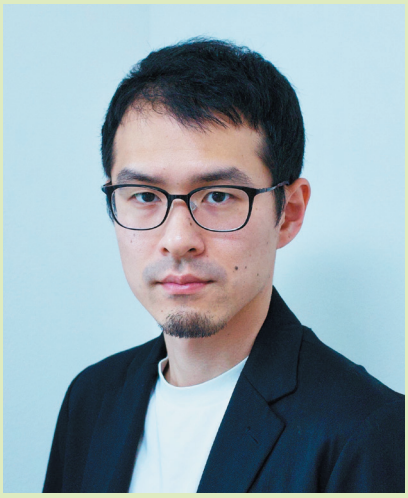
Fast Sparse Modeling Technology Opening Up the Future with Ultra-high-dimensional Data

Yasutoshi Ida
Distinguished Researcher,
NTT Computer and Data Science Laboratories



Abstract

As one pillar of NTT's IOWN (Innovative Optical and Wireless Network) vision, Digital Twin Computing aims to construct the world in a digital space. To this end, it is essential that data be obtained from people and things by sensors. However, advances in sensing technologies have created a new issue in which an increase in the number of data dimensions makes processing time longer, which makes it difficult to analyze and use collected data within a realistic period of time. We asked NTT Distinguished Researcher Yasutoshi Ida to tell us about "fast sparse modeling technology" as a means of solving this problem.

Keywords: sparse modeling, deep learning, high-dimensional data

Creation of a new technique for overcoming increased processing time of ultra-high-dimensional data

—Dr. Ida, what exactly is "sparse modeling" that you are now researching?

Sparse modeling is a technology that applies sparsity to data usage in the sense that "the amount of information needed is only a small part of all information obtained—most of the remaining information is unnecessary." Here, "sparse" has the meaning of "thin" or "scattered" based on the idea that "we should not make more assumptions than necessary to explain a certain matter" (Occam's razor) put forth by William of Ockham, a 14th-century English theologian.

To give an example of sparse modeling, let's consider the case of predicting tomorrow's temperature in Tokyo from weather data consisting of temperature, wind direction, pressure, etc. throughout Japan. In this case, the hypothesis that "data related to Tokyo temperature prediction is only that from some of the neighboring prefectures" can be postulated taking geographical relationships into account. This hypothesis is a basic precondition for using sparse modeling expressed as "important data is only a small part of the entire amount of data—all other data is unnecessary." Incorporating this prior knowledge called sparsity into our analysis enables us to specify those prefectures related to the prediction of tomorrow's temperature in Tokyo.

Example of high-dimensional data

In a global observation system consisting of ships, buoys, airplanes, weather satellites, etc., the number of dimensions in weather data collected by a sensing network can reach from several tens of thousands to several hundreds of millions or more.

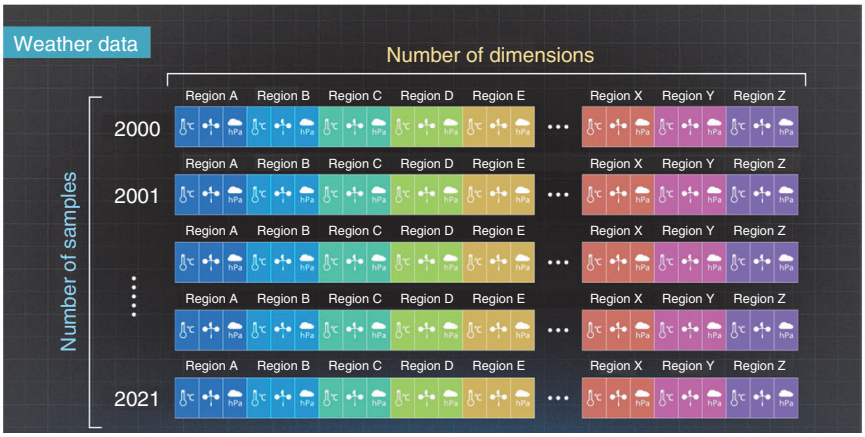


Fig. 1. Number of dimensions and number of samples in weather data.

