Nonparametric Regression for Complex Data

博士論文（要約）

Nonparametric Regression for Complex Data
(複雑データのためのノンパラメトリック回帰)

今泉 允聡

Masaaki Imaizumi

2016/11/28

博士論文（要約）
Nonparametric Regression for Complex Data

Masaaki Imaizumi

2016/11/28

Chapter 1

Introduction

1.1 Overview

A purpose of this thesis is to investigate and propose a methodology for analyzing the complex data using the nonparametric statistics. In this chapter, we briefly introduce some basic ideas of the complex data and the nonparametric regression. Furthermore, we review the main concept of this thesis. Finally, we provide reviews for the rest of this thesis.

The complex data appear in various application fields with modern data science, such as medical analysis and finance data analysis. A methodology for the complex data is still a developing problem. Among several types of the complex data, we mainly focus on the tensor data and the functional data which are typical cases of the complex data.

The approach with nonparametric method is the statistical methodology which allows the statistical models to have an infinite dimensional parameter. Since the nonparametric statistics can reduce the bias from the model misspecification problem, the approach with the nonparametric method is intensively studied. However, it is known that the performance of the nonparametric methods is bad or unknown since the highly complicated structure of the complex data.

To propose the framework with the nonparametric statistics for the complex data, we mainly work on evaluating and reducing the complexity of the data. Namely, we define an estimator from a less complex hypothesis set to improve the speed of convergence of the estimator, while preserving the misspecification bias small. Especially, we focus on the smoothness property of the complex data or the statistical model. Smoothness often appears in the real data, and it is relatively easy for the statisticians to evaluate the effect theoretically. Our theoretical evaluation shows that the accuracy is improved by the complexity reduction, and experimental results guarantee the theoretical claim.

1.2 Complex Data

There are many types of the complex data, and we mainly consider the tensor data and the functional data.

1.2.1 Tensor Data

We let *tensor* denote a K-dimensional array. More formally, a K-mode tensor is an element of the tensor product of K linear spaces; e.g., the space of K-mode tensors is a

subspace of

$$\mathbb{R}^{I_1 \times I_2 \times \cdots \times I_K},$$

where I_k denotes a dimensionality of the k-th mode of tensors. Also, *tensor data* denotes a structure of data which can represent a higher order relation between several elements, and it is regarded as a generalization of the matrices. Figure 1.1 shows an image of the tensor data.

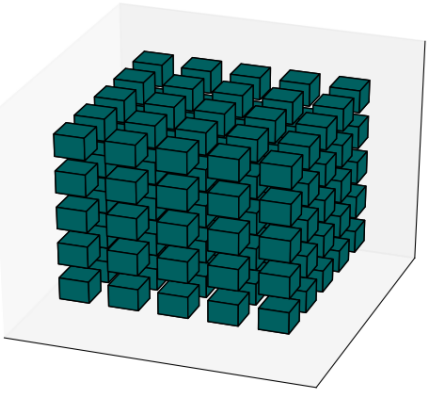

Figure 1.1: An image of 3-mode tensor data which belong to $\mathbb{R}^{5\times5\times5}$.

The tensor data appear in various modern application fields, such as medical analysis, finance, data-mining and biostatistics. For instance, in the fields of medical analysis, functional magnetic resonance imaging (fMRI) measures a change of blood flow in a brain as a 3-dimensional voxels, such as X-axis $\times$ Y-axis $\times$ Z-axis $\times$ *time*. Another typical example of a recommendation system in the field of data-mining. Stored data in the recommendation system contain scores for each good, user, and context. Thus, each of the scores is indexed by the three elements, then a set of the scores has the tensor structure.

