 & 0 & 2 & 2 & 3 & 3 & 3 & 3 & 4 \end{pmatrix};$$

and for $r = \frac{1}{15}$

$$\mathbf{P}_{15 \times 15} = \frac{1}{54} \begin{pmatrix} 6 & 6 & 6 & 6 & 6 & 6 & 6 & 4 & 3 & 2 & 2 & 1 & 0 & 0 & 0 \\ 6 & 6 & 6 & 6 & 6 & 6 & 6 & 4 & 3 & 2 & 2 & 1 & 0 & 0 & 0 \\ 5 & 6 & 6 & 6 & 5 & 4 & 4 & 4 & 4 & 3 & 3 & 2 & 1 & 1 & 0 \\ 4 & 5 & 6 & 6 & 6 & 5 & 4 & 4 & 3 & 3 & 2 & 2 & 2 & 2 & 0 \\ 3 & 4 & 5 & 6 & 6 & 6 & 5 & 4 & 3 & 3 & 3 & 2 & 2 & 2 & 0 \\ \\ 3 & 3 & 3 & 4 & 4 & 6 & 6 & 4 & 4 & 4 & 3 & 3 & 3 & 2 & 2 \\ 2 & 2 & 3 & 3 & 4 & 6 & 6 & 6 & 4 & 4 & 4 & 3 & 3 & 2 & 2 \\ 3 & 3 & 3 & 3 & 3 & 4 & 5 & 6 & 5 & 4 & 3 & 3 & 3 & 3 & 3 \\ 2 & 2 & 3 & 3 & 4 & 4 & 4 & 6 & 6 & 6 & 4 & 3 & 3 & 2 & 2 \\ 2 & 2 & 3 & 3 & 3 & 4 & 4 & 4 & 6 & 6 & 4 & 4 & 3 & 3 & 3 \\ \\ 0 & 2 & 2 & 2 & 3 & 3 & 3 & 4 & 5 & 6 & 6 & 6 & 5 & 4 & 3 \\ 0 & 2 & 2 & 2 & 2 & 3 & 3 & 4 & 4 & 5 & 6 & 6 & 6 & 5 & 4 \\ 0 & 1 & 1 & 2 & 3 & 3 & 4 & 4 & 4 & 4 & 5 & 6 & 6 & 6 & 5 \\ 0 & 0 & 0 & 1 & 2 & 2 & 3 & 4 & 6 & 6 & 6 & 6 & 6 & 6 & 6 \\ 0 & 0 & 0 & 1 & 2 & 2 & 3 & 4 & 6 & 6 & 6 & 6 & 6 & 6 & 6 \end{pmatrix}.$$

Figure 2.3 compares the detection probabilities of four different sampling designs. We set the sampling rate at $r = 1/6$ in the left plot, $1/10$ for the middle, and $1/15$ in the right. The x -axis represents the extent of contamination proportion p per *floor*, and the y -axis represents the average detection probability. The expected *floor* detection probability of the systematic sampling scenario based on simulation is represented by solid blue lines, that of a Markov chain design by red short-dashed lines, one-per-stratum design in black long-dashed lines, and simple random sampling by solid gray lines. For a Markov chain design, we use the transition probability matrix as shown above for the respective

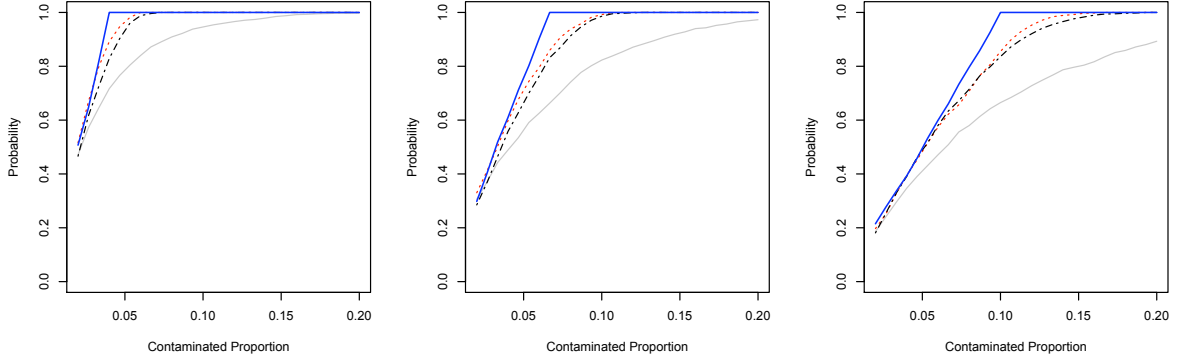


Figure 2.3: Comparing detection probability of four sampling strategies at three different sampling rates $r = 1/6$ (left), $1/10$ (middle), and $1/15$ (right). The x -axis represents the extent of contamination p in a given *floor*. On average, the systematic sampling scenario (solid blue) is better than a Markov chain design (short red dash), one-per-stratum design (long black dash), and SRS (solid gray).

sampling rate r .

Among the class of spatially stratified designs, systematic sampling has the largest chance of hotspot detection due to its periodic location sampling. Systematic sampling achieves the highest detection probability because it reduces the redundancy of sampling nearby locations and never leaves a sampling area greater than the size of a stratum uncovered. As shown in Lemma 2.3.1 the systematic sampling plan floor detection probability (in solid blue) increases linearly as the contamination cluster size pN increases up to a stratum size. Markov chain design and one-per-stratum design are comparable in detection probabilities when the Markov chain design transition probability matrix imposes a relatively even distribution of probabilities across $\mathbf{P}(i, \cdot)$'s for the row index $i = 1, \dots, N_l$ as in the example given above for $\mathbf{P}_{15 \times 15}$. Their detection probabilities are both less than a systematic sampling design and are significantly greater than a simple random design. The simple random sampling *floor* detection probability is the lowest, and its detection probability increases at the slowest rate with respect to the size of the hotspot because it ignores the spatially clustered nature of hotspot data. The average

detection probability is as shown in equation (2.1). Through simulation and analytical derivations of the detection probabilities of simple random and systematic sampling in equations (2.1) and (2.2), we verify that the systematic sampling design is the most efficient in *floor* detection among the four stochastic designs often considered for spatial sampling.

2.3.3 Two-stage sampling to maximize detection probability

In sampling literature, a two-stage sampling is often used synonymously with subsampling where the first-stage consists of sampling the components of interest, and the second-stage involves selecting and measuring one or more aliquots from each sampled component. The proposed two-stage sampling design is not a type of subsampling since we are interested in locating hotspots instead of estimating the mean or the total level of contaminant dispersal. The focus is on sampling as wide of a coverage area as possible adaptively, and we expect the overall two-stage sampling frequency to be greater than that of one-stage sampling using the same sample size.

An adaptive multi-stage design has an advantage over one-stage design in obtaining more accurate information of sampling domain. For example, Thompson (1990) proposes a two-stage design where in the first stage probability sampling is employed, and then in the second stage cluster sampling is performed nearby the sites of first-stage detections. As shown in example, it is sensible to use the ensuing stage sampling resources efficiently. The trade-off of a multi-stage design is that it requires planning and allotting appropriate time and sampling resources for each stage. Most often the additional effort in planning and coordination should be worth the additional information. However, when some measurement readings take a long time for lab analysis and time is an important factor, a multi-stage design would not be preferred over a one-stage design. In practice, it is judicious to curtail the benefit of a multi-stage design to a two-stage design given that extra detailed information is more valuable than savings of the sampling resources by

performing a one-stage design.

Consider a two-stage design where each stage sampling strategy is simple random sampling. As we focus on maximizing the *floor* detection probability, second-stage sampling is required where there is no hotspot detected in first-stage sampling. Note that there is no difference in detection probability between a one-stage simple random sampling of size n over N sampling sites and a multi-stage simple random sampling design of the same size. See Appendix A (A.1) for the detailed proof. Let the first-stage sample size be n_1 , the second-stage n_2 , $n_1 + n_2 = n$, and b represent the size of a hotspot in sampling units. The two-stage simple random sampling *floor* detection probability is

$$Pr(D_{SRS, n_1}) + Pr(D_{SRS, n_2}) = 1 - \frac{\binom{N-b}{n}}{\binom{N}{n}}. \quad (2.3)$$

The detection probability of two-stage simple random sample whose sample sizes are split in any combination of n_1 and n_2 that sum to n in (2.3) equals the detection probability of simple random sample of the same total sample size n in (2.1). Therefore, when the Department of Energy Technical Standard (2005) proposes a simple random sample of first $n_1 = 15$ and adaptively an additional $n_2 = 15$, this in fact yields the same *floor* detection probability as the simple random sample of $n = 30$.

We are interested in a two-stage sampling method that has a higher contaminated *floor* detection probability than a one-stage sampling design. We adopt a systematic sampling plan in order to maximize the detection probability as shown in Section 2.3.2 Figure ???. This design requires partitioning the sample and the sampling area for a sequential two-stage sampling; each *floor* should be divided into two parts, as well the sample. A key idea enters by relieving a part of the “planned” second-stage sampling, i.e. reducing the total number of sampling sites N to say $N' < N$ and naturally increasing the overall effective sampling rate r from n/N to n/N' . The locations released from second-stage sampling are of size $N - N'$, whose *floors* have hotspots detected in first-

stage sampling. Ideally, we would maintain the overall sampling plan like that found in a one-stage systematic sampling design, so that the detection probability is kept the highest as shown in Figure 2.3. In reality, there is a deviation from the expectation given the best guess of the relative size of contamination p to the *floor* area and the proportion c of *floors* contaminated.

As mentioned in the proposal of then two-stage sampling, the sample splitting proportion α between the first-stage and second-stage sample need to be determined, as well as the *floor* splitting proportion τ . For a systematic sampling each *floor* is divided into a stratum of size $\tau N/(\alpha n)$ sampling locations, assuming $\tau N/(\alpha n)$ is an integer. We sample one location at random from the first stratum on one end, and then select every $\tau N/(\alpha n)^{\text{th}}$ location thereafter the $\alpha n/T^{\text{th}}$ sample on each *floor* is used. Since there is some control over the sample size n , one should adjust n and α so that the number of sample on each *floor* $\alpha n/T$ is an integer. The first-stage sampling area on each *floor* should span over τR rooms. We also make an operative decision on τ such that it is close to the optimal τ^* and satisfy that the number of sampling sites τR is an integer. In second-stage sampling, ideally, one should continue sampling every $\tau N/(\alpha n)^{\text{th}}$ location from the last sample until $(1 - \alpha)n$ sample is depleted. In reality, an adjustment needs to be made. After covering τR locations in first-stage sampling, the remaining sampling locations would be $(1 - \tau)R(T - D_1)$ determined by the number of *floors* D_1 detected in the first stage. The remaining sample size for second-stage sampling is $(1 - \alpha)n$. Hence, the number of sampling units in a second-stage stratum should be the largest integer less than or equal to $(1 - \tau)R(T - D_1)/\{(1 - \alpha)n\}$.

Aiming for an overall systematic sampling design, we equate the first-stage sampling rate to the second-stage sampling rate computed under the expected *floor* detection scenario of first-stage sampling $E(D_1)$, since first-stage sampling is yet to be administered in the planning stage. Given a sample size n and the total area of sampling N with R rooms per each of T floors, we set up an equilibrium equation to find the first-stage sample

proportion α and its corresponding *floor* splitting proportion τ given the contamination proportions p and c relative to R and T .

$$\frac{\alpha n}{\tau R T} = \frac{(1 - \alpha)n}{(1 - \tau)R(T - E(D_1))} \quad (2.4)$$

Under the Assumptions in Section 2.2, the expected number of contaminated *floor* detections in first-stage sampling is $E(D_1; p, c, \alpha, n, T) = c\alpha n/R$. Solving for τ in equation (2.4), we get

$$\tau = \frac{(T - E(D_1))\alpha}{T - \alpha E(D_1)} = \frac{\alpha(1 - pcr)}{1 - \alpha^2 pcr}. \quad (2.5)$$

In determining the optimal α , we run simulations under different settings of p , c , and T . See Section 2.4 for findings from numerical studies. We find as a general rule to set $\alpha^* = 0.45$ as the optimal sample splitting proportion. In Figure 2.4, we plot optimal first-stage *floor* sampling proportion τ versus sample proportion α . The relationship is slightly slanted toward α in that $\alpha \geq \tau$ for any fixed α except for $\alpha = 0$ or 1 , at which point the two-stage setting reverts to one-stage sampling. From three settings of p and c : where the black line represents the case of low *floor* detection probability due to a small size of contamination $p = 0.04$ yet with 50% chance of each *floor* being contaminated, the red line is for a large size of contamination $p = 0.4$ and the same 50% chance of each *floor* being contaminated, and the blue line where $p = 0.4$ and $c = 0.8$, we see that as each p and c becomes larger and hint at a higher *floor* detection probability, it is recommended to use a smaller first-stage *floor* splitting proportion τ than when there is a small *floor* detection probability.

2.3.4 Theoretical properties of two-stage sampling

We use the result from Lemma 2.3.1 to prove that our proposed two-stage systematic sampling design has a higher detection probability than a one-stage systematic sampling design of equal size. In this section, we assume that there is at most one contaminated *room* per *floor* and that there is no uncertainty in declaring a *room* contaminated to keep

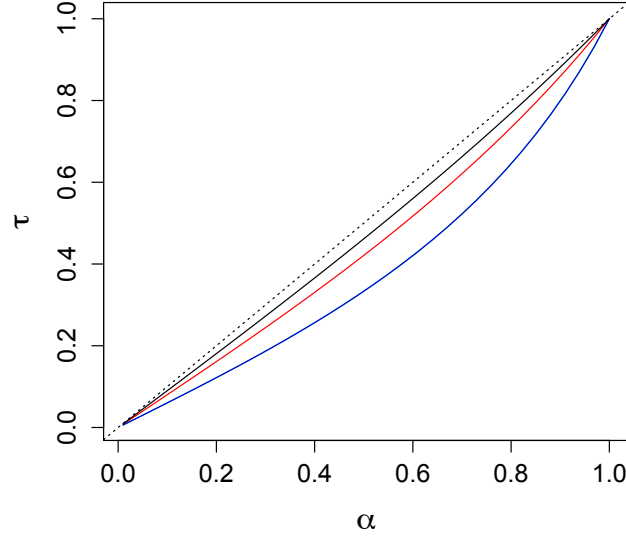


Figure 2.4: First-stage *floor* sampling proportion τ versus sample proportion α in three scenarios: (i) $p = 0.04$ and $c = 0.5$ (black); (ii) $p = 0.4$ and $c = 0.5$ (red); and (iii) $p = 0.4$ and $c = 0.8$ (blue). There is a more-or-less linear relationship between τ and α . From the color-coded lines we see that when the relative size of contamination p is greater, it is recommended to use a smaller first-stage *floor* splitting proportion τ than α .

the proof concise. The same result holds with more complex contamination dispersal scenarios, which are discussed in Section 2.2.

Theorem 2.3.2 *Let D_0 be the number of contaminated floors detected via one-stage systematic sampling. Let D_1 and D_2 be the number of contaminated floors detected in first- and second-stage sampling respectively in a two-stage adaptive sampling for given sample splitting proportion α and floor splitting proportion τ . Let T be the total number of floors, R be the number of rooms per floor, and n be the sample size, and r the overall sampling rate. Under Assumptions 1 - 6 in Section 2.2, we have*

$$E(D_1 + D_2) \geq E(D_0)$$

as n and N are at a fixed rate $r = n/N$.

Proof of Theorem 2.3.2: Let the size of contamination be at most one sampling unit large, i.e. $p = 1/R$ per contaminated *floor*. By Lemma 2.3.1, the detection probability of a one-stage systematic design is $r = n/N$. The expected number of *floor* detections in a one-stage sampling plan is

$$E(D_0) = E(E(D_0|C)) = E(Cd_0) = Tcd_0 = Tc\frac{n}{RT} = \frac{cn}{R}.$$

For a two-stage sampling design, the first-stage detection probability is $d_1 = \alpha n/(\tau RT)$ and the second-stage detection probability is $d_2 = (1 - \alpha)n/(1 - \tau)R(T - E(D_1))$. The expected number of *floor* detections in a two-stage plan is

$$\begin{aligned} E(D_1 + D_2) &= E_C(E_{C_1}(E(D_1|C_1, C))) + E_C(E_{C_1}(E_{D_1}(E(D_2|D_1, C_1, C)))) \\ &= E_C(E_{C_1}(C_1d_1|C)) + E_C\left(E_{C_1}\left((C - C_1)\frac{(1 - \alpha)n}{(1 - \tau)R}\left(E_{D_1}\left(\frac{1}{T - D_1}|C_1, C\right)\right)\right)\right) \\ &= E_C(C\tau d_1) + E_C\left(C(1 - \tau)\frac{(1 - \alpha)n}{(1 - \tau)R}E_{C_1}\left(E_{D_1}\left(\frac{1}{T - D_1}|C_1, C\right)\right)\right) \\ &= Tc\tau\frac{\alpha n}{RT} + Tc\frac{(1 - \alpha)n}{R}E_C\left(E_{C_1}\left(E_{D_1}\left(\frac{1}{T - D_1}|C_1, C\right)\right)\right) \\ &= \frac{c\alpha n}{R} + \frac{c(1 - \alpha)n}{R}E_C\left(E_{C_1}\left(E_{D_1}\left(\frac{T}{T - D_1}|C_1, C\right)\right)\right) \\ &\geq \frac{cn}{R} = E(D_0). \end{aligned}$$

A two-stage design detection probability is greater than or equal to a one-stage design detection probability for all pairs of α and τ . \square

2.4 Numerical results

In Theorem 2.3.2, a two-stage design has an advantage over a one-stage design in detecting contaminated *floors*. Now, it remains to determine the optimal sample and *floor* splitting proportions α and τ so that the detection probability is maximized. We numerically identify the optimal proportions α^* and τ^* because p and c are the parameters of the contamination data model on which α and τ jointly depend as in equation (2.5).

Simulations were run under the following conditions. We fix $T = 60$ *floors*. On each *floor* there are $R = 120$ sampling units, i.e. *rooms*. The total sampling units N is 7200. We place one contamination cluster per *floor*, and the placement is chosen at random from $R - b + 1$ number of contiguous *rooms*, as the size of the contamination cluster is set to b -sampling units, ranging from 1 to 15 by an increment of 1. We ran a two-stage systematic sampling plan at three sampling rates $r = 0.1, 0.2$, and 0.25 . We varied the number of *floors* contaminated at nine levels from 6 to 54 *floors* by an increment of 6 *floors*. Lastly, we experimented with the levels of α from 0.05 to 0.95 by an increment of 0.05 and a corresponding level of τ , calculated in equation (2.5) along with the appropriate values for p, c and r substituted. Note that when $\alpha = 0$ or 1 , the proposed sampling design resorts to a one-stage systematic design. As we have a probability sampling design, we simulated 9 (building contamination levels) $\times 15$ (*floor* contamination sizes) $\times 3$ (sampling rates) $\times 19$ (levels of two-stage design parameter) ≈ 7700 settings each 1000 times to obtain *floor* detection probabilities for two-stage sampling and compared to the corresponding one-stage designs.

In Figure 2.5, we list three plots to summarize the sampling simulation results. We use sampling rate $r = 0.2$. In each plot, we assume a different size of contamination; from left to right we set p being 0.05, 0.1, and 0.2, i.e. the size of contamination $b=3, 6$, and 12 respectively. Each plot contains four profiles of the expected number of contamination detected *floors* in which $C = 12, 24, 36$, and 48 *floors* exactly contain contamination (corresponding to the expected scenario of $c=0.2, 0.4, 0.6$, and 0.8) from bottom to top. The first-stage sample proportion α is marked on the x -axis below each plot. The y -axis shows the expected number of *floors* detected. The solid line represents the average of the total number of *floors* detected for a two-stage design given an α . The long-dashed horizontal lines correspond to the expected total number of *floors* detected under the same r, c , and b , when the value is constant across α as $E[Tcr]$. This is the same as setting $\alpha = 0$ or 1 , which is a one-stage design. Describing the optimal α^* in terms of

the proportion c of contaminated *floors*, we have a four-tier explanation: when $c < 0.35$, there is very little difference between the one-stage and two-stage detection probability because of low probability of overall *floor* detection, and hence, the estimated α^* is unreliable and its variability is high; when $0.35 < c < 0.6$, α^* is around 0.4; when $0.6 < c < 0.7$, α^* is between 0.4 and 0.7; and when $c > 0.7$, α^* is about 0.6. The pattern of high-to-low-to-high α^* , as c grows from 0.35 to 0.9, is a bit unexpected. We can explain the relatively large α^* of 0.6 when the probability of *floor* detection is high due to either large p or c because detecting as many contaminated *floors* as possible in the first-stage helps us increase the second-stage sampling rate in comparison to the first-stage.

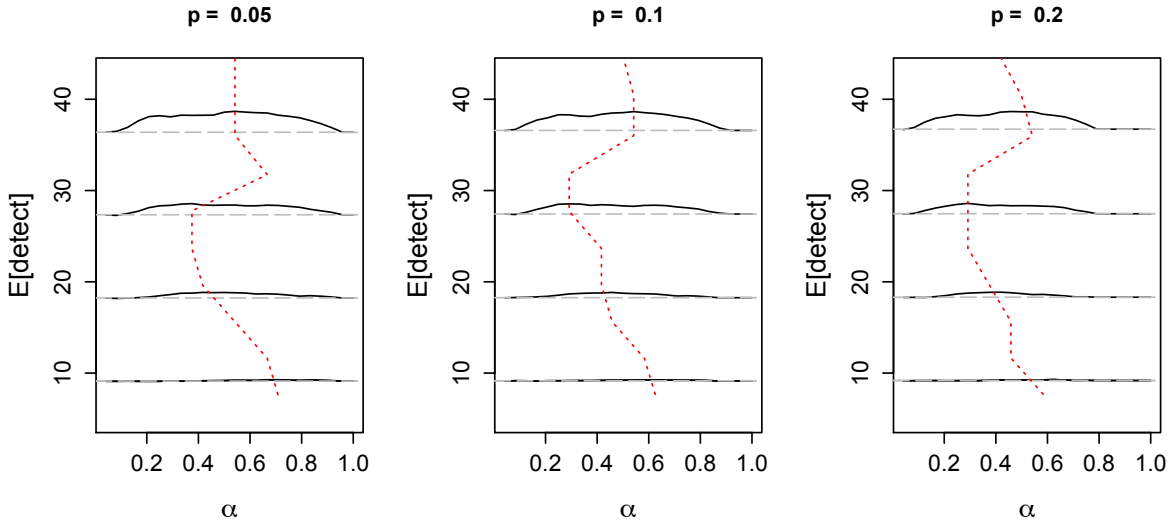


Figure 2.5: Black solid lines are the expected *floor* detection count profiles of two-stage sampling where $T = 60$ *floors* and $R = 120$ *rooms* per *floor* as the sample proportion α varies from 0 to 1. The gray long-dashed lines are the expected *floor* detections of one-stage systematic sampling. The red short-dashed lines trace the maximum for each *floor* detection profiles and mark the optimal α^* (given p and C). The jaggedness in the optimal α profiles is due to the discreteness of the sample selection probability.

You will see from expected number of *floor* detection profile plots in Figure 2.5 that when the proportion of contaminated *floors* is small ($c \leq 0.3$), there is little difference in

the detection probability between one- and two-stage systematic sampling because the detection probability is relatively low. When the proportion of *floors* contaminated c increases and the size of contamination p is greater than 0.1, a two-stage design with large α (> 0.8) gives us the same detection probability as in a one-stage design. Since detection is more likely when c and p are large, a two-stage design has marginally little to offer over a one-stage design.

In practice, one may have prior knowledge of the extent of contamination on each *floor* and the sample size. This allows a sampling design planner to refer to the results in Figure 2.5 and select an optimal two-stage sample splitting proportion for the proposed two-stage design. When one has little basis for making a good guess, we suggest a conservative approach of setting $\alpha^* = 0.5$ because it is the median of the estimated α^* from the detection probability profiles.

2.5 Beryllium clean-up study at Ames Laboratory

In this section, we detail a study that utilizes the proposed two-stage sampling design. This study is motivated by the recent decontamination efforts of surfaces containing beryllium dust at Ames Laboratory, a U.S. Department of Energy (DOE) facility operated under contract by Iowa State University. Beryllium is a metal that was widely used within the DOE complex for a variety of purposes including as moderators or reflectors in nuclear reactors and as reactor fuel element cladding. Inhalation of beryllium dust or particles can cause chronic beryllium disease or beryllium sensitization. In the 1940s and early 1950s, beryllium was regularly used in uranium and thorium purification processes developed at the Laboratory in support of the Manhattan Project. Although beryllium usage decreased significantly in subsequent years, legacy beryllium contamination exists in primarily inaccessible areas of each research building. Accordingly, it is necessary to characterize research-generated beryllium dispersal in order to ensure that current

employees are not at risk for exposure. Hence, we are interested in devising a spatial sampling plan to establish the state of surface contamination. We propose a two-stage systematic sampling design, which consistently detects problematic areas with higher probability than a one-stage systematic sampling design of equivalent sample size.

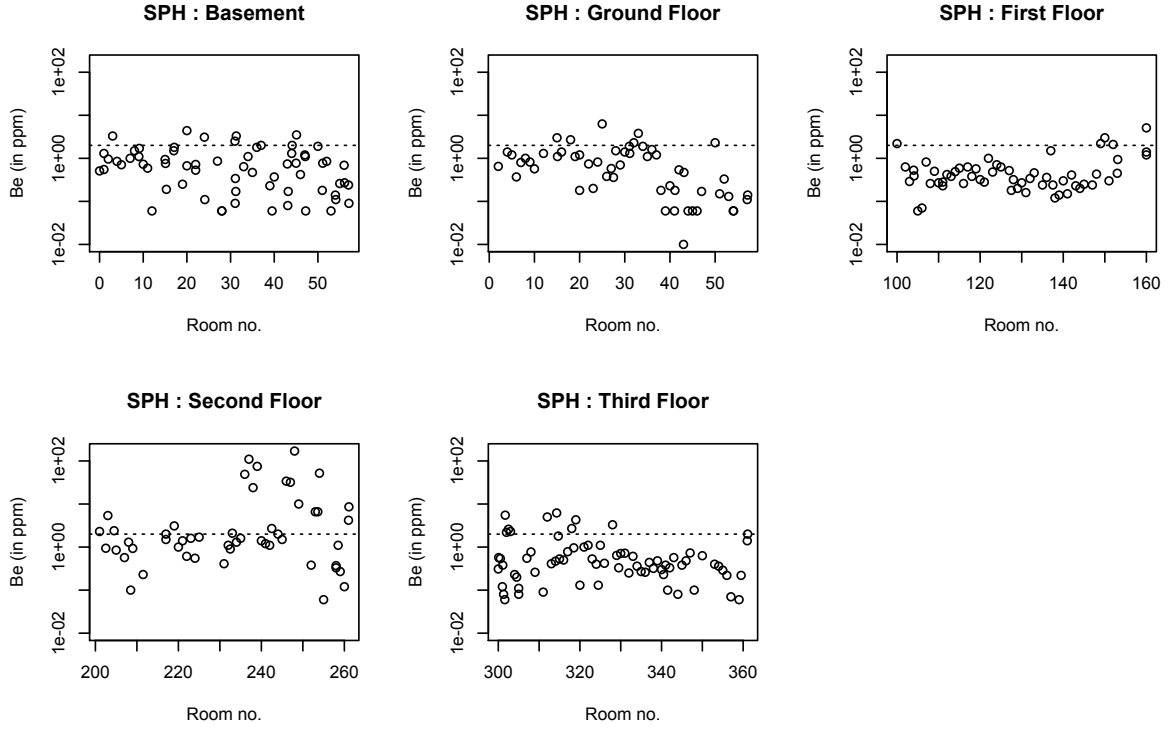
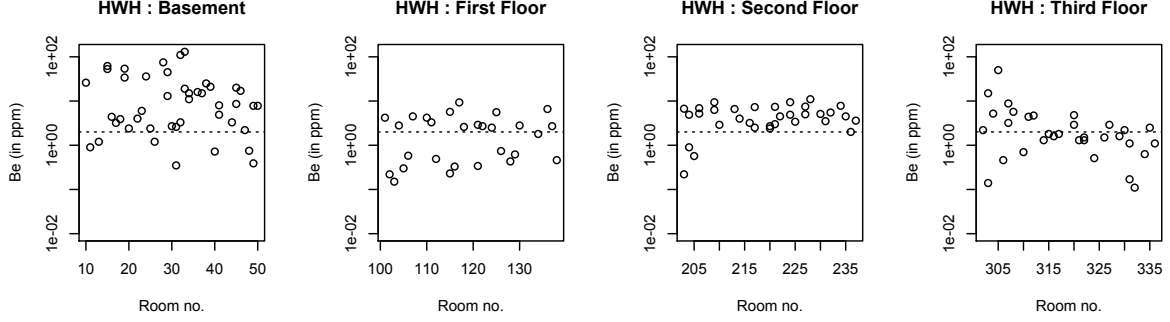


Figure 2.6: Five-story Spedding Hall (SPH) door-top beryllium census in log scale. The dashed line represents a derived background limit.

