

Bayesian Nonparametric Methods for Analyzing Ever-increasing Infinite Data

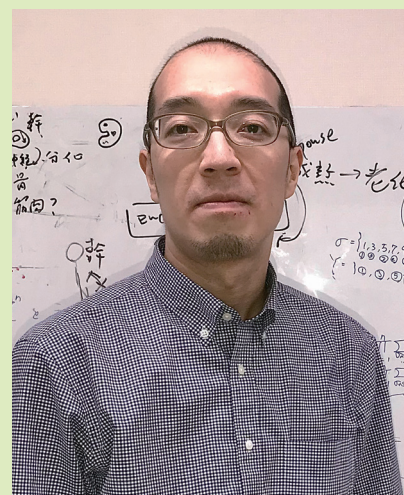
Masahiro Nakano

***Distinguished Researcher, NTT
Communication Science Laboratories
and NTT Basic Research Laboratories***

Abstract

One field of research that is gathering attention today is statistical machine learning that aims to apply analytical data to real-world problems. Conventional analysis has mainly used methods that manually set the scale and parameters of a statistical model or training model based on the data given and that involve costly tuning. These methods, however, incur massive amounts of time and labor, so using them to analyze data that will only continue to expand from here on is difficult. To solve this problem, there is a strong demand for a new technique that automatically adjusts model scale and parameters according to the quality and volume of data. In this article, we talked with NTT Distinguished Researcher Masahiro Nakano about his research on Bayesian nonparametric methods for solving problems faced by today's data society.

Keywords: statistical machine learning, Bayesian nonparametric methods, extremal combinatorics



Establishing a new technique for analyzing relational data of infinite size

—Mr. Nakano, could you first tell us what kind of technology is “Bayesian nonparametric relational data analysis”?

The volume and scale of data in today's society continue to increase on a yearly basis while the types of data being processed are becoming increasingly diverse. In such an environment, the conventional approach has been to manually set appropriate model parameters each time and to perform various types of processing. This approach, however, requires a mas-

sive amount of time and labor, which will make it hard for this approach to deal with further development of the data society. Bayesian nonparametric methods that I introduce here is a means of solving this problem. This technology adjusts the scale and parameters of a statistical model or training model in a data-driven manner (automatically according to the quality and volume of data) thereby reducing the massive amount of time and labor incurred by manual means that has been a problem in the past.

A technology called “relational data analysis” can be offered as one example of applying Bayesian nonparametric methods. Let's consider the problem of “indefinite complaints” in modern society as an

Example: discovery of the relationship between genes and disease

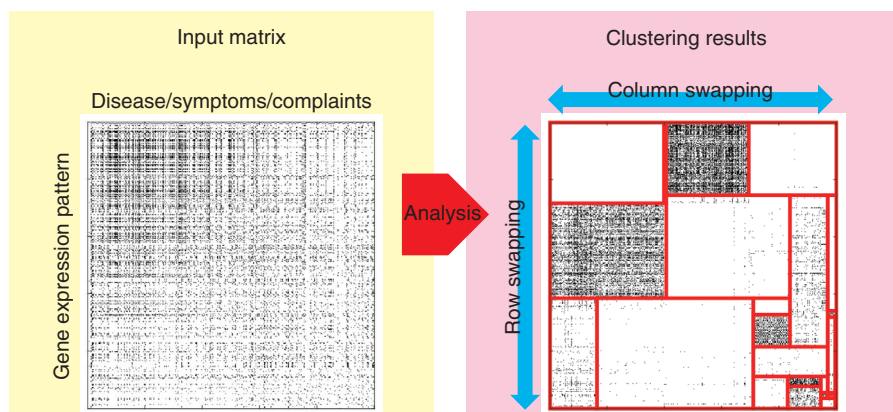


Fig. 1. Overview of Bayesian nonparametric relational data analysis.

application example. An indefinite complaint corresponds to a state in which a medical exam performed in response to subjective symptoms describing a physical or mental disorder is unable to identify the underlying disease or determine an effective method of treatment. Such symptoms are wide ranging and may appear in a composite manner as in “I’m feeling sluggish,” “I can’t stop feeling tired,” and “My arms feel dull,” which makes it difficult to identify disease in units of individual patients, presenting a social problem. However, if relational data between such a large number of patients and symptoms could be collected and analyzed, it would be possible to clarify correlations among individual items of data and uncover some sort of common features through data science.