Tensor decomposition is one of the standard methods for the tensor data analysis. Since the tensors contain a large number of elements, i.e. $\prod_{k=1}^{K} I_k$, there are some difficulties in analyzing the tensor data directly. Then, a representation of the tensor data by orthonormal basis vectors and their tensor products is applied. Let us define $\{e_j^{(k)} \in \mathbb{R}^{I_k}\}_j$ as a set of orthonormal vectors for each $k = 1, \ldots, K$. Then, by the decomposition with the basis, the tensor data $X \in \mathbb{R}^{I_1 \times \cdots \times I_K}$ are expressed as

$$X = \sum_{j_1=1}^{J_1} \cdots \sum_{j_K=1}^{J_K} \lambda_{j_1 \ldots j_K} e_{j_1}^{(1)} \otimes \cdots \otimes e_{J_K}^{(K)},$$

where $\{\lambda_{j_1 \ldots j_K}\} j_1 \ldots j_K = 1^{J_1 \ldots J_K}$ denotes a set of coefficients with the parameters $(J_1, \ldots, J_K)$. Here, $\otimes$ denotes the tensor product. Based on the form of the tensor decomposition, several methods are developed, and they are intensively used in the tensor data analysis. Coppi and Bolasco (1989), Bro (1997) and Kolda and Bader. (2009) provide a survey for the method and its variation.

1.2.2 Functional Data

Functional data denotes a structure of data which is regarded as a realized random function or stochastic process. Practically, the processes as the functional data are observed on infinite grids or finite but huge number of grids. Such data are observed in many application fields, such as time series analysis, biology, finance, and chemistry. Generally, the functional data are expressed by a random function $X(t)$ with index $t \in I$ on a compact interval $I \subset \mathbb{R}$. It is also viewed as a realization of a stochastic process in some functional space, such as $L^2(I)$. Figure 1.2 provides an image of the functional data.

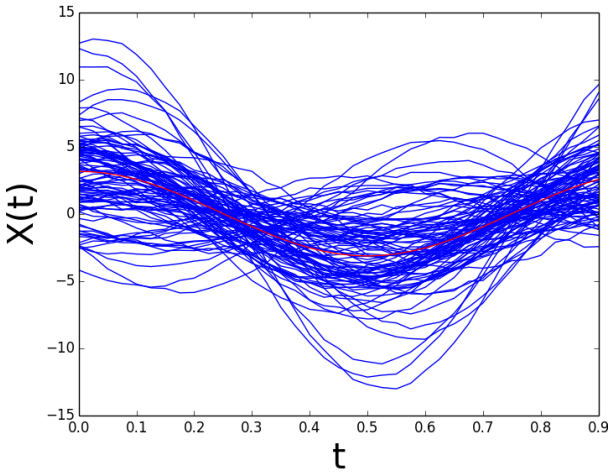

Figure 1.2: Plots of functional data. Each blue line denotes a realized path of $X(t)$ with index $t \in [0.0, 0.9]$. The red line is a mean $E[X(t)]$.

Since the functional data are expressed as an infinite dimensional vector or the high dimensional vector, a methodology and theory to handle the dimensionality are necessary. For analyzing the functional data, one of the key factors is the smoothness property which represents a kind of continuity of the functional data and is a rich source of information of the data. The methods for the functional data can avoid the high-dimensional problem of the functional data and obtain better performance by using the smooth property. Ramsey and Silverman (2005), Wang et al. (2016) and other literature provide introduction and survey for the methodology and applications.

1.2.3 Behavioural Data

Behavioural data are types of data observed from a behavior of individuals. Such data are collected in the fields of economics, marketing, robotics, biology, and many other areas. The behavior of individuals is observed as various forms. Thus there are a large number of frameworks for analyzing the behavioral data.

One of the difficulties of analyzing the behavioral data to understand decision makings of the individuals. In practice, the decision makings are affected by numberless information, for example, their current situations, future predictions, and past histories. For the analysis of the behavioral data, it is essential for an analyst to extract the relation between the decision making process and other environmental factors through modeling.

Markov decision process (MDP) is one of the representative methods to analyze the behavioral data. MDP is a common tool in the field of the reinforcement learning, and

it is a fundamental framework is summarized in Sutton and Barto (1998). In MDP, the decision making of individuals has represented as a maximizing a sum of future rewards which is affected by their current decision and situation. The framework is applied to the statistical analysis which are interested in the inference for parameters of MDP. Rust (1987) proposed the statistical framework with an asymptotic analysis for an estimator for the parameter of MDP.