In beryllium sampling reports from the National Energy Technology Laboratory-Albany (Bond 2008) and of the United States Enrichment Corporation's Portsmouth gaseous diffusion plant, the main sampling strategy is a combination of a simple random sample and a judgement sample. In the DOE Technical Standard (2005) a simple random sample of first $n_1 = 15$ and adaptively an additional $n_2 = 15$ is suggested for a surface scoping survey. Though sampling is over two stages, there is no difference from a one-stage simple random sample of 30 in terms of detection probability as shown in (2.3)



(a) Wilhelm Hall

Figure 2.7: Four-story Wilhelm Hall (HWH) door-top beryllium census in log scale. The dashed line represents a derived background limit.

with the details in the Appendix (A.1). Rondeau (2009, p. 71-72) also describes a surface sampling using a simple random sample. Simple random sampling is less efficient if the contaminations are spatially clustered. In industrial hygiene practice a commonly used design for contaminant detection is simple random sampling, as seen in the DOE Technical Standard (2005) and Rondeau and et al. (2009).

In this case study, we use beryllium door-top concentration data, which our client collected from the troughs of every door top, to validate our method. We simulate three sampling strategies on two Ames Laboratory buildings, Spedding Hall and Wilhelm Hall, and calculate the actual detection probabilities among the sampling plans. We take a *floor* as the unit of remediation and sample each *floor* over two stages. In Spedding Hall, each *floor* contains 51 to 72 rooms, and in Wilhelm Hall there are 28 to 44 rooms per *floor*. From Spedding Hall data, we have found that the spatial correlation among the beryllium dust data is strong only in neighboring door top measurements and very weak among rooms across the hallway. To model the observations in a one-dimensional setting, we string out the sampling sites around the hallway. In Wilhelm Hall, there is strong spatial correlation in all directions. Here we order the rooms in a zigzag pattern, alternating across hallways.

In Figure 2.7, we show the door-top data of three representative *floors* in two buildings mentioned above, with the horizontal dotted line representing the DOE site-specific derived background beryllium concentration. The vertical axis is in logarithmic scale. Spedding Hall is mostly free from research-derived beryllium, while Wilhelm Hall has a higher level of beryllium dust detected.

In Table 2.1, we report the *floor-by-floor* and overall building beryllium detection probability and each of their simulation standard errors in percentage. We set the sampling rate at $r = 0.15$. The first column has the number of rooms/door-top measurements per *floor*. The second column reports the percentage of beryllium measurements above the site-specific derived background concentration per *floor*. The third column contains the simple random sampling (SRS) detection probability. The *floor* detection probability is not dependent on the beryllium dispersal scenario but only dependent on the sampling rate $r = 0.15$, so we use equation (2.1) to calculate it. Also, the detection probability of a one-stage systematic sampling, denoted as ‘One-stage’ in Table 2.1, is obtained using a direct numerical calculation. The detection probability of a two-stage sampling plan, denoted as ‘Two-stage’ in Table 2.1, may depend on the choice of α and τ . We use $\alpha = 0.6$ and τ from equation (2.5). In order to use equation (2.5), we need to know pR the contiguous contamination size or a dispersal scenario and the proportion c of *floors* with contamination. We set $p = 0.05$ and $c = 0.7$ since the Laboratory expected a small area to be contaminated per *floor* yet quite a few spots scattered throughout the buildings. In other words, more than half of the *floors* should contain at least one sampling location with a high level of beryllium.

Recall from Section 2.3.1 that a simple random sampling detection probability is lower than that of a systematic design when beryllium dispersal is concentrated, i.e. spatially correlated. In Table 2.1, one-stage and two-stage systematic sampling plans have similar or higher *floor* contamination detection probabilities than a simple random sampling. In Wilhelm Hall, we see almost no difference in detection probability among the sampling

Table 2.1: *Floor* by *floor* contamination detection percentage from 1000 simulation runs using three different sampling strategies. The two-stage sampling plan combined two buildings as one. Sampling rate was fixed at 15%. The first-stage sample splitting promotion was 60%. The number of rooms T and the percentage of contamination in each *floor* are given in the first two columns. The next three columns show the average (and the standard deviation of) hotspot-*floor* detection probabilities.

Spedding Hall	T	Hotspot (%)	SRS	One-stage	Two-stage
Basement	65	9.2	64.9 (5.11)	74.3 (4.59)	76.7 (4.05)
Ground Floor	54	11.1	63.7 (3.68)	61.1 (4.50)	72.3 (3.91)
First Floor	58	8.6	58.4 (4.97)	58.6 (4.89)	68.1 (5.23)
Second Floor	51	37.3	98.3 (1.33)	100 (0)	100 (0)
Third Floor	72	12.5	79.6 (3.27)	81.9 (3.29)	88.4 (4.84)
Wilhelm Hall	T	Hotspot (%)	SRS	One-stage	Two-stage
Basement	44	84.1	100 (0)	100 (0)	100 (0)
Ground Floor	28	53.6	96.5 (1.56)	100 (0)	100 (0)
First Floor	33	87.9	100 (0)	100 (0)	100 (0)
Second Floor	32	43.8	93.9 (1.97)	96.8 (1.38)	99.8 (0.43)

Table 2.2: *Floor* detection percentages for each building separately and collectively using three different sampling strategies.

Building	SRS	One-stage	Two-stage
Spedding	73.00	76.02	81.12
Wilhelm	97.60	99.22	99.85
Total	83.93	85.88	89.48

plans because the proportion of contamination is so high that any sampling plan would detect the *floors* contaminated. However, when the *floor* contamination proportion is small as it is in Spedding Hall, a two-stage sampling plan displays a higher detection probability than the other two. It shows that a two-stage systematic sampling plan makes more strategic use of sample.

2.6 Conclusion

We are interested in maximizing the detection probability of contaminated *floors*. Simple random sampling is a common practice of surface sampling in industry, which is not the most efficient use of sampling resources. We recommend a two-stage systematic sampling design to achieve a higher *floor* detection probability than any one-stage sampling designs. Two-stage sampling is more effective than one-stage sampling because in two-stage sampling a set of *floors* detected in the first-stage is excluded in the second-stage sampling and this leads to an increased sampling frequency over the area that is to be sampled.

The optimal first-stage sample proportion is but is not sensitive to the contamination size and the fraction of *floors* contaminated. In practice it can be chosen via a Bayesian approach or by optimizing the minimax risk of detection. One could plan an adaptive multi-stage design beyond two stages, but it has a diminishing return in detection probability. For an additional stage of sampling, extensive prior knowledge of the sampling domain is required, as well as a greater amount of planning effort and implementation time. Therefore, we recommend the use of the two-stage design, which achieves a balance of implementation complexity and efficiency.

CHAPTER 3. DIFFERENCE-BASED VARIANCE FUNCTION ESTIMATION OF A ONE-DIMENSIONAL NONSTATIONARY PROCESS

3.1 Introduction

3.1.1 Motivation

Many spatial processes exhibit nonstationary features, such as non-constant mean, variance, and varying covariance structures. We often encounter data with these features in ecology, geology, meteorology, astronomy, and in sociology. More specifically, natural phenomena possess these characteristics in species and mineral abundance, wind fields, crop yields (Hu and Mo (2011)) , and the Cosmic Microwave Background maps (Inman et al. (1997)). Human activities on the aggregate level also display nonstationary spatial patterns such as an Internet search query pattern associated with a geo-referenced code (Kessler and Shnerb (2009)), a real estate price map (Helbich et al. (2014) and Gelfand et al. (2004)), and an air pollution map (Briggs et al. (1997)), to name a few. With the development of modern technology especially in communication, the prevalent use of hand-held devices and the capacity for large data storage have brought about high demand for spatial data analysis. It is not only important to estimate the process mean map but also useful to construct reasonable interval estimates of the mean process and spatial prediction intervals.

We are interested in estimating the variance function of a one-dimensional spatial

process where the mean and the variance functions are smooth and have additive correlated errors. We assume a fixed equidistant design in one dimension and consider a mixed domain asymptotic framework to develop the asymptotic properties of a variance function estimator. Our estimator starts with the same differencing idea as Brown and Levine (2007) and Wang et al. (2008). We use Gasser-Müller kernel for smoothing, which helps to simplify the theoretical derivation as is the case in the latter of the two papers. A development from their approach, however, is that we extend the scenario to a correlated nonstationary process and also discuss the cross-validation idea for bandwidth selection. Our estimator requires estimating the correlation structure embedded in the data, as well as adjusting the scale of the difference-based estimator using the correlation information. The adjustment scale is the variogram value at a difference lag distance, i.e. one minus the correlation between a set of lagged observations.

3.1.2 Literature Review

Two common approaches to variance estimation are the likelihood-based method and the method-of-moments. Anderes and Stein (2011) have presented a likelihood-based approach to estimate the parameters of a nonstationary spatial process. A series of likelihood is constructed in a cascading form using the nearest observation from the location of estimation to then nearby observations once the neighborhood size is increased. The final likelihood function is constructed by heavily weighting the likelihoods formed by nearby observations and discounting the weights on the farther observations. Such weighting schemes marginalize the influence of far away observations and strengthens the idea of local stationarity. This method deals well with irregularly spaced data. Also, the smoothing kernel applied across the domain should produce a smooth parameter functional estimation. A few drawbacks are the computational burden of inverting covariance matrices at every location for variance estimation, especially when using their bandwidth proposal ideas; lack of statistical efficiency in the risk measure; and a rigid

Gaussian distributional assumption of the process for analytical tractability.

Another common and general approach to variance estimation is the method-of-moments estimation. This approach often requires estimating a mean structure, while a difference-based approach does not require estimating the mean. Also, the variance estimation suffers less from the bias generated by estimating the mean (Seifert et al. (1993)). Using differences of successive observations, von Neumann et al. (1941) have proposed variance estimation of independent and identically distributed errors. Gasser et al. (1986) have used second order differences to estimate the variance for non-fixed designs. Gasser et al. (1985) developed kernels, which we use, for nonparametric curve estimation. Brown and Levine (2007) have introduced a difference-based kernel estimator for a non-constant variance process in one dimension. They assume that a nonstationary process has a smooth mean and variance functions and that the errors are independent and identically distributed. The variance estimator is defined as the local polynomial regression estimator based on the squares of the pseudo-residuals. They show the optimal convergence rate of risk and a corresponding bandwidth. The asymptotic variance of the estimator is affected by the choice of the difference sequence, but the asymptotic bias is not affected. For bandwidth selection, Levine (2006) proposes a K -fold cross-validation type method.

Adopting a difference-based method without estimating the mean, a large-scale effect, suggests that there should be an effect of the smoothness of mean function to that of variance function. Hall and Carroll (1989) have discussed the effect of relative smoothness of mean function to the smoothness of variance function on the order of risk of a variance function estimator. Wang et al. (2008) have followed up with the minimax risk rate of convergence and have found that the rate is the same as in a nonparametric regression setting, whose convergence rate of risk of a functional estimator is $O(n^{-\beta/(2\beta+1)})$ in one-dimensional estimation, where β is the degree differentiability of the estimated function. If the degree differentiability of the mean function is less than $1/4$, then the convergence

rate of risk is larger than the common form.

As we follow the tradition of nonparametric estimation, we develop the method further and also contrast it with the likelihood approach. In Section 3.2, we define the local variogram as a product of the variance function of location and the traditional variogram function of lag. Then we explain the rationales behind some definitions. In Section 3.3, we discuss the theory and the method of local variogram function estimation and variance function estimation. In Section 3.4, the estimation algorithm is presented. In Section 3.5, we illustrate our method through a simulation study. In Section 3.6, I discuss the advantages of the difference-based variance function estimator in comparison to a likelihood-based estimator.

3.2 Model and Definition

Consider a nonstationary continuous process model

$$Z(s) = \mu(s) + \sigma(s)X_s \quad (3.1)$$

on $0 \leq s \leq 1$ with a smooth mean function $\mu(s)$ and an additive noise function as a product of a smooth standard deviation function $\sigma(s)$ and a second-order stationary error process $\{X_s\}_{s \in \mathcal{S}}$ where $E(X_s) = 0$, $\text{var}(X_s) = 1$, and $\text{cov}(X_s, X_{s'}) = \rho(|s - s'|; \theta)$ for $s \neq s'$. We assume that X_s is isotropic and the correlation function is defined as

$$\rho(|s - s'|; \theta) = \begin{cases} 1 & s = s' \\ 1 - \theta |s - s'|^\alpha + o(|s - s'|^2) & s \neq s' \end{cases} \quad (3.2)$$

where $\theta > 0$ and $0 < \alpha < 2$ for a valid correlation structure. There are several correlation function models that are readily available such as linear, spherical, Matérn and exponential, and these satisfy the condition (3.2). We assume a fixed equally spaced design such that $s_{i,n} = \frac{2i-1}{2n}$ where the location is indexed by $i = 1, \dots, n$. When we deal with a general n , we may drop the second index and express $s_{i,n} = s_i$. The following

shorthand is also used.

$$Z_i = Z(s_i) \quad \rho_{h,n} = \rho(h/n),$$

$$\mu_i = \mu(s_i), \quad \sigma_i = \sigma(s_i),$$

and for a parametric correlation function, we use $\rho_{h;\theta} = \rho(h/n; \theta)$.

Definition 3.2.1 Let $c_1, c_2 > 0$. Denote $q' \doteq q - \lfloor q \rfloor$ where $\lfloor q \rfloor$ is the largest integer less than q . We say that the function $f(x) \in \Lambda_q(c_f)$ if for all $x, y \in (0, 1)$, $|f^{(\lfloor q \rfloor)}(x) - f^{(\lfloor q \rfloor)}(y)| \leq c_1 |x - y|^{q'}$, $|f^{(k)}(x)| \leq c_2$ for $k = 0, \dots, \lfloor q \rfloor$, and $c_f = \max(c_1, c_2)$.

Definition 3.2.2 If a function $f(x)$ is in class $\Lambda_q(c_f)$ and there exists $\delta > 0$ such that $f(x) > \delta$ for all $x \in [0, 1]$, we say the function is in $\Lambda_q^+(c_f)$.

In this paper, we consider $\mu(s) \in \Lambda_q, q \geq 0$ and $\sigma^2(s) \in \Lambda_\beta^+, \beta \geq 2$, which are continuously differentiable Lipschitz functions.

We borrow the idea of a variogram, which is widely used in geostatistics, to one-dimensional nonstationary processes defined in (3.1). Mathernon (1962) introduced the term variogram for a second-order stationary random field $\{X_{\mathbf{s}}\}$ to represent $2\gamma(\|\mathbf{h}\|) = \text{var}(X(\mathbf{s} + \mathbf{h}) - X(\mathbf{s}))$ for any pair of observations separated by \mathbf{h} . In our data model the process is heteroscedastic. The differenced process also contains heteroscedasticity, and the central location of the pair of observations from which the differencing is taken contains information. For a fixed space design of n sample in one dimension, the variance of a simple order lag- h differenced process centered about s is,

$$\begin{aligned} & \text{var} \left(Z \left(s - \frac{h}{2n} \right) - Z \left(s + \frac{h}{2n} \right) \right) \\ &= 2\sigma^2(s) (1 - \rho_{h,n}) + 2(\sigma^{(1)}(s))^2 (1 + \rho_{h,n}) \left(\frac{h}{2n} \right)^2 + o(n^{-2}) \end{aligned} \quad (3.3)$$

when expanded about s where $\sigma^{2(j)}(s) = d^j \sigma^2(x)/dx^j|_{x=s}$. See equation (A.2) in Appendix A for details.

Definition 3.2.3 *The local variogram $2\gamma_L(s, h)$ is defined as the leading term of (3.3), i. e.*

$$\gamma_L(s, h) = \sigma^2(s)(1 - \rho(h/n)). \quad (3.4)$$

The local variogram (3.4) is a product of a heteroscedastic variance function and a variogram function where the lag size h is relatively small in comparison to n . Therefore, for variance function estimation at location s we need to estimate the local variogram at location s and lag h and the correlation structure from the data at lag h . We proceed by defining a differencing sequence to be used in the local variogram estimator.

Definition 3.2.4 *A one dimensional order l differencing filter has an l order binomial expansion coefficients $c_j = (-1)^j \binom{l}{j}$ as a coefficient for the $(j+1)^{th}$ term involved in the filter where $j = 0, \dots, l$. We define a squared order l difference process at lag- h as*

$$\{D_{i,h}^2\}_{i=1}^{n-hl} = \frac{\{(\sum_{j=0}^l c_j Z(s_{i+jh}))^2\}_{i=1}^{n-hl}}{\sum_{j=0}^l c_j^2}.$$

where $d_j = c_j / \sqrt{\sum_{j=0}^l c_j^2}$.

Remark For any positive integer l , $\sum_{j=0}^l d_j = 0$ and $\sum_{j=0}^l d_j^2 = 1$ since $\sum_{j=0}^l c_j = 0$. If $\{Z_i\}$ is an independent and identically distributed error process with mean 0, this implies the sequence of $\{D_{i,h}\}$ is an identically distributed error process with $E(D_{i,h}) = 0$ and $var(D_{i,h}) = E(D_{i,h}^2) = 1$ and not independent. Hence, in the literature, $D_{i,h}$ are often called pseudo-residuals.

In constructing a local variogram estimator, we use a lag- h , first order, normalized squared difference sequence of fixed-design data. That is, $\{D_{i,h}^2\} = \{(Z_i - Z_{i+h})^2 / 2\}_{i=1}^{n-h}$, which has a direction connection to the definition of a local variogram. The pseudo-residuals in the simplest terms offer the most compact form at fixed a lag and introduces the smallest bias in local variogram estimation among the class of lag- h difference filters. We suggest using lag-1 simple difference sequence because the variance of the squared lag-1 sequence is smaller than that of larger lags.

Definition 3.2.5 Let a kernel function $K(\cdot)$ be supported on $[-1, 1]$. It is called a kernel of order m if it satisfies the following four conditions:

1. $\int_{-1}^1 K(x)dx = 1$,
2. $\int_{-1}^1 K(x)x^i dx = 0$ for $i = 1, \dots, m-1$,
3. $\int_{-1}^1 K(x)x^m dx > 0$, and
4. $\int_{-1}^1 K^2(x)dx < \infty$.

Nonparametric smoothing often has its problems at the boundaries of estimating domain. Local polynomial regression with an odd degree kernel is a common solution to remove the boundary effect. Gasser et al. (1985) provides an asymmetric m -order kernel function that removes a boundary effect.

Definition 3.2.6 Let kernel $K^B(x)$ be a boundary kernel for a lower boundary $0 \leq s \leq \lambda$. For some $0 \leq b < 1$, $s = b\lambda$, and we require that

1. $\int_{-1}^b K^B(x)dx = 1$,
2. $\int_{-1}^b K^B(x)x^i dx = 0$ for $i = 1, \dots, m-1$,
3. $\int_{-1}^b K^B(x)x^i dx > 0$ for $i = m$, and
4. $\int_{-1}^b (K^B(x))^2 dx < \infty$.

Remark A boundary kernel for an upper boundary $1 - \lambda < s \leq 1$ has the limits from $-b$ to 1 satisfying the conditions 1 to 4 above as the upper boundary is $1 - s = b\lambda$ for $0 \leq b < 1$.

Definition 3.2.7 Let λ be a bandwidth parameter. Define an m -order Gasser-Müller kernel function $K_{\lambda,i}(s)$ (Gasser et al., 1985) for the i^{th} term weight as

$$K_{\lambda,i}(s) = \begin{cases} \int_{(s_i+s_{i-1})/2}^{(s_i+s_{i+1})/2} \frac{1}{\lambda} K\left(\frac{s-u}{\lambda}\right) du & s \in (\lambda, 1-\lambda) \\ \int_{(s_i+s_{i-1})/2}^{(s_i+s_{i+1})/2} \frac{1}{\lambda} K^B\left(\frac{s-u}{\lambda}\right) du & s \in [0, \lambda] \\ \int_{(s_i+s_{i-1})/2}^{(s_i+s_{i+1})/2} \frac{1}{\lambda} K^B\left(-\frac{s-u}{\lambda}\right) du & s \in [1-\lambda, 1] \end{cases}$$

where $0 < \lambda < 1/2$ and $0 \leq s \leq 1$ for $i = 2, \dots, n-2$.

Remark For $i = 1$, the limits of the integral of $K^B((s - u)/\lambda)$ are from 0 to $(s_1 + s_2)/2$; and for $i = n-1$, the limits are from $(s_{n-1} + s_n)/2$ to 1. For any $0 \leq s \leq 1$, $\sum_{i=1}^{n-1} K_{\lambda,i}(s) = 1$.

For fixed design data, a local polynomial regression and the above kernel adjust the boundary estimation problem at the same rate. However, the latter has an advantage of dealing with random design data and induces a cleaner asymptotic expansion. Gasser-Müller kernel smoothing also gives the same asymptotic properties as the local polynomial smoothing does (Fan and Gijbels (1992)).

3.3 Theoretical Results

3.3.1 Local variogram estimator

We define a Gasser-Müller kernel estimator of local variogram as

$$\hat{\gamma}_{L\lambda}(s, h) = \sum_{i=1}^{n-h} K_{\lambda, i+h/2}(s) D_{i,h}^2 \quad (3.5)$$

where $D_{i,h}^2$ is a simple normalized square difference of an observed process, and $K_{\lambda,i}$ is a Gasser-Müller kernel of order $\beta > 2$. Without loss of generality, assume that the domain of a variance function is from 0 to 1.

Remark Note that in the local variogram estimator (3.5) the i^{th} difference square, $D_{i,h}^2$, is associated with the Gasser-Müller kernel weight indexed by $i + h/2$. This index represents the kernel centering location, and it is aligned with the weight center of $D_{i,h}^2$. So, for example, when $h = 1$, then the kernel $K_{\lambda, i+1/2}$ integration limits are s_i and s_{i+1} . If the kernel weight had been $K_{\lambda,i}$, then the integration limits would have been $(s_{i-1} + s_i)/2$ and $(s_i + s_{i+1})/2$.

Let

$$D_{i,h}^2 = \frac{(Z_i - Z_{i+h})^2}{2}$$

$$\delta_{i,h} = \mu_i - \mu_{i+h} \text{ and}$$

$$g_{i,h} = \sigma_i^2 + \sigma_{i+h}^2 - 2\sigma_i\sigma_{i+h}\rho_{h,n}$$

for $i = 1, \dots, n-h$. As the data model is set up in (3.1), $E(D_{i,h}^2) = \frac{1}{2}(\delta_{i,h}^2 + g_i)$. For an asymptotic expansion of the local variogram estimator (3.5), we need the following results. Under the condition $\mu(\cdot) \in \Lambda_q, q \geq 1$, a Taylor expansion of $\delta_{i,h}$ about location s is

$$\begin{aligned} \delta_{i,h} &= \sum_{k=1}^{\lfloor q \rfloor} \frac{\mu_s^{(k)}}{k!} \{(s_i - s)^k - (s_{i+h} - s)^k\} + O(|s_i - s|^q + |s_{i+h} - s|^q) \\ &= -\frac{h}{n} \sum_{k=1}^{\lfloor q \rfloor} \frac{\mu_s^{(k)}}{k!} \sum_{a=0}^{k-1} (s_i - s)^a (s_{i+h} - s)^{k-1-a} + O(|s_i - s|^q + |s_{i+h} - s|^q). \end{aligned} \quad (3.6)$$

When $0 \leq q < 1$,

$$\delta_{i,h} = c \left(\frac{i}{n} \right)^q - c \left(\frac{i+h}{n} \right)^q = cn^{-q} \{i^q - (i+h)^q\} = O(n^{-q}). \quad (3.7)$$

As for $g_{i,h}$, we rewrite $g_{i,h} = \sigma_i(\sigma_i - \sigma_{i+h}\rho_{h,n}) + \sigma_{i+h}(\sigma_{i+h} - \sigma_i\rho_{h,n})$ and expand each two-term factor under the condition that $\sigma^2(\cdot) \in \Lambda_\beta$ for $\beta \geq 2$.

$$\begin{aligned} \sigma_i - \sigma_{i+h}\rho_{h,n} &= \sigma_i - \left(\sigma_i + \sigma_i^{(1)} \frac{h}{n} + \frac{\sigma_i^{(2)}}{2} \frac{h^2}{n^2} + o(n^{-2}) \right) \rho_{h,n} \\ &= \sigma_i (1 - \rho_{h,n}) - \sigma_i^{(1)} \frac{h}{n} \rho_{h,n} - \frac{\sigma_i^{(2)}}{2} \frac{h^2}{n^2} \rho_{h,n} + o(\rho_{h,n} n^{-2}), \\ \sigma_{i+h} - \sigma_i \rho_{h,n} &= \left(\sigma_i + \sigma_i^{(1)} \frac{h}{n} + \frac{\sigma_i^{(2)}}{2} \frac{h^2}{n^2} + o(n^{-2}) \right) - \sigma_i \rho_{h,n} \\ &= \sigma_i (1 - \rho_{h,n}) + \sigma_i^{(1)} \frac{h}{n} + \frac{\sigma_i^{(2)}}{2} \frac{h^2}{n^2} + o(n^{-2}). \end{aligned}$$

Then, we see the $2\gamma_L = 2\sigma_i^2(1 - \rho_{h,n})$ appearing in the leading term of the expansion of $g_{i,h}$ about s_i .

$$g_{i,h} = \sigma_i^2 + \sigma_{i+h}^2 - 2\sigma_i\sigma_{i+h}\rho_{h,n}$$

$$\begin{aligned}
&= \sigma_i(\sigma_i - \sigma_{i+h}\rho_{h,n}) + \sigma_{i+h}(\sigma_{i+h} - \sigma_i\rho_{h,n}) \\
&= \sigma_i^2(1 - \rho_{h,n}) - \sigma_i\sigma_i^{(1)}\frac{h}{n}\rho_{h,n} - \frac{\sigma_i\sigma_i^{(2)}}{2}\frac{h^2}{n^2}\rho_{h,n} + o(\rho_{h,n}n^{-2}) \\
&\quad + \left(\sigma_i + \sigma_i^{(1)}\frac{h}{n} + \frac{\sigma_i^{(2)}}{2}\frac{h^2}{n^2} + o(n^{-2}) \right) \left\{ \sigma_i(1 - \rho_{h,n}) + \sigma_i^{(1)}\frac{h}{n} + \frac{\sigma_i^{(2)}}{2}\frac{h^2}{n^2} + o(n^{-2}) \right\} \\
&= \sigma_i^2(1 - \rho_{h,n}) - \sigma_i\frac{h}{n} \left(\sigma_i^{(1)} + \frac{\sigma_i^{(2)}}{2}\frac{h}{n} \right) \rho_{h,n} + o(\rho_{h,n}n^{-2}) \\
&\quad + \sigma_i^2(1 - \rho_{h,n}) + \sigma_i\frac{h}{n} \left(\sigma_i^{(1)} + \frac{\sigma_i^{(2)}}{2}\frac{h}{n} \right) (2 - \rho_{h,n}) + \left(\sigma_i^{(1)} \right)^2 \frac{h^2}{n^2} + o(n^{-2}) \\
&= 2(1 - \rho_{h,n}) \left(\sigma_i^2 + \sigma_i\sigma_i^{(1)}\frac{h}{n} + \sigma_i\sigma_i^{(2)}\frac{h^2}{n^2} \right) + \left(\sigma_i^{(1)} \right)^2 \frac{h^2}{n^2} + o(n^{-2})
\end{aligned}$$

A Taylor expansion of $g_{i,h}$ about location s is

$$\begin{aligned}
g_{i,h} &= 2(1 - \rho_{h,n}) \left(\sigma_s^2 + \sigma_s\sigma_s^{(1)}\frac{h}{n} + \frac{\sigma_s\sigma_s^{(2)}}{2}\frac{h^2}{n^2} \right) + \left(\sigma_s^{(1)}\frac{h}{n} \right)^2 \\
&\quad + 2(1 - \rho_{h,n}) \sum_{j=1}^{\lfloor \beta \rfloor} \left\{ \frac{(\sigma_s^{(2)})^{(j)}}{j!} + \frac{(\sigma_s^{(2)})^{(j+1)}}{2(j+1)!} \left(1 + \frac{h}{n} \right) \frac{h}{n} \right\} (s_i - s)^j \\
&\quad + \frac{h^2}{n^2} \sum_{k=1}^{\lfloor \beta \rfloor} \sum_{j=1}^k c_{j,k} \sigma_s^{(j)} \sigma_s^{(k-j+2)} (s_i - s)^k + O(|s_i - s|^\beta)
\end{aligned} \tag{3.8}$$

where c_k is a constant that is independent of n .

