Nonparametric Regressions in Econometrics:

Unbiasedness, Consistency, Efficiency, and Interpretability

Jingwei Ma[1], Wenqi Jiang[2]

[1]Johns Hopkins University

[2]Technical University of Berlin

November 30, 2022

Abstract

This paper consists of four different chapters, concentrating on unbiasedness, consistency, efficiency, and interpretability.

The first chapter serves as introduction. In the second chapter, a prototype of nonparametric regression is proposed. This prototype is simple; nonetheless, it can be proven to be the root of complicated nonparametric regressions.

The following chapters are designed to generalize conclusions of the prototype. The third chapter focuses on estimation and optimization error: a new estimation for nonlinear regressions is established, which can be proven to eliminate optimization error, minimize variances of unbiased estimators and prevent overfittings. In the last chapter, we focus on approximation error and interpretability: we prove that Deep Neural Network is interpretable as it can be re-structured into one-hidden-layer Neural Network. Meanwhile, some of one-hidden-layer Neural Network, with proper basis functions, can be proven to be unbiased, consistent, efficient, and interpretable. This paper only discusses about numerical data.

Keywords: function approximations, nonparametric regressions, interpretability, Minimum Variance Unbiased Estimator (MVUE), overfittings, Deep Neural Networks

Introduction

All data, functions, samples, and populations in this paper are literally numeric. This paper doesn't involve any comments on pictures, texts, voices, moods, cognitions and et ectara. Expansion explanations and analogy explanations to the following contents lead to mistakes.

1.1 Fixed versus flexible: why nonparametric regressions are necessary?

Traditional econometric regressions typically refer to parametric regressions. Parametric regressions focus on estimating true functions, whose forms are already pre-established before performing regressions. In linear regressions, for instance, form of true function must be linear, thus fitting function is restricted to be linear as well.

Generally speaking, form of fitting function must be identical with that of true function; otherwise, estimations are biased. In traditional econometrics, although forms of true functions are proved or defined in some studies; in many cases, true functions' forms are assumed. Thus, true functions and fitting functions may have different forms. As the fitting functions fail to capture complicated non-linearity of true function, underfittings happen. Underfittings are featured with biased estimators and small variances; thus, both estimators and statistical inferences are distorted.

In nonparametric regressions, to the opposite, forms of true functions not under strong assumptions. Shapes of fitting functions are more flexible and are able to successfully capture non-linearity, as long as true functions meet some requirements.

1.2 Forms and parameters: why function approximations?

As forms of fitting functions (and of true functions) are not assumed in nonparametric regressions, we need to put effort into estimating both forms and parameters of true functions.

Function approximations can be of help in nonparametric regressions. Function approximation is a technique, independent from regressions. This technique can decompose a (true) function into a combination of basis functions. Basis functions are typically in identical forms. Connections among basis functions are pre-established. By changing parameters of basis functions, forms and parameters of true functions would change simultaneously.

We can understand nonparametric regressions from another perspective. Before performing regressions, forms of true functions are already known as combinations of basis functions. As connections among basis functions have been pre-established, our only task is to estimate parameters of basis functions. Hence, boundary between parametric and nonparametric methods blurs, because in both situations we only need to consider estimating parameters.

1.3 Three errors: typical problems in nonparametric regressions

If an estimation is unbiased, total error of a nonparametric regression can be decomposed into three types of errors: approximation error, estimation error, and optimization error:

$$\underbrace{f_{true} - f_{\widetilde{optımzıed}}}_{total\ error} = \underbrace{\left[f_{true} - f_{\breve{approxımatıon}}\right]}_{approximation\ error} + \underbrace{\left[f_{\breve{approxımatıon}} - f_{\widehat{estımated}}\right]}_{estimation\ error} + \underbrace{\left[f_{\widehat{estımated}} - f_{\widetilde{optımzıed}}\right]}_{optimization\ error}$$

Function approximation are typically under conditions, that number of basis functions tends to infinity. Equivalently, given limited number of basis functions, function approximations are different from true functions. This difference is defined as approximation error, which can also be considered as a proportion of estimation biases.

Similar to parametric regressions, estimators in nonparametric regressions are also different from true parameters, hence, estimation error happens. Overfitting, a common problem in nonparametric regressions, is one of the most famous special cases of estimation error. Overfitting can be defined as including irrelevant orthogonal terms, expected values of which are 0 (statistical insignificancy).

