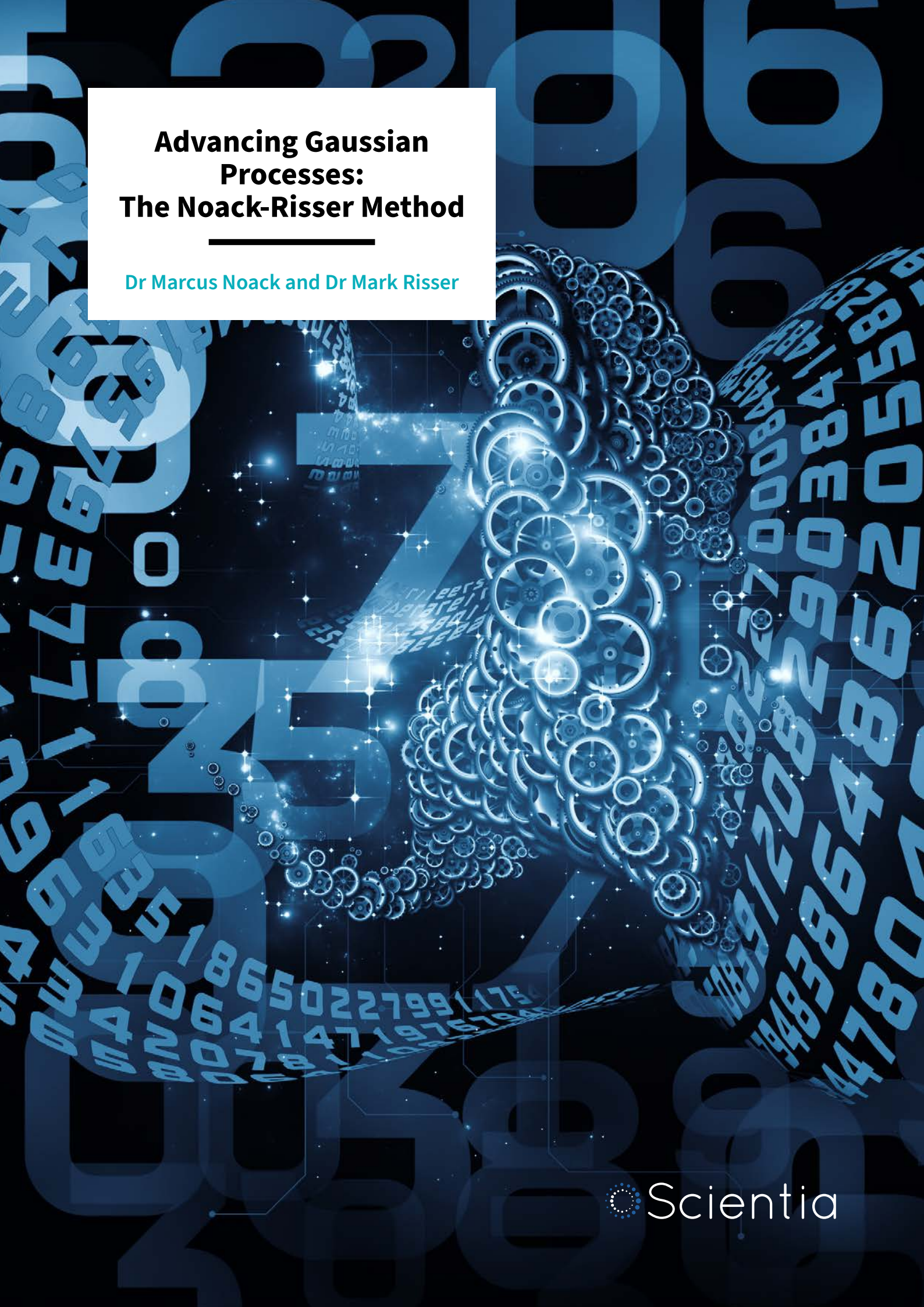
Advancing Gaussian Processes: The Noack-Risser Method

Dr Marcus Noack and Dr Mark Risser



Scientia

ADVANCING GAUSSIAN PROCESSES: THE NOACK-RISSER METHOD

Dr Marcus Noack and **Dr Mark Risser**, researchers at Lawrence Berkeley National Laboratory, have recently proposed a significant advancement in the area of machine learning and data science that promises significant computational improvements: the enhancement of exact Gaussian Processes for large datasets, significantly improving data analysis capabilities for samples even beyond 5 million data points.