Here, we mention asymptotic properties of the Gasser-Müller kernel. Note that

$$\begin{aligned}
\sum_{i=1}^{n-1} K_{\lambda,i}(s) &= \sum_{i=1}^{n-1} \int_{s_{i-1/2}}^{s_{i+1/2}} \frac{1}{\lambda} K\left(\frac{s-u}{\lambda}\right) du = 1 \\
&= \sum_{i=\lfloor ns-\lambda-1 \rfloor}^{\lfloor ns+\lambda-1 \rfloor} \int_{s_{i-1/2}}^{s_{i+1/2}} \frac{1}{\lambda} K\left(\frac{s-u}{\lambda}\right) du.
\end{aligned}$$

This implies

$$K_{\lambda,i}(s) = O\left(\frac{1}{n\lambda}\right). \tag{3.9}$$

Using the above fact, the sum of quadratic terms of the kernel is

$$\sum_{i=1}^{n-h} K_{\lambda,i+\frac{h}{2}}^2(s) = O(n\lambda) O\left(\frac{1}{(n\lambda)^2}\right) = O\left(\frac{1}{n\lambda}\right), \tag{3.10}$$

and

$$\sum_{i=1}^{(n-h-2)} \sum_{j>i}^{(n-h-1)} K_{\lambda, i+\frac{h}{2}}^2(s) K_{\lambda, j+\frac{h}{2}}^2(s) = O((n\lambda)^2) O\left(\frac{1}{(n\lambda)^2}\right) = O(1). \quad (3.11)$$

When the higher order terms in the expansions of $\delta_{i,h}$ and $g_{i,h}$, respectively in equations (3.6) and (3.8), are convolved with a Gasser-Müller kernel of order m ,

$$\begin{aligned} \sum_{i=1}^{n-h} K_{\lambda, i+\frac{h}{2}}(s) (s_i - s)^j &= \sum_{i=1}^{n-h} \int_{(s_{i+\frac{h}{2}}+s_{i+\frac{h}{2}-1})/2}^{(s_{i+\frac{h}{2}}+s_{i+\frac{h}{2}+1})/2} \frac{1}{\lambda} K\left(\frac{s-x}{\lambda}\right) dx (s_{i+\frac{h}{2}} - s)^j \\ &= \sum_{i=1}^{n-h} \int_{s_{i+(h-1)/2}}^{s_{i+(h+1)/2}} \frac{1}{\lambda} K\left(\frac{s-x}{\lambda}\right) \left\{ (s_{i+\frac{h}{2}} - s)^j - (x-s)^j \right\} dx \\ &= \sum_{i=1}^{n-h} \int_{s_{i+(h-1)/2}}^{s_{i+(h+1)/2}} \frac{1}{\lambda} K\left(\frac{s-x}{\lambda}\right) j \xi_i^{j-1} (s_{i+\frac{h}{2}} - x) dx, \end{aligned}$$

and the Mean Value theorem is used in the last equality where $\xi_i + s$ is in the interval $(s_{i+(h-1)/2}, s_{i+(h+1)/2})$. Let $u = (s-x)/\lambda$ and $u_i = (s-s_{i+h/2})/\lambda$.

$$\begin{aligned} \left| \sum_{i=1}^{n-h} \int_{s_{i+(h-1)/2}}^{s_{i+(h+1)/2}} \frac{1}{\lambda} K\left(\frac{s-x}{\lambda}\right) j \xi_i^{j-1} (s_{i+\frac{h}{2}} - x) dx \right| &\leq \sum_{i=1}^{n-h} \frac{j |\xi_i|^{j-1}}{n} \left| \int_{u_i}^{u_{i+1}} K(u) du \right| \\ &\leq \frac{j}{n} = O(n^{-1}). \end{aligned}$$

By the property of a Gasser-Müller kernel of order m , the kernel keeps the terms in the asymptotic expansions of $\delta_{i,h}$ and $g_{i,h}$ with factors $(s_i - s)^j$ for $j = 1, \dots, m-1$ at $O(n^{-1})$. Applying a Gasser-Müller kernel of order m to $(s_i - s)^\beta$ where $\beta \geq m$ in the expansion of $\delta_{i,h}^2$ and σ_i^2 about s ,

$$\begin{aligned} \sum_{i=1}^{n-h} K_{\lambda, i+\frac{h}{2}}(s) |s_{i+\frac{h}{2}} - s|^m &\leq \sum_{i=\lfloor n(s-\lambda) \rfloor}^{\lfloor n(s+\lambda) \rfloor + 1} \left| K_{\lambda, i+\frac{h}{2}}(s) \right| \left| s_{i+\frac{h}{2}} - s \right|^m \\ &\leq \sum_{i=\lfloor n(s-\lambda) \rfloor}^{\lfloor n(s+\lambda) \rfloor + 1} \left| K_{\lambda, i+\frac{h}{2}}(s) \right| \left(\lambda + \frac{1}{n} \right)^m \\ &= O(\lambda^m). \end{aligned}$$

3.3.2 Bias of the estimator

The expected value of the local variogram estimator is

$$\begin{aligned} E(\hat{\gamma}_{L\lambda}(s, h)) &= \sum_{i=1}^{n-h} K_{\lambda, i+\frac{h}{2}}(s) E(D_{i,h}^2) \\ &= \frac{1}{2} \sum_{i=1}^{n-h} K_{\lambda, i+\frac{h}{2}}(s) \{(\mu_i - \mu_{i+h})^2 + \sigma_i^2 + \sigma_{i+h}^2 - 2\sigma_i\sigma_{i+h}\rho_{h,n}\}. \end{aligned}$$

The bias of the local variogram estimator is

$$\begin{aligned} \text{bias}(\hat{\gamma}_{\lambda}(s, h)) &= E(\hat{\gamma}_{\lambda}(s, h)) - (1 - \rho_{h,n})\sigma^2(s) \\ &= \sum_{i=1}^{n-h} K_{\lambda, i+\frac{h}{2}}(s) \left\{ \frac{1}{2}(\delta_{i,h}^2 + g_{i,h}) - (1 - \rho_{h,n})\sigma^2(s) \right\}. \end{aligned} \quad (3.12)$$

Note that $(1 - \rho_h) = O(n^{-\alpha})$ and $0 < \alpha < 2$.

Theorem 3.3.1 *Assume a data model (3.1) and (3.2). The process functions $\mu(s)$ and $\sigma^2(s)$ are continuously differentiable Lipschitz functions (see Definitions 3.2.1 and 3.2.2) where $\mu(s) \in \Lambda_q, q \geq 0$ and $\sigma^2(s) \in \Lambda_{\beta}^+, \beta \geq 2$. The difference-based local variogram m -order Gasser-Müller kernel estimator (3.5) at location s and lag h has an asymptotic bias of order*

$$\text{bias}(\hat{\gamma}_{\lambda}(s, h)) = \begin{cases} O(n^{-2} + n^{-2q} + n^{-\alpha-1}) & \text{where } q, \beta < m \\ O(n^{-2} + n^{-2q} + n^{-\alpha-1}) + O(n^{-\alpha}\lambda^m) & \text{where } q < m \leq \beta \\ O(n^{-2} + n^{-2q} + n^{-\alpha-1}) + O(\lambda^m) & \text{where } m \leq q. \end{cases} \quad (3.13)$$

Proof To calculate an asymptotic bias we split (3.12) into two parts. The first term is $\delta_{i,h}^2$ whose expansion is in (3.6) for $q \geq 1$ and in (3.7) for $0 < q < 1$. Convolved with a Gasser-Müller kernel of order m (see Definition 3.2.5 - 3.2.7), the higher order terms in $\delta_{i,h}^2$ is cancelled when the number of derivatives of the mean function $q \leq m$.

$$\sum_{i=1}^{n-h} K_{\lambda, i+\frac{h}{2}}(s) \delta_{i,h}^2 = \begin{cases} O(n^{-2}) + O(n^{-2q}) & \text{where } q < m \\ O(n^{-2}) + O(n^{-2q}) + O(\lambda^m) & \text{where } q \geq m. \end{cases} \quad (3.14)$$

The second part of the bias is $\frac{1}{2}g_{i,h} - \sigma^2(s)(1 - \rho_{h,n})$. In equation (3.8), the leading term in $g_{i,h}$ expansion about s is the local variogram $\sigma^2(s)(1 - \rho_{h,n})$. Applying a Gasser-Müller kernel to the remaining high order terms in (3.8), we get the following:

$$\begin{aligned}
& \sum_{i=1}^{n-h} K_{\lambda, i+\frac{h}{2}}(s) \left\{ \frac{1}{2}g_{i,h} - \sigma^2(s)(1 - \rho_{h,n}) \right\} \\
&= \sum_{i=1}^{n-h} K_{\lambda, i+\frac{h}{2}}(s) \left\{ (1 - \rho_{h,n}) \left(\sigma_s \sigma_s^{(1)} \frac{h}{n} + \frac{\sigma_s \sigma_s^{(2)} h^2}{2 n^2} \right) + \frac{1}{2} \left(\sigma_s^{(1)} \frac{h}{n} \right)^2 \right\} \\
&+ \sum_{i=1}^{n-h} K_{\lambda, i+\frac{h}{2}}(s) (1 - \rho_{h,n}) \sum_{j=1}^{\lfloor \beta \rfloor} \left\{ \frac{(\sigma_s^2)^{(j)}}{j!} + \frac{(\sigma_s^2)^{(j+1)}}{2(j+1)!} \left(1 + \frac{h}{n} \right) \frac{h}{n} \right\} (s_i - s)^j \\
&+ \sum_{i=1}^{n-h} K_{\lambda, i+\frac{h}{2}}(s) \frac{h^2}{2n^2} \sum_{k=1}^{\lfloor \beta \rfloor} \sum_{j=1}^{k+1} c_k \sigma_s^{(j)} \sigma_s^{(k-j+2)} (s_i - s)^k + \sum_{i=1}^{n-h} K_{\lambda, i+\frac{h}{2}}(s) O(|s_i - s|^\beta) \\
&= \begin{cases} O(n^{-\alpha-1}) + O(n^{-2}) & \text{where } \beta < m \\ O(n^{-\alpha-1}) + O(n^{-2}) + O(n^{-\alpha} \lambda^m) & \text{where } \beta \geq m. \end{cases} \tag{3.15}
\end{aligned}$$

Combining derivations in (3.14) and (3.15), the bias is summarized in (3.13). \square

Remark The asymptotic bias has an order dependent on the data smoothness parameter α and the degree differentiability of the mean and variance functions, which are q and β respectively in comparison to the order m of the kernel. Assume that the first case is true, i.e. the order of kernel is greater than the degree of differentiability of both mean and variance functions. Then, when $\alpha < 1$ and $\frac{\alpha+1}{2} < q \leq 1$, the bias is in the order of $n^{-\alpha-1}$; when $\alpha < 1$ and $2q \leq \alpha + 12$, the bias is in the order of n^{-2q} ; and when $\alpha \geq 1$ and $q \geq 1$, the order of bias is n^{-2} . The setting of $\alpha < 1$ translates to the data process being less smooth than a process with an exponential correlation structure whose $\alpha = 1$. The first two settings indicate a rough error process with a less smooth mean process, while the latter setting suggests a smooth error process with the mean function with at least one derivative. Assume that the second case is true and that $\lambda = O(n^{-x})$ where $0 < x < 1$. Then $O(n^{-\alpha} \lambda^m)$ is the order of bias in the following three settings: (1) $1 \leq q$, $\alpha \leq 1$, and $x < 1/m$; (2) $q \geq 1$, $\alpha \geq 1$, and $x < (2 - \alpha)/m$; and $\alpha < 1$, $2q < \alpha + 1$,

and $x < (2q - \alpha)/m$. All other settings should be referred to the first case. Assume that the third case is true. Then the bias is $O(\lambda^m)$ in the following three settings: (1) $q \geq 1$, $\alpha \geq 1$, and $2/m > x$; (2) $q < 1$, $2q < \alpha + 1$, and $x < 2q/m$; (3) $\alpha < 1$, $\alpha + 1 < 2q$ and $x < (\alpha + 1)/m$.

Roughly speaking, the greater the order of the kernel is (or as long as the order of the kernel is greater than q and β), the smaller the asymptotic bias term is. In reality, we do not know q and β in advance, but it is still better to choose a high order kernel function.

3.3.3 Variance of the estimator

The variance of the local variogram estimator at location s and lag h is

$$\text{var}(\hat{\gamma}_\lambda(s, h)) = \sum_{i=1}^{n-h} \sum_{j=1}^{n-h} K_{\lambda, i+\frac{h}{2}}(s) K_{\lambda, j+\frac{h}{2}}(s) \text{cov}(D_{i,h}^2, D_{j,h}^2). \quad (3.16)$$

Recall $D_{i,h} = (\delta_i + \sigma_i X_i - \sigma_{i+h} X_{i+h}) / \sqrt{2}$ where X_i is a Gaussian process with mean 0, variance 1, and a fixed correlation function $\rho_\theta(h) = \text{cov}(X_i, X_{i+h})$. Then, we have $(\sigma_i X_i - \sigma_{i+h} X_{i+h})$ distributed $Normal(0, g_{i,h})$ and $E(\sigma_i X_i - \sigma_{i+h} X_{i+h})^4 = 3g_{i,h}^2$. For the sake of simplicity in notation, from here on we use g_i for $g_{i,h}$, δ_i in place of $\delta_{i,h}$, and ρ_h for $\rho_{h,n}$.

$$\begin{aligned} \text{var}(D_{i,h}^2) &= E(D_{i,h}^4) - E^2(D_{i,h}^2) \\ &= \frac{1}{4} \left\{ \delta_i^4 + 6\delta_i^2 g_i + 3g_i^2 - (\delta_i^2 + g_i)^2 \right\} \\ &= \delta_i^2 g_i + \frac{1}{2} g_i^2 \end{aligned}$$

The covariance between the normalized and squared differences centered at location $s_{i+h/2}$ and $s_{j+h/2}$ is

$$\begin{aligned} \text{cov}(D_{i,h}^2, D_{j,h}^2) &= \frac{1}{4} \left\{ E((Z_i - Z_{i+h})^2 (Z_j - Z_{j+h})^2) - (\delta_i^2 + g_i)(\delta_j^2 + g_j) \right\} \\ &= \delta_i \delta_j \{ \rho_{|i-j|}(\sigma_i \sigma_j + \sigma_{i+h} \sigma_{j+h}) - \rho_{|i-j-h|} \sigma_i \sigma_{j+h} - \rho_{|i-j+h|} \sigma_{i+h} \sigma_j \} \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{2} \{ (\rho_{|i-j|} \sigma_i \sigma_j - \rho_{|i-j-h|} \sigma_i \sigma_{j+h})^2 + (\rho_{|i-j+h|} \sigma_{i+h} \sigma_j - \rho_{|i-j|} \sigma_{i+h} \sigma_{j+h})^2 \} \\
& + (\rho_{|i-j|}^2 + \rho_{|i-j-h|} \rho_{|i-j+h|}) \sigma_i \sigma_{i+h} \sigma_j \sigma_{j+h} - \rho_{|i-j|} \sigma_i \sigma_{i+h} (\rho_{|i-j+h|} \sigma_j^2 + \rho_{|i-j-h|} \sigma_{j+h}^2) \\
& = \delta_i \delta_j P_{ij} + \frac{1}{2} P_{ij}^2.
\end{aligned}$$

where $P_{ij} = \rho_{|i-j|} (\sigma_i \sigma_j + \sigma_{i+h} \sigma_{j+h}) - \rho_{|i-j-h|} \sigma_i \sigma_{j+h} - \rho_{|i-j+h|} \sigma_{i+h} \sigma_j$ for $i \neq j$. Note that when $i = j$, the expression $P_{ij} = g_{i,h}$. The Taylor expansion of $P_{i,j}$ about s_i for any $i \neq j$ is

$$P_{ij} = \frac{h^2}{n^2} \left(\sigma_i^{(1)} \right)^2 - \frac{2h^2}{(n\theta)^2} \sigma_i^2 + o(n^{-3}). \quad (3.17)$$

See Appendix A (A.3) for the derivation. We are interested in the asymptotic rate of convergence of (3.16).

Theorem 3.3.2 *Assume the same conditions as in Theorem 3.3.1. The asymptotic variance of local variogram estimator $\hat{\gamma}_{L,\lambda}$ in 3.5 is of the order*

$$\text{var}(\hat{\gamma}_\lambda(s, h)) = O\left(\frac{1}{n\lambda}\right) O(n^{-2q-\alpha} + n^{-2\alpha}). \quad (3.18)$$

Proof By plugging in (3.6), (3.8), and (3.17), a Taylor expansion of the local variogram estimator about s and lag h in (3.16) is

$$\begin{aligned}
\text{var}(\hat{\gamma}_\lambda(s, h)) &= \sum_{i=1}^{n-h} K_{\lambda, i+\frac{h}{2}}^2(s) \left(\delta_i^2 g_i + \frac{g_i^2}{2} \right) + 2 \sum_{i>j=1}^{n-h-1} K_{\lambda, i+\frac{h}{2}}(s) K_{\lambda, j+\frac{h}{2}}(s) \left(\delta_i \delta_j P_{ij} + \frac{P_{ij}^2}{2} \right) \\
&= 2 \sum_{i=1}^{n-h} K_{\lambda, i+\frac{h}{2}}^2(s) \{ \delta_i^2 (1 - \rho_h) O(1) + (1 - \rho_h)^2 O(1) \} \\
&\quad + 2 \sum_{i>j=1}^{n-h-1} K_{\lambda, i+\frac{h}{2}}^2(s) K_{\lambda, j+\frac{h}{2}}^2(s) \{ \delta_i \delta_j O(n^{-2}) + O(n^{-4}) \} \\
&= 2(1 - \rho_h) \sum_{i=1}^{n-h} K_{\lambda, i+\frac{h}{2}}^2(s) \{ O(n^{-2} + n^{-2q}) + (1 - \rho_h) O(1) \} \\
&\quad + 2 \sum_{i>j=1}^{n-h-1} K_{\lambda, i+\frac{h}{2}}(s) K_{\lambda, j+\frac{h}{2}}(s) O(n^{-4}) \quad (3.19)
\end{aligned}$$

Note that $P_{ij} = O(n^{-2})$. Using the results in (3.10) and (3.11) in (3.19), we have (3.18).

□

Let us define

$$\eta_h(i, j) = 2\rho(|i - j|) - \rho(|i - j - h|) - \rho(|i - j + h|). \quad (3.20)$$

If $\sigma(\cdot) = 1$ a constant function, then $P_{ij} = \eta_h(i, j)$. More generally,

$$\eta_h(i, i + k) = 2\rho(|k|) - \rho(|k + h|) - \rho(|k - h|).$$

Assume that the correlation function is exponential, i.e. $\rho_{h,\theta} = \exp\left(-\frac{h}{n\theta}\right)$. Then, when

$k \geq h$:

$$\begin{aligned} \eta_h(i, i + k) &= \exp\left(-\frac{k}{n\theta}\right) \left(2 - \exp\left(-\frac{h}{n\theta}\right) - \exp\left(\frac{h}{n\theta}\right)\right) \\ &= -2 \exp\left(-\frac{k}{n\theta}\right) \sum_{i=1}^{\infty} \frac{1}{(2i)!} \left(\frac{h}{n\theta}\right)^{2i} = o(n^{-2}); \end{aligned}$$

and when $k < h$:

$$\begin{aligned} \eta_h(i, i + k) &= 2 \exp\left(-\frac{k}{n\theta}\right) - \left(\exp\left(-\frac{h-k}{n\theta}\right) - \exp\left(-\frac{h+k}{n\theta}\right)\right) \\ &= 2 \exp\left(-\frac{k}{n\theta}\right) - \exp\left(-\frac{h}{n\theta}\right) \left(\exp\left(\frac{k}{n\theta}\right) + \exp\left(-\frac{k}{n\theta}\right)\right) \\ &= 2 \exp\left(-\frac{k}{n\theta}\right) - \exp\left(-\frac{h}{n\theta}\right) \sum_{i=0}^{\infty} \frac{2}{(2i)!} \left(\frac{k}{n\theta}\right)^{2i} \\ &= -2 \left\{ \sum_{i=1}^{\infty} \frac{1}{(2i-1)!} \left(\frac{k}{n\theta}\right)^{2i-1} + \sum_{i=1}^{\infty} \frac{1}{i!} \left(-\frac{h}{n\theta}\right)^i \sum_{i=0}^{\infty} \frac{1}{(2i)!} \left(\frac{k}{n\theta}\right)^{2i} \right\} \\ &= o(n^{-1}). \end{aligned}$$

We use the shorthand notation η_h for $\eta_h(i, j)$ when it is clear from the context which two indices are picked for correlation measurement.

Remark Let the index difference match the lag size, i.e. $|i - j| = h$. When the underlying process is correlated via the exponential correlation function, we have $\eta_h = 2\rho_h - \rho_0 - \rho_{2h} = 2\rho_h - 1 - \rho_h^2 = -(1 - \rho_h)^2$. When the underlying process is independent, we have $\eta_h = 2\rho_h - \rho_0 - \rho_{2h} = -1$. Hence, for a stationary process the correlation between D_i^2 and D_{i+1}^2 is stronger when the process is independent rather than when the process is correlated.

We define $\dot{\mu}_\delta(i, j) = \delta_i \delta_j n^2 / h^2$. The correlation between $D_{i,h}^2$ and $D_{j,h}^2$ of a lag- h nonstationary difference squared process is

$$\begin{aligned}
& \text{Cor}(D_{i,h}^2, D_{j,h}^2) \\
&= \frac{\text{Cov}(D_{i,h}^2, D_{j,h}^2)}{\sqrt{\text{var}(D_{i,h}^2) \text{var}(D_{j,h}^2)}} \\
&= \frac{\delta_i \delta_j P_{ij} + \frac{1}{2} P_{ij}^2}{\sqrt{(\delta_i^2 g_i + \frac{1}{2} g_i^2)(\delta_j^2 g_j + \frac{1}{2} g_j^2)}} \\
&= \frac{\frac{h^4}{n^4} \left[\frac{2}{\theta^2} \sigma_i^2 \left\{ \frac{\sigma_i^2}{\theta^2} - \left(\sigma_i^{(1)} \right)^2 - \dot{\mu}_\delta(i, j) \right\} + \left\{ \dot{\mu}_\delta(i, j) + \frac{1}{2} \left(\sigma_i^{(1)} \right)^2 \right\} \left(\sigma_i^{(1)} \right)^2 + o(n^{-1}) \right]}{\sqrt{(\delta_i^2 g_i + \frac{1}{2} g_i^2) (\delta_j^2 g_j + \frac{1}{2} g_j^2)}} \\
&= \frac{O(n^{-4})}{O(n^{-2(2-\alpha)})} = O(n^{-2(2-\alpha)}).
\end{aligned}$$

Note that the correlation between the squared pseudo-residuals $D_{i,h}^2$ and $D_{j,h}^2$ converges asymptotically to 0 for $i \neq j$. With the infill asymptotic the differencing not only removes the feature of a mean function but also drastically reduces the correlated nature of the data. Assuming that $\dot{\mu}_\delta(i, j)$ is negligible, which comes from δ_i and δ_j being negligible, or in other words $\mu(s) \in \Lambda_\epsilon$ where ϵ is small, the third line of equality above is reduced to

$$\text{Cor}(D_{i,h}^2, D_{j,h}^2) = \frac{h^4 \left\{ \frac{2\sigma_i^2}{\theta^2} - \left(\sigma_i^{(1)} \right)^2 \right\}^2 + o(n^{-1})}{n^4 g_i g_j} = \frac{P_{ij}^2}{g_i g_j}.$$

The above is trivially true when $\mu(\cdot)$ is constant. The asymptotic rate of convergence for the correlation is $O(n^{-2(2-\alpha)})$ whether the mean is constant or smoothly varying.