Expressions of function approximations are typically complex. In such situation, optimizing criterion (loss) functions can be complex as well. If gradient descents don't successfully minimize criterion functions, optimization error takes place.

1.4 Requirements of regressions: unbiasedness, consistency, efficiency, and interpretability

Unbiasedness, consistency, and efficiency are also required in traditional parametric regressions.

Besides traditional requirements, in multivariate nonparametric regressions with number of variables n, efficient estimators should also be independent from n, which typically makes convergence rates disappointingly low.

Interpretability is an additional requirement. By legislations, in areas of, for example, mortgage assessments, interpretability is mandatory. Different from other viewpoints, as what will be discussed in the last chapter, nonparametric regressions, including DNN, can be interpreted. The only difference is that, whether it is easy to interpret or not. Easiness to interpret nonparametric regressions is affected by complexity, of both connections among basis functions and calculations within basis functions.

There is a straightforward way to understand interpretability. In some cases, calculations among and within basis functions are complex, even implicit. Albeit complexity, such calculations are still known before regression. In addition, after performing regression, all estimators in all basis functions are indicated. Hence, the regression is, at least, implicitly known after regression. To some extent, we can regard the whole regression as an implicit function. From another perspective, a nonparametric regression, where calculations among and within basis functions are straightforward, would be more interpretable.

Polynomial Regression Based on *Weierstrass First Theorem*: A prototype

In this chapter, we would establish a fundamental nonparametric regression. This regression is based on *Weierstrass First Theorem*, whose expression is simple.

However, as what will be proved in Chapter 4, *Weierstrass First Theorem* is the root of other complex function approximations, and this regression serves as blocks and mortars of complicated DNN.

2.1 Introduction

As prementioned, complexity, of both connections among basis functions and calculations within basis functions, affect interpretability. Nonparametric regression introduced in this chapter is simple, because connections among basis functions are linear, and, calculations within basis functions are multiplication. Regression process can be concluded as three convergences:

1. Averages of estimators are nearly normally distributed. Taking advantage of this property, as long as estimators are unbiased, average of estimators would converge to true parameters.
2. *Lagrange Interpolation*, an algebraic function going through every point, *may* be a special case of overfitting function. However, gradients are not necessary for solving *Lagrange Interpolation*, because it is an equation.
3. *Lagrange Interpolation* generates algebraic polynomial functions. To compare polynomial functions with true function, we need to use *Bernstein Polynomials* as a function approximation, to decompose true function into a combination of algebraic polynomials.

2.2 Theorem of Bernstein Polynomial Regression

1) Collect k groups of data, independently and randomly sampled from a population. Each group contains $m + 1$ numbers. Total amounts of samples are $k(m + 1)$. Suppose that $k \gg 0$

and $m \gg 0$. k and m are integers.

2) For $group\ 1$, using *Lagrange Interpolation*, a polynomial function can be generated, which can go through $m + 1$ points in dataset 1:

$$y_1 = a_{10} + a_{11}x^1 + a_{12}x^2 + \cdots + a_{1m}x^m$$

Likewise, all other groups,

$$y_2 = a_{20} + a_{21}x^1 + a_{22}x^2 + \cdots + a_{2m}x^m$$

$$\ldots$$

$$y_k = a_{k0} + a_{k1}x^1 + a_{k2}x^2 + \cdots + a_{km}x^m$$

3) Finally, calculate averages of coefficients collected from different groups. These averages will converge to true parameters.

$$\widehat{\alpha_0} = \frac{a_{10} + a_{20} + a_{30} + \cdots + a_{k0}}{k}$$

$$\widehat{\alpha_1} = \frac{a_{11} + a_{21} + a_{31} + \cdots + a_{k1}}{k}$$

$$\ldots$$

$$\widehat{\alpha_m} = \frac{a_{1m} + a_{2m} + a_{3m} + \cdots + a_{km}}{k}$$

Hence,

$$\hat{y} = \widehat{\alpha_0} + \widehat{\alpha_1}x^1 + \widehat{\alpha_2}x^2 + \cdots + \widehat{\alpha_m}x^m$$

2.3 Proof

In the following contents, we concentrate on proving convergences, where m and k tend to infinity.

2.3.1 Assumptions

1. Denote the true function as $f(x)$. Suppose that $f(x)$ is uniformly continuous.

2. Denote a *Bernstein Polynomial* function as $B_m(f, x)$, whose highest power is m. Or equivalently, there are $m + 1$ parameters in a *Bernstein Polynomial* function.

