Performance Analysis for Sequential Statistical Discrimination of Gaussian Random Fields

Ivan Leonov
Belarusian State University
Faculty of Applied Mathematics
and Computer Science
Minsk, Belarus
ivanleonov.eu@gmail.com

***Abstract*—Stochastic processes stand as a versatile tool for modeling and understanding random phenomena. These processes describe system evolution while taking into account randomness and uncertainty. Nowadays, stochastic processes find application in a variety of domains, such as modeling financial markets, monitoring manufacturing procedures and predicting disease spread. This paper proposes a sequential procedure for testing hypotheses concerning the correlational structure of random fields and their trends.**

Keywords—Random Field, Sequential Decision Rule, Gaussian Random Field, Sequential Analysis, Gaussian Random Fields with Trends

I. Introduction

In the realm of applied mathematics, the use of random processes leads to solving challenges in a variety of industries, i.e. telecommunications, computer science, biotechnology, health, risk management, etc. In pursuit of modeling and understanding complex phenomena, the role of stochastic processes has been foundational. Still, the classical definition of random processes limits the dimension of parameter space to one. Usually, this parameter represents time. However, when we encounter spatial or spatial-temporal dependencies --- e.g. geology/geostatistics --- we require parameters of higher dimensions. Random fields generalize the notion of stochastic processes. In addition, it allows to capture of spatial dependencies. Moreover, in a data-driven era where information is often gathered across a variety of spatial points, random fields act as a framework capable of accommodating the intricacies of the collected data as well as representing inherited relationships between these data points.

There are two fundamental approaches to statistical inference: classical hypothesis testing and sequential analysis. In the classical approach, the dataset size is known in advance. This method offers a straightforward procedure for hypothesis testing. However, it may not be the most efficient technique when resources are limited. On the contrary, sequential statistical analysis is a dynamic and adaptive procedure, in which the data is collected only when necessary. This feature makes it effective since the number of observations is a random variable itself.

Random fields pose significant challenges in the realm of statistical inference due to their complex spatial correlational structure. Unlike i.i.d. random variables, random fields exhibit spatial correlations. meaning that nearby values are interdependent. The interdependence violates one of the fundamental assumptions of statistical tests. In this paper, we address the problem of sequential hypothesis testing concerning Gaussian Random Fields with trends.

II. Model

A Gaussian Random Field (GRF) with a trend refers to a spatial or spatiotemporal random field where the primary variation is modeled using a trend component, typically a deterministic function, in addition to a Gaussian random component. This combination allows for the modeling of spatial or temporal data that exhibits both systematic trends or patterns and random fluctuations.

The trend component represents the underlying, often non-random, behavior or structure in the data. It is typically specified based on prior knowledge or domain expertise and can take various functional forms, such as linear, quadratic, exponential, or more complex functions, depending on the nature of the trend in the data. The trend component helps capture the overall behavior of the data and provides a way to model long-term or large-scale patterns.

The Gaussian random component, on the other hand, introduces stochastic variability or noise into the model. This component is a Gaussian random field and represents the smaller-scale, random fluctuations that cannot be accounted for by the trend alone. The Gaussian assumption is often made for simplicity and mathematical tractability.

GRFs with trends are commonly used in various fields, including geostatistics, environmental science, and spatial epidemiology, to model data with spatial or temporal dependencies and to separate deterministic trends from random variability. This modeling approach allows researchers to better understand and analyze complex datasets that exhibit both structured patterns and random noise, enabling more accurate predictions and statistical inference.

Let x_t be an observation of a random field with a trend, such that:

$$x_t = f(t) + \varepsilon_t$$

where $f(t)$ is a deterministic function representing the non-random behavior of a field, ε_t a Gaussian random field with zero mean. There are two simple hypotheses: null hypothesis H_0, alternative H_1. With $\varphi_i(t)$ lets denote our assumption for the trend function. In addition, we can formulate a hypothesis for the correlational structure of Gaussian random field ε_t. To illustrate, let $\rho(t)$ denote the correlation function of $\varepsilon_t (t \in \mathbb{R}^2)$, then a sample hypothesis can be formulated as:

$$\rho(t) = \exp\{-(\frac{\tau_1^2}{\theta_1^2} + \frac{\tau_2^2}{\theta_2^2})\}$$

where θ_1, θ_2 represent length of correlational dependencies along the corresponding axis.