1.3 Nonparametric Regression

1.3.1 Introduction

We introduce the nonparametric statistical models. Consider a family of probability distributions

$$\{P_\theta : \theta \in \Theta\},$$

where θ is a parameter of the distribution and Θ is a parameter space. For the statistical analysis, we estimate a true parameter in Θ from the observation from the distribution by statistical methods, and also provide the statistical inference. When Θ is an infinite dimensional space, we call the family of the distribution as the nonparametric statistical model.

Using the definition of the nonparametric statistical model, we introduce the nonparametric regression model. Consider that we have n independent and identically distributed pairs of random variables $\{(X_i, Y_i)\}_{i=1}^n$. Also, we assume that the set of pairs is generated from the following model

$$Y_i = f^*(X_i) + \epsilon_i, i = 1, \ldots, n,$$

where f is some function and ϵ_i is a random noise variable for each $i = 1, \ldots, n$. Here, the joint distribution of (X, Y) is written as the regression model

$$\{P_f : f \in \mathcal{F}\},$$

where $\mathcal{F}$ is a parameter space as a hypothesis set containing f^*. When $\mathcal{F}$ is an infinite dimensional space, e.g., (X, Y) is a random variable in $\mathbb{R}^2$ and $\mathcal{F}$ is a set of all $\alpha > 0$ times differentiable function, the regression model is regarded as a nonparametric regression model.

1.3.2 Estimation and Convergence Analysis

For the statistical analysis, we aim to find a true $f^* \in \mathcal{F}$ by an estimator $\widehat{f_n}$. A large number of estimators are suggested for the analysis. One of the standard methods is the Nadaraya-Watson estimator which is defined by the kernel function. Let (X, Y) take values in $[0, 1] \times \mathbb{R}$. Suppose we have a set of n observations $\{(X_i, Y_i)\}_{i=1}^n$, and let $k_h : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be the kernel function with the bandwidth $h > 0$. Then, the Nadaraya-Watson estimator is defined as

$$\widehat{f_n^{NW}}(x) := \frac{\sum_{i=1}^n k(x, X_i) Y_i}{\sum_{i=1}^n k(x, X_i)},$$

for all $x \in [0,1]$ with $\sum_{i=1}^n k(X_i, x) \neq 0$. The convergence of the estimator $\widehat{f}_n^{NW}$ is measured by order of the risk with respect to n. When we set $h = h_n \asymp n^{-1/(2\alpha+1)}$, we have the convergence rate of the Nadaraya-Watson estimator as

$$E\left[\|\widehat{f}_n^{NW} - f^*\|_2^2\right] = O\left(n^{-2\beta/(2\beta+1)}\right),$$

by Theorem 1.6 in Tsybakov (2003).

Another standard method is the series estimator. Let $\{\phi_j\}_{j=1}^\infty$ be an orthonormal basis in $\mathcal{F}$. Thus, we can consider the decomposition of f as $f^*(x) = \sum_{j=1}^\infty \theta_j^* \phi_j(x)$ with the coefficients $\{\theta_j\}_{j=1}^\infty$. By estimating the coefficients, we can define the estimator as

$$\widehat{f}_n^S(x) := \sum_{j=1}^m \widehat{\theta}_j \phi_j(x),\ \widehat{\theta}_j := \frac{1}{n}\sum_{i=1}^n Y_i \phi_j(X_i),\quad \forall j = 1, \ldots, m,$$

where m is an integer which denotes a properly selected truncation number. The convergence of the series estimator is also evaluated. When we select $m \asymp n^{1/(2\beta+1)}$, we have the following risk bound

$$E\left[\|\widehat{f}_n^S - f^*\|_2^2\right] = O\left(n^{-2\beta/(2\beta+1)}\right),$$

by Theorem 1.9 in Tsybakov (2003).