However, there is a major problem in carrying out such an analysis. Specifically, if the number of data items that can be collected is (nearly) infinite, dealing with such a large volume of data is difficult. For example, the number of subjective symptoms is quite large including symptoms that differ from person to person, and the number of patients with indefinite complaints is potentially immeasurable. As such, there is the possibility that the relational data of patients and symptoms could become infinitely large along the rows and columns of a table, making data analysis extremely difficult.

To potentially analyze such data, technology that applies “uncertainty” to the infinite possibilities lying behind data in a data-driven manner is called “Bayesian nonparametric methods,” which was created in 2000 as a field in machine learning. A specific result

of this technology was the “discovery of the relationship between genes and disease.” This was achieved by partitioning the data into rectangles when analyzing that data based on clustering (the grouping of data based on similarities between data) and constructing a new probabilistic generative model that could generate all possible combinatorial patterns for those grouping candidates (**Fig. 1**). Additionally, adjusting and clustering optimal rectangular partitions in accordance with the given data achieved an efficient data analysis method. This approach made it possible to analyze all sorts of “nearly infinite” data according to the quality and volume of data, which could not be done by conventional techniques.

—What other technologies are you working on in your research?

In parallel with the technology I just described, I have been researching “super-Bayesian relational data analysis” since 2022. In simple terms, this technology relaxes the feature of “dealing with infinity” in Bayesian nonparametric relational data analysis.

In actuality, the greatest barrier to achieving Bayesian nonparametric relational data analysis has been the construction of an inference algorithm. Let me explain why this is so. By proactively using “infinity” on a computer, the construction of an inference algorithm for use in data analysis will be accompanied by “infinity,” and as a result, the possibility of some sort of infinite loop must be avoided, which makes it all the more difficult to design a model for performing analysis. This is a problem stemming from the use of

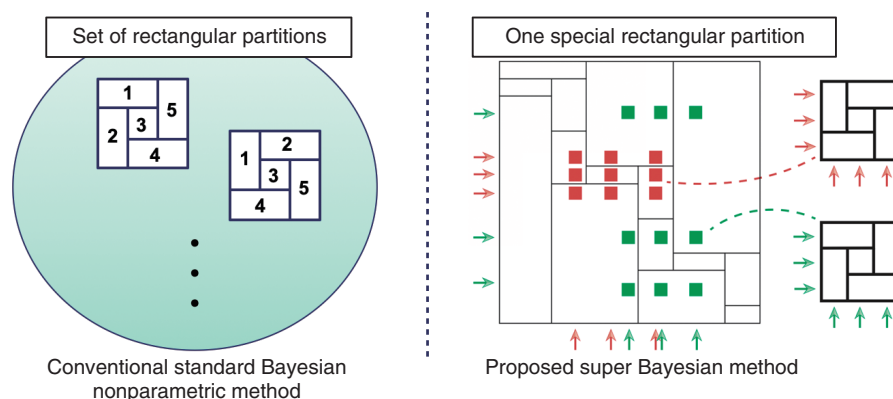


Fig. 2. Comparison of Bayesian nonparametric relational data analysis and super Bayesian method.

stochastic operations by the inference algorithm in Bayesian nonparametric relational data analysis. In general, when talking about probability, the total probability of the occurrence of candidate events must always be 100%. However, with an infinite number of cases, even if total probability is 100%, the probability of each number of cases itself is 0%. In other words, by making a partial collection of an “infinite number of cases” from the entire “infinite number of cases,” we would have a positive probability. Consequently, if one should carelessly think “I can ignore this event since its probability is zero,” the probability of that neglected event could become positive without realizing it, which means that there is a risk of a breakdown without total probability becoming 100%.

Soon after I started out on my research path, the sales pitch then being given for Bayesian nonparametric relational data analysis was that it was “capable of handling infinity in a virtual manner.” For me, this was a magic-like quality that I found very appealing. This was the 2000s, the time of the first boom of Bayesian nonparametric methods, and the emergence of new techniques on practically a daily basis was very stimulating to me as a researcher. However, around 2012, excitement began to grow around artificial intelligence (AI) in parallel with the third boom in neural networks, and at the same time, the excitement surrounding Bayesian nonparametric methods appeared to calm down relative to its early days. Personally, I believe the reason for this, in a few words, is that “it was not technology that could be applied in practice.” Given that the key feature of Bayesian nonparametric methods is the capability of “handling infinity in a virtual manner,” they turn out

to be models that are quite difficult to handle on computers. From the perspective of actual applications, the level of difficulty of constructing Bayesian nonparametric relational data analysis is often high, and it’s no exaggeration to say that first trying out a method based on existing deep learning technology is the most promising approach.