3.3.4 Risk of Local Variogram Estimator

A point-wise risk of the local variogram estimator is the sum of the squared bias in (3.12) and variance in (3.19). The asymptotic point-wise risk, using the results in

equations (3.13) and (3.18), is:

$$\begin{aligned}
& Risk(\hat{\gamma}_\lambda(s, h), \gamma(s, h)) \\
&= bias(\hat{\gamma}_\lambda(s, h))^2 + var(\hat{\gamma}_\lambda(s, h)) \\
&= \begin{cases} O(n^{-4} + n^{-4q} + n^{-2\alpha-2}) + O\left(\frac{1}{n\lambda}\right) O(n^{-2\alpha} + n^{-2q-\alpha}) & \text{where } m > q, \beta \\ O(n^{-4} + n^{-4q} + n^{-2\alpha-2} + n^{-2\alpha}\lambda^{2m}) + O\left(\frac{1}{n\lambda}\right) O(n^{-2\alpha} + n^{-2q-\alpha}) & \text{where } \beta \geq m > q, \\ O(n^{-4} + n^{-4q} + n^{-2\alpha-2} + \lambda^{2m}) + O\left(\frac{1}{n\lambda}\right) O(n^{-2\alpha} + n^{-2q-\alpha}) & \text{where } q \geq m. \end{cases}
\end{aligned} \tag{3.21}$$

Theorem 3.3.3 Consider a one-dimensional nonstationary process local variogram estimation problem described in Section 3.2 with a data model (3.1) and (3.2) and the local variogram estimator described as in (3.5). We assume that $\mu(s) \in \Lambda_q, q \geq 0$, $\sigma^2(s) \in \Lambda_\beta, \beta \geq 2$, $\rho_\theta(h) = 1 - \theta(h/n)^\alpha + o((h/n)^2)$ for $0 < \alpha < 2$, and that the bandwidth $\lambda = O(n^{-x})$ where $0 < x < 1$.

When the order of Gasser-Müller kernel m is greater than both q and β , the point-wise risk of the Gasser-Müller kernel estimator of local variogram and the asymptotic convergence rate of bandwidth are

$$Risk(\hat{\gamma}_\lambda(s, h)) = \begin{cases} O(n^{-4q}) & \text{where } \lambda \asymp n^{-1-\alpha+2q} \\ O(n^{-4}) & \text{where } \lambda \asymp n^{3-2\alpha} \end{cases} \tag{3.22}$$

given $\alpha < 2q < \min(\alpha + \frac{1}{2}, 2)$ and given $q \geq 1$ and $\alpha > \frac{3}{2}$, respectively.

When the order of Gasser-Müller kernel m is greater than either $q > 1$ or β , the

point-wise risk of the Gasser-Müller kernel estimator of the local variogram is

$$\text{Risk}(\hat{\gamma}_\lambda(s, h)) = \begin{cases} O(n^{-2m(1+2\alpha)/(1+2m)}) & \text{where } \lambda \asymp n^{-(1+2\alpha)/(1+2m)} \\ O(n^{-2\alpha-2m/(1+2m)}) & \text{where } \lambda \asymp n^{-1/(1+2m)} \end{cases} \quad (3.23)$$

where $\alpha < \min\left(2q, \frac{3}{2}\right)$ and $\alpha < 2q$ respectively.

Proof We assume $m > q, \beta$. Then, in terms of the asymptotic risk, we are concerned with the first case in (3.21). Let us assume that $q \geq 1$. Then, the asymptotic order of variance in (3.18) should be $O(n^{-2\alpha-1}\lambda^{-1})$ as $2\alpha < 2q + \alpha$ as for $q \geq 1$. When $\alpha \geq 1$, the corresponding asymptotic order of bias in (3.13) is $O(n^{-2})$, and when $\alpha < 1$, the asymptotic order of bias is $O(n^{-\alpha-1})$. The latter results in an unsuitable rate of a bandwidth λ , but the former gives $\lambda \asymp n^{3-2\alpha}$. The order of a bandwidth is assumed to be $\lambda = O(n^{-x})$ where $0 < x < 1$, hence $\alpha > 3/2$.

Assume that $q < 1$. When $\alpha \geq 2q - 1$, the asymptotic order of bias is $O(n^{-2q})$. Or else, it is $O(n^{-\alpha-1})$. The latter case has been taken care of above, so we concentrate on the former scenario. When $\alpha \geq 2q - 1$, the asymptotic order of variance is $O(n^{-\alpha-2q-1}\lambda^{-1})$. Equating the doubled order of bias and the order of variance gives $\lambda \asymp n^{-1-\alpha+2q}$.

When $q < m \leq \beta$, we focus in on line 2 of (3.21), where the lowest order term of the squared asymptotic bias is $O(n^{-2\alpha}\lambda^{2m})$. When $m \leq q$ as in line 3 of (3.21), the lowest order term of bias is $O(\lambda^{2m})$. In both cases, the asymptotic order of variance is $O(n^{-2\alpha-1}\lambda^{-1})$ where $\alpha < 2q$ as we assume $q > 1$.

In the former case, there is a condition for the kernel order m , i.e. $\frac{m}{1+2m} < 2 - \alpha \Rightarrow m > \frac{\alpha-2}{3-2\alpha} = \frac{1}{2} \left(\frac{1}{2\alpha-3} - 1 \right)$, which is not a restriction at the end as long as $q < m \leq \beta$. Here, the latter case, we need to check (i) $\frac{m(1+2\alpha)}{1+2m} < 2$ and (ii) $\frac{m(1+2\alpha)}{1+2m} < 1 + \alpha$. Since $\frac{2+2\alpha}{2m} > \frac{1+2\alpha}{2m} > \frac{1+2\alpha}{1+2m}$, (ii) holds true for any valid α

and m . As for (i) $\Rightarrow m < \frac{2}{2\alpha - 3}$, and for any positive $2\alpha - 3$, we could find m that satisfies (i). Therefore, we have $\alpha > \frac{3}{2}$. \square

Remark In the bottom half of Theorem 3.3.3, the rate of risk $O(n^{-2m(1+2\alpha)/(1+2m)})$ and $O(n^{-2\alpha-2m/(1+2m)})$ are, in fact, very similar. In both cases, the implied range of the smoothness parameter is $\alpha \lesssim 3/2$.

Remark There is no convergence of risk when $q \geq \beta$ and the process is very smooth with $\alpha \gtrsim 3/2$ since the variance function is masked by the mean process.

Given that we set $m = \beta$, as $\alpha \rightarrow 0$ (a process becoming less smooth and independent), the risk converges to $O(n^{-2\beta/(1+2\beta)})$ in all three cases, which is consistent with the nonparametric literature. As the Lipschitz differentiability of a variance function increases, i.e. $\beta \uparrow$, a larger bandwidth is preferred, which is consistent with the results in the nonparametric literature. The greater the degree differentiability, or Lipschitz differentiability, a mean function has than that of a variance function, the smaller the smoothing bandwidth. This is because the locally changing variance information is sufficiently retrieved from a small scale neighborhood.

3.4 Algorithm and Bandwidth Selection

We are interested in estimating the variance function embedded in a nonstationary spatial process where the mean and the variance functions are smooth and the standardized spatial process is isotropic. The estimation of local variogram function at a given location is formed by smoothing a squared lag- h difference process using a Gasser-Müller kernel. We use a high order kernel to keep the bias small and perform cross-validation to select an appropriate bandwidth.

Here is the algorithm.

1. Fix lag size $h = 1$ and create a set of J bandwidths $\{\lambda_j\}_{j=1}^J$ between 0 and $1/2$.
For each λ_j and $h = 1$ the local variogram estimation $\hat{\gamma}_{L,\lambda_j}(s, h)$ is obtained as in (3.5) for any location in $[0, 1]$.
2. Select bandwidth via cross-validation (see below for the details of new notations):

$$\hat{\lambda} \leftarrow \arg_{\lambda} \min_{\lambda_j} \sum_{i=1}^{n-h} \left(\frac{dev_{\rho^{-1}}^2(s_i)}{1 - M\left(i + \frac{h}{2}, i + \frac{h}{2}\right)} \right)$$

3. $\{Z_i^*\}_{i=1}^n \leftarrow \frac{\{Z_i\}_{i=1}^n}{\sqrt{\hat{\gamma}_{\hat{\lambda}_j}(s_i, h)}}$
4. Select an appropriate correlation model for $\{Z_i^*\}$ and estimate its parameters Θ .
5. $\hat{\sigma}^2(s) \leftarrow \frac{\hat{\gamma}_{L,\hat{\lambda}_j}(s; h)}{1 - \hat{\rho}(h; \hat{\Theta})}$

Algorithm 1: Variance Function Estimation at a Point

In step 2 of Algorithm 1, $dev_{\rho^{-1}}(s_i)$ is a de-correlated deviance $C^{-1/2}\hat{\epsilon}_i$ where $\hat{\epsilon}_i = D_{i,h}^2 - \hat{\sigma}_{i+\frac{h}{2}}^2$ is a raw deviance and $C = C(i, j; h) = \sigma_i^2 \sigma_j^2 \rho(|s_i - s_j|; \theta)$ is the covariance matrix of $D_{i,h}^2$. In the denominator $M\left(i + \frac{h}{2}, i + \frac{h}{2}\right)$ is the i^{th} diagonal of the smoothing matrix of $D_{i,h}^2$ where $M_{\left(i+\frac{h}{2}, j+\frac{h}{2}\right)}(\lambda) = K\left(\frac{s_{i+\frac{h}{2}} - s_{j+\frac{h}{2}}}{\lambda}\right) = K\left(\frac{i-j}{n\lambda}\right)$.

A cautionary tale in local variogram or variance function estimation is to guarantee that it is positive. We may encounter negatively estimated values in step 1 of Algorithm 1 when we run a kernel smoothing with small bandwidth. Most often negative values occur near the boundaries. We fix this problem by increasing the bandwidth size near the boundary, then the edge effect should not skew nor drive the functional estimation near the boundary.

It is well known in nonparametric statistics literature that when underlying data are correlated, bandwidth selection requires an adjustment either to the data or to a penalty term. Hart (1991) For bandwidth selection in nonparametric regression two common practices are cross-validation and finding optimal smoothness in the estimating

function. Opsomer et al. (2001) compiles of several proposals. Altman (1990) proposes to adapt the weights of residuals. Han and Gu (2008) simultaneously select the bandwidth and estimate the correlation parameters after adding a penalty term to the likelihood function to adjust for the correlation.

Even though the data are correlated, the differencing greatly reduces the correlation between the differenced process and keeps it small. See Remark in Section 3.3.3. We estimate the covariance parameters assuming a suitable parametric correlation model for the process. We obtain the variance estimate for location s by dividing the local variogram estimate at the specified location with the value of variogram function at a fixed lag- h derived from the estimated correlation function.

In terms of computing time for estimating a realization of size $n = 1000$, the difference-based method including the correlation parameter estimation takes 1/10 of time as the likelihood-based method with the known correlation parameter values plugged-in. Bandwidth selection adds much greater computing cost for the likelihood-based method since Anderes and Stein (2011) suggest a simulation based approach which requires the inversion of a correlation matrix as many number of times as simulation is required. The difference-based method also requires the inverse of a estimated data correlation structure in Step 2 of Algorithm 1 but it is just once.

3.5 Simulation Study

3.5.1 Set-up

From the data model $Z_s = \mu(s) + \sigma(s)X_s$ in (3.1), we set $\mu(s) = 0$ and test our method on a small-scale local spatial process, which would be broadly termed correlated errors. We set a stationary error process X_s to be a Gaussian random field for two reasons. One is due to the ease of simulation, and the other is due to the distribution's analytical tractability, which is welcomed when adopting a likelihood approach. The

dependent structure is generated using an exponential correlation function with the range parameter $\theta = 0.01$ and 0.1 , which translates to a practical range of $3\theta = 0.03$ and 0.3 , where the correlation becomes $.05$. We also generate an independent error process as the third level of dependency investigated. The process is generated on an equally spaced grid over a unit interval $0 \leq s \leq 1$. The sample size is set at four levels: $n = 100, 200, 500$, and 1000 . In terms of variance function, we have the following four standard deviation functions:

(a) an infinitely-differentiable sinusoidal function: $\sigma(s) = 2 \sin(s/0.15) + 2.8$,

(b) a quadratic function: $\sigma(s) = 8(s - 0.5)^2 + 0.5$

(c) a piece-wise-differentiable function: (a hockey stick) $\sigma(s) = \mathbb{1}_{\{0 \leq s \leq 1/3\}} + 3s \mathbb{1}_{\{1/3 < s \leq 1\}}$,

(d) a discontinuous step function: $\sigma(s) = 1 + \mathbb{1}_{\{1/3 < s \leq 1\}}$.

The functional smoothness, or differentiability, changes from infinitely smooth to discontinuous from (a) to (d). Since our method assumes a smooth variance function, we should detect a change point (i.e. a point of discontinuity) in (d) and estimate the variance functions separately about the change point. Nevertheless, it is worth noticing the effect of this violation of the assumption, and so we include the function (d).

3.5.2 Discussion of Results

We display and discuss the results of variance function estimation in this section. In evaluating functional estimation, we focus on the following two criteria:

(i) Discretized integrated squared error (*DMSE*): $\sum_{i=1}^n \{\hat{\sigma}_{i,\lambda} - \sigma_i\}^2 / n$

(ii) Maximum deviation: $\max_i \{|\hat{\sigma}_{i,\lambda}^2 - \sigma_i^2| : \text{for } i = 1, \dots, n\}$.

From the *DMSE* we can evaluate the average variance of the variance function estimator.

From the maximum deviation, we are able to assess the worst point estimation result

across the estimation domain. Yet, we do acknowledge that the variance is greater for the larger functional estimation values, and the approximate location of the worst estimation result is well gauged. For the ease of computation, we evaluate the estimated functions at 100 equally-spaced points on $[0,1]$.

We compare the results from our method to the likelihood-based method proposed by Anderes and Stein (2011) for a nonstationary process and to the difference-based method proposed by Brown and Levine (2007) for an independent error process using the bandwidth selection idea presented in Levine (2006). As Brown and Levine (2007) assume an independent error data model, Levine’s bandwidth selection idea uses this property to randomly leave out a training set in K -fold cross validation. Therefore, we do not expect a reasonable result when correlated errors are presented. Nonetheless we test Levine’s bandwidth selection idea against the oracle and our bandwidth selection proposal for a nonstationary correlated error process.

Figure 3.1 shows the results of the step standard deviation function, Model (d), estimation using a difference-based method and Anderes and Stein (2011)’s likelihood-based method both with an oracle bandwidth. We assume that the true covariance model and the parameters are known when producing oracle bandwidth selected results. Figure 3.2 has the results of sinusoidal standard deviation function, Model (a), estimation as Figure 3.1 and additionally contains a difference-based estimation with regular bandwidth selection and the estimated covariance parameter. In both figures, the thick solid red line represents the true standard deviation function $\sigma(\cdot)$ and the thin gray lines are estimation results from 11 innovation processes. The $DMSE$ is smaller, i.e. the variance of the estimator grows smaller as the sample size n grows. Also, the maximum deviation grows smaller, as the sample size grows. Notice in Figures 3.1 and 3.2 that the local-likelihood approach renders an estimation result that is less smooth and more ragged than a local polynomial smoothing, which can be seen by comparing the plots in the first row to the last. Relying on the oracle bandwidth selection criterion of the

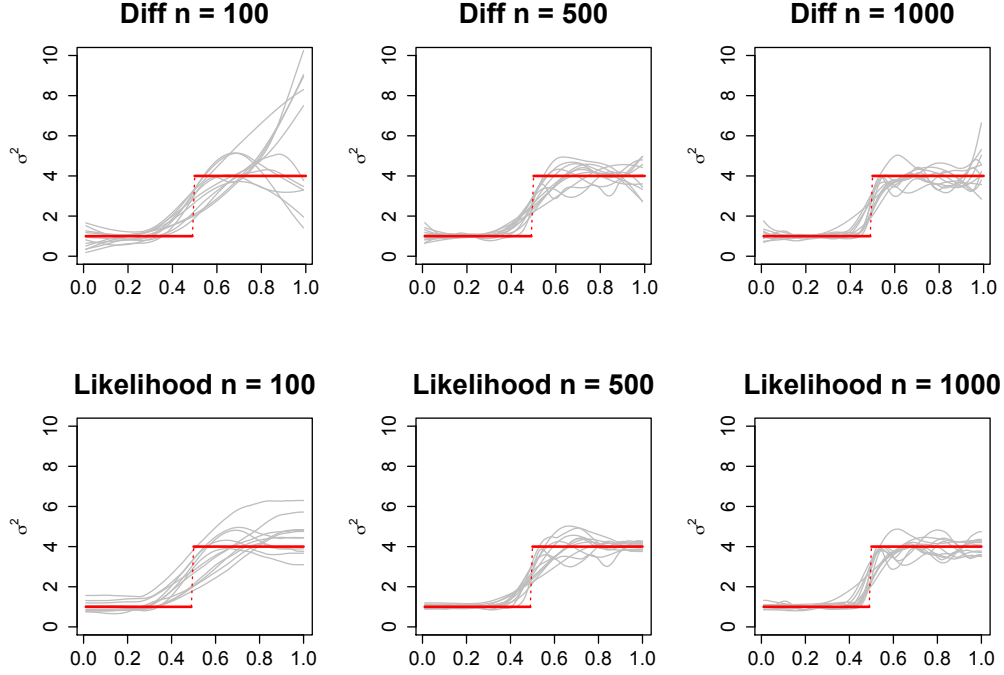


Figure 3.1: Piece-wise continuous function estimation: $\sigma(s) = 1 + \mathbb{1}_{\{1/3 < s \leq 1\}}$. The thick solid red line represents the true $\sigma(\cdot)$. The thin gray lines are examples of estimation results from 11 innovation processes.

minimized $DMSE$ gives the likelihood-based estimation results to be less than ideal. As n grows, a similar qualitative difference remains between the two estimation methods. For a local-likelihood approach it is better to choose a larger bandwidth than for the oracle selection, yielding a smoother estimation result, which is closer to the true form of the variance function. We, then, deduce that the $DMSE$ should be larger than the minimized $DMSE$, which was the criterion of the oracle bandwidth selection, and also than the $DMSE$ of a difference-based method.

Figure 3.3 illustrates the difference in estimation summary between the likelihood-based method (blue boxplots) and the difference-based method (red boxplots) for different sample sizes $n = 100$ (left), 500 (middle) and 1000 (right). We test four dependency range parameters $\theta = 0.05, 0.1, 0.2$, and 0.3 . Within each plot, there are four pairs of red and blue boxplots with a corresponding label. When the sample size is

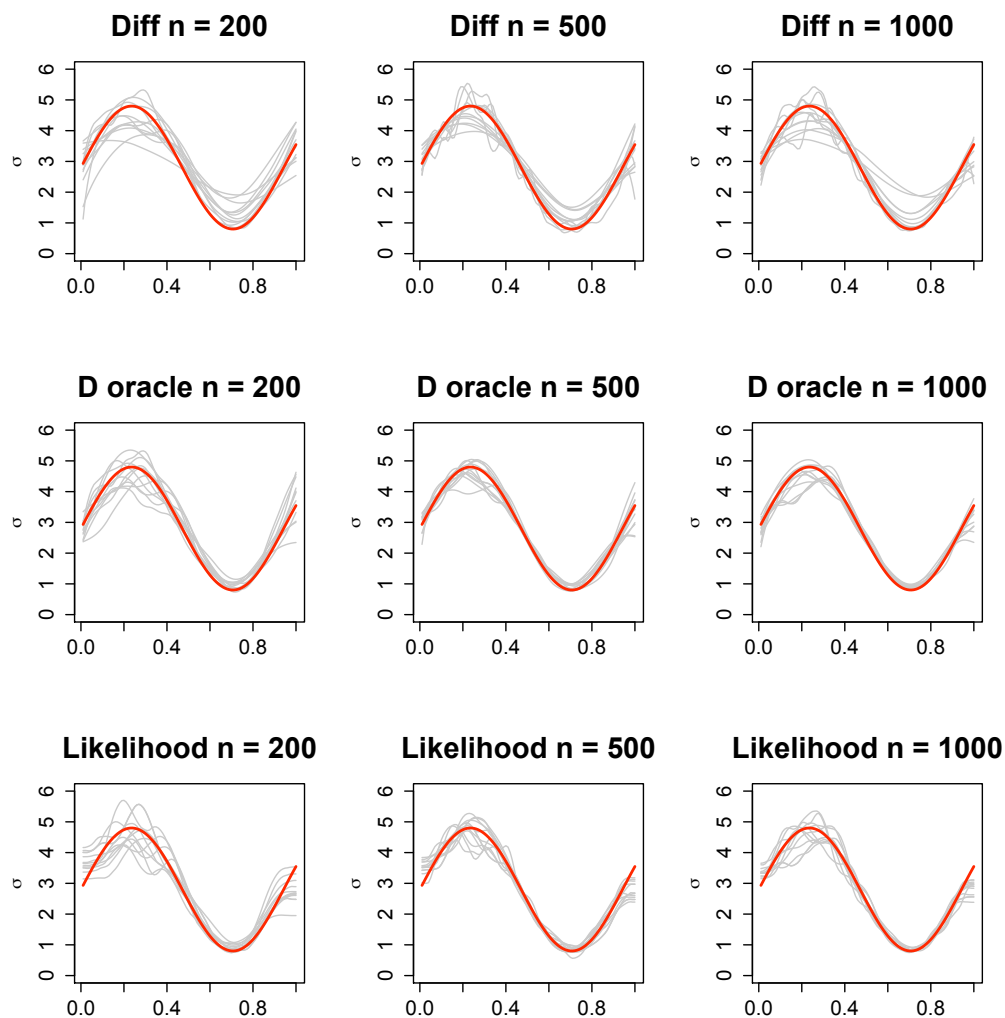


Figure 3.2: Smooth variance function estimation : $\sigma(s) = 2 \sin(s/0.15) + 2.8$. The thick red line represents the true $\sigma(\cdot)$. The thin gray lines are examples of estimation results from 11 innovation processes. The top row shows a difference-based estimation with bandwidth selection. The middle row also uses the difference-based method but with an oracle bandwidth. The bottom row has a local-likelihood-based estimation with an oracle bandwidth. From left to right, the columns reflect an increase in the process sample size, as indicated in the title of each graph.

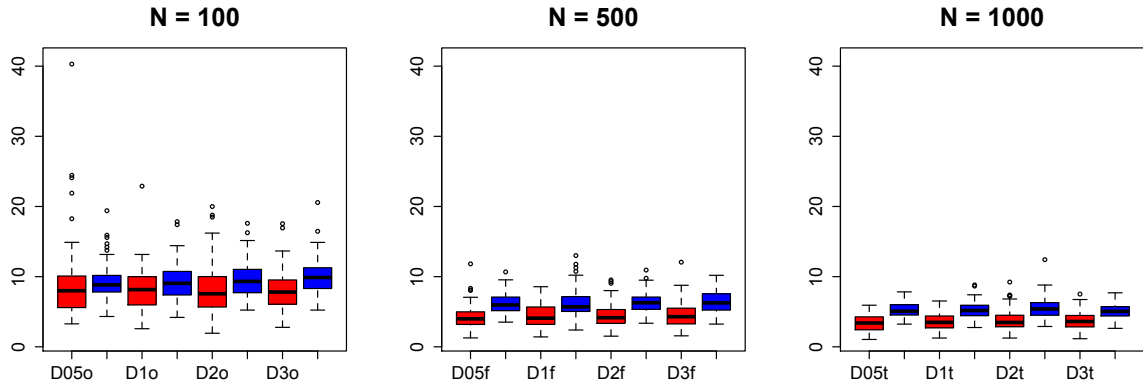
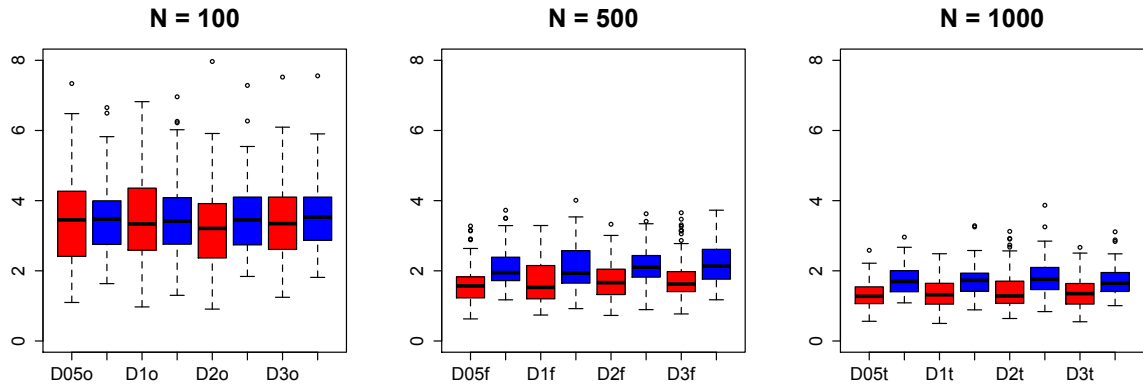
(a) L_∞ (b) $DMSE$

Figure 3.3: Comparing $\sigma(\cdot)$ Model (a) estimation summaries of the difference-based method (red boxplots) and the likelihood-based method (blue boxplots) using oracle bandwidths in both settings.

small to moderate, say $n \leq 200$, the two methods perform similarly. When the sample size is large, say $n > 200$, the difference-based method has a significantly smaller variance than the local-likelihood method. In the right plots, the differenced-based method boxplots have a Q3 that nearly matches or is less than the likelihood-based method estimation summary Q1 in Figure 3.3 (b). In maximum deviation, the difference is even more pronounced in that the difference-based method performs better than the likelihood-based method where $n > 200$. Tables 3.1 and 3.2 show the estimation results of standard deviation functions (a) and (b) respectively. There are three main comparisons to be made: first, between a difference-based estimation using oracle bandwidth (Diff-oracle) and our difference-based estimation with all parameters estimated (Diff-proposed); secondly, between Diff-oracle and the difference-based method of Brown and Levine (2007) assuming independent errors (Diff-Levine); and lastly, between Diff-oracle and the likelihood-based estimation of Anderes and Stein (2011) using oracle bandwidth (Likelihood). As mentioned above, we take the true covariance model and the parameter values as given for the estimation with oracle bandwidth so that the comparison of Diff-oracle and Likelihood(-oracle) is on equal footing. In the columns of each table below the heading of $DMSE$ and L_∞ , there are three levels of dependency in the data: $\theta = 0.1, 0.01$ and independent. The numerical summaries in Tables 3.1 and 3.2 show little difference between a difference-based estimation and a likelihood-based estimation both with an oracle bandwidth when n is relatively small like $n=100$ and 200 . However, the numerical summaries often overlook some qualitative difference in estimation summary, and we have discussed the selection of smaller than perfect bandwidths in this likelihood-based estimation, which results in a slightly ragged estimation as seen in the bottom row of Figures 3.1 and 3.2.

Table 3.3 contains the optimal bandwidth summary for the settings listed above. Notice that the oracle bandwidths tend to be small for the likelihood-based method. The kernel used for both cases are different; a higher order kernel is used for the local-

likelihood method. Yet the comparison is not completely ungrounded because in both cases, bounded support kernels are used for smoothing. It is worth noting that the oracle bandwidth tends to be smaller for a dependent process variance estimation but that our bandwidth selection gives the contrary result. The difference in the bandwidth sizes for dependent processes and for independent processes becomes more drastic as n grows. Also, the case is more obvious for Model (b), a quadratic standard deviation function, than for Model (a), a sinusoidal standard deviation function. When estimating the quadratically-shaped function, Levine's method works only when the errors are independent; otherwise it fails to provide an accurate estimation as the method is based on N independent error model assumption.

We note that the differentiability of the variance function affects the estimation bias and variance of both difference-based and likelihood-based methods. The strength of the dependency, i.e. the size of autocorrelation at a fixed lag, does not affect the asymptotic result.