3. In multivariate case, denote number of variables as n.

4. There are k groups of data and each of them contains $m+1$ data points. All data is independently and randomly selected from the true population.

2.3.2 The first Convergence

In $group\ 1$, there are $(m+1)$ data points, denote as (x_{10}, y_{10}),(x_{11}, y_{11}) … (x_{1m}, y_{1m}). According to *Lagrange Interpolation Theorem*, a polynomial function can be generated, which can go through every point in $dataset\ 1$:

$$y_1 = \sum_{j=0}^{m} y_j\, L_j$$

$$where\ L_j = \prod_{i=0, i \neq j}^{m} \frac{x - x_i}{x_j - x_i} = \frac{x - x_0}{x_j - x_0} \cdots \frac{x - x_{j-1}}{x_j - x_{j-1}} \frac{x - x_{j+1}}{x_j - x_{j+1}} \cdots \frac{x - x_m}{x_j - x_m}$$

Coefficients are not important at this step. Denote as $a_{10}, a_{11} \dots a_{1m}$

$$y_1 = a_{10} + a_{11}x^1 + a_{12}x^2 + \cdots + \ a_{1m}x^m$$

Likewise, using *Lagrange Interpolation Theorem* in $dataset$ 2, $dataset$ 3 … $dataset\ k$.

$$y_2 = a_{20} + a_{21}x^1 + a_{22}x^2 + \cdots + \ a_{2m}x^m$$

$$\dots$$

$$y_k = a_{k0} + a_{k1}x^1 + a_{k2}x^2 + \cdots + \ a_{km}x^m$$

According to *Weierstrass Theorem*, as true function $f(x)$ is uniformly continuous, it is approximate to a combination of algebraic polynomials:

$$B_m(f, x) = \sum_{l=0}^{m} f\left(\frac{l}{m}\right)\binom{m}{l} x^l (1-x)^{m-l}$$

Coefficients can be *re-structured* and be denoted as $b_0, b_1 \dots b_m$:

$$B_m(f, x) = b_0 + b_1x^1 + b_2x^2 + \cdots + \ b_mx^m$$

$B_m(f, x)$ uniformly converges to $f(x)$ as m increases. Convergence rate of $B_m(f, x)$ to $f(x)$ is:

$$|B_m - f| \leq cW_f\left(m^{-\frac{1}{2}}\right)$$

In multivariate regressions with number of variables n, convergence rate of approximation is:

$$|B_m - f| \sim m^{-\frac{1}{2n}}$$

2.3.3 The Second Convergence

Approximation error is out of consideration at this stage. In each sub-regression, it follows that:

$$Y = X\beta + u$$

where Y, β and u are $(m \times 1)$ vectors, and X is a $(m \times m)$ matrix. Detailed discussions on sub-regressions are in the next chapter.

Running this regression, using least squares as criterion function,

$$Loss\ function = u^T \cdot u = (Y - X\beta)^T \cdot (Y - X\beta)$$

$$= Y^T Y - 2Y^T X\beta + (X\beta)^T \cdot (X\beta)$$

By minimizing criterion function, estimator is

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

In this case, criterion function can decrease to 0, if:

$$Y = X\beta \Leftrightarrow Y^T Y - 2Y^T X\beta + (X\beta)^T \cdot (X\beta) = 0$$

Please notice that loss function decreases to 0 is an extreme case of “overfitting”. However, this “overfitting” is not frustrating, as it doesn’t result from including too many parameters; instead, it is because of small size of data in each group.

According to *Lagrange Interpolation Theorem*, a polynomial function, with highest power of m, can goes through m+1 data points and satisfies:

$$Y = X\beta$$

Hence, the polynomial function generated from *Lagrange Interpolation*, is an extreme case of overfitting regressions.

2.3.4 The Third Convergence

Using *Lagrange Interpolation* in $group\ 1$ to k,

$$y_1 = a_{10} + a_{11}x^1 + a_{12}x^2 + \cdots + a_{1m}x^m$$

$$\dots$$

$$y_k = a_{k0} + a_{k1}x^1 + a_{k2}x^2 + \dots + a_{km}x^m$$

Due to *the second convergence*, all these equations can be considered as (overfitting) regressions, where $a_{10}, \dots a_{km}$ are unbiased estimators. Since data points are independently and randomly sampled from true population, estimators, for example, $a_{10}, \dots a_{k0}$, are mutually independent as well. According to *Central Limitation Theorem*, as $k \gg 0$, averages of $a_{10}, \dots a_{k0}$, or, $a_{11} \dots a_{k1}$ et cetera will approximately be normally distributed. Thus, convergence rate of estimation only depends on $k^{-\frac{1}{2}}$. When k tends to infinity, averages of estimators converge to true parameters.