The convergence results can be extended to the case when X is d-dimensional random variable, namely, X takes a value in $[0,1]^d$. In the case, the above estimators with a proper setting satisfy

$$E\left[\|\widehat{f}_n - f\|_2^2\right] = O\left(n^{-2\beta/(2\beta+d)}\right),$$

where $\widehat{f}_n \in \{\widehat{f}_n^{NW}, \widehat{f}_n^S\}$. Here, we can easily check that the smoothness property via β improves the convergence, and the dimensionality of the input d makes the convergence worse. It is shown that this convergence rate satisfies the minimax optimality. These results are summarized in Tsybakov (2003).

1.4 Concept of Thesis

A purpose of this thesis is to develop an estimation problem which can obtain better performance for statistical analysis. For the purpose, we aim to improve the convergence rate of the estimator by reducing the complexity of the hypothesis set $\mathcal{F}$. We firstly review the existing theories which explain how the complexity of $\mathcal{F}$ affects the convergence. Then, we provide an overview of the idea of this thesis.

1.4.1 Review : Hypothesis Complexity and Convergence

An intuition from the convergence result is that the rate of the convergence is determined by a complexity of the parameter space $\mathcal{F}$ as the hypothesis set. Here, we use the term *complexity* as the *covering number* $N(\epsilon, \Theta, \|\cdot\|)$ which is a smallest number of ϵ-balls to cover Θ with respect to the norm $\|\cdot\|$. Namely, it is defined as

$$N(\epsilon, \mathcal{F}, \|\cdot\|) := \min\left\{k : \exists\{f_j\}_{k=1}^K \subset \mathcal{F} : F \subset \cup_{j=1}^k B(f_j, \epsilon)\right\},$$

where $B(f, \epsilon) := \{f' \in T : \|f - f'\| \leq \epsilon\}$. Similarly, the idea of the *packing number* represents an idea of the maximum number of the well-separated elements in $\mathcal{F}$. It is defined as

$$D(\epsilon, \mathcal{F}, \|\cdot\|) := \max\big\{k : \exists\{f_j\}_{k=1}^{K} \subset \mathcal{F} : \min_{1\leq j,j'\leq k} \|f_j - f_{j'}\| \geq \epsilon\big\}.$$

Note that there exists a relation between the above numbers as $N(\epsilon, \mathcal{F}, \|\cdot\|) \leq D(\epsilon, \mathcal{F}, \|\cdot\|) \leq N(\epsilon/2, \mathcal{F}, \|\cdot\|)$, for all $\epsilon > 0$. The set of elements $\{f_j\}_{j=1}^{K}$ measures the complexity of $\mathcal{F}$ and the covering and packing numbers are the upper bound and the lower bound of the measure. For instance, assume that $\mathcal{F}$ be the class of all continuous functions $f : [0,1]^d \to \mathbb{R}$ with β-th derivative. Then, Theorem 2.7.1 in van der Vaart and Wellner (1996) shows that

$$\log N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq C_k \left(\frac{1}{\epsilon}\right)^{d/\beta},$$

where $C_k > 0$ is a finite constant.

The simple relation between the convergence rate and the complexity of the hypothesis set is provided by Yang and Barron (1999). Let the norm $\|f - f'\|$ be bounded the Kullback-Leibler divergence, and $\overline{\mathcal{F}}_n$ be a set of estimators from the n observed samples $\{(X_i, Y_i)\}_{i=1}^{n}$. With additional proper conditions, Corollary 1 in Yang and Barron (1999) shows that

$$\min_{\widehat{f}_n \in \overline{\mathcal{F}}_n} \max_{f^* \in \mathcal{F}} E\left[\|\widehat{f}_n - f^*\|^2\right] \asymp \epsilon_n^2,$$

where the sequence ϵ_n satisfies

$$\epsilon_n^2 = \log D(\epsilon_n, \mathcal{F}, \|\cdot\|)/n.$$

By the result, we can check that the more complexity hypothesis set $\mathcal{F}$ makes the convergence rate slower, and the convergence rate is optimal from the aspect of the minimax theory. Using this result, we can also see that the convergence rates in the examples about the Nadaraya-Watson and the series estimator provides are derived by the bound for covering number for the differentiable functions. These ideas are summarized in van der Vaart (2000) and Gine and Nickl (2015).