III. Sequential Probability Ratio Test

Sequential Probability Ratio Tests (SPRT) are a class of statistical tests used for sequential decision-making. These tests are designed to analyze data as it is collected in a sequential manner, allowing for the early termination of data collection when there is sufficient evidence to make a

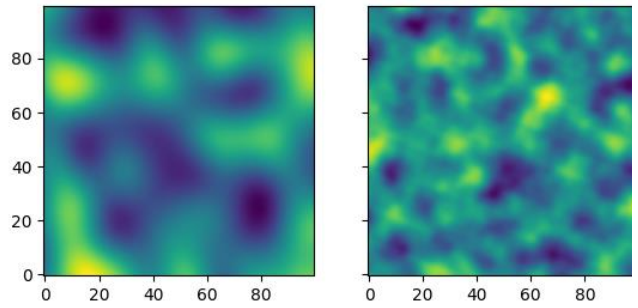

Fig. 1 Example of a random field visualization.

decision[1]. SPRTs are particularly valuable in situations where resources or time are limited, as they can lead to more efficient data collection.

Wald continues to demonstrate that the SPRT is the most powerful test with a given sample size. Conversely, the SPRT also requires a smaller sample size to achieve a given a. This smaller sample size can be referred to as a sample size savings. Wald and Wolfowitz provide proof of the optimal characteristic of the SPRT in their paper "Optimum Character of the Sequential Probability Ratio Test"[2].This proof shows the generalization that of all tests with the same power, the sequential probability ratio test requires on average the fewest observations. This result is imperative in its selection as the optimal test method and validates the statement that the SPRT provides significant savings over other hypothesis testing methods.

For any positive integer m, let p_{im} denote the probability that the sample was obtained under hypothesis $H_i, i \in \{0,1\}$. The sequential probability ratio test for testing simple hypotheses is defined as follows: Two positive thresholds A, B are chosen $(B < A)$. At every step m of the procedure, the probability ratio $\frac{p_{1m}}{p_{0m}}$ is computed. The ratio is used to make a decision whether or not to stop the procedure:

$$SPRT = \begin{cases} \text{continue experiments} & \text{if } B < \frac{p_{1m}}{p_{0m}} < A \\ \text{reject } H_0 & \text{if } \frac{p_{1m}}{p_{0m}} \geq A \\ \text{accept } H_0 & \text{if } \frac{p_{1m}}{p_{0m}} \leq B \end{cases}$$

The threshold A, B are determined so that the test will have desired strength (α, β), where α, β are respectively probabilities of first and second type. The exact determination of the values A, B is a complex task. Thus, in practice, we put $A = (1-\beta)/\alpha$, $B = \beta/(1-\alpha)$. Wald showed that the resulting probabilities of the first and second kind (α', β') satisfy the following inequality [1]:

$$\alpha' + \beta' \leq \alpha + \beta$$

Since we are working with Gaussian Random Fields, the vector containing observed values follows the multivariate normal distribution:

$$p_{im} = \frac{exp\left(-\frac{1}{2}\left(x_{(m)} - \mu_i\right)^T \Sigma_i^{-1}\left(x_{(m)} - \mu_i\right)\right)}{\sqrt{(2\pi)^k |\Sigma_i|}}, i \in \{0,1\}$$

In order to calculate the covariance matrix Σ_i we use the corresponding correlation function to the i-th hypothesis. The mean is calculated using φ_i since we assumed that the random field ε_t has zero mean.