2.4 Potential drawbacks of new model: inefficiency

In multivariate cases, when using *Bernstein Polynomials*, convergence rate of approximation depends on $m^{-\frac{1}{2n}}$. Hence, performance of *Bernstein Polynomial Regression* would be satisfying in univariate regressions, while convergence rates of approximation would be frustrating in multivariate regressions.

Despite of approximation error*,* when using *Lagrange Interpolation*, both estimation and optimization error can be proved to be minimized, and overfitting can be prevented. Using *Lagrange Interpolation* can be considered as a special case of *Exact-Identification Ensemble (EIE)*, a method of estimation to be brought up in the next chapter.

On one hand, if we only care about whether a regression can be approximate to the true function or not, performing polynomial regressions is sufficient. On another hand, most function approximations are under the condition that m, numbers of parameters, tend to infinity; while sizes of data, typically, cannot tend to infinity. In order to perform regressions, number of parameters has to be much smaller than size of data.

Hence, another question must be considered: what is difference between true function and estimated function? That is why convergence rates of approximation, estimation, and optimization do matter. In details, we would consider:

1. How can we diminish, even minimize all these three errors? In multivariate regressions, can they be independent from n?
2. Are these three convergence rates related?
3. Are properties of basis functions would affect convergence rates?
4. Why structures of connections among basis functions matter? Why would we use DNN?

All these questions will be discussed in the following chapters.

MVUE for regressions: minimum estimation and optimization error

This chapter firstly discuss about standards of *Minimum Variance Unbiased Estimator (MVUE)*. Then, we would claim *OLS, MM, MLE, GMM* and *2SLS* can all be regarded as special cases of *MVUE.* Then, a new estimation, *Exact-Identification Ensemble (EIE),* is brought up, which can perform (asymptomatically) unbiased estimations with minimum variances while eliminating optimization error and preventing overfitting. Finally, statistical inferences are used to evaluate performances of *MVUE* (including *EIE*), which serves as an alternative to regularizations.

3.1 Standards for Minimum Variance Unbiased Estimator (MVUE)

At this stage, we discuss parametric differentiable nonlinear regressions. Undifferentiable basis functions, like *MAXOUT*, can be estimated and interpreted in another way, as their underlying logics are similar to *local regressions*.

3.1.1 Assumption

For m numbers of mutually independent variables $x_1, x_2, \dots\ x_m$ from an implicit true function $g(x_m, \theta) = 0$, a regression satisfies that:

$$g(x_m, \theta) = u$$

where u is an irrelevant residual term, with 0 expected value and unknown distribution as well as variance. Or, equivalently, u is orthogonal to the true function $g(x_m, \theta)$.

In order to be more generalized, true function is defined in form of implicit function. In this paper, true functions only refer to uniformly continuous and differentiable functions. x_m represents that true function contains m numbers of variables and parameters of the true function is denoted as θ.

Such assumptions are not as flexible as expected. The assumptions require that relationships among variables can be expressed by a fixed function. Residual terms must be orthogonal to true functions. If, and only if assumptions are fulfilled, regressions can be performed. Regressions are not the answer for everything and hence shall not be abused and misused.

Corollary 3.1 Only true function can minimize SSE to $var(u)$.

3.1.2 Standards for MVUE: from Hansen's GMM

At this stage, we consider directly performing regressions. Panel data and *Ensemble Algorithm* will be discussed afterwards.

Lemma 3.1 $\sum_{i=1}^{m}\frac{\partial g(x_m,\hat{\theta})}{\partial\theta}=0$ is sufficient but not necessary condition of (asymptotically) unbiased estimations.