Table 3.1: A sine $\sigma(\cdot)$ estimation result summary

\sim	sine	$DMSE$			L_∞		
n	Methods	$\theta = 0.1$	$\theta = 0.01$	indep.	$\theta = 0.1$	$\theta = 0.01$	indep.
100	Diff-oracle	0.32 (0.19)	0.33 (0.22)	0.37 (0.24)	1.20 (0.45)	1.43 (0.65)	1.47 (0.66)
	Diff-proposed	1.09 (2.13)	0.55 (0.74)	0.53 (0.33)	1.84 (0.93)	1.56 (0.69)	1.71 (0.78)
	Diff Levine	4.93 (0.41)	1.00 (0.35)	0.76 (0.47)	3.62 (0.21)	2.14 (0.55)	1.98 (0.75)
	Likelihood	0.32 (0.14)	0.30 (0.15)	0.28 (0.13)	1.40 (0.33)	1.37 (0.35)	1.34 (0.37)
200	Diff-oracle	0.15 (0.08)	0.16 (0.09)	0.21 (0.12)	0.96 (0.34)	0.98 (0.39)	1.18 (0.44)
	Diff-proposed	0.71 (0.79)	0.31 (0.18)	0.30 (0.17)	1.48 (0.58)	1.21 (0.47)	1.28 (0.56)
	Diff Levine	6.09 (0.21)	1.69 (0.26)	0.52 (0.28)	3.90 (0.12)	2.32 (0.27)	1.55 (0.50)
	Likelihood	0.19 (0.09)	0.18 (0.08)	0.18 (0.08)	1.13 (0.29)	1.12 (0.26)	1.11 (0.28)
500	Diff-oracle	0.08 (0.05)	0.07 (0.04)	0.11 (0.06)	0.70 (0.28)	0.70 (0.24)	0.85 (0.28)
	Diff-proposed	0.54 (0.52)	0.23 (0.19)	0.14 (0.08)	1.23 (0.51)	0.90 (0.33)	0.87 (0.34)
	Diff Levine	7.38 (0.10)	3.37 (0.21)	0.26 (0.19)	4.18 (0.06)	2.99 (0.14)	1.08 (0.35)
	Likelihood	0.10 (0.05)	0.10 (0.05)	0.10 (0.05)	0.88 (0.20)	0.85 (0.20)	0.90 (0.22)
1000	Diff-oracle	0.05 (0.03)	0.04 (0.02)	0.06 (0.03)	0.58 (0.23)	0.54 (0.19)	0.62 (0.24)
	Diff-proposed	0.51 (0.54)	0.22 (0.23)	0.07 (0.04)	1.11 (0.55)	0.81 (0.32)	0.64 (0.26)
	Diff Levine	8.03 (0.05)	4.82 (0.13)	0.11 (0.08)	4.35 (0.03)	3.47 (0.08)	0.75 (0.29)
	Likelihood	0.06 (0.03)	0.06 (0.02)	0.06 (0.02)	0.74 (0.18)	0.74 (0.15)	0.70 (0.17)

Table 3.2: A quadratic $\sigma(\cdot)$ estimation result summary

\smile	Quadratic	DMSE			L_∞		
n	Methods	$\theta = 0.1$	$\theta = 0.01$	indep.	$\theta = 0.1$	$\theta = 0.01$	indep.
100	Diff-oracle	0.07 (0.06)	0.08 (0.07)	0.10 (0.09)	0.90 (0.37)	0.83 (0.43)	0.90 (0.47)
	Diff-proposed	0.18 (0.14)	0.12 (0.07)	0.15 (0.11)	1.01 (0.40)	0.93 (0.41)	1.10 (0.66)
	Diff Levine	0.83 (0.09)	0.16 (0.07)	0.19 (0.15)	1.97 (0.19)	1.16 (0.36)	1.12 (0.57)
	Likelihood	0.06 (0.03)	0.06 (0.03)	0.07 (0.04)	0.79 (0.23)	0.76 (0.24)	0.81 (0.27)
200	Diff-oracle	0.03 (0.02)	0.04 (0.03)	0.06 (0.05)	0.54 (0.27)	0.58 (0.27)	0.70 (0.33)
	Diff-proposed	0.16 (0.39)	0.08 (0.06)	0.08 (0.05)	0.84 (0.44)	0.70 (0.31)	0.82 (0.40)
	Diff Levine	1.03 (0.05)	0.24 (0.06)	0.11 (0.09)	2.05 (0.12)	1.28 (0.24)	0.83 (0.40)
	Likelihood	0.04 (0.02)	0.04 (0.02)	0.04 (0.02)	0.62 (0.17)	0.62 (0.19)	0.63 (0.18)
500	Diff-oracle	0.02 (0.02)	0.02 (0.01)	0.03 (0.02)	0.43 (0.20)	0.39 (0.18)	0.51 (0.23)
	Diff-proposed	0.17 (0.13)	0.17 (0.13)	0.17 (0.11)	0.75 (0.27)	0.77 (0.26)	0.93 (0.31)
	Diff Levine	1.26 (0.03)	0.52 (0.05)	0.05 (0.05)	2.18 (0.07)	1.53 (0.11)	0.58 (0.28)
	Likelihood	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)	0.49 (0.14)	0.46 (0.13)	0.47 (0.15)
1000	Diff-oracle	0.011 (0.010)	0.009 (0.008)	0.013 (0.011)	0.34 (0.18)	0.32 (0.14)	0.37 (0.17)
	Diff-proposed	0.09 (0.10)	0.04 (0.04)	0.017 (0.011)	0.65 (0.36)	0.48 (0.23)	0.42 (0.22)
	Diff Levine	1.381 (0.017)	0.781 (0.031)	0.021 (0.012)	2.26 (0.04)	1.79 (0.07)	0.42 (0.22)
	Likelihood	0.012 (0.006)	0.012 (0.005)	0.012 (0.006)	0.40 (0.13)	0.40 c(0.12)	0.40 (0.12)

Table 3.3: Bandwidth selection summary of sine and quadratic $\sigma(\cdot)$ estimation

n	Bandwidth	\sim			\smile		
	Methods	$\theta = 0.1$	$\theta = 0.01$	indep.	$\theta = 0.1$	$\theta = 0.01$	indep.
100	Diff- λ^O	0.203 (.054)	0.206 (.059)	0.209 (.052)	0.218 (.071)	0.222 (.084)	0.229 (.076)
	Diff- λ^*	0.262 (.074)	0.281 (.079)	0.266 (.069)	0.405 (.126)	0.415 (.087)	0.434 (.074)
	Levine	0.356 (0.297)	0.455 (0.274)	0.420 (0.281)	0.360 (.304)	0.467 (.267)	0.418 (.289)
	Like- λ^O	0.165 (.054)	0.168 (.055)	0.154 (.033)	0.137 (.032)	0.138 (.030)	0.133 (.030)
200	Diff- λ^O	0.170 (.034)	0.171 (.037)	0.177 (.046)	0.191 (.050)	0.185 (.060)	0.203 (.066)
	Diff- λ^*	0.240 (.090)	0.218 (.108)	0.190 (.119)	0.381 (.126)	0.336 (.143)	0.289 (.163)
	Levine	0.234 (0.248)	0.380 (0.224)	0.347 (0.229)	0.248 (.249)	0.369 (.230)	0.334 (.217)
	Like- λ^O	0.131 (.034)	0.129 (.028)	0.127 (.021)	0.113 (.025)	0.113 (.024)	0.112 (.023)
500	Diff- λ^O	0.140 (.027)	0.141 (.031)	0.154 (.037)	0.154 (.042)	0.152 (.042)	0.158 (.047)
	Diff- λ^*	0.217 (.107)	0.205 (.117)	0.180 (.111)	0.357 (.143)	0.329 (.147)	0.260 (.159)
	Levine	0.186 (0.186)	0.256 (0.164)	0.232 (0.165)	0.192 (.193)	0.264 (.152)	0.240 (.166)
	Like- λ^O	0.098 (.016)	0.098 (.016)	0.100 (.016)	0.091 (.019)	0.090 (.016)	0.094 (.017)
1000	Diff- λ^O	0.120 (.026)	0.121 (.026)	0.133 (.023)	0.131 (.033)	0.125 (.033)	0.148 (.038)
	Diff- λ^*	0.209 (.121)	0.186 (.117)	0.170 (.109)	0.329 (.159)	0.300 (.157)	0.255 (.165)
	Levine	0.180 (0.155)	0.289 (0.118)	0.174 (0.094)	0.199 (.157)	0.288 (.123)	0.191 (.092)
	Like- λ^O	0.086 (.013)	0.084 (.011)	0.086 (.013)	0.078 (.015)	0.076 (.013)	0.078 (.014)

3.6 Discussion

We have developed a nonparametric variance function estimator for a one-dimensional nonstationary process whose correlation structure is isotropic. We have investigated mixed-domain asymptotic properties of the local variogram estimator and have shown that the asymptotic rate of convergence is dependent on the relative smoothness of mean function to the smoothness of variance function and the mean square differentiability of a data process.

We have shown through a simulation study that difference-based estimation has a smaller bias and variance than a local-likelihood approach. Boundary bias can be fixed by adjusting the objective function of a nonparametric estimation whereas the local-likelihood method should introduce additional innovation terms to solve the boundary bias problem. Another contrast between the two approaches is in computing time. A difference-based method needs no matrix inversion and reduces the computing time by $O(1/n^2)$ to that of a likelihood-based method, where n is the size of the data process. The bandwidth selection idea by Anderes and Stein (2011) also requires a global covariance matrix inversion and increases the computing time by $O(mn^2)$ where m is the number of simulations of a stationary process to test against the observed nonstationary process. While their bandwidth selection ideas are insightful and useful when there is a specific data model which can be simulated, it is still costly to perform likelihood-based estimation in terms of computing time and power.

Under certain regularity conditions we directly estimate a variance function, applying a difference filter to the data, instead of estimating a large-scale model component or the marginal mean function before the variance function. In signal processing, a band-pass filter could also provide local variance estimation assuming that the marginal mean function is changing slowly. First, carefully select a filter and the passband, then pass signals through the filter to reduce the effect of signals outside of the preferred range

of frequencies, and lastly those filtered spectra of noise are converted into the variance estimation. However, there are some disadvantages to this approach. It works well under second-order stationarity of the error process and not under nonstationarity, and this limits the range of variance functions to be estimated. Also, the shape of a filter and the passband should interact with the underlying data process and have an impact on the estimation result, and the number of points of consideration exceeds that of time-domain smoothing. Lastly, depending on the variance function to be estimated, the estimation output from the band-pass filter may introduce bias from the frequency to time domain conversion. Therefore, the difference-based smoothing in the time-domain gives more precise and accurate estimation of the variance function.

In the following chapter we extend the variance function estimation via a difference-based method to a two-dimensional nonstationary random field. There are many more difference filters to consider, and therefore we add more conditions to the linear filters and discuss the properties of the filters.

CHAPTER 4. VARIANCE FUNCTION ESTIMATION OF TWO-DIMENSIONAL NONSTATIONARY PROCESS

4.1 Introduction

We extend the differencing idea for variance function estimation of one-dimensional nonstationary process to a two-dimensional nonstationary random field. A random field takes values in an Euclidean space and is a stochastic process with a non-zero correlation function. Examples of random fields are weather variable maps, areas of abundance of ecological resources in a continuous domain, the topography of an area, etc. These examples are broadly characterized by a large-scale trend, which is often referred to as a mean process. After removing a large-scale trend from an observed stochastic process, there still remains local variations which we refer to as a meso-scale trend. A temperature map of the US, for example, shows a large-scale trend of warm temperature in the south and cool temperature in the north and a meso-scale trend of the temperature field around the Great Lakes which is significantly cooler than the northwest or northeast corner of the US in the same longitude. The large-scale pattern and the meso-scale variation together capture the important features of a map. The remaining variations, which we refer to as an error process, may still contain correlation and nonstationarity. We are interested in estimating the variance function of a nonstationary correlated error process.

Gasser et al. (1986), Müller and Stadtmüller (1987), Buckley et al. (1988), and Hall and Carroll (1989) have considered one-dimensional differencing scenarios for variance function estimation. A nonparametric estimation of variance should be reasonable when a variance function is of a smooth function. Gasser et al. (1986) and Müller and

Stadtmüller (1987) have also been interested in estimating the derivatives of the mean function, which could be interpreted as a variance function, and have directly estimated them using a differencing idea. Smoothing the difference filter applied data can be used either as a preliminary stage of exploration or as a final stage of the estimation of the underlying function of interest. When there is insufficient covariate information to build a parametric model for a regression analysis, a nonparametric approach helps to bring out significant features of the data. Often it is a case in spatial data analysis that we observe heteroscedasticity and that we do not know the target parametric form of the variance function. Hence, it is convenient to adopt the difference-based idea for removing the trend and to estimate the covariance parameters of a nonstationary process.

Hall et al. (1991) discuss the two-dimensional optimal difference filter configurations that achieve minimum variance in variance estimation. Their application is for image processing, and they assume independent and identically distributed errors for the observed process. They find that averaging over a number of different linear filters besides the rotation of filters helps reduce the variance of the estimator in the order of $N^{-1}l^{-2}$ where N is the number of observations in an increasing domain and l is the the number of filter configurations. Zhu and Stein (2002) use difference filters for estimating the fractal dimension of fractional Brownian fields and introduce a generalized variogram.

A similar idea is also used to estimate a generalized local variogram introduced in Section . For applying a difference filter and a smoothing kernel to the data, we implicitly assume local stationarity when nonstationarity is present in a random field. Our data model for a random field assumes a smooth and non-constant variance function and an isotropic correlation function over the parameter space. In order to apply a two-dimensional difference filter, we explore several configurations and compare two different weighting options: a symmetric weighting scheme and a Hall-Kay-Titterington weighting scheme from Hall et al. (1991). Our difference-based variance function estimator is dependent only on the set of points in two-dimensional space and should not rely on any

assumptions about the lattice. The lattice type, e.g. a triangular lattice versus a square, rectangular or hexagonal lattice, can determine the number of directions to average, resulting in the greater statistical efficiency in variance estimation with a greater number of directions. For the ease of data simulation, we use the data on a rectangular lattice.

In Section 4.2.1, I describe the correlated data model and define the variance function estimator and its basic properties. In Section 4.2.2 I explain the difficulty of a fully theoretical derivation of choice filters and report a numerical study to provide insight into the proposed method. Section 4.3 details the difference weight sequences and discusses the optimal choice of the weighting schemes. Examples of optimal sequences are given for several filter configurations, such as a line, box, cross, Y, and rotations of a line and Y-configurations. In Section 4.4, I examine the filter performance depended on the weight schemes, directional rotation and averaging, and the filter scale with respect to the degree of dependence in the data and fineness through a simulation study. Finally, in Section 4.5 I conclude the chapter with comments and discussions of the simulation study.

4.2 Data Model and Method

Consider a two-dimensional regular lattice $\mathbf{Z}_n = \{Z(\mathbf{s}_i)\}_{i \in \mathcal{R}_n}$ where \mathbf{s}_i is a location indexed by $\mathbf{i} \in \mathcal{R}_n \subseteq \mathbb{Z}^2$. Using the same data model setup as in Chapter 3 with the exception of the observations being in two-dimensional Euclidean space, we let $Z(\mathbf{s}) = \mu(\mathbf{s}) + \sigma(\mathbf{s})X(\mathbf{s})$ be a nonstationary random field on a unit grid $[0, 1]^2$ in \mathbb{R}^2 where $X(\mathbf{s})$ is a stationary Gaussian random field with mean 0, variance 1, and $\text{cor}(X(\mathbf{s}), X(\mathbf{s}')) = c\|\mathbf{s} - \mathbf{s}'\|^\alpha$ where $0 < \alpha < 2$ for some constant $0 < c < 1$. We assume for some $C > 0$ and $0 < \epsilon < \frac{1}{3}$, $|\mu(\mathbf{s}_i) - \mu(\mathbf{s}_j)| \leq C\|\mathbf{s}_i - \mathbf{s}_j\|^{1/2(1-3\epsilon)}$ for any $\mathbf{i}, \mathbf{j} \in \mathbb{Z}^2$ and $\sigma^2(\mathbf{s}) \in \Lambda^\beta$ where $\beta \geq 2$.

4.2.1 Notations and Definitions

A linear filter function L is defined by a set of weights $A = \{a_j : j \in \mathcal{J}\}$ associated with a set of relative locations $\mathcal{J} = \{\mathbf{p}_j = (p_{1j}, p_{2j}) \in \mathbb{Z}^2 : \sum_j (\mathbf{p}_j - \mathbf{p}_0) = \mathbf{0}\}$.

Let's define some attributes of a difference filter and difference filtered data:

- L : linear difference filter function
- l : index of a filter represented by an integer or the filter configuration
- $R = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix}$: a rotation matrix for a two-dimensional filter. On a square lattice data, we use $\phi = 0, \frac{\pi}{4}, \frac{\pi}{2}$, and $\frac{3\pi}{4}$.
- $\mathcal{J}^{(l)} = \{\mathbf{p}_j = (p_{j1}, p_{j2}) : \sum_j \mathbf{p}_j = \mathbf{0}\}$, a set of relative node locations of filter L_l
- ν_l : the number of nodes in filter L_l
- h_l : the minimum spacing among the nodes in \mathcal{J}_l of filter L_l . The scale factor is represented as a multiple of one-dimensional grid spacing.
- D_i^2 : shorthand of $D^2(\mathbf{s}_i; L)$ when the filter L is apparent from context
- $A^{(l)} = \{a_j^{(l)} : j \in \mathcal{J}_l\}$, a set of difference weights for filter L_l
- \mathcal{R} : a set of location indices of lattice data
- \mathcal{R}^B : a set of boundary location indices of \mathcal{R}
- $\mathcal{R}_l = \{k : \text{for every } j \in \mathcal{J}^{(l)}, j + k \in \mathcal{R}\}$, a set of location indices of \mathbf{D}_L^2
- $N_l = |\mathcal{R}_l|$: the cardinality of \mathcal{R}_l , i.e. number of pseudo-residuals
- $\varrho_L(h_l)$: L -filter autocorrelation at lag h , the closest node pair distance
- α : smoothness parameter of the data

$$\text{Condition 1. } \sum_{j \in \mathcal{J}} a_j = 0 \quad \implies \quad E(\sum_{j \in \mathcal{J}} a_j X_{i+j}) = 0$$

$$\text{Condition 2. } \sum_{j \in \mathcal{J}} a_j^2 = 1 \quad \implies \quad E \left(\sum_{j \in \mathcal{J}} a_j X_{i+j} \right)^2 = \varrho_L(h)$$

$$\text{Condition 3. } \sum_{j \in \mathcal{J}} a_j \mathbf{p}_j = (0, 0) \quad \implies \quad L(Z_{\mathbf{s}}) \text{ gives a pseudo-residual at } \mathbf{s}.$$

- θ : data covariance function parameter

Definition 4.2.1 Define an L -filter variogram at scale h as

$$\varrho_L(h) = 1 - 2 \sum_{j \in \mathcal{J}_l} \sum_{\substack{k \neq j \\ k \in \mathcal{J}_l}} a_j a_{j+k} \rho(h \|k\|).$$

We use \mathbf{i} as the location index in \mathcal{R} and reserve indices k, k_1, k_2 and m, m_1, m_2 for node indices \mathcal{J} of a filter. We occasionally use j for the data location. Let $L(Z(\mathbf{s})) = \sum_{j \in \mathcal{J}} a_j Z(\mathbf{s} + \mathbf{p}_j)$ represent a filter L applied to a random field \mathbf{Z} about location \mathbf{s} . As a shorthand, we use the following.

$$\begin{aligned} Z(\mathbf{s} + \mathbf{p}_j) &= Z_{\mathbf{s}+j}, & Z(\mathbf{s} + h\mathbf{p}_j) &= Z_{\mathbf{s}+jh}, & Z(\mathbf{s}_i + \mathbf{p}_j) &= Z_{i+j}, \\ \rho(\|\mathbf{s}_i - \mathbf{s}_j\|) &= \rho_{\|\mathbf{i}-\mathbf{j}\|}, & \mathbf{p}_j^{(l)} \in \mathcal{J}_l &\Rightarrow j \in \mathcal{J}_l \end{aligned}$$

In Chapter 3, note that we have used the notation $D_{i,h}$ for the i^{th} pseudo-residual lag- h difference. For two-dimensional random field, we use \mathbf{D}_L^2 as a square of $\{L(\mathbf{Z}(\mathbf{s}))\}_{\mathbf{s} \in R_n}$.

The \mathbf{i}^{th} value is

$$D^2(\mathbf{s}_i; L(h)) = \left(\sum_{j \in \mathcal{J}} a_j Z_{\mathbf{i}+jh} \right)^2$$

which is referred to as the \mathbf{i}^{th} pseudo-residual in two-dimensional random field.

The following conditions are imposed on filter L :

Each condition has an implication which makes the filter applied data to be called as a pseudo-residual because of mean 0 (Condition 1) and because of the expected value of the squared pseudo-residual being an L -filter variogram (Condition 2).

Remark Note that Condition 1 and 2 above imply $a_* \sum_{j \in \mathcal{J} \setminus \{*\}} a_j = a_*(1 - a_*)$. Then the sum of all cross-terms of filter weights are

$$\sum_{i \in \mathcal{J}} \sum_{j \neq i, j \in \mathcal{J}} a_i a_j = \sum_{i \in \mathcal{J}} \left(a_i \sum_{j \in \mathcal{J} \setminus \{i\}} a_j \right) = \sum_{i \in \mathcal{J}} -a_i^2 = -1.$$

Therefore, when X_i is a mean 0, variance 1, and an isotropic process, we have $E \left(\sum_{j \in \mathcal{J}} a_j X_{i+j} \right)^2 = \sum_{j \in \mathcal{J}} a_j^2 - \sum_{i \in \mathcal{J}} \sum_{j \neq i, j \in \mathcal{J}} a_i a_j \rho(\|i - j\|)$. We conceptually regard $E \left(\sum_{j \in \mathcal{J}} a_j X_{i+j} \right)^2$ as a variogram.

Definition 4.2.2 Let $K(\cdot)$ and $K^B(\cdot)$ be defined as in 3.2.5-3.2.6. For a vector of bandwidth parameter $\Lambda = (\lambda_x, \lambda_y) \in (0, \frac{1}{2})^2$, define a $2m$ -order Gasser-Müller kernel function $K_\Lambda(\mathbf{i}, \mathbf{s})$ as a product of two m -order Gasser-Müller kernel functions each centered at each coordinate of $\mathbf{s} = (s_x, s_y)$. $K_\Lambda(\mathbf{i}, \mathbf{s}) = K_{\lambda_x \mathbf{i}}(s_x) K_{\lambda_y \mathbf{i}}(s_y)$ where

$$K_{\lambda_x \mathbf{i}}(s_1) = \begin{cases} \int_{(s_{x\mathbf{i}} + s_{x\mathbf{i}-1})/2}^{(s_{x\mathbf{i}} + s_{x\mathbf{i}+1})/2} \frac{1}{\lambda_x} K\left(\frac{s_x - u}{\lambda_x}\right) du & \text{when } s_1 \in (\lambda_x, 1 - \lambda_x) \\ \int_{(s_{x\mathbf{i}} + s_{x\mathbf{i}-1})/2}^{(s_{x\mathbf{i}} + s_{x\mathbf{i}+1})/2} \frac{1}{\lambda_x} K^B\left(\frac{s_x - u}{\lambda_x}\right) du & \text{when } s_1 \in (0, \lambda_x) \\ \int_{(s_{x\mathbf{i}} + s_{x\mathbf{i}-1})/2}^{(s_{x\mathbf{i}} + s_{x\mathbf{i}+1})/2} \frac{1}{\lambda_x} K^B\left(-\frac{s_x - u}{\lambda_x}\right) du & \text{when } s_1 \in (1 - \lambda_x, 1) \end{cases}$$

and $K_{\lambda_y \mathbf{i}}(s_y)$ is defined in the same fashion as $K_{\lambda_x \mathbf{i}}(s_1)$ only on a different coordinate with bandwidths λ_y centered at s_2 . The notations in the limits of the integral represent $s_{x,\mathbf{i}+1} = s_{x,\mathbf{i}} + 1/n$, $s_{x,\mathbf{i}-1} = s_{x,\mathbf{i}} - 1/n$, and likewise for $s_{y,\mathbf{i}+1}$ and $s_{y,\mathbf{i}-1}$ for any $\mathbf{i} \in \mathcal{R}_l \setminus \mathcal{R}^B$. For $\mathbf{i} \in \mathcal{R}^B$ the limits of the integral $K^B\left(\frac{s_x - u}{\lambda_x}\right)$ are from 0 to $s_{x\mathbf{i}+1}/2$, and the limits of the integral $K^B\left(-\frac{s_x - u}{\lambda_x}\right)$ are from $s_{x\mathbf{i}-1}/2$ to 1. Then, for any $\mathbf{s} \in [0, 1]^2$, $\sum_{\mathbf{i} \in \mathcal{R}_l} K_\Lambda(\mathbf{i}, \mathbf{s}) = 1$.

Let $b_{x,\mathbf{i}j} = s_{x,\mathbf{i}+j} - s_{x,\mathbf{i}}$ and $b_{y,\mathbf{i}j} = s_{y,\mathbf{i}+j} - s_{y,\mathbf{i}}$, so $(\mathbf{s}_{\mathbf{i}+j} - \mathbf{s}_{\mathbf{i}}) = (b_{x,\mathbf{i}j}, b_{y,\mathbf{i}j})$. The expected value of the \mathbf{i}^{th} squared pseudo-residual is

$$\begin{aligned}
& E(D^2(\mathbf{s}_i; L(h))) \\
&= E\left(\sum_{j \in \mathcal{J}_l} a_j Z_{\mathbf{i}+jh}\right)^2 \\
&= \sum_{j \in \mathcal{J}_l} \sum_{k \in \mathcal{J}_l} a_j a_k \sigma_{\mathbf{i}+jh} \sigma_{\mathbf{i}+kh} \rho_{h\|\mathbf{j}-\mathbf{k}\|} \\
&= \sigma_{\mathbf{i}}^2 \left(1 - \sum_{j \neq k} a_j^2 a_k^2 \rho_{h\|\mathbf{j}-\mathbf{k}\|}\right) + \nabla \sigma_{\mathbf{i}}^2 \sum_{j \in \mathcal{J}_l} a_j^2 (\mathbf{s}_{\mathbf{i}+jh} - \mathbf{s}_{\mathbf{i}}) \\
&\quad + \sum_{j \in \mathcal{J}_l} a_j^2 \left\{ \frac{1}{2} \frac{\partial^2 \sigma_{\mathbf{i}}^2}{\partial s_x^2} b_{x,ij}^2 + \frac{\partial^2 \sigma_{\mathbf{i}}^2}{\partial s_x \partial s_y} b_{x,ijh} b_{y,ijh} + \frac{1}{2} \frac{\partial^2 \sigma_{\mathbf{i}}^2}{\partial s_y^2} b_{y,ijh}^2 + o(\|\mathbf{s}_{\mathbf{i}+jh} - \mathbf{s}_{\mathbf{i}}\|^2) \right\} \\
&\quad + \sum_{j \in \mathcal{J}_l} \sum_{k \in \mathcal{J}_l} a_j a_k \left\{ \nabla \sigma_{\mathbf{i}}(\mathbf{s}_{\mathbf{i}+jh} - \mathbf{s}_{\mathbf{i}}) \nabla \sigma_{\mathbf{i}}(\mathbf{s}_{\mathbf{i}+kh} - \mathbf{s}_{\mathbf{i}}) \rho_{h\|\mathbf{j}-\mathbf{k}\|} + o(\|\mathbf{s}_{\mathbf{i}+jh} - \mathbf{s}_{\mathbf{i}}\|^2) \right\} \quad (4.1)
\end{aligned}$$

If $\sigma_{\mathbf{i}}^2(\cdot)$ is constant, then equation (4.1) reduces down to $\sigma_{\mathbf{i}}^2(1 - \sum_{j \neq k} a_j^2 a_k^2 \rho_{h\|\mathbf{j}-\mathbf{k}\|})$. If $\sigma_{\mathbf{i}}^2(\cdot)$ has a small degree differentiability, then, again, the equation (4.1) becomes dominated by the first term.