IV. Computer Modeling

We encountered several challenges associated with modeling random fields. Random fields inherently exhibit spatial or spatiotemporal dependencies, which necessitate the careful specification of covariance structures to capture the underlying spatial relationships. Additionally, the high dimensionality of random field data can strain computational resources, making it essential to explore dimension reduction techniques. Addressing measurement error, and uncertainty, and dealing with irregularly sampled or sparse data further complicates the modeling process. Successfully tackling these challenges is fundamental to advancing our understanding of random fields and improving the reliability of statistical hypothesis testing within this domain.

In the context of spatial-statistical research and hypothesis testing of random fields, the GSTools library emerges as a valuable tool for modeling complex spatial data. This framework allows a user-friendly interface for spatial data modeling. A pivotal step in this process entails defining a covariance model, a critical component capturing the spatial structure of the data. There are a variety of predefined models, such as Gaussian, exponential, cubic, and circular covariance models. With the covariance model established, the subsequent phase involves the creation of a random field object, specifying the spatial grid or mesh for field generation.

Using Python visualization libraries, we can create informative plots of generated random fields. We used two types of plots: heatmap and 3D plot. The heat map is used to visualize random fields e_t. Since we assumed zero mean, we

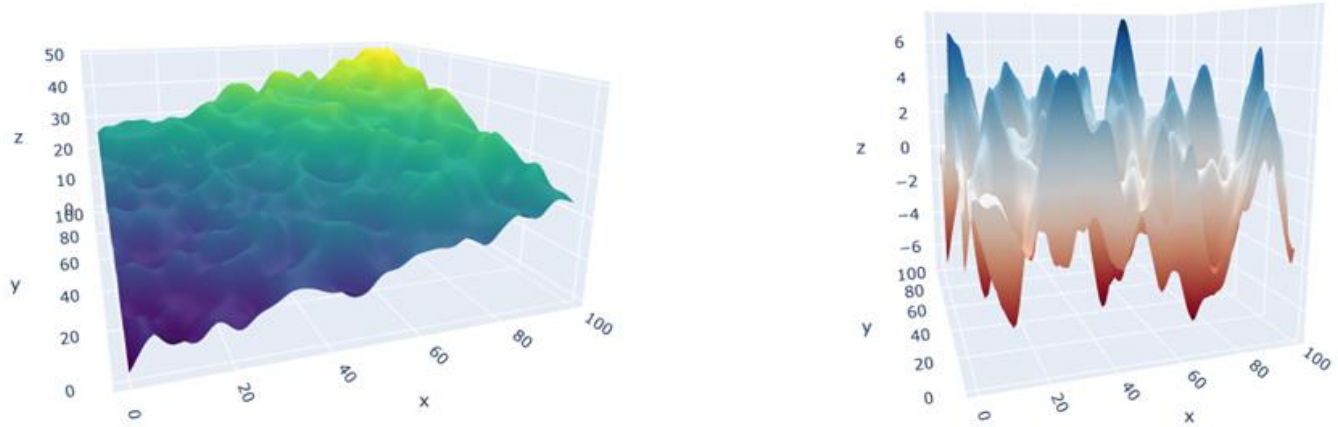

Fig. 2. Example of a random field with Trend.

can perform side by side comparison of the correlational structure. (“Fig. 1”) Whereas, 3D plots better represent random fields with trends. (“Fig. 2”)

By definition, random fields are infinite, however, modeling an infinite object of a computer is a challenging task. Therefore, in this paper, we used finite random fields for the analysis of the proposed procedure. To generate a new observation, we sampled two numbers for the corresponding uniform distribution and then took the value of the random field modeled with GSTools.

To validate the procedure, we calculated the probabilities of error. Calculating probabilities of error is a fundamental concept in statistics, especially in the context of hypothesis testing. We iterated the procedure on the same set of hypotheses. After all the data was collected, we calculated the error probabilities of the first and second types.

V. Modeling Results

Let N_{iter} be number of iterations of the SPRT needed to calculate α^*, β^* observed error probabilities of type 1, type 2 respectively. We proceed by defining the hypotheses and target error probabilities in table 1.