The background to the need for this technique lies in the increase in the number of data dimensions that can be obtained due to recent advances in sensing technologies. In recent weather analysis, for example, the number of items of weather information (number of dimensions) that can be observed in regions throughout the world can reach several tens of thousands to several hundreds of millions with respect to the number of years (number of samples) on the vertical axis (**Fig. 1**). However, machine learning in recent years has widely adopted the approach that using a massive number of samples in training improves accuracy, but handling data in which the number of samples is relatively smaller than the number of dimensions is difficult. The aim of sparse modeling is to therefore solve this problem at least in part by using sparsity to select only those dimensions needed for analysis.

Model compression in deep learning is another example of how sparse modeling can be used. In deep learning, which can be called a fundamental technology of modern-day artificial intelligence (AI), accuracy has continuously improved by increasing the number of model parameters. However, a massive number of parameters increase memory consumption and processing time, so applying AI to edge-side hardware having limited memory capacity, for example, can be difficult. Under these conditions, incorporating sparsity in deep learning in the sense that "important parameters are only a small part of all parameters—all other parameters are unnecessary" decreases the number of AI parameters and the amount of memory consumed.

—What are the strengths of "fast sparse modeling technology" compared with existing technology?

Fast sparse modeling that I'm researching is achieving speeds up to 35 times faster than existing sparse modeling technology, which requires much processing time to compute a score expressing the importance of each dimension. Fast sparse modeling, on the other hand, replaces this score with the score's upper and lower bounds that can be computed at high speeds. In this way, fast sparse modeling has been successful in greatly shortening processing time while preventing degradation in accuracy. This approach is influenced by the high-speed technique used in database research by NTT Distinguished Researcher Yasuhiro Fujiwara, who was my research mentor when I entered NTT. By combining sparse modeling with databases, which are seemingly unrelated, I was able to achieve high speeds with an original technique not found in the existing research field of sparse modeling (**Fig. 2**).

Armed with this approach, I have my sights on the processing of ultra-high-dimensional data of several hundreds of millions of dimensions or more of which there are not many examples in existing sparse modeling technology. In 2019, the fast algorithm I just

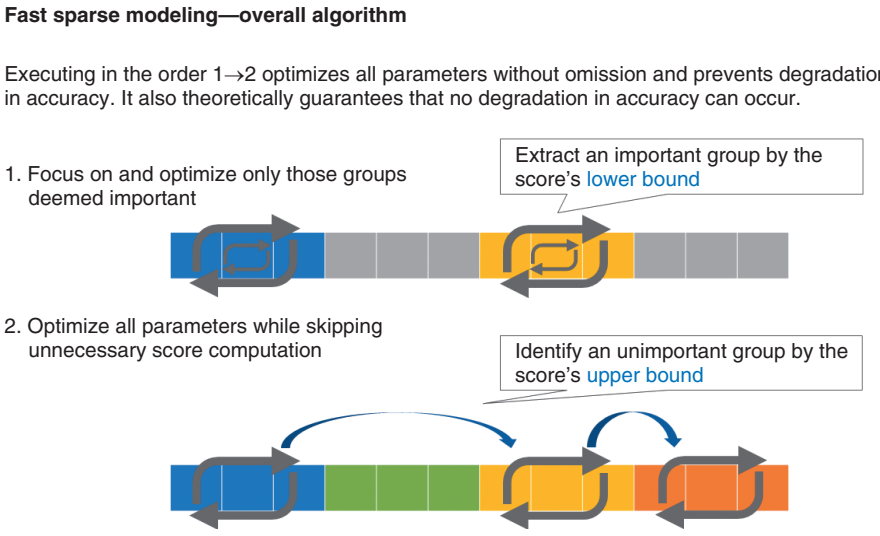


Fig. 2. Computational algorithm of fast sparse modeling.

mentioned increased processing speed by as much as 35 times, and the paper describing this achievement was accepted at Neural Information Processing Systems (NeurIPS), a leading conference in this field. This technology was subjected to tests under a variety of environments over a period of about one year so that it could be provided to NTT Group companies, and it was found that the algorithm could increase speeds in many analysis examples.