Gaussian Processes

Gaussian Processes (GPs) represent a fundamental concept in data science and predictive analytics. One of their main applications is as a mathematical framework for the approximation of stochastic functions in the form of Gaussian Process Regression (GPR), which employs a Gaussian probability distribution over a function space.

The success of GPs is due to their analytical tractability, robustness and versatility. Further, GPs provide Bayesian uncertainty quantification on top of the function approximations, so every estimate comes with a measure of confidence through no additional effort.

Unfortunately, GPs have traditionally been constrained by computational inefficiencies: in fact, they exhibit cubic scaling in computation, meaning the computational complexity increases at a cubic rate with the number of data points, and is quadratic in terms of storage. This makes them prohibitively expensive or even untractable for large datasets. This limitation has been a significant barrier to the wider adoption of GPs, especially in fields where large datasets are increasingly common.

A new approach

Leading the charge in the enhancement of GPs are Dr Marcus Noack and Dr Mark Risser, both based at Lawrence Berkeley National Laboratory. Dr Noack's expertise revolves around uncertainty quantification and function optimisation, with a particular focus on Gaussian processes. Dr Risser is an expert in non-stationary spatial process modelling.

Working with colleagues, Dr Noack and Dr Risser developed a methodology that allows GPs to scale efficiently to large datasets, overcoming the traditional computational bottlenecks. While most existing methods rely on various forms of approximation, the method proposed in the paper maintains the integrity and exactness of the GP model while still enabling applications to large datasets. This is based on leveraging a mathematical property exhibited by most datasets on which GPs are commonly used: sparsity, when only a few data points are significantly related or influential, while the rest have little to no impact on the overall model. Current implementations of GPs do not allow for the leverage of this characteristic, leading to significant inefficiencies in the analysis of the data.

Building Block 1: Non-Stationary, Ultra-Flexible, and Compactly-Supported Kernel Functions

The first cornerstone of Dr Noack and Dr Risser's approach was the development of innovative kernel functions. A kernel function plays a crucial role in defining its behaviour and capabilities. Essentially, it is a mathematical function used to interpret the relationships between different data points in the dataset being analysed, and it helps to determine how similar or correlated different points in the input space are. This similarity is a key factor in predicting the value of a new data point based on the values of known data points. The kernel function calculates the covariance between pairs of points in the input space, essentially shaping the GP's understanding of the data structure.

The choice of kernel has a significant impact on the GP's performance, as it influences how the model interprets the data. For example, a linear kernel might imply that points closer together are more similar, while a more complex kernel could capture non-linear relationships or periodic patterns in the data. Most approximate methods employ kernels that are not equipped to recognise and,

therefore, discover sparsity; instead, sparsity is induced synthetically before the kernel is brought into play to minimise the impact of performance problems.

Unlike traditional kernels, the ones devised by Dr Noack and Dr Risser are designed to adapt and change, reflecting the diverse and dynamic nature of real-world data. They possess the ability to discern both the correlations and the lack thereof among data points. This adaptability is crucial in identifying naturally occurring sparse structures within large datasets, allowing for significant reductions in computational complexity. The compact support of these kernels ensures that existing as well as non-existing correlations can be learned.

Building Block 2: High-Performance Computing for Sparse Kernels

The second building block involves the strategic use of high-performance computing to harness the potential of these newly developed kernels. Given the extensive size of datasets in modern research, computing the covariance matrix (a critical component in GPs and the source of most computational bottlenecks) becomes a challenge, both in terms of time and memory requirements. To address this, Dr Noack and Dr Risser's method involves distributing the computation of this matrix across multiple computing resources. This distribution not only tackles the issue of memory constraints by splitting the data but also accelerates the computation process.

Importantly, this new approach makes it feasible to apply their advanced GPs to datasets of sizes that were previously unmanageable, thus significantly expanding the scope and applicability of GPs in practical research scenarios.

Building Block 3: Augmented and Constrained Optimisation

As mentioned, the method above was developed by the researchers to perform optimally on sparse datasets, having recognised the prevalence of those in standard applications. The final building block deals with cases when the dataset may not be sufficiently sparse, enabling the GPs to correct their inefficiency by switching to an approximate model when a set resource threshold is hit.