Necessary condition has been proved by Hansen (1982). Ignore weight matrix in GMM:

$$(\hat{\theta}-\theta)=-\left(\frac{\partial^2\left(g(x_m,\theta\,)^T g(x_m,\theta\,)\right)}{\partial\theta\partial\theta^T}\right)^{-1}\left(\frac{\partial\left(g(x_m,\theta)^T g(x_m,\theta)\right)}{\partial\theta}-\frac{\partial\left(g\left(x_m,\hat{\theta}\,\right)^T g\left(x_m,\hat{\theta}\,\right)\right)}{\partial\theta}\right)$$

When$\sum_{i=1}^{m}\frac{\partial g(x_m,\hat{\theta})}{\partial\theta}=0$, $\frac{\partial\left(g(x_m,\hat{\theta}\,)^T g(x_m,\hat{\theta}\,)\right)}{\partial\theta}=\ 0.$ It follows that:

$$(\hat{\theta}-\theta)=-\left(\frac{\partial^2\left(g(x_m,\theta\,)^T g(x_m,\theta\,)\right)}{\partial\theta\partial\theta^T}\right)^{-1}\left(\frac{\partial\left(g(x_m,\theta)^T g(x_m,\theta)\right)}{\partial\theta}-0\right)$$

Asymptotically, value of $\frac{\partial\left(g(x_m,\theta)^T g(x_m,\theta)\right)}{\partial\theta}$ follows that:

$$\frac{\partial\left(g(x_m,\theta)^T g(x_m,\theta)\right)}{\partial\theta}=2\left[\frac{1}{m}\sum_{i=1}^{m}\frac{\partial g(x_m,\theta)}{\partial\theta}\right]^T\left[\frac{1}{m}\sum_{i=1}^{m}g(x_m,\theta)\right]$$

$$\overset{p}{\rightarrow}2\mathrm{E}\left(\frac{\partial g(x_m,\theta)}{\partial\theta}\right)^T\mathrm{E}\left(g(x_m,\theta)\right)\overset{p}{\rightarrow}0$$

Hence, asymptotically, expected value of $\left(\hat{\theta}-\theta\right)$ is 0:

$$\mathrm{E}\left(\hat{\theta}-\theta\right)\overset{p}{\rightarrow}0$$

Subsequently, variance of $\hat{\theta}$ is:

$$var(\hat{\theta}-\theta)=\frac{1}{m}\left(\mathrm{E}\left(\frac{\partial g(x_m,\theta)}{\partial\theta}\right)^T\left(\mathrm{E}(g(x_m,\theta)g(x_m,\theta)^T)\right)^{-1}\mathrm{E}\left(\frac{\partial g(x_m,\theta)}{\partial\theta}\right)\right)^{-1}$$

Lemma 3.2 $\sum_{i=1}^{m}\frac{\partial g(x_m,\hat{\theta})}{\partial\theta}=0$ is necessary and sufficient condition of (asymptotically) unbiased estimation with minimum variance.

In the formula:

$$(\hat{\theta}-\theta)=-\left(\frac{\partial^2\left(g(x_m,\theta)^T g(x_m,\theta)\right)}{\partial\theta\partial\theta^T}\right)^{-1}\left(\frac{\partial\left(g(x_m,\theta)^T g(x_m,\theta)\right)}{\partial\theta}-\frac{\partial\left(g(x_m,\hat{\theta})^T g(x_m,\hat{\theta})\right)}{\partial\theta}\right)$$

When including irrelevant terms, with 0 expected value:

$$\sum_{i=1}^{m}\frac{\partial g(x_m,\hat{\theta})}{\partial\theta}=u^*$$

$$where\ expected\ value\ of\ u^*=0, distribution\ and\ variance\ are\ unknown$$

Similar with before:

$$\left(\frac{\partial\left(g(x_m,\theta)^T g(x_m,\theta)\right)}{\partial\theta}-\frac{2}{m^2}u^*\sum_{i=1}^{m}g(x_m,\hat{\theta})\right)\xrightarrow{p}-\frac{2}{m^2}\left(u^*\sum_{i=1}^{m}g(x_m,\hat{\theta})\right)$$

As $\hat{\theta}$ is different from θ, $\sum_{i=1}^{m}g(x_m,\hat{\theta})\neq 0$. By *Continuous Mapping Theorem*, it follows that,

$$-\frac{2}{m^2}\mathrm{E}\left(u^*\sum_{i=1}^{m}g(x_m,\hat{\theta})\right)=0$$

Asymptotically, $\left(\frac{\partial^2\left(g(x_m,\theta)^T g(x_m,\theta)\right)}{\partial\theta\partial\theta^T}\right)^{-1}$ would converge to a fixed matrix. Hence,

$$\mathrm{E}(\hat{\theta}-\theta)\xrightarrow{p}0$$

Hence, it is still asymptotically unbiased. However, the variance is:

$$var(\hat{\theta}-\theta)=var\left(\left(\frac{\partial^2\left(g(x_m,\theta)^T g(x_m,\theta)\right)}{\partial\theta\partial\theta^T}\right)^{-1}\left(\frac{\partial\left(g(x_m,\theta)^T g(x_m,\theta)\right)}{\partial\theta}-\frac{2}{m^2}u^*\sum_{i=1}^{m}g(x_m,\hat{\theta})\right)\right)$$