1.4.2 Our Approach : Reduce Complexity

The main concept of this thesis is to improve the performance of the nonparametric regression model by proposing a less complex hypothesis set. We firstly decompose the risk of estimator which is our interest. Suppose that there exists an original hypothesis set $\mathcal{F}$, and the unknown true parameter $f^* \in \mathcal{F}$. Our main interest is to propose an estimator $\widehat{f}_n$ and evaluate the risk

$$E\left[\|\widehat{f}_n - f^*\|^2\right].$$

To propose the estimator $\widehat{f}_n$, consider we have the less complex hypothesis set $\widetilde{\mathcal{F}}$, and we assume that $\widetilde{\mathcal{F}}$ and the original hypothesis set $\mathcal{F}$ satisfy

$$\log N(\epsilon, \widetilde{\mathcal{F}}, \|\cdot\|) \leq \log N(\epsilon, \mathcal{F}, \|\cdot\|).$$

Then, we define the estimator $\widehat{f}_n$ from the parameter space $\widetilde{\mathcal{F}}$. Also, let us define an element of $f^0 \in \widetilde{\mathcal{F}}$ which satisfies

$$f^0 = \arg\min_{f \in \widetilde{\mathcal{F}}} \|f - f^*\|.$$

Then, we evaluate the risk as

$$E\left[\|\widehat{f}_n - f^*\|^2\right] \lesssim \|f^0 - f^*\|^2 + E\left[\|\widehat{f}_n - f^0\|^2\right].$$

Then, we regard the term $\|f^0 - f^*\|^2$ as the bias from the model misspecification by $\widetilde{\mathcal{F}}$, and also regard $E\left[\|\widehat{f}_n - f^0\|^2\right]$ as the effect from the estimator.

According to the above discussion about the complexity and convergence, the latter term is decreased when the complexity of $\mathcal{F}$ is sufficiently smaller than that of $\mathcal{F}$. Thus, we can improve the speed of convergence of the estimator with respect to n. The misspecification bias $\|f^0 - f^*\|^2$ is also our interest, however, there is no general method to evaluate the term. We handle the bias term by considering hypothesis sets $\widetilde{\mathcal{F}}$ properly, using characteristics of the various complex data. One of the representative ways is to use the smoothness property which appears in many types of complex data.

1.5 Bibliography

Bro, R. (1997). PARAFAC. Tutorial and applications. *Chemometrics and intelligent laboratory systems*, **38**(2), 149-171.

R. Coppi and S. Bolasco, eds., *Multiway Data Analysis*, North-Holland, Amsterdam, 1989.

Gine, E., and Nickl, R. (2015). Mathematical foundations of infinite-dimensional statistical models (Vol. 40). Cambridge University Press.

Kolda, T. G., and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, **51**(3), 455-500.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. 2nd Edition. Springer.

Rust, J. (1987). Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher. *Econometrica*, 55(5), 999-1033.

Sutton, R. S., and Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge: MIT press.

Tsybakov, A.B. (2003). *Introduction to Nonparametric Estimation*. Springer.

van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge university press.

van Der Vaart, A. W., and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer New York.

Wang, J. L., Chiou, J. M., and Mller, H. G. (2016). Functional Data Analysis. *Annual Review of Statistics and Its Application*, **3**, 257-295.

Yang, Y., and Barron, A. (1999). *Information-theoretic determination of minimax rates of convergence*. The Annals of Statistics, 27(5), 1564-1599.

要約

本博士論文の 2 章および 3 章について、以下の媒体で刊行されるため、非公表とする。

- Journal of Machine Learning Research W & CP series, volume 48 (ICML 2016), pp. 727-736.
- Journal of Machine Learning Research W & CP series, volume 70 (ICML 2017), To appear.

また残りの章について、近い将来に刊行（5 年以内に出版予定）される期待があるため、非公表とする。