This is achieved by augmented and constrained optimisation, a technique that prioritises sparsity in the GP model. This focus on sparsity is not just a matter of efficiency but also accuracy, as it allows the model to concentrate computational resources on the most informative parts of the data. Dr Noack and Dr Risser employ a specialised optimisation process that balances the need for a sparse representation against the overall accuracy of the model, which ensures that the GP does not become overly complex or computationally demanding, especially when dealing with the enormous datasets prevalent in contemporary research.

This careful balancing act between sparsity and accuracy is exactly what allows their approach to maintain the integrity of the GP model while making it viable for large-scale applications.

An Example: Weather Data

To demonstrate a practical application of their work, Dr Noack and Dr Risser applied their GP model to large-scale climate data, modelling daily maximum temperatures across the continental USA – a task that encompassed vast amounts of data due to the geographical expanse and temporal depth of the dataset.

Their approach adeptly handled the dataset of more than five million data points, showcasing the method's capacity to process and analyse extensive climate data efficiently, and also achieving the successful execution of the largest exact Gaussian process to date. In the context of climate science, where understanding and predicting patterns are crucial for policy-making and environmental management, the ability to efficiently process large datasets is invaluable. The effective application of their method to this area highlights the broad potential impact of their research, extending beyond theoretical advancements to tangible contributions in addressing pressing environmental and climatic challenges.

What's Next?

Dr Noack and Dr Risser will continue refining and expanding the capabilities of their GP methodology. One key area of focus is the further optimisation of kernel designs. By enhancing the flexibility and adaptability of these kernels, they will make their method even more effective in discerning the underlying patterns and relationships in complex datasets, and further refinements may enable more tailored applications to specific fields.

Meet the Researchers

Dr Marcus Noack and Dr Mark Risser
Lawrence Berkeley National Laboratory
Berkeley, California
USA

Dr Marcus Noack earned his PhD in Applied Mathematics from the University of Oslo, where he specialised in theoretical and numerical wave propagation and mathematical optimisation. He later joined Lawrence Berkeley National Laboratory, initially as a postdoctoral researcher and subsequently as a research scientist. Dr Noack is renowned for his work in stochastic processes, random fields, and stochastic function approximation, significantly contributing to autonomous and optimal data acquisition methods. He was pivotal in developing a world-record-breaking Gaussian process algorithm and received the Lawrence Berkeley National Laboratory Early-Career-Achievement Director's Award in 2022. Dr Noack plays a prominent role in furthering discussions about science, and in 2021, established the Community for Autonomous Scientific Experimentation to support the free exchange of ideas and knowledge in a friendly and constructive environment.

CONTACT

E: marcusnoack@lbl.gov
W: www.marcusnoack.com

KEY COLLABORATORS

Dr Eli Rothenberg, Dr Alexander Hexemer (The Advanced Light Source, Lawrence Berkeley National Laboratory)
Dr Kevin Yager, Dr Masafumi Fukuto (Brookhaven National Laboratory)
Dr Peter Ercius (Molecular Foundry, Lawrence Berkeley National Laboratory)
Dr Mark Risser (Earth Science Department, Lawrence Berkeley National Laboratory)
Dr Hengrui Luo (Department of Statistics, Rice University)

FUNDING

We want to acknowledge the Center for Advanced Mathematics for Energy Research Applications (CAMERA), which is jointly funded by the Advanced Scientific Computing Research (ASCR) and Basic Energy Sciences (BES) within the United States Department of Energy's Office of Science, under Contract No. DE-AC02-05CH11231. We want to further acknowledge the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under Department of Energy Contract No. DE-AC02-05CH11231. This work was further supported by the Regional and Global Model Analysis Program of the Office of Biological and Environmental Research in the Department of Energy Office of Science under contract number DE-AC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231 using NERSC award m4055-ERCAP0020612

FURTHER READING

M Noack, H Krishnan, MD Risser, KG Reyes, Exact Gaussian processes for massive datasets via non-stationary sparsity-discovering kernels, *Scientific Reports*, 2023, 13(1), 3155. DOI: https://doi.org/10.1038/s41598-023-30062-8

BERKELEY LAB

gp2Scale