Yet, it was also found that high speeds could not be attained for a very small number of use cases. On conducting a detailed analysis as to why this occurred, we found that there were cases in which the overhead (pre-computation) for computing the score's upper bound in the fast algorithm was too large depending on the properties of the target data. We are now developing an algorithm with the aim of reducing this overhead as far as possible, and if successful, we expect a twofold increase in speed, which means that we should be able to improve processing speed by 35×2, or in other words, by a maximum of 70 times.

In parallel with these activities, we are studying an approach that combines sparse modeling with deep learning that achieves high accuracy with large-scale data. Basically speaking, deep learning requires a large number of samples, so we are researching how deep learning could be applied to use cases with a small number of samples. Combining the knowledge gained from this research with sparse modeling, we are studying technology that can achieve high accuracy even for high-dimensional data in which the number of samples is small. This is research with a long-term view, and my aim here is to improve the accuracy of sparse modeling by 10%.

—What kind of world can we expect to see through fast sparse modeling?

Fast sparse modeling can be applied mainly to tasks that perform analyses and predictions from ultra-high-dimensional data. In the Industrial Internet of Things targeting factories, for example, applying fast sparse modeling to the preprocessing of analysis for identifying time slots affecting an increase/decrease in production can speed up the PDCA (Plan–Do–Check–Act) cycle in data analysis and shorten the lead time to decision-making. This increase in speed also allows for the handling of time-series data spanning even longer periods thereby enabling high-speed analyses that have so far not been possible. Similarly, in genome-wide association studies, fast sparse modeling can be applied to the preprocessing of analysis for identifying genetic factors, or single nucleotide polymorphisms (SNPs), related to cancer and other diseases. Increasing processing speed in this way will enable the analysis of SNP data on an even larger scale. Moreover, in a nuclear fusion reactor, fast sparse modeling can be applied to the preprocessing of analysis for identifying control operations and sensors related to plasma disruption and maintenance. This capability should contribute to the explanation of phenomena related to the stable operation of

nuclear fusion reactors.

In addition, Digital Twin Computing in NTT's IOWN (Innovative Optical and Wireless Network) vision aims to create digital twins using large amounts of sensor data collected from sensors attached to people and things. Amid this initiative toward the future, sensing technologies are progressing and sensor data is becoming increasingly high dimensional, so I think that the opportunities for using sparse modeling that is especially adept at high-dimensional data analysis are increasing. Moreover, since the processing time devoted to sparse modeling will be increasing as data becomes increasingly high dimensional, I think that establishing technology for making sparse modeling even faster will become all the more important.

—How do you think fast sparse modeling will evolve going forward?

From here on, a key issue will be spatial complexity, that is, "To what extent can the amount of memory consumption be minimized?" The speed-up factor of fast sparse modeling is top class even from a global perspective, and in terms of shortening processing time, it's reaching milestones. It must be kept in mind, though, that memory consumption naturally increases as data takes on more dimensions. For example, when performing technical testing of fast sparse modeling to support the implementation of developed technologies in services, it may happen that memory consumption becomes too large. As a result, the number of dimensions at which processing can be performed can hit a ceiling even if no problems are occurring in the speed-up factor. Of course, bolstering the amount of memory through additional facility investment can solve this problem, but depending on the type of service being targeted, this method may prevent a profit from being made thereby creating a new problem. To resolve these issues, we are searching for a new method of fast sparse modeling that can hold down memory consumption while maintaining the speed-up factor.

Another direction that I can envision for fast sparse modeling is a "quantum leap in speed." As data becomes increasingly high dimensional together with advances in sensing technologies and limits to increasing speed are reached whatever the algorithm, I believe that we will have to consider new integrated approaches to achieving higher speeds that include distributed processing platforms and the use of advanced hardware. To tackle this research issue, collaborating with experts in other fields is essential—we must carefully search out a solution taking a long-term perspective. At NTT Computer and Data Science Laboratories, the research and development of advanced computers is also underway, so we are now investigating whether even faster sparse modeling could be achieved by making good use of such computers.

Determined to create practical technology after experiencing a "valley of death"

—What do you think is an important attitude to adopt in the work of research and development?