On the other hand, similar to the way that deep learning rose to prominence from 2012 on thanks to the third boom in neural networks, I began to wonder whether there was a method that could satisfy both ease of construction and practical use in relational data analysis. A breakthrough in my search occurred when I encountered “extremal combinatorics” in 2021–2022.

Extremal combinatorics is a field of study that deals with order arising from a certain kind of redundancy. For example, when considering a tremendously long numeric sequence (more precisely, a uniform random permutation), it is known that all kinds of short permutations, such as 14523 or 8245361, have a high probability of appearing as a partial series of that sequence. An important insight obtained from extremal combinatorics is that “if you prepare a massive thing with sufficient redundancy, more than enough things of various types can be expressed.” In short, in the case of Bayesian nonparametric relational data analysis, I noticed that an alternative to something thought to require “infinity” might be possible by replacing it with a “sufficiently redundant and large thing” and proposed the “super Bayesian method” (Fig. 2). Additionally, since no infinity is involved in such a “sufficiently redundant and large thing,” it holds the possibility of making construction of an inference algorithm much easier. By incorporating

Concept: wearable/visible stethoscope

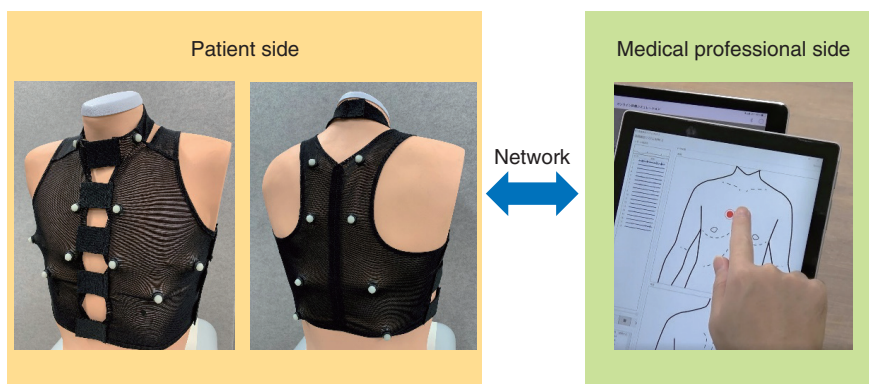


Fig. 3. "Telestethoscope" initiative.

insight on extremal combinatorics, I expect the super Bayesian method to become a technology that can be applied in practice while having the power of Bayesian nonparametric relational data analysis, and as such, to lead the way to future developments.

The super Bayesian method shows much potential as a new branch of Bayesian nonparametric relational data analysis, and taking advantage of the opportunity I had in making an oral presentation (top 4%) at AISTATS 2022, a major conference in this field, I have been putting effort into this method since then. Machine learning methods that proactively use such extremal combinatorics was still unexplored territory at that time, so I was fortunate in receiving a positive response from reviewers and the audience too, all of which encouraged me to continue with my research. In contrast to all the excitement surrounding the machine learning and AI field at present, research laboratories that are working on Bayesian nonparametric methods are relatively few in number. Research is something that makes progress through the efforts of mankind on the whole. By therefore continuing with this type of research and attracting more attention in this field, I hope to increase the size of my research team with an eye to further progress as a joint effort.

Making use of knowledge gained to contribute to NTT's Medical and Health Vision

—What is the outlook for your research going forward?

Looking to the future, I would like to make some

sort of contribution to NTT's Medical and Health Vision, which aims for a future in which everyone is always healthy with a hopeful outlook. At NTT, Digital Twin Computing, one of the key elements of the Innovative Optical and Wireless Network (IOWN), will be used to make a detailed mapping of a person's physical and mental state (bio-digital twin). Additionally, by predicting a person's physical and mental state sometime in the future, bio-digital twins will contribute to a future vision of medical care. I myself feel that I can contribute to making NTT's vision a reality by making good use of the knowledge I have gained through my work to date on Bayesian nonparametric relational data analysis and the super Bayesian method.