Definition 4.2.3 Define a generalized local variogram $\Gamma_{\Lambda}(\mathbf{s}, L(h))$ for a two-dimensional nonstationary random field at location \mathbf{s}_0 and L -filter variogram at lag-scale size h as the leading term in the expected value of $D^2(\mathbf{s}_i; L)$:

$$\Gamma_{\Lambda}(\mathbf{s}; L(h)) = \sigma(\mathbf{s})^2 \left(1 - \sum_{\substack{j \neq k \\ j, k \in \mathcal{J}_L}} a_j^2 a_k^2 \rho_{h\|\mathbf{j}-\mathbf{k}\|}\right). \quad (4.2)$$

4.2.2 Method

We propose the method-of-moments generalized local variogram estimator of $\Gamma_{\Lambda}(\cdot; L(h))$ for a nonstationary isotropic process. Assuming local stationarity for smoothing, we apply a Gasser-Müller kernel as in Definition 4.2.2 to the squared pseudo-residuals D_L^2 of the observations and estimate the generalized local variogram at location \mathbf{s}_0 :

$$\hat{\Gamma}_{\Lambda}(\mathbf{s}_0; L(h)) = \sum_{\mathbf{i} \in \mathcal{R}_l} K_{\Lambda}(\mathbf{i}, \mathbf{0}) D_{L(h)}^2(\mathbf{s}_{\mathbf{i}}). \quad (4.3)$$

Then, define a variance estimator at location \mathbf{s}_0 as

$$\hat{\sigma}_\Lambda(\mathbf{s}_0) = \frac{\hat{\Gamma}_\Lambda(\mathbf{s}_0; L(h))}{1 - \varrho_L(h; \hat{\theta})}, \quad (4.4)$$

averaging the squared pseudo-residuals first as in equation (4.3) and then scaling appropriately by the L -filter variogram, instead of averaging all the scaled pseudo-residual squares, because of gaining robustness in the scale of variance and the correlation parameter estimation. In detail, the numerator is a generalized local variogram estimator as in (4.3) and the denominator is an L -filter variogram as explained in Definition 4.2.1.

In the expectation of the local variogram estimator (4.1), the higher-order terms of the directly squared pseudo-residual are in the first line, and in the following lines are the higher-order terms of the cross-terms of the pseudo-residual. The bias of generalized local variogram estimator (4.3) contains these terms, but the odd order terms are canceled out when the filter shape and weights are symmetric about \mathbf{p}_0 . For example, a $\nu = 3$ -node filter with $\mathcal{J} = \{j_1 = (-1, 0), j_2 = (0, 0), j_3 = (1, 0)\}$ and $A = \{1/\sqrt{6}, -2/\sqrt{6}, 1/\sqrt{6}\}$ has the first-order term in the bias $\sum_{j \in \mathcal{J}} a_j^2 \nabla \sigma_0^2(\mathbf{s}_{i+j} - \mathbf{s}_i)$ expanded about \mathbf{p}_0 as $\nabla \sigma_0^2(\mathbf{s}_{i+(1,0)} - \mathbf{s}_i + \mathbf{s}_{i+(-1,0)} - \mathbf{s}_i) = 0$. The bias arises from small cross-terms.

We study the effect of filter weights for the estimator bias and variance via a numerical method due to a number of terms increasing as the order of the terms increases. For bias we need to consider second-order terms of a random field, and for variance and covariance the fourth-order terms are involved. When the underlying standard deviation function is constant or linear, a normalized linear filter does not introduce any or negligible bias. When the underlying variance function is non-linear, a linear filter interacts with the underlying function and creates a small order bias whose shape closely reflects the filter weight distribution. The size of the bias is negligible, under appropriate conditions, and can be smoothed out using a high-order kernel, whose order should match the degree differentiability of the underlying variance function as seen in Theorem 3.3.3.

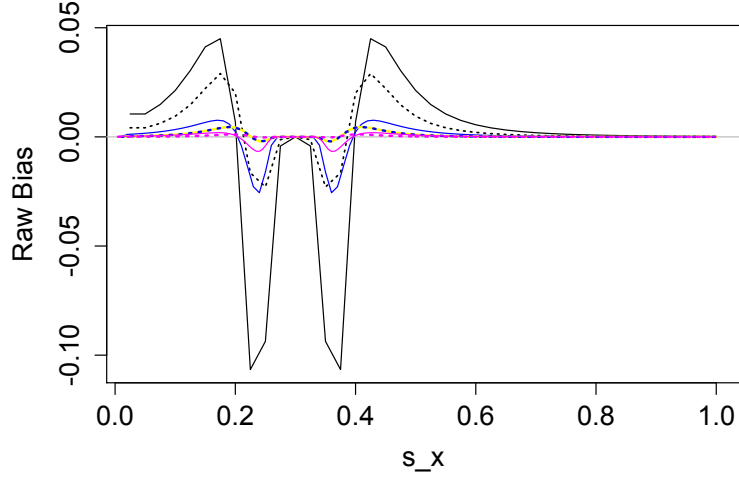


Figure 4.1: Bias in estimation when using a symmetric weight, line configuration filter where $A = \{(1, -2, 1)/\sqrt{6}\}$.

We use Model C in Section 4.4 and equation (4.1) for a numerical calculation of bias and have sliced the bias surface function where $s_y = 0.3$. The size of observations $N = n \times n$ are set to $n = 40$ (in black), 100 (in blue), and 200 (in red). For each sample size, the dashed line represents the bias of a variance function estimation with strong correlation, and the line the case of an independent error process. for a cross-section of raw bias for estimating the variance function which plateaus at $(s_x, s_y) = (0.3, 0.3)$. As the number of sample points increases, the bias goes to 0. The shape of the bias reflects the weight assigned to the filter nodes; with filter weight $(1, -2, 1)/\sqrt{6}$, the bias shows a peak, a trough, a flat 0barea, a trough, and a peak. The flat 0 area about $s_x = 0.3$ is due to the true underlying function having 0 or very small value derivative for a radius of 0.04 about $(s_x, s_y) = (0.3, 0.3)$. As there is no interaction effect in estimation between a linear filter and a linear underlying function. The true variance close to $s_x = 0.18$ is 13, and the raw bias is no larger than 0.05 when $n = 40$. The variance of the variance function also shows the same shape as the bias, and the scale is much more dependent on the underlying function scale.

Table 4.1: Number of Z 's fourth order terms in the expression of $(D_i^2)^2$

Terms	Coefficient	# of Combinations	Condition
Z^4	a_j^4	ν	$\nu \geq 1$
$Z_j^2 Z_k^2$	$6a_j^2 a_k^2$	$\binom{\nu}{2}$	$\nu \geq 2$
$Z_j^3 Z_k$	$4a_j^3 a_k$	$2\binom{\nu}{2}$	$\nu \geq 2$
$Z_j^2 Z_k Z_m$	$12a_j^2 a_k a_m$	$3\binom{\nu}{3}$	$\nu \geq 3$
$Z_i Z_j Z_k Z_m$	$24a_i a_j a_k a_m$	$\binom{\nu}{4}$	$\nu \geq 4$

The non-centralized fourth moment of a general pseudo-residual D_i is

$$\begin{aligned}
E \left((D_i^2)^2 \right) &= E \left(\sum_{j,k} a_j a_k Z_{i+j} Z_{i+k} \right)^2 \\
&= 3 \sum_{j \in \mathcal{J}} a_j^4 \sigma_{i+j}^4 + 3 \sum_{j \neq k \in \mathcal{J}} a_j^2 a_k^2 \{1 + 2\rho_{\|j-k\|}^2\} \sigma_{i+j}^2 \sigma_{i+k}^2 \\
&\quad + 6 \sum_{j \neq k} a_j^3 a_k \rho_{\|j-k\|} \sigma_{i+j}^3 \sigma_{i+k} \\
&\quad + 6 \sum_{j \neq k \neq m} a_j^2 a_k a_m \{2\rho_{\|j-k\|} \rho_{\|j-m\|} + \rho_{\|k-m\|}\} \sigma_{i+j}^2 \sigma_{i+k} \sigma_{i+m} \\
&\quad + \sum_{j \neq k \neq l \neq m} a_j a_k a_l a_m \{ \rho_{\|j-k\|} \rho_{\|k-l\|} + \rho_{\|l-m\|} \rho_{\|j-l\|} \\
&\quad \quad + \rho_{\|k-m\|} \rho_{\|l-m\|} \} \sigma_{i+j} \sigma_{i+k} \sigma_{i+l} \sigma_{i+m}.
\end{aligned} \tag{4.5}$$

The expectation of $D_i^2 D_j^2$ when $i \neq j$ is a bit more complex because the relative position of i and j should be taken into account when combining the filter applied terms.

Table 4.2: Number of Z 's fourth order terms in $D_i^2 D_j^2$ when $Z_{i+\mathcal{J}}$ and $Z_{j+\mathcal{J}}$ are overlapping on one node

Terms	# of Combinations
Z^4	1
$Z_j^2 Z_k^2$	$\nu^2 - 1$
$Z_j^3 Z_k$	$2(\nu - 1)$
$Z_j^2 Z_k Z_m$	$(\nu - 1)(\nu^2 + \nu - 3)$
$Z_i Z_j Z_k Z_m$	$(\nu - 1)^2(\nu - 2) \left(\frac{\nu+2}{4}\right)$

$$\begin{aligned}
& E(D_i^2 D_j^2) \\
&= E \left(\sum_{k_1, k_2 \in \mathcal{J}} a_{k_1} a_{k_2} Z_{i+k_1} Z_{i+k_2} \rho_{\|k_1-k_2\|} \sum_{m_1, m_2 \in \mathcal{J}} a_{m_1} a_{m_2} Z_{j+m_1} Z_{j+m_2} \rho_{\|m_1-m_2\|} \right) \\
&= 3 \sum_{k-m=j-i} a_k^2 a_m^2 \sigma_{i+k}^4 + 3 \sum_{k-m \neq j-i} a_k^2 a_m^2 (1 + 2\rho_{\|i-j+k-m\|}^2) \sigma_{i+k}^2 \sigma_{j+m}^2 \\
&\quad + 3 \sum_{k \neq m} \rho_{\|i-j+k-m\|} (a_k^3 a_m \sigma_{i+k}^3 \sigma_{j+m} + a_k a_m^3 \sigma_{i+k} \sigma_{j+m}^3) \\
&\quad + 6 \sum_{\substack{m_1 \neq m_2 \\ k-m_1 \neq j-i \\ k-m_2 \neq j-i}} a_k^2 a_{m_1} a_{m_2} (2\rho_{\|i-j+k-m_1\|} \rho_{\|i-j+k-m_2\|} + \rho_{\|m_1-m_2\|}) \sigma_{i+k}^2 \sigma_{j+m_1} \sigma_{j+m_2} \quad (4.6) \\
&\quad + 6 \sum_{\substack{k_1 \neq k_2 \\ k_1-m \neq j-i \\ k_2-m \neq j-i}} a_{k_1} a_{k_2} a_m^2 (2\rho_{\|j-i+m-k_1\|} \rho_{\|j-i+m-k_2\|} + \rho_{\|k_1-k_2\|}) \sigma_{i+k_1} \sigma_{i+k_2} \sigma_{j+m}^2 \\
&\quad + \sum_{k_1 \neq k_2, m_1 \neq m_2} a_{k_1} a_{k_2} a_{m_1} a_{m_2} (\rho_{\|k_1-k_2\|} \rho_{\|m_1-m_2\|} + \rho_{\|i-j+k_1-m_1\|} \rho_{\|i-j+k_2-m_2\|} \\
&\quad \quad + \rho_{\|i-j+k_1-m_2\|} \rho_{\|i-j+k_2-m_1\|}) \sigma_{i+k_1} \sigma_{i+k_2} \sigma_{j+m_1} \sigma_{j+m_2}.
\end{aligned}$$

The number of unique fourth order terms $Z_i Z_j Z_k Z_m$ for any possible arrangement of indices i, j, k , and m , which belong to \mathcal{J} , is shown in Table 4.1 for the fourth order terms of D_i . In Table 4.2 the number of fourth order terms are shown for the covariance between D_i^2 and D_j^2 where the filter applied observations $Z_{i+\mathcal{J}}$ and $Z_{j+\mathcal{J}}$ overlap on a

Table 4.3: Number of Z 's fourth order terms in $D_i^2 D_j^2$ when $Z_{i+\mathcal{J}}$ and $Z_{j+\mathcal{J}}$ are overlapping on two nodes

Terms	# of Combinations
Z^4	2
$Z_j^2 Z_k^2$	$\nu^2 - 1$
$Z_j^3 Z_k$	$4(\nu - 1)$
$Z_j^2 Z_k Z_m$	$2(\nu - 2)(2\nu - 3)$
$Z_i Z_j Z_k Z_m$	$(\nu - 2)(3\nu - 7)$

single point. Table 4.3 shows the same information as Table 4.2 but the filter applied observations overlap on two points. Depending on ν , the order (size) of a filter, the total number of unique terms is determined. From these three tables, we see that there are many error terms to rising in variance function estimation. An analytical study becomes complicated as ν increases and a filter becomes less “sparse” (i.e. $Z_{i+\mathcal{J}}$ and $Z_{j+\mathcal{J}}$ overlap at most one node), and we rest to simulation study to investigate the risk of the variance function estimation.

4.3 Properties of Difference Filter

As noted in Conditions 1 and 2 of Section 4.2.1, we require the weights of a linear filter L to sum to zero and the sum of squared weights to be one. Then, for any stationary random field $X_{\mathbf{s}}$ with a constant variance σ^2 , the h -scale difference filtered process of $X_{\mathbf{s}}$ has mean zero and the variance as the L -filter-variogram, $1 - \varrho_L(h)$. With Condition 3 in Section 4.2.1, we expect the weight center of L to be within the closed periphery of L so that a minimal shift bias is introduced from the center of L should the filter have a symmetric weight distribution. As we have discussed in Section 4.2.2, the variance of our variance estimator is dependent on the design of filter L . To achieve statistical efficiency in the variance of the estimator at a fixed location or in the functional estimation, we should explore different filter configurations and weight options.

4.3.1 Configuration of Difference Filter

We investigate five configurations of difference filters to estimate a nonstationary process variance function. Four of these five are suggested by Hall et al. (1991) as two-dimensional difference filters for independent and identically distributed error variance estimation. They recommend compact, linear, or sparse configurations, where sparsity means the overlap between two off-set configurations occurs at most on one node.

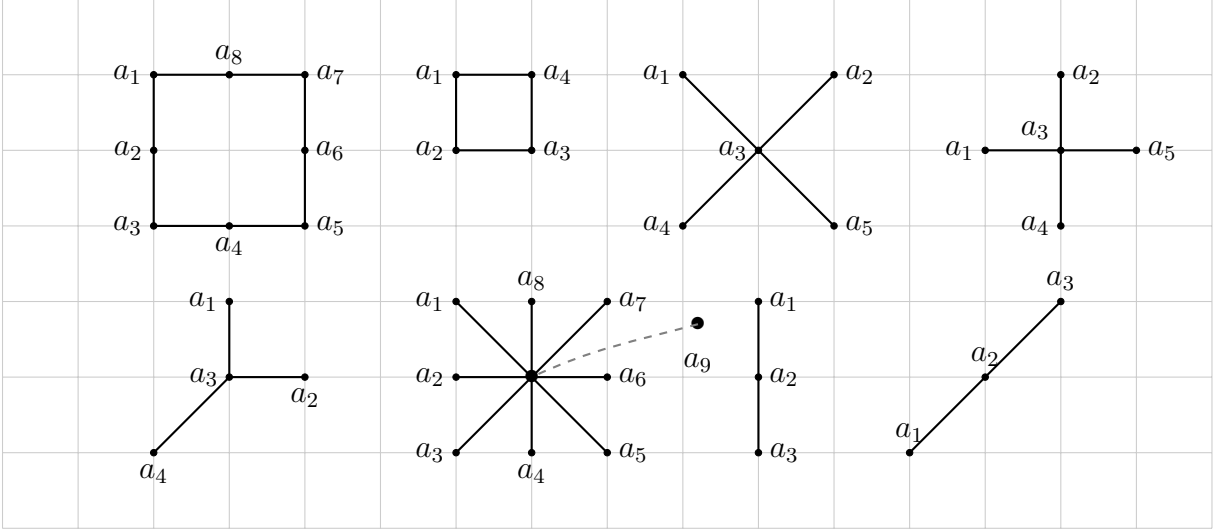


Figure 4.2: Five different configurations. In the top row from left to right, there are Square3, Square2, two crosses of $h = \sqrt{2}$ and 1. In the bottom row from left to right, there are Y, four rotated line filters making a star, and a single linear filter of $h = 1$ and $\sqrt{2}$.

Figure 4.2 shows the filters we investigate. There are two square-shaped filters on the top left, which are in fact two different configurations. Both have $h = 1$, but the leftmost one, name it Square3, spans $\frac{2}{n} \times \frac{2}{n}$ with $\nu = 8$, whereas the second one, name it Square2, spans $\frac{1}{n} \times \frac{1}{n}$ and has $\nu = 4$. Following these to the right are two cross-shaped filters with $\nu = 5$. These are the same configurations except for the rotation and the scale h : the \times -configuration has $h = \sqrt{2}$, and the $+$ configuration $h = 1$. On the bottom left, the Y configuration is asymmetrical about a_3 . For such asymmetric filters, in order to reduce a potential directional bias introduced by the interaction with the underlying estimating

function, we apply a rotation matrix R where ϕ 's are at 90° , 180° , and 270° angles about a_3 . For the final estimation, we average the four directional Y estimation results. For the line configuration at the bottom right, the same rotation idea applies, but the diagonal lines need the scaling of $h = \sqrt{2}$ like the \times configuration. Applying the rotation matrix R to \mathcal{J}_l again at 45° with $h = \sqrt{2}$, 90° , and 135° with $h = \sqrt{2}$ and combining those line filters. From both Y and line filter rotations, we get the star configuration filter as in the middle of the second row. However, they have different weights applied to each of the star node.

4.3.2 Determining Weights

The conditions imposed on a differencing filter do not determine the weights uniquely unless the number of conditions matches the order ν of a filter. The Conditions 1, 2, and 3 in Section 4.2.1 contain four conditions for the weights because Condition 3 has two equations for the weights one each in the x and y directions. Hence, the weights for any filter with $\nu \leq 4$ are uniquely paired with a weight center \mathbf{p}_0 under Conditions 1-3. For $\nu \geq 5$ we impose an additional condition:

Condition 4. Let the weights be symmetrically distributed about $(0,0)$.

We call a filter *symmetric* if it satisfies Conditions 1-4. If a filter does not satisfy Condition 4, for example a Y configuration in the lower left of Figure 4.2, then we apply a rotation matrix R to \mathcal{J} and combine the fully rotated filters so that the resulting averaged filter should meet Condition 4. For a Y configuration, we have $\mathcal{J} = \{(0,1), (1,0), (0,0), (-1,-1)\}$ and $A_Y = \{1,1,-3,1\}/\sqrt{12}$, whose weights are centered at $(0,0)$, but not all the grid points at distance 1 or $\sqrt{2}$ from $(0,0)$ having the weight $1/\sqrt{12}$. Hence, we apply a rotation matrix of degree 90° , 180° , and 270° to the Y configuration and complete the symmetry. For Square2, Square3, and $+$ configurations, there is no need to apply a rotation matrix due to their symmetry in configuration unless a rectangular lattice requires an adjustment to the filter configuration. In Appendix

A.2.1, we list the five configurations of Figure 4.2 with symmetric weights. The weight derivations for Y, +, and Square3 filters are shown in Appendix A. Since Square3 has $\nu = 8$, it does not have a unique set of weights satisfying Conditions 1-3. However, using the last symmetry condition and the non-zero weight implication from $\nu = 8$, we get $A = \{-1, 1, -1, 1, -1, 1, -1, 1\}/\sqrt{8}$.

We refer to the filter weights as *Hall-Kay-Titterington* weights, in short HKT weights, when they are computed to minimize the variance of a difference-based variance estimator. Hall et al. (1991) assume independent and identically distributed errors and derive the weights analytically for filters with $\nu \leq 4$ and numerically for $\nu \geq 5$. When the underlying error process is independent and identically distributed, the fourth order terms of the pseudo residuals have a relatively concise expression compared to those in equations (4.5) and (4.6) since any odd order of combination Z_j 's would render 0 for the expectation of the fourth order terms. Assuming we can find a stationary process $\{X_s\}$ for a mean 0 nonstationary $\{Z_s = \sigma(s)X_s\}$, we impose the following condition for HKT weights:

Condition 5. Let the weights minimize
$$\sum_{j_1 \neq j_2, k_1 \neq k_2 \in \mathcal{J}} a_{j_1} a_{j_2} a_{k_1} a_{k_2} E(X_{j_1} X_{j_2} X_{k_1} X_{k_2}).$$

For independent error processes, the optimal HKT weights are shown in Appendix A.2.2. The weight centers of these filters are loaded on one end of the configuration, with numerical rounding error shifting the center slightly away from each filter, as marked with '×'. For correlated processes, the optimal weights must be different from the HKT weights for independent and identically distributed processes due to all combinations of j_1, j_2, k_1 and k_2 $E(X_{j_1} X_{j_2} X_{k_1} X_{k_2})$ being non-zero for a mean-zero correlated process $\{X_j\}$. Instead of deriving conditions we use the HKT weights for an independent and identically distributed process in the simulation study and compare the variance estimation results against the case for symmetric weights in Section 4.4.

4.3.3 L -filter variogram

An L -filter variogram (Definition 4.2.1) is determined by a set of weights A , the relative locations \mathcal{J} of a filter L , and the underlying correlation structure $\rho(\cdot)$ of the data. As it summarizes the dispersion of multi-dimensional correlated data by applying a difference filter L and squaring, the name contains L -filter and variogram. The expectation of the cross-terms in a squared pseudo-residual depends on the filter L . In Table 4.4, assuming an exponential correlation structure, we provide the L -filter variogram values for both symmetric and HKT weights for $h = 1$ of the five filters showcased in Figure 4.2 and for \times shape with $h = \sqrt{2}$. Table 4.5 shows the regular variogram values with an exponential correlation function. They are in a range similar to that of the symmetric L -filter variograms in Table 4.4. When the weights are symmetrically distributed about $(0,0)$ and the correlation is approximately $\rho(h)$ between cross-terms, the sum of symmetric filter cross-terms is approximately $-\rho(h)$ since the coefficients of the cross-terms should sum to negative one as mentioned in Remark 4.2.1.

Like a variogram the larger the range parameter of a correlation function is, the smaller the L -filter variogram value is at a fixed h/n . Also, by comparing the scales of the $+$ and \times shapes in the last two columns or by implicitly changing the scale as we vary n from 40 to 100, we see that when the scale h/n is larger, the L -filter variogram value is larger. We also note that the HKT weights have a larger L -filter variogram at a fixed h/n . In symmetric filters, the weights are evenly distributed among the nodes on the periphery, so the cross-terms between the central node, which is the weight center, and the peripheral nodes take large negative coefficients while the cross-terms generated among the peripheral nodes take small positive coefficients. In HKT filters, a weight considered in an absolute scale is loaded at one end; therefore the cross-terms have coefficients of the same scale with negative signs except for the pairs with the loaded node. This results in larger values of L -filter variograms for HKT weights than symmetric weights.

Table 4.4: L -filter variogram at the smallest possible scale h for each filter with two weight options mentioned in Section 4.3.2






Weight	n	θ						
Symmetric	40	$\theta = 0.1$	0.16	0.18	0.16	0.14	0.16	0.22
		$\theta = 0.01$	0.89	0.91	0.87	0.86	0.88	0.96
	100	$\theta = 0.1$	0.07	0.07	0.06	0.06	0.06	0.09
		$\theta = 0.01$	0.56	0.59	0.53	0.51	0.54	0.68
HKT	40	$\theta = 0.1$	0.31	N/A	0.28	0.37	0.25	0.40
		$\theta = 0.01$	0.96		0.96	0.99	0.94	0.97
	100	$\theta = 0.1$	0.14	N/A	0.12	0.17	0.11	0.20
		$\theta = 0.01$	0.75		0.73	0.85	0.67	0.82

Table 4.5: Exponential variogram $1 - \rho\left(\frac{1}{n}; \theta\right)$

$n \backslash \theta$		
	0.1	0.01
40		
	0.22	0.92
100		
	0.10	0.63

Here we provide the formulae of L -filter variograms for the five configurations in Figure 4.2. Instead of our usual notation $\rho_{h,n}$ or ρ_h standing for $\rho\left(\frac{h}{n}\right)$, we use $\rho(h)$ for the readability of the scale parameter.

(a) Square2, 2×2 square ($\nu = 4$):

$$\varrho_L(h) = 1 + 2(a_1a_3 + a_2a_4)\rho_\theta(\sqrt{2}h) + 2(a_1 + a_3)(a_2 + a_4)\rho_\theta(h)$$

(b) Square3, 3×3 square ($\nu = 8$):

$$\begin{aligned} \varrho_L(h) = & 1 + 2(a_1a_2 + a_2a_3 + a_3a_4 + a_4a_5 + a_5a_6 + a_6a_7 + a_7a_8 + a_8a_1)\rho_\theta(h) \\ & + 2\{a_2(a_5 + a_7) + a_4(a_7 + a_1) + a_6(a_1 + a_3) + a_8(a_3 + a_5)\}\rho_\theta(\sqrt{5}h) \\ & + 2(a_2 + a_6)(a_4 + a_8)\rho_\theta(\sqrt{2}h) + 2(a_1a_5 + a_3a_7)\rho_\theta(2\sqrt{2}h) \\ & + 2(a_1 + a_5)(a_3 + a_7)\rho_\theta(2h) \end{aligned}$$

(c) + configuration ($\nu = 5$):

$$\varrho_+(h) = 1 - 2(a_3)^2\rho_\theta(h) + 2(a_1 + a_5)(a_2 + a_4)\rho_\theta(\sqrt{2}h) + 2(a_1a_5 + a_2a_4)\rho_\theta(2h)$$

(d) \times configuration ($\nu = 5$): Scale h in (c) by $\sqrt{2}$

(e) Y-configuration ($\nu = 4$):

$$\varrho_Y(h) = 1 + 2a_1(a_2 + a_3)\rho_\theta(h) + 2(a_1a_2 + a_3a_4)\rho_\theta(\sqrt{2}h) + 2(a_1 + a_2)a_4\rho_\theta(\sqrt{5}h)$$

(f) Line configuration ($\nu = 3$):

$$\varrho_*(h) = 1 - 2(a_2)^2\rho_\theta(h) + 2a_1a_3\rho_\theta(2h)$$

4.4 Simulation Study

We have undertaken a simulation study to measure relative filter efficiency and to circumvent the complicated derivation of filter bias and variance as the analytical formulae show in equations (4.1), (4.5), and (4.6) along with the number of the fourth order terms summarized in Tables 4.1, 4.2 and 4.3.

4.4.1 Data Model and Measures of Estimation

We simulate a Gaussian random field with an exponential correlation function on a regular square lattice design over a unit square, $[0, 1]^2$. The numbers of sampling points per square lattice are $N = 40 \times 40$ and 100×100 . The dependence structure has three levels: independent, weak correlation with the exponential function range parameter set to $\theta = 0.01$, and strong correlation with the range parameter $\theta = 0.1$. The following three standard deviation functions are multiplied to stationary error processes to generate heteroscedastic processes with mean 0.