TABLE I. SPRT PARAMETERS

Hypotheses parameters			
#	***Variance***	***Correlation Function***	***Trend Function***
Null Hypothesis			
1	1	$\exp(-(\tau^2 \setminus 5))$	0
2	5	$\exp(-(\tau^2 \setminus 15))$	$-0.05t_1 + 0.05t_2$
3	3	$\exp(-(\tau \setminus 5))$	$-0.05t_1 + 0.05t_2$
4	3	$\exp(-(\tau \setminus 5))$	$t_1 + 0.05t_2$
Alternative Hypothesis			
1	1	$\exp(-(\tau^2 \setminus 25))$	0
2	15	$\exp(-(\tau^2 \setminus 5))$	0
3	3	$\exp(-(\tau \setminus 5))$	$-0.1t_1 + 0.05t_2$
4	3	$\exp(-(\tau \setminus 5))$	$-0.1t_1 + 0.05t_2$
Test Parameters			
#	N_{iter}	***α, Type I Error***	***β, Type II Error***
1	1000	0.01	0.01
2	1000	0.01	0.01
3	10000	0.01	0.1
4	10000	0.2	0.2

The main metrics we are interested in are the error probabilities and the average sample number.t The average sample number (ASN) is a critical concept that measures the expected number of observations or samples required to reach a decision or stopping point in the test. ASN is essential in sequential analysis since it helps optimize the use of resources, such as time, and illustrates cost and time efficiency.

We run the proposed sequential probability ration test on 11th Gen Intel(R) i5-1135G7 2.4 GHz with 16 GB of RAM. The results of the experiments are shown in table 2.

The first experiment demonstrates that the proposed test works on random fields without trend. The experiments illustrate that the procedure can successfully distinguish between fields with and without trends. The third example shows that we can test hypotheses for two Gaussian Random Fields with trend. Finally, yet importantly, the fourth experiment demonstrates that there is indeed a dependency between the observed errors and theoretical.

TABLE II. SPRT RESULTS

Results			
#	***ASN***	α^****, Type I Error***	β^****, Type II Error***
1	11.1	0.01	0.0
2	7.14	0.0	0.003
3	5.23	0.017	0.0
4	3.29	0.112	0.0

The results show that the largest sample size is needed in the case of stationary fields. It is expected since the trend provides additional information to the test. Also, worth noticing that the decision rule performs better than the target error rates since the classical sequential probability ratio test was proposed for independent observations.

VI. Conclusion

In summary, this research paper has aimed to develop a sequential procedure for hypothesis testing of Gaussian random fields. One of the most notable outcomes of this study is a sequential decision rule suitable for solving real-world problems. This procedure has the potential to be applied to a vast variety of applied problems concerning spatial-temporal dependencies.

It is crucial to acknowledge the limitations of this research. The described statistical test only works for Gaussian Random Fields. Moreover, the computer simulation is limited to finite Random Fields. These limitations provide opportunities for future research to delve deeper into this area and address these gaps.

In conclusion, this research contributes to the existing body of knowledge by defining an easy-to-implement procedure of hypothesis testing for Gaussian Random Fields. The insights gained from this study can be valuable for geologists and others, working with spatial data. They provide a foundation for further exploration in this field. As we move forward, it is essential to continue exploring statistical tests for non-Gaussian Random Fields.

Ultimately, this research adds to the understanding of the random fields and offers a valuable resource for researchers, practitioners, and policymakers interested in statistical inference from spatial data.

References

[1] A. Wald. Sequential Analysis. New York, John Wiley and Sons, 1947, 212 p.

[2] A. Wald, J. Wolfowitz, “Optimum character of the sequential probability ratio test.” The Annals of Mathematical Statistics, 19, 1948, 326-339.

[3] Müller, S., Schüler, L., Zech, A., and Heße, F., “GSTools v1.3: a toolbox for geostatistical modelling in Python”, Geosci. Model Dev., 15, pp 3161–3182

[4] T. Lai. “Sequential analysis: Some classical problems and new challenges,” StatisticaSinica, 2001, vol. 11, pp 303–408.