As a specific initiative in this regard, the Biomedical Informatics Research Group that I currently belong to is researching a device called a "telestethoscope" (Fig. 3). Based on the concept of a wearable/visible stethoscope, this technology transmits the data of a patient wearing the stethoscope to a remote location. To begin with, this capability enables auscultation by a medical professional with no risk of infection from the COVID-19 virus. Additionally, it enables early detection of any abnormalities when assessing the urgency of patient care for a remotely located patient while also promoting an understanding and awareness of one's own body on the patient's side.

Of course, research of bio-digital twin does not stop here. At NTT Communication Science Laboratories that I belong to, we are working on a long-term research goal of constructing human organs as digital computer models through a co-creation project with

Osaka University called “PRIME.” In this way, by searching for and implementing a variety of applications in addition to the telestethoscope, I will move forward with my research with the aim of contributing to NTT’s Medical and Health Vision and creating a world in which many people can live a highly enriched life.

—Mr. Nakano, please leave us with a message for researchers, students, and business partners.

I have been blessed with many connections even from the time prior to entering NTT. Professor Shigeki Sagayama, my academic advisor at the laboratory that I belonged to in graduate school, was originally from NTT, and I had the opportunity of receiving guidance from NTT Senior Distinguished Researcher Hirokazu Kameoka during joint research with NTT while I was a graduate student. Known globally from that time as a top research institute in the field of Bayesian nonparametric relational data analysis, NTT Communication Science Laboratories (in particular, Visiting Fellow Dr. Naonori Ueda and Takeshi Yamada, former head of NTT Communication Science Laboratories) made important contributions to the field even in those early days, so it was an attractive research institute for me from my student days onward.

Additionally, after entering NTT, I came to feel that NTT laboratories were an environment that, instead of restricting a researcher to only short-term choices or concentrations, actually welcomed the selection of diverse themes and challenges tailored to that person’s enthusiasm and passion. For example, around 2011 when I entered NTT, the third neural network boom was heavily influencing computer science around the world. However, NTT was able to avoid getting swept up in that tide, and I myself was able to

continue with my research on Bayesian nonparametric relational data analysis. Of course, it’s important to grasp global happenings and research trends, but on the other hand, a researcher may obtain unexpected results through research involving diverse and wide-ranging fields. As a researcher, I find this NTT environment that welcomes a variety of themes as a medium- and long-term investment to be very appealing.

At present, moreover, I am fortunate in having many occasions to serve as a mentor to junior researchers. In the course of giving all sorts of guidance, I have observed junior researchers launch research themes from up close, so based on this experience, I would be happy to give advice to any young researchers that happen to be reading this article. Specifically, in research, it is important to try out “something that you may think to be a waste” many times. Of course, it must be said here that the correct way of conducting research and boosting motivation differs from person to person. In recent years, there has been much emphasis on “cost performance” and “time performance,” which is opposite to my way of thinking.

However, when attempting to draw a new picture on a totally blank canvas, for example, it may happen that your brush does not move at all when faced with white paper having not a single mark. At this time, anything is fine, so once you dive right in and mark up the paper in some way, you will immediately feel better and your hand will start to move. Research is much the same. Even if you should erroneously mark up a sheet of paper at first, you will stop worrying about those marks as you come to overwrite them with a picture. Furthermore, as in the “sunk cost effect” that is often talked about, there is psychology at work here in that a person who invests any type of cost such as money, time, or effort into some sort of project will look back and find value in doing so. If you are having a problem in which your research is not progressing at all, my advice is to “start with small things.” Putting yourself to work beginning with what is familiar to you and from what you are capable of doing will expand possibilities. In this way, I would like you to open up the way to a new future!



■ Interviewee profile

Masahiro Nakano received his M.E. from the Graduate School of Information Science and Technology, The University of Tokyo in 2011 and entered Nippon Telegraph and Telephone Corporation (NTT) in the same year. He is currently assigned to NTT Communication Science Laboratories and has been concurrently working at NTT Bio-Medical Informatics Research Center since 2020. He is engaged in the study of statistical machine learning and its application to biomedical information processing using stochastic processes (stochastic models using infinite-dimensional parameter space) and universal objects in extremal combinatorics.