Based on what I learned from participating in a venture company during my university days, I place much importance on research and development focused on practical use. It was in 2010 during my third year of undergraduate studies that I first encountered machine learning. At that time, I was researching "topic modeling" technology that can visualize what kinds of topics occur in a document, and I felt strongly that I would like to incorporate this technology into an actual service. As a result, I participated in a venture company on the invitation from a friend. Once there, however, I was confronted with a number of problems not present at the research stage, such as "the mixing of data I had not expected with training data," "the model consequently behaving in unexpected ways," "output results that could not be interpreted by humans," and "a scale of data that was so large overall that training could not be completed." As a result, we never got to the point of implementing the technology in an actual service. The barrier between research and services is sometimes called the "valley of death." This was my first experience with this "valley of death" in which I was not able to

implement machine learning into some kind of a service. From this experience, my idea of wanting to create machine-learning technology that could solve problems in real-world settings took root. Today, I am committed to "developing practical algorithms that can be applied to actual services" and am researching such algorithms on a day-to-day basis while exchanging information with people at actual work sites.

Of course, there's more to research than simply adopting a practical approach. If a researcher is to produce research results that are truly groundbreaking and original, I believe it's also necessary to take up long-term research that is somewhat removed from practical considerations. A researcher who pursues both short-term and long-term research may find that the knowledge gained from practical research can also be useful in long-term research, and conversely, that some of the ideas being pondered in long-term research can be useful in practical research. In short, a researcher can take on real-world problems in practical research and feed back the knowledge gained there to long-term research, and vice versa, can share ideas nurtured in long-term research with practical research to solve problems in the real world. Long-term research may also lead to technologies or papers that can produce groundbreaking results. These two types of research with different properties can mutually interact with each other or create a feedback loop, which I think is an ideal form of research.

—Dr. Ida, could you leave us with a message for other researchers, students, and business partners?

Yes, of course. I feel that research and development at NTT spans a wide range of fields and has a portfolio of unique research fields even on a global basis. This is especially advantageous in problem solving. For example, to speed up deep learning, severe targets can be imposed on the speed-up factor. NTT, however, has a lineup of experts across many areas from algorithms to hardware and interconnects, so by simply talking to a colleague sitting right next to you, it may be possible to discover an integrated solution in no more than ten minutes. I believe that this ability of coming up with an approach to a difficult problem in a short period of time is something that only the NTT environment can provide thanks to its original portfolio and diverse lineup of research personnel.

In addition, I think that NTT Computer and Data Science Laboratories that I belong to is an organization that researches a broad layer of technology related to data science in terms of both the breath of fields and short-term/long-term perspectives. For example, data analysis technology developed with a focus on real-world problems is researched on the premise that it will be introduced into business relatively soon, which I think contributes to added value in NTT services. On the other hand, I think that the research and development of advanced computers and searching out applications for them play an important role in the long-term development of data science.

Finally, if working as a corporate researcher, surrounding conditions can suddenly change due, for example, to an organizational restructuring. At this time, if you significantly change your direction of research in an attempt to adapt to changes in the environment, you will not be able to use the technical abilities that you have so far developed and results will not be forthcoming, all of which can lead to stress and a loss of confidence. In my experience, however, if you have a universal, unshakeable core character and refuse to compromise, I think that you will be able to adapt to your environment more often than not without sacrificing your good qualities. This way of thinking can be expressed as "gentle in appearance but tough in spirit," or in other words, "my core research philosophy is uncompromising, but apart from that, I will adapt in a flexible manner to my environment." To students reading this who are looking to become a corporate researcher, I would like them to enjoy research while discovering and treasuring one's core character.

■ Interviewee profile

Yasutoshi Ida received his M.E. from Waseda University in 2014 and entered NTT in the same year. He received his Ph.D. in informatics from Kyoto University in 2021. He has been a distinguished researcher at NTT Computer and Data Science Laboratories since 2022. He is engaged in the research of fast-and-accurate sparse modeling and supports the implementation of developed technologies in services. He administers a liaison meeting that brings together AI-related engineers in the NTT Group for information exchange (NTT Deep Learning Liaison Meeting). His papers on developed technologies in this field have been accepted at many leading conferences (NeurIPS/ICML/AAAI/IJCAI/AISTATS).