Model A. $\sigma(s_x, s_y) = s_x + 2s_y + 1$

$$\text{Model B. } \sigma(s_x, s_y) = \begin{cases} 1 & \text{for } 0 \leq s_x < 1/2 \\ 4s_x - 1 & \text{for } 1/2 \leq s_x < 3/4 \\ 2 & \text{for } 3/4 \leq s_x \leq 1 \end{cases}$$

$$\text{Model C. } \sigma(s_x, s_y) = 4 - \exp\left(-\frac{0.12^2}{(s_x - 0.3)^2 + (s_y - 0.3)^2}\right)$$

Figure 4.3 displays the standard deviation function Models A, B, and C from left to right, and the top row shows two-dimensional heat maps and the bottom row has three-dimensional perspective drawings. The three standard deviation functions represent Figure 4.4 showcases a realization of a random field under the Model A set-up where the top row has $n = 100$ and the bottom row $n = 40$. Column-wise from left to right, the data process is generated using an independent model, weak correlation, and strong correlation.

In terms of the range of values, Model A $\sigma(\cdot)$ has a range between 1 and 4, Model B between 1 and 2, and Model C between 3 and 4. When they are transformed to variance functions, the range widens. For any estimating location \mathbf{s} , the larger the signal $\sigma^2(\mathbf{s})$, the larger the estimation error. Thus, we define a standardized deviance $\epsilon(\mathbf{s})$ at any

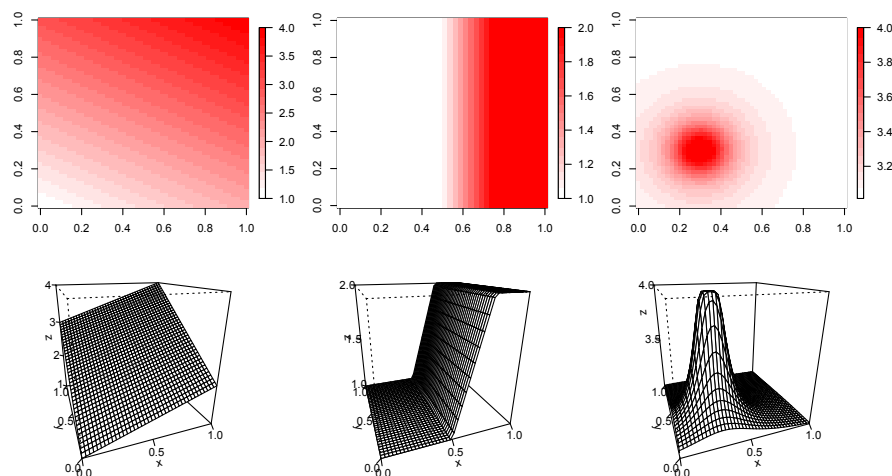


Figure 4.3: Three $\sigma(s_x, s_y)$ functions in a heat-map (top row) and a 3-dimensional perspective drawing (bottom row). The first column shows Model A, the middle Model B, and the last Model C.

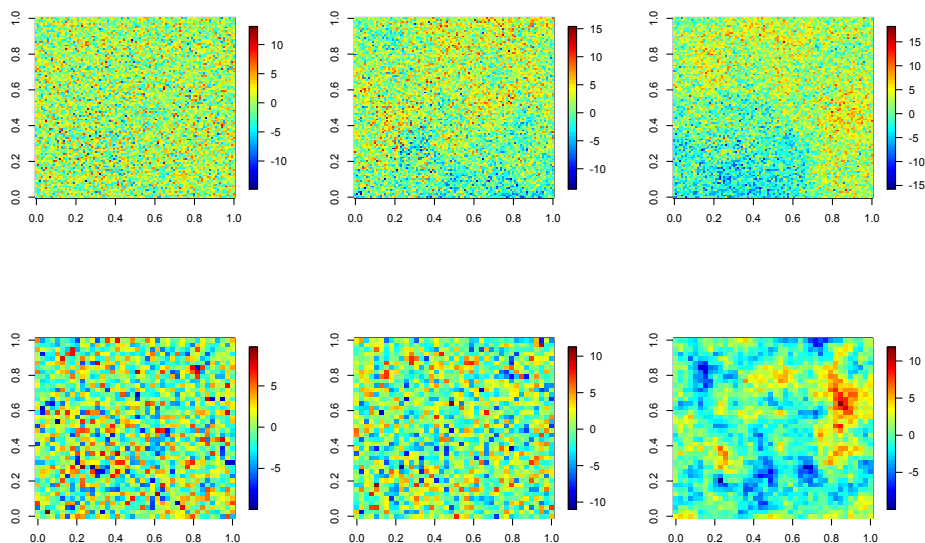


Figure 4.4: Examples of nonstationary data with $\sigma(s_x, s_y)$ of Model A. The top row has 100×100 points and bottom row has 40×40 points. Column-wise, from left to right, we see independent error (left), exponential correlation function with the range $\theta = 0.01$ (middle), and $\theta = 0.1$ (right).

location $\mathbf{s} \in [0, 1]^2$

$$\hat{\epsilon}_\Lambda(\mathbf{s}) = \frac{\hat{\sigma}_\Lambda^2(\mathbf{s}) - \sigma^2(\mathbf{s})}{\sigma^2(\mathbf{s})} \quad (4.7)$$

as a scaled the raw error, $\hat{\sigma}_\Lambda^2(\mathbf{s}) - \sigma^2(\mathbf{s})$. It is a reasonable error term when comparing statistical efficiency of the variance estimation at two different locations.

The following measures are used for summarizing estimation errors on a lattice where the estimating variance function is relatively flat.

- $DMSE_\Lambda(L_\nu) = \frac{1}{N_l} \sum_{\mathbf{i} \in \mathcal{R}_l} (\hat{\sigma}_\Lambda^2(\mathbf{i}) - \sigma^2(\mathbf{i}))^2$
- $MAD_\Lambda(L_\nu) = \frac{1}{N_l} \sum_{\mathbf{i} \in \mathcal{R}_l} |\hat{\sigma}_\Lambda^2(\mathbf{i}) - \text{median}_{\mathbf{i} \in \mathcal{R}_l} (\hat{\sigma}_\Lambda^2(\mathbf{i}) - \sigma^2(\mathbf{i}))|$
- $MAX_\Lambda(L_\nu) = \max_{\mathbf{i} \in \mathcal{R}_l} |\hat{\sigma}_\Lambda^2(\mathbf{i}) - \sigma^2(\mathbf{i})|$

A nonstationary process often has a wide range of values of a variance function, so we use the standardized deviances $\hat{\epsilon}(\mathbf{s})$ defined in (4.7) in place of the estimated raw errors and call them *relative DMSE* ($rDMSE$), *relative MAD* ($rMAD$), and *relative MAX* ($rMAX$).

- $rDMSE_\Lambda(L_\nu) = \frac{1}{N_l} \sum_{\mathbf{i} \in \mathcal{R}_l} \hat{\epsilon}_\Lambda(\mathbf{s}_i)^2$
- $rMAD_\Lambda(L_\nu) = \frac{1}{N_l} \sum_{\mathbf{i} \in \mathcal{R}_l} |\hat{\epsilon}_\Lambda(\mathbf{s}_i) - \text{median}_{\mathbf{i} \in \mathcal{R}_l} \hat{\epsilon}_\Lambda(\mathbf{s}_i)|$
- $rMAX_\Lambda(L_\nu) = \max_{\mathbf{i} \in \mathcal{R}_l} |\hat{\epsilon}_\Lambda(\mathbf{s}_i)|$

Multiplying 100 to the standardized deviances should show the overall percentage of deviation from the true signal when plugged into the relative summary measures above. Note that the MAX and $rMAX$ measures are different in that the location-by-location measures do not correlate linearly unless the surface of estimation is relatively flat. To be precise, the relative measures scale the raw deviations by the true variances, and the percentage difference in the error provides new information. Briefly we note that the Model B estimation summary is in Table 4.6, which is further explained in the

next subsection, and in the last column of *MAX* the absolute scale measure is more reasonable as the summary mean is 1.17 with the standard error of 0.3 than the *rMAX* with the average about 233% and the standard error of 80%. These large *rMAX* values must be from where the true functional value is small yet the deviation is relatively large. In Figure 4.3 Model B, in the middle column, which has two levels of constant functions and the steep plane to join them, the large *rMAX* occurred near the steep plane on the low constant surface, while the large *MAX* occurred on either levels of constant surface. When we collapse the two-dimensional structure of the data to a one-dimensional measure, we lose sight of where the maximum deviation occurs and what the relative size of the spread is. Hence, it is worth considering the originating location of the summary measures by dividing sections on the lattice.

4.4.2 Results

There are three main conclusions we draw from the simulation study. First, the dependency structure of the data affects the weighting options of linear filters. HKT weights are the most efficient for independent and identically distributed error variance estimator under certain regularity conditions on a mean function as shown in Hall et al. (1991) through the minimization of the estimator variance. When we assume independent yet changing levels of errors as in the leftmost side-by-side boxplot of Figure 4.5, the HKT filter weights (in blue) still show better statistical efficiency for nonstationary independent error random field variance function estimation than the symmetric weights (in white) do. However in the rightmost side-by-side boxplots where the correlation is strong, the symmetric weight filters display greater statistical efficiency in the estimation than the HKT weight filters. When the correlation structure is present but weak, there is not a clear choice of a weighting scheme for a statistical efficient difference filter. The answer depends heavily on the configuration of a filter.

In Figure 4.6 we compare the two weighting schemes specifically for three configura-

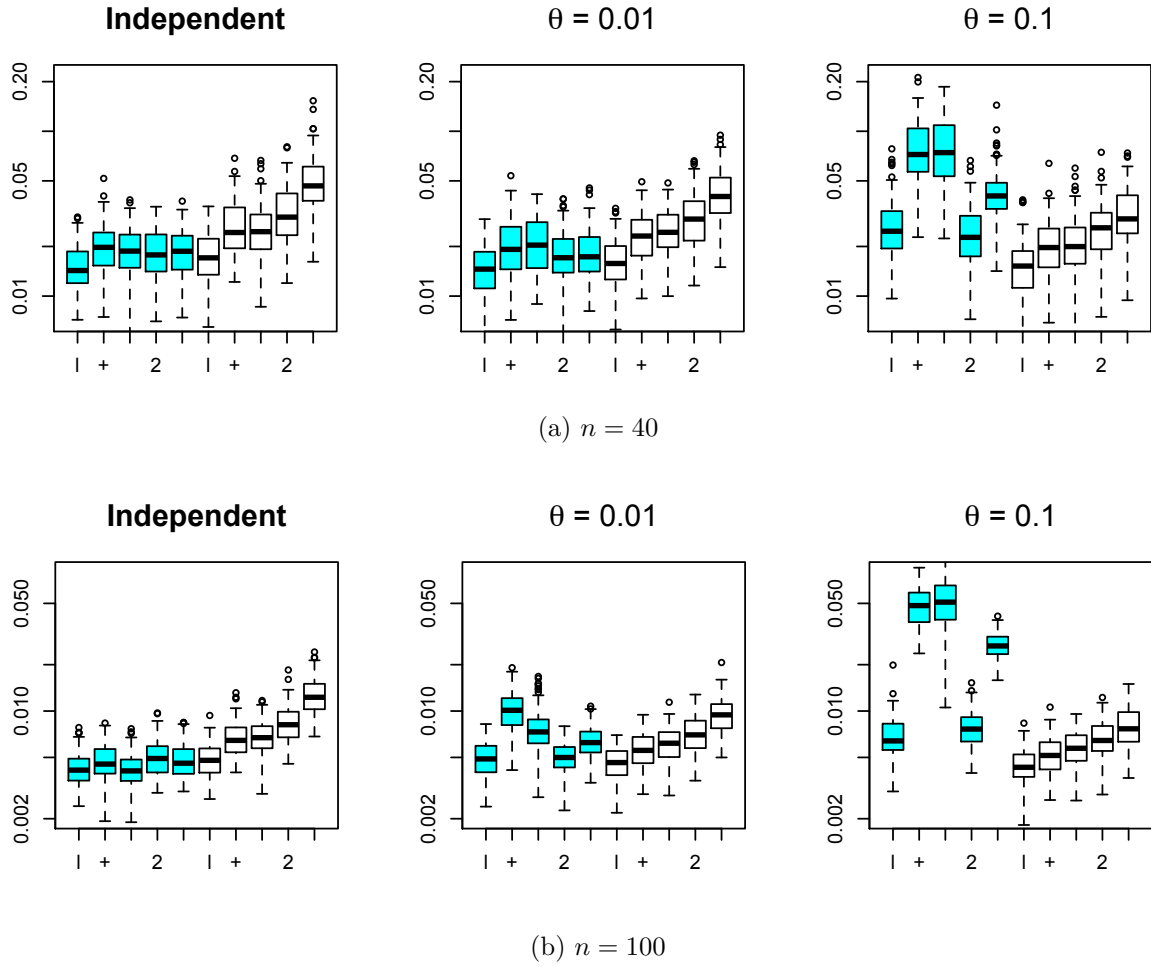


Figure 4.5: HKT weight (in light blue) versus Laplace weight (in white) filter estimation comparison using all six filters in Figure 4.2 in the order of: line, +, \times , Square2, and Square3. The summary measure is the $rMSE$ for Model C.

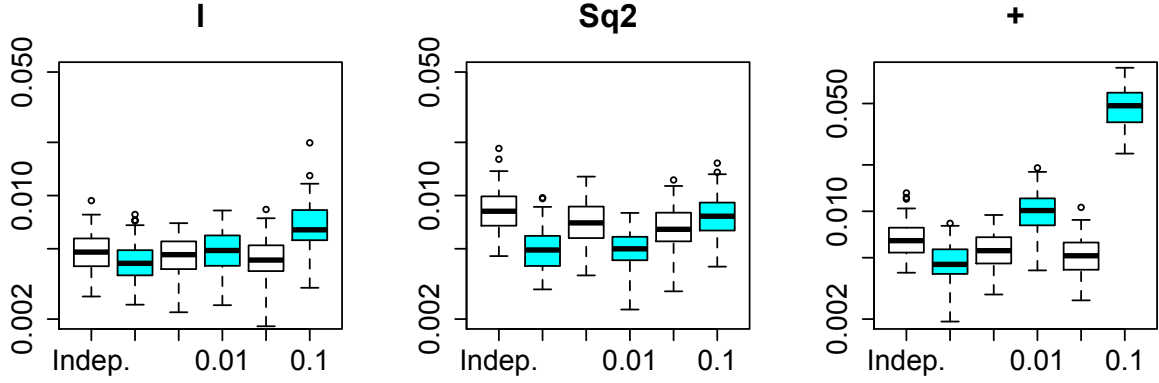


Figure 4.6: Symmetric (in white) and HKT (in light blue) weight filter compared using three configurations: directionally-averaged line, Square2, and $+$. The summary measure is $rMSE$; the underlying $\sigma(\cdot)$ is Model C; and $n = 100$. In each plot there are three levels of dependency presented: independent error from the first pair, $\theta = 0.01$ in the middle, and $\theta = 0.1$ in the third pair.

tions: directionally-averaged line configuration in the left plot, Square2 in the middle, and $+$ in the right plot. For a directionally-averaged line configuration, both weighting schemes show a similar range of $rDMSE$. Yet, when the correlation is strong (i.e. $\theta = 0.1$), the symmetric weight filter performs quite better than the HKT weight filter. Notice that the Square2 configuration provides a similar story but with a clearer contrast when the correlation is weak to none. The HKT weight shows a smaller $rDMSE$ for independent and weakly correlated error random fields, but when the dependency becomes strong, both weighting schemes perform similarly. For the $+$ configuration, when a random field is correlated (i.e. $\theta \geq 0.01$), the contrast is obvious: the symmetric weight filter has a smaller $rDMSE$ than the HKT weights. Taking the estimation results of the \times and Square3 configurations from Figure 4.5, we also see the same phenomena as the $+$ configuration.

The second conclusion regards the effect of directional rotation and averaging. The line filter estimation averaged over four directions gives the smallest $rDMSE$ among all other configurations of filters regardless of the weighting scheme. The estimation sum-



maries reveal that the line and Y configurations that are directionally-averaged perform the best. Directional rotation and averaging helps the efficiency of estimation more than applying a symmetric difference filter. Since the symmetric weight filters perform better than HKT filters, or at least not worse for a Square2 configuration under strong correlation, we compare the performances among symmetric weight filters and investigate filter choices for a random field with $\theta = 0.1$. The summaries of Model B estimation are in Table 4.6. The raw and relative scale summaries of $DMSE$, MAD , and MAX are shown. In parentheses are the standard error of each summary measure from simulation.

Table 4.7 provides a five-point-summary of $rMAD$ in percentage for the Model C standard deviation function estimation where $n = 100$ using the line and Y configurations, and in the last two columns are the mean performances of a single filter and the directionally averaged filter. Comparing the top three estimation summaries of the line configuration to the bottom three of the Y configuration without directional rotation and averaging, we see that a single Y filter performs better than a single line filter with $\nu = 3$. The rotation angle does not affect the estimation performance for both configurations due to the fineness of the observation on lattice ($n = 100$) in comparison to the changing levels of the variance function. In other words, all four directions for the Y configuration (i.e. 90° , 180° and 270°) and the line configuration for the line configuration (45° , 90° , and 135°) show more or less the same range of estimation summary values. When directional rotation and averaging is performed, however, the estimation performances of the two filters become very similar as we see in the last column of Table 4.7. In other words, a directionally averaged Y configuration estimation dose not improve on the $rMAD$ measure as much as the line configuration does. Two other variance functions estimation give the same result. The underlying variance function and the filter configuration have an interaction, and it is necessary to have a spread-out configuration of a filter when constructing pseudo-residuals instead of a line configuration, as the filters should capitalize on the isotropic property of a random field. Hence, we recommend

Table 4.6: Comparing six symmetric weight filters via discretized mean square error, median absolute deviation, maximum absolute deviation of Model B where the correlation range parameter $\theta = 0.1$

		Filters							
	n	\mid	\nearrow	\square	\square	$+$	\times		
$DMSE$	Original	40	0.15 (0.06)	0.16 (0.07)	0.18 (0.08)	0.19 (0.10)	0.21 (0.10)	0.24 (0.09)	
		100	0.04 (0.01)	0.04 (0.01)	0.05 (0.01)	0.05 (0.01)	0.06 (0.02)	0.07 (0.02)	
	Relative %	40	4 (1)	4 (1)	4 (2)	4 (1)	6 (2)	6 (2)	
		100	1 (0.2)	1 (0.3)	1 (0.3)	1 (0.4)	2 (0.4)	59 (20.9)	
	MAD	Original	40	0.25 (0.04)	0.25 (0.05)	0.26 (0.05)	0.27 (0.05)	0.29 (0.06)	0.31 (0.06)
			100	0.13 (0.02)	0.13 (0.02)	0.14 (0.02)	0.14 (0.02)	0.15 (0.02)	0.16 (0.02)
Relative %		40	15 (3)	15 (3)	16 (3)	17 (3)	18 (4)	19 (4)	
		100	8 (1)	9 (1)	9 (1)	10 (1)	0.10 (1)	28 (7)	
MAX		Original	40	1.49 (0.50)	1.63 (0.60)	1.76 (0.66)	1.81 (0.72)	1.95 (0.67)	2.05 (0.75)
			100	0.95 (0.20)	0.98 (0.21)	1.02 (0.23)	1.05 (0.24)	1.11 (0.27)	1.17 (0.30)
	Relative %	40	88 (21)	81 (22)	85 (26)	85 (23)	104 (31)	97 (32)	
		100	49 (11)	50 (13)	51 (13)	57 (15)	54 (14)	233 (80)	

Table 4.7: $rMAD$ comparison of line versus Y configurations. Without directional rotation and averaging in the 5-point summary, Y gives a smaller $rMAD$ (in %) across the range of dependence structure in the data. With directional rotation and averaging shown in the last column, both the line and Y configurations perform similarly.

Shape	θ	$rMAD$ (%)						Dir. Avg.
		Min.	Q1	Median	Q3	Max.	Mean (Stdev.)	
	0	5.10	7.10	8.20	9.20	12.20	8.20 (1.40)	6.52 (1.02)
	0.01	5.00	7.10	7.80	8.60	11.80	7.80 (1.20)	6.13 (0.87)
	0.1	5.70	7.20	7.90	8.60	12.10	8.00 (1.20)	6.11 (1.10)
	0	4.80	6.10	7.00	7.80	9.30	6.90 (1.10)	6.69 (1.08)
	0.01	3.80	6.10	6.60	7.20	9.50	6.70 (1.00)	6.26 (0.95)
	0.1	4.50	5.80	6.60	7.30	10.10	6.60 (1.20)	6.17 (1.06)

a directionally averaged line and Y configurations or a + configuration to capture the locally stationary neighborhood characteristics.

Figure 4.7 re-presents the information summarized in Table 4.6 via side-by-side boxplots where the left plot contains the MAD summaries of Model B standard deviation function estimation and the right plot contains the MAX summaries. In each plot, the first set of six boxplots summarizes the estimation results for $n = 40$, and the second set of six boxplots contains the summaries for $n = 100$; each set contains directionally averaged line, directionally averaged Y, + configuration, 45° rotated + at $h = \sqrt{2}$, Square2, and Square3 and of symmetric weights. Given that Model B $\sigma(\cdot)^2$ ranges from 1 to 4, it is reasonable that the raw scaled MAD is between 0.15 and 0.45 for $n = 40$ and the same measure ranges between 0.07 and 0.18 for $n = 100$. The MAX is much larger in scale especially when $n = 40$ as the simulation result shows values between 0.9 and 4, and when $n = 100$, the MAX is between 0.8 and 1.5. The ranges of summary measures vary depended on the underlying functions of estimation, but the relative standing of each configuration remains the same as shown in Figure 4.7. The first three configurations (including directionally rotated and averaged and scaled for rotation in + configuration) are preferred over the last two, which are Square2 and Square3, and the difference be-

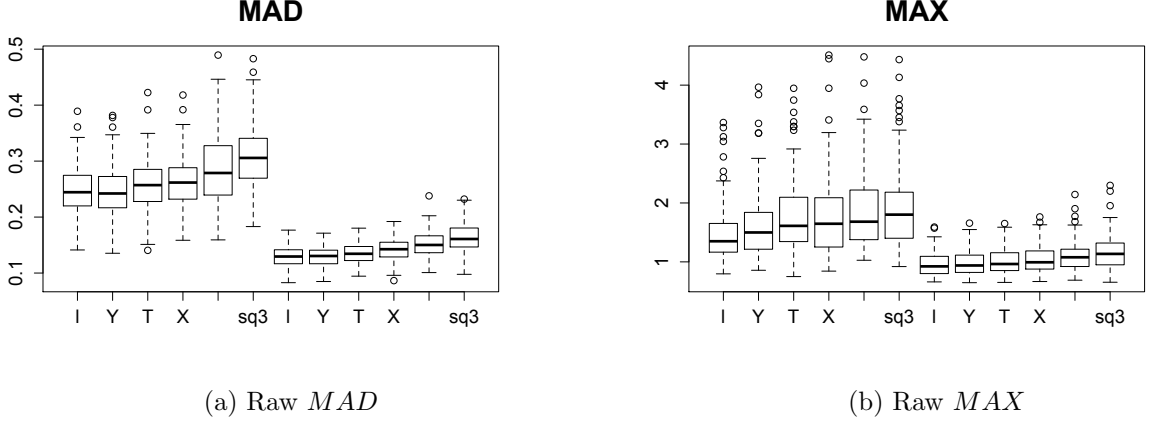


Figure 4.7: Side-by-side boxplots of six symmetric weight filters' estimation summary. The first six are for data with $N = 40 \times 40$. The second six are for $N = 100 \times 100$. We use Model B $\sigma(\cdot)^2$ and set $\theta = 0.1$

Table 4.8: $rMAD$ % comparison of $\nu = 2$ simple differencing filters on the shortest direction vs. diagonal direction where $n = 100$.

		$rMAD$ (%)					
	θ	Min.	Q1	Median	Q3	Max.	Mean (Stdev.)
independent	—	4.49	6.28	6.99	7.96	10.22	7.16 (1.17)
	/	4.95	6.62	7.17	7.75	10.10	7.23 (1.10)
dependent ($\theta = 0.1$)	—	5.36	6.99	7.81	8.93	12.45	7.94 (1.41)
	/	6.34	8.21	8.98	10.13	14.26	9.26 (1.65)

tween the divisions is whether the point of estimation is included in the pseudo-residual (the former group of configurations) or not (the latter group).

Lastly, the more compact a filter is, the more precise the estimation result is. In Figure 4.7 and Table 4.6, between $h = 1$ and $h = \sqrt{2}$ of the + configuration, the compact case of $h = 1$ has slightly smaller summary measures when $n = 100$. Table 4.8, additionally, summarizes the $rMAD$ of a simple $\nu = 2$ filter estimating Model A at two scales $h = 1$ and $\sqrt{2}$. Since Model A is a linearly increasing standard deviation function, the bias of the estimator is negligible, and the $rMAD$ measure should be a good proxy for the spread of the variance estimator. In the top two rows of Table 4.8

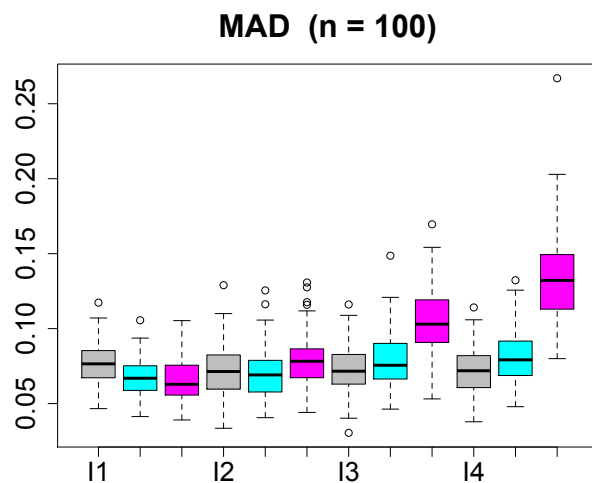


Figure 4.8: Comparison of the line filter scale effect on the estimation depending on the data dependency via the MAD summary of model A $\sigma^2(\cdot)$ estimation using a 3-point linear filter. $N = 100 \times 100$ and the data are independent (gray), $\theta=0.01$ (light blue), and $\theta=0.1$ (pink) for every set of three, when the scale $h=1,2,3$, to 4.

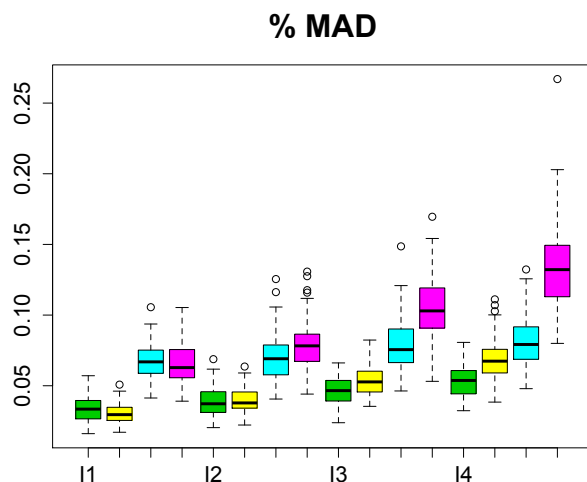


Figure 4.9: Study of the filter scale effect on the estimation depending on the correlation structure of the data and grid size (n) via the MAD summary of model A $\sigma^2(\cdot)$ estimation using a 3-point linear filter. There are four settings on each filter scale $h = 1, 2, 3$ and 4. The green boxplot represents $n = 200$ and $\theta = 0.01$; the yellow boxplot $n = 200$ and $\theta = 0.1$; the light blue boxplot $n = 100$ and $\theta = 0.01$; and the pink boxplot $n = 100$ and $\theta = 0.1$.

where the simulated errors are independent, the performance of the estimators are the same for both scales. In other words, in a small range the scale of a filter does not have an impact on the spread of the variance estimator. However, in the bottom two rows where the errors are strongly correlated, the nearest neighbor differencing of $h = 1$ filter shows a smaller measure of error than the diagonal differencing filter with $h = \sqrt{2}$. These two cases do not imply that in every filter selection, the most compact and the smallest h filter should guarantee the smallest error measures. The data size, as well as the estimating function, should affect the estimation result.

In order to investigate the effect of scale h on the estimation, h is set at 1, 2, 3, and 4 for the line $\nu = 3$ configuration. The simulated data contains Model A in the standard deviation function, and there are three levels for the dependent structure: independent (gray), weak correlation (light blue), and strong correlation (pink). Figure 4.8 contains the estimation summary. From left to right, there are four sets of three side-by-side boxplots displaying $rMAD$. The three colors display independent, weakly correlated, and strongly correlated data, and the four sets are marked by the scale: l1 for $h = 1$, l2 for $h = 2$, l3 for $h = 3$, and l4 for $h = 4$. In each set, except for l1 ($h = 1$), as the correlation becomes stronger, the $rMAD$ becomes larger; and the slope of increase increases as the filter scale h becomes larger. When $h = 1$, the dependency structure may not seem to affect the performance of the estimator because of the fineness of the observation $n = 100$.

For further comparison, two levels of observation $n=100$ and 200 are used, each containing two levels of correlation $\theta = 0.01$ (weak) and 0.1 (strong). Four scales $h = 1, 2, 3$ and 4 of the line filter is applied to estimate Model A. Figure 4.9 contains the estimation summary in four sets of four side-by-side boxplots. Each set of $rMAD$ measures has two boxplots colored in light blue and pink, which are the same ones as in Figure 4.8. Green and yellow boxplots are the estimation of weakly and strongly correlated data respectively with $n = 200$. When the correlation is weak, the estimation result gives smaller

$rMAD$ than strong. When n is large, the estimation result gives smaller $rMAD$ than n small. Between the fineness of grid size and the strength of correlation, the impact of the lattice size is pronounced when the scale is small ($h = 1$) while the impact of correlation on estimation is strongly exhibited as the scale grows. From the simulation results displayed in Figures 4.8, 4.9 and Table 4.8 we see the interaction between the observation scale and the dependent structure.

4.5 Discussion

As the underlying variance function is non-constant over the parameter space, local averaging is necessary with a difference filter of a small scale h . Since we estimate the variance function of a nonstationary correlated process, we recommend using a symmetric weight filter for initial estimation. When the correlation is weak, the HKT filter should work equally well or slightly better, but there is potential for large variance when the function for estimation is a high-order polynomial function. In short, the weight applied to the observed random field should be balanced to reduce variance of the estimator. When the weight distribution is uneven, the leading direction in terms of filter weight should not only interact with the underlying variance function to be estimated but also introduce bias.

Throughout this chapter, we assume that the data is recorded on a square lattice. In practice, the geo-referenced data may not be recorded in such a format. In those situations, given that the points of observation are quite regular, the filter shape can be more flexible to adapt to the data, or an interpolation can be used to map the observations on a rectangular lattice. We have avoided discussing the practical problem of bandwidth selection of the kernel function $K_\Lambda(\cdot)$. The bandwidth controls the scope of data averaging in the neighborhood of the estimation point and affects the overall surface estimation result greatly. We recommend taking several nearby cross-sections of

the data for both x - and y -axes and performing the bandwidth selection on the square pseudo-residuals as in Chapter 3. This could potentially lead to another discussion where an adaptive bandwidth selection is more promising, so we conclude the current discussion here.

CHAPTER 5. SUMMARY AND DISCUSSION

I have dealt with two topics of different flavors. The first topic focuses on an effective spatial sampling design, and the second topic is concerned with a parameter estimation of spatial data where we often encounter nonstationary process features.

In Chapter 2, I propose a two-stage systematic sampling design that has a higher hotspot detection probability than any one-stage design under a fixed budget constraint. For spatial sampling design systematic sampling is known to give small variance for several parameter estimations. It is also efficient in hotspot detection. The proposed sampling method innovates the well-known sampling method by transitioning the sampling framework from estimation to a game theoretic approach under the objective of maximizing the hotspot detection probability. Let us suppose that there are two intelligent agents knowing the location of hotspots in each one half of all strata respectively. Both agents also know in advance the sampling strategies each other has planned. In order to maximize the detection probability given the sampling plan of the other agent, the corresponding agent should allocate sampling resource in a complimentary fashion for each half of stratum. When two agents cooperate and use the sampling resources stochastically efficiently, the hotspot detection probability is no less than a one-stage design. In other words, the proposed two-stage sampling plan hinges on two agents cooperating by knowing each other's plan. For implementation, the sampling occurs in two stages. Whether there is or not the prior knowledge of the hotspot dispersion scenario, the sequential sampling gives the analogous story of cooperative sampling.

In Chapter 3, I discuss a difference-based variance function estimation for a one-

dimensional nonstationary process and contrast it with difference-based estimation under an independent error process and with a likelihood-based method. For a difference-based method I introduce a local varogram, which is theoretically the multiplication of a variance function and a function involving correlation. A nonparametric idea is useful in the variance function estimation because many real data could not be easily represented in an analytical distribution form for a parametric method. Due to the correlated structure of the data, it is crucial to estimate the correlation structure and adjust for it in variance function estimation. Due to the correlated data process, bandwidth selection also needs some adjustments in cross-validation.

In Chapter 4, I extend the difference-based variance function estimation to a two-dimensional nonstationary random field. Using a simulation-based approach, several difference filters in terms of configurations and weight, in addition to scaling and rotation, have been investigated. Symmetry in weight distribution either by directional rotation and averaging or using symmetric weight about the point of estimation is one of the most important features. In variance function estimation a symmetric weight filter performs better than a non-symmetric weight filter especially when the data are strongly correlated. However, when the errors are independent, the HKT weight filter whose weight is loaded on one end performs better as it has been derived to minimize the variance of the estimator. Filters that include the node at the point of estimation gives a more precise estimation. To be specific, a line, Y, and cross configurations are preferred over a square-shaped Square2 and Square 3 configurations. This suggest a line filter with directional rotation and averaging is the most efficient filter for the variance function estimation among the filter configurations we have tested.

APPENDIX A. ADDITIONAL MATERIAL

A.1 Derivations

A.1.1 Proof of Equation (2.3) in Chapter 2

Denote the total number of sampling sites as N , the number of sample as n , first-stage sample size as n_1 , second-stage sample size as n_2 , and the number of sampling sites for a hotspot as b . Using the same number of sample for a one-stage and two-stage designs, we have $n_2 = n - n_1$. The event of hotspot detection through two-stage simple random sampling (SRS) of total n sample is denoted as $D_{SRS,two}$ regardless of the sample size split, the event of hotspot detection through one-stage simple random sampling (SRS) of size n is denoted as $D_{SRS,n}$, and the complement event of $D_{SRS,n}$, i.e. non-detection of hotspot through a one-stage simple random sampling (SRS) is denoted as $\overline{D_{SRS,n}}$.

$$\begin{aligned}
 Pr(D_{SRS,two}) &= Pr(D_{SRS,n_1}) + Pr(\overline{D_{SRS,n_1}}) Pr(D_{SRS,n_2}) \\
 &= 1 - \frac{\binom{N-b}{n_1}}{\binom{N}{n_1}} + \frac{\binom{N-b}{n_1}}{\binom{N}{n_1}} \left\{ 1 - \frac{\binom{N-n_1-b}{n_2}}{\binom{N-n_1}{n_2}} \right\} \\
 &= 1 - \frac{\binom{N-b}{n_1} \binom{N-n_1-b}{n_2}}{\binom{N}{n_1} \binom{N-n_1}{n_2}} \\
 &= 1 - \frac{\frac{(N-b)!}{n_1!(N-n_1-b)!} \frac{(N-n_1-b)!}{n_2!(N-n_1-n_2-b)!}}{\frac{N!}{n_1!(N-n_1)!} \frac{(N-n_1)!}{n_2!(N-n_1-n_2)!}}
 \end{aligned}$$

$$\begin{aligned}
&= 1 - \frac{\binom{N-b}{n}}{\binom{N}{n}} \\
&= Pr(D_{SRS, one})
\end{aligned} \tag{A.1}$$

□

A.1.2 Derivations for Chapter 3

Here is the detailed expansion of the variance of an h -lagged nonstationary process with smooth mean and variance functions. This shows (3.3) in deriving the local variogram (3.4) as the main term of the expansion.

$$\begin{aligned}
&var \left(Z \left(s - \frac{h}{2n} \right) - Z \left(s + \frac{h}{2n} \right) \right) \\
&= Var \left(Z \left(s - \frac{h}{2n} \right) \right) + Var \left(Z \left(s + \frac{h}{2n} \right) \right) - 2Cov \left(Z \left(s - \frac{h}{2n} \right), Z \left(s + \frac{h}{2n} \right) \right) \\
&= \sigma^2 \left(s - \frac{h}{2n} \right) + \sigma^2 \left(s + \frac{h}{2n} \right) - 2\rho_{h,n} \sigma \left(s - \frac{h}{2n} \right) \sigma \left(s + \frac{h}{2n} \right) \\
&= 2\sigma^2(s) + 2 \frac{(\sigma^2(s))^{(2)}}{2!} \left(\frac{h}{2n} \right)^2 + o(n^{-2}) \\
&\quad - 2\rho_{h,n} \left\{ \sigma^2(s) - (\sigma^{(1)}(s))^2 \left(\frac{h}{2n} \right)^2 + \frac{\sigma(s)\sigma^{(2)}(s)}{2} \left(\frac{h}{2n} \right)^2 + o(n^{-2}) \right\} \\
&= 2\sigma^2(s) (1 - \rho_{h,n}) + 2 (\sigma^{(1)}(s))^2 (1 + \rho_{h,n}) \left(\frac{h}{2n} \right)^2 + o(n^{-2})
\end{aligned} \tag{A.2}$$

Detailed calculation of (3.17) in Chapter 3.

$$\begin{aligned}
P_{ij} &= \rho_{|i-j|} (\sigma_i \sigma_j + \sigma_{i+h} \sigma_{j+h}) - \rho_{|i-j-h|} \sigma_i \sigma_{j+h} - \rho_{|i-j+h|} \sigma_{i+h} \sigma_j \\
\sigma_i \sigma_j &= \sigma_i \left\{ \sigma_i + \sigma'_i \frac{j-i}{n} + \sigma''_i \frac{1}{2} \left(\frac{j-i}{n} \right)^2 + o(n^{-2}) \right\} \\
\sigma_{i+h} \sigma_{j+h} &= \left\{ \sigma_i + \sigma'_i \frac{h}{n} + \sigma''_i \frac{1}{2} \left(\frac{h}{n} \right)^2 + o(n^{-2}) \right\} \\
&\quad \times \left\{ \sigma_i + \sigma'_i \frac{j+h-i}{n} + \sigma''_i \frac{1}{2} \left(\frac{j+h-i}{n} \right)^2 + o(n^{-2}) \right\}
\end{aligned}$$

$$\begin{aligned}
\sigma_i \sigma_{j+h} &= \sigma_i \left\{ \sigma_i + \sigma'_i \frac{j+h-i}{n} + \sigma''_i \frac{1}{2} \left(\frac{j+h-i}{n} \right)^2 + o(n^{-2}) \right\} \\
\sigma_{i+h} \sigma_j &= \left\{ \sigma_i + \sigma'_i \frac{h}{n} + \sigma''_i \frac{1}{2} \left(\frac{h}{n} \right)^2 + o(n^{-2}) \right\} \\
&\quad \times \left\{ \sigma_i + \sigma'_i \frac{j-i}{n} + \sigma''_i \frac{1}{2} \left(\frac{j-i}{n} \right)^2 + o(n^{-2}) \right\}
\end{aligned}$$

Rewriting P_{ij} with the Taylor expansions of the exponential functions and $\sigma_i \sigma_j$ expansions gives:

$$\begin{aligned}
P_{ij} &= \frac{h^2}{n^2} \left(\sigma_i^{(1)} \right)^2 - \sigma_i^2 \Upsilon_{ij} - \sigma_i \sigma_i^{(1)} \frac{j+h-i}{n} \Upsilon_{ij} \\
&\quad + \frac{\sigma_i \sigma_i^{(2)}}{2} \left[\left(\frac{j+h-i}{n} \right)^2 \left\{ (s_i - s_j)^\alpha - \left(s_i - s_j - \frac{h}{n} \right)^\alpha \right\} \right. \\
&\quad \left. + \left\{ \left(\frac{h}{n} \right)^2 + \left(\frac{j-i}{n} \right)^2 \right\} \left\{ (s_i - s_j)^\alpha - \left(s_i - s_j + \frac{h}{n} \right)^\alpha \right\} \right] \\
&\quad + \left(\sigma_i^{(2)} \right)^2 \left[\frac{h(j-i)}{n^2} \left\{ (s_i - s_j)^\alpha - \left(s_i - s_j + \frac{h}{n} \right)^\alpha \right\} + \left(\frac{h}{n} \right)^2 (s_i - s_j)^\alpha \right] + O(n^{-3-\alpha})
\end{aligned} \tag{A.3}$$

$$\text{where } \Upsilon_{ij} = \theta \left\{ 2(s_i - s_j)^\alpha - \left(s_i - s_j - \frac{h}{n} \right)^\alpha - \left(s_i - s_j + \frac{h}{n} \right)^\alpha \right\}.$$

A.2 Filter Weights

A.2.1 Simple Differencing: Symmetric Weight

$$\begin{array}{ccc}
\begin{array}{cc} \bullet & \bullet \\ \times & \\ \bullet & \bullet \end{array} & \text{weights:} & \begin{array}{cc} -a & a \\ a & -a \end{array}
\end{array}
\quad a = \pm \frac{1}{2}$$

(a) 2×2 square ($\nu = 4$) : $\varrho_L(h) = 1 + \rho_\theta \left(\frac{\sqrt{2}h}{n} \right) - 2\rho_\theta \left(\frac{h}{n} \right)$

(b) 3×3 square ($\nu = 8$):

$$\varrho_L(h) = 1 - 2\rho_\theta \left(\frac{h}{n} \right) + \frac{3}{2}\rho_\theta \left(\frac{2h}{n} \right) - 2\rho_\theta \left(\frac{\sqrt{5}h}{n} \right) + \frac{1}{2}\rho_\theta \left(\frac{2\sqrt{2}h}{n} \right) + \rho_\theta \left(\frac{\sqrt{2}h}{n} \right)$$

Condition 1: (S1) $a_8 = a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7$

Condition 1 & 2: (S2) $\sum_{i=1}^7 \sum_{j=1}^7 a_i a_j = 1/2$

Condition 3: (S3) $a_1 + a_2 + a_3 = a_5 + a_6 + a_7$ and $a_1 + a_7 + a_8 = a_3 + a_4 + a_5$.

Plugging (S1) into (S3): (S4) $2(a_1 + a_7 + a_8) + a_2 + a_6 = 0$ and $2(a_1 + a_2 + a_3) + a_4 + a_8 = 0$.

$$\begin{array}{ccc} \bullet & \bullet & \bullet \\ & & -a \quad a \quad -a \end{array}$$

$$\begin{array}{ccc} \bullet & \times & \bullet \\ & & \text{weights:} \quad a \quad \quad a \quad \quad a = \pm \frac{1}{\sqrt{8}} \end{array}$$

$$\begin{array}{ccc} \bullet & \bullet & \bullet \\ & & -a \quad a \quad -a \end{array}$$

$$(c) + \text{shaped cross } (\nu = 5): \varrho_+(h) = 1 - \frac{8}{5}\rho_\theta \left(\frac{h}{n} \right) + \frac{2}{5}\rho_\theta \left(\frac{\sqrt{2}h}{n} \right) + \frac{1}{5}\rho_\theta \left(\frac{2h}{n} \right)$$

$$\begin{array}{ccc} & & a \\ \bullet & & \end{array}$$

$$\begin{array}{ccc} \bullet & \times & \bullet \\ & & \text{weights:} \quad a \quad -4a \quad a \quad \quad a = \pm \frac{1}{\sqrt{20}} \end{array}$$

$$\begin{array}{ccc} & & a \\ \bullet & & \end{array}$$

$$(d) \times \text{shaped cross } (\nu = 5): \varrho_\times(h) = 1 - \frac{8}{5}\rho_\theta \left(\frac{\sqrt{2}h}{n} \right) + \frac{2}{5}\rho_\theta \left(\frac{2h}{n} \right) + \frac{1}{5}\rho_\theta \left(\frac{2\sqrt{2}h}{n} \right)$$

$$\begin{array}{ccc} \bullet & & \bullet \\ & & a \quad \quad a \end{array}$$

$$\begin{array}{ccc} & \times & \\ & & \text{weights:} \quad -4a \quad \quad a = \pm \frac{1}{\sqrt{20}} \end{array}$$

$$\begin{array}{ccc} \bullet & & \bullet \\ & & a \quad \quad a \end{array}$$

Condition 1: (X1) $a_3 = a_1 + a_2 + a_4 + a_5$

Condition 1 & 2: (X2) $2(a_1^2 + a_2^2 + a_4^2 + a_5^2) + 2(a_1a_2 + a_1a_4 + a_1a_5 + a_2a_4 + a_2a_5 + a_4a_5) = 1$

Condition 3: (X3) $a_1 = a_5$ and $a_2 = a_4$.

Plugging (X3) into (X2) gives: (X4) $2(a_1 + a_2)^2 + a_1^2 + a_2^2 = 1/2$. This quadratic equation does not have a unique solution of (a_1, a_2) . Therefore, we could find infinite pairs of (a_1, a_2) such

that (X1), (X2), and (X3) hold true. If we choose $a_1 = a_2$, then $A^{(+/\times)} = \{(a_1, a_2, a_3, a_4, a_5) = (a, a, -4a, a, a) \text{ where } a = 1/\sqrt{20}\}$.

(e) Y-configuration ($\nu = 4$): $\varrho_Y(h) = 1 - \frac{1}{3}\rho_\theta\left(\frac{\sqrt{2}h}{n}\right) - \rho_\theta\left(\frac{h}{n}\right) + \frac{1}{3}\rho_\theta\left(\frac{\sqrt{5}h}{n}\right)$

• a

• ✕ weights: a -3a $a = \pm \frac{1}{\sqrt{12}}$

• a

Condition 1: (Y1) $a_3 = a_1 + a_2 + a_4$

Condition 1 & 2: (Y2) $2(a_1^2 + a_2^2 + a_4^2) + 2(a_1a_2 + a_1a_4 + a_2a_4) = 1$

Condition 3: (Y3) $a_1 = a_4$ and $a_2 = a_4 \Rightarrow a_1 = a_2 = a_4$. Plugging (Y3) into (Y2) gives: (Y4) $6a_1^2 = 1/2 \Rightarrow a_1 = 1/\sqrt{12}$. Therefore, $A^{(Y)} = \{(a_1, a_2, a_3, a_4) = (a, a, -3a, a) \text{ where } a = 1/\sqrt{12}\}$.

(f) Star-shape averaging over 4 directions with ($\nu = 3$) linear filters:

$$\varrho_*(h) = 1 - \frac{4}{3}\rho_\theta\left(\frac{h}{n}\right) + \frac{1}{3}\rho_\theta\left(\frac{2h}{n}\right)$$

• • • a

• ✕ • weights: -2a $a = \pm \frac{1}{\sqrt{6}}$

• • • a

A.2.2 Variance Minimization under Independent and Identically Distributed Errors: Hall-Kay-Titterington Weight

The following weights are obtained from Hall et al. (1991). They assume independent and identically distributed errors and derive a variance of the variance estimator.

(A) 2×2 square ($\nu = 4$) : $\varrho_L(h) = 1 - \frac{1}{3}\rho_\theta\left(\frac{\sqrt{2}h}{n}\right) - \frac{2}{3}\rho_\theta\left(\frac{h}{n}\right)$

• •

• × weights: $\begin{matrix} -3a & a \\ a & a \end{matrix} \quad a = \frac{1}{\sqrt{12}}$

(B) 3×3 square ($\nu = 8$):

$$\varrho_L(h) = 1 - 2 \left(0.182\rho_\theta\left(\frac{h}{n}\right) + 0.145\frac{3}{2}\rho_\theta\left(\frac{2h}{n}\right) + 0.102\rho_\theta\left(\frac{\sqrt{5}h}{n}\right) + 0.125\rho_\theta\left(\frac{2\sqrt{2}h}{n}\right) - 0.068\rho_\theta\left(\frac{\sqrt{2}h}{n}\right) \right)$$

• • •

• • -0.147 -0.114 -0.133

• • × weights: -0.114 -0.147

-0.133 -0.147 0.935

(C) + shaped cross ($\nu = 5$):

$$\varrho_+(h) = 1 - 2 \left(0.053\rho_\theta\left(\frac{h}{n}\right) + 0.490\rho_\theta\left(\frac{\sqrt{2}h}{n}\right) - 0.044\rho_\theta\left(\frac{2h}{n}\right) \right)$$

• 0.231

×• • • weights: 0.263 0.167 -0.892

• 0.231

(D) Star-shape averaging over 4 directions with ($\nu = 3$) linear filters:

$$\varrho_\times(h) = 1 - \frac{1}{2} \left(\rho_\theta\left(\frac{h}{n}\right) + \rho_\theta\left(\frac{2h}{n}\right) \right)$$

• • × weights: $\frac{\sqrt{5}+1}{4} \quad -\frac{1}{2} \quad -\frac{\sqrt{5}-1}{4}$

BIBLIOGRAPHY

- Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *Journal of American Statistical Association*, 85(411):749–759.
- Anderes, E. B. and Stein, M. L. (2011). Local likelihood estimation for nonstationary random fields. *Journal of Multivariate Analysis*, 102:506–520.
- Breidt, F. J. (1995). Markov chain designs for one-per-stratum sampling. *Survey Methodology*, 21(1):63–70.
- Briggs, D. J., Collins, S., Elliott, P., Fischer, P., Kingham, S., Lebre, E., Pryl, K., Van-Reeuwijk, H., Smallbone, K., and VanderVeen, A. (1997). Mapping urban air pollution using gis: a regression-based approach. *International Journal of Geographical Information Science*, 11:699–718.
- Brown, L. D. and Levine, M. (2007). Variance estimation in nonparametric regression via the difference sequence method. *The Annals of Statistics*, 35(5):2219–2232.
- Buckley, M. J., Eagleson, G. K., and Silverman, B. W. (1988). The estimation of residual variance in nonparametric regression. *Biometrika*, 75(2):189–199.
- Christman, M. C. (2003). Adaptive two-stage one-per-stratum sampling. *Environmental and Ecological Statistics*, 10(1):43–60.

- Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, 20(4):2008–2036.
- Gasser, T., Müller, H.-G., and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *Journal of Royal Statistical Society. Series B*, 47(2):238–252.
- Gasser, T., Sroka, L., and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, 73(3):625–633.
- Gelfand, A. E., Schmidt, A. M., Banerjee, S., and F., S. C. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *TEST*, 13(2):1–50.
- Gilbert, R. O. (1982). Some statistical aspects of finding hot spots and buried radioactivity. *TRAN-STAT: Statistics for Environmental Studies*, PNL-SA-10274(19).
- Gilbert, R. O. (1987). *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold.
- Hall, P. and Carroll, R. J. (1989). Variance function estimation in regression: the effect of estimating the mean. *Journal of Royal Statistical Society. Series B*, 51(1):3–14.
- Hall, P., Kay, J. W., and Titterton, D. M. (1991). On estimation of noise variance in two-dimensional signal processing. *Advanced Applied Probability*, 23:476–495.
- Han, C. and Gu, C. (2008). Optimal smoothing with correlated data. *Sankhya: The Indian Journal of Statistics*, 70-A(1):38–72.
- Hart, J. D. (1991). Kernel regression estimation with time series errors. *Journal of Royal Statistical Society. Series B*, 53(1):173–187.

- Helbich, M., Brunauer, W., Vaz, E., and Nijkamp, P. (2014). Spatial heterogeneity in hedonic house price models: the case of Austria. *Urban Study*.
- Hu, S. and Mo, X. (2011). Interpreting spatial heterogeneity of crop yield with a process model and remote sensing. *Ecological Modelling*, 222(14):2530–2541.
- Inman, C. A., Cheng, E. S., Cottingham, D. A., Fixsen, D. J., Kowitt, M. S., Meyer, S. S., Page, L. A., Puchalla, J. L., Ruhl, J. E., and Silverberg, R. F. (1997). A Cosmic Microwave Background radiation measurement reproduced: a statistical comparison of MSAM1-94 to MSAM1-92. *The Astrophysical Journal*, 478:L1–L4.
- Kessler, D. A. and Shnerb, N. M. (2009). The effect of spatial heterogeneity on the extinction transition in stochastic population dynamics. *New Journal of Physics*, 11:<http://iopscience.iop.org/1367-2630/11/4/043017/fulltext/>.
- Levine, M. (2006). Bandwidth selection for a class of difference-based variance estimators in the nonparametric regression: A possible approach. *Computational Statistics and Data Analysis*, 50(12):3405–3431.
- Müller, H.-G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics*, 15(2):610–625.
- Opsomer, J., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science*, 16(2):134–153.
- Parkhurst, D. F. (1984). Optimal sampling geometry for hazardous waste sites. *Environmental Science Technology*, 18:521–523.

- Rondeau, G. L. and et al. (2009). *Beryllium: Environmental Analysis and Monitoring*. The Royal Society of Chemistry, Editors: Brisson, Mike J. and Ekechukwu, Amy A.
- Seifert, B., Gasser, T., and Wolf, A. (1993). Nonparametric estimation of residual variance revisited. *Biometrika*, 80(2):373–383.
- Singer, D. A. (1972). ELIPGRID: A Fortran IV program for calculating the probability of success in locating elliptical targets with square, rectangular and hexagonal grids. *Geocom Programs*, 4:1–16.
- Standard, D. T. (2005). *Management of items and areas containing low levels of beryllium*. Department of Energy, Washington, D.C.
- Thompson, S. K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85(412):1050–1059.
- von Neumann, J., Kent, R. H., Bellinson, H. R., and Hart, B. I. (1941). The mean square successive difference. *The Annals of Mathematical Statistics*, 12(2):153–162.
- Wang, L., Brown, L. D., Cai, T. T., and Levine, M. (2008). Effect of mean on variance function estimation in nonparametric regression. *The Annals of Statistics*, 36(2):646–664.
- Zhu, Z. and Stein, M. L. (2002). Parameter estimation for fractional brownian surfaces. *Statistica Sinica*, 12:863–883.
- Zirschky, J. and Gilbert, R. O. (1984). Detecting hot spots at hazardous-waste sites. *Chemical Engineering*, 91:97–100.