Due to non-negativity of variance of $\left(\frac{2}{m^2}u^* \sum_{i=1}^{m} g\left(x_m, \hat{\theta}\right)\right)$:

$$var\left(\left(\frac{\partial^2\left(g(x_m,\theta\,)^T g(x_m,\theta\,)\right)}{\partial\theta\partial\theta^T}\right)^{-1}\left(\frac{\partial\left(g(x_m,\theta)^T g(x_m,\theta)\right)}{\partial\theta}-\frac{2}{m^2}u^*\sum_{i=1}^{m} g\left(x_m,\hat{\theta}\right)\right)\right)$$

$$> var\left(\frac{\partial^2\left(g(x_m,\theta\,)^T g(x_m,\theta\,)\right)}{\partial\theta\partial\theta^T}\right)^{-1}\left(\frac{\partial\left(g(x_m,\theta)^T g(x_m,\theta)\right)}{\partial\theta}\right)$$

unless variance of $\left(\frac{2}{m^2}u^* \sum_{i=1}^{m} g\left(x_m, \hat{\theta}\right)\right)$ is 0.

Hence, $\sum_{i=1}^{m}\frac{\partial g(x_m,\hat{\theta})}{\partial\theta} = 0$ is necessary and sufficient condition of asymptotical *MVUE*.

The proving process of this lemma is analogous to proof of *BLUE* in *OLS*, where we minimize a function (variance of $\hat{\theta}$), subjecting to a constraint (unbiasedness).

Corollary 3.2 Include statistically insignificant estimators is equivalent to$\sum_{i=1}^{m}\frac{\partial g(x_m,\hat{\theta})}{\partial\theta} = u^*$.

When including statistically insignificant estimators, fitting function is $\mathrm{E}[g^*(x^*,\theta^*)] = 0$. Set

$$\frac{\partial\left(g^*\left(x^*,\widehat{\theta^*}\right)^T g^*\left(x^*,\widehat{\theta^*}\right)\right)}{\partial\theta^*} = 2\left[\frac{1}{m^*}\sum_{i=1}^{m}\frac{\partial g^*\left(x^*,\widehat{\theta^*}\right)}{\partial\theta^*}\right]^T\left[\frac{1}{m^*}\sum_{i=1}^{m} g^*\left(x^*,\widehat{\theta^*}\right)\right] = 0$$

As $\widehat{\theta^*}$ is different from θ^*, $\sum_{i=1}^{m} g^*\left(x^*,\widehat{\theta^*}\right) \neq 0$. Hence, when $\frac{\partial\left(g^*\left(x^*,\widehat{\theta^*}\right)^T g^*\left(x^*,\widehat{\theta^*}\right)\right)}{\partial\theta^*} = 0$, it follows that:

$$\sum_{i=1}^{m}\frac{\partial g^*\left(x^*,\widehat{\theta^*}\right)}{\partial\theta^*} = 0$$

Now consider $\mathrm{d}\left(g^*(x^*,\theta^*) - g(x_m,\theta)\right) = -u^*\,\partial\theta$. Then,

$$\left[\sum_{i=1}^{m}\frac{\partial g^*\left(x^*,\widehat{\theta^*}\right)}{\partial\theta^*}\right]^T\partial\theta^* - \left[\sum_{i=1}^{m}\frac{\partial g\left(x_m,\hat{\theta}\right)}{\partial\theta}\right]^T\partial\theta = -u^*\,\partial\theta$$

$$\sum_{i=1}^{m}\frac{\partial g^*\left(x^*,\widehat{\theta^*}\right)}{\partial\theta^*} = \left[\left(\left[\sum_{i=1}^{m}\frac{\partial g\left(x_m,\hat{\theta}\right)}{\partial\theta}\right]^T\partial\theta - u^*\,\partial\theta\right)(\partial\theta^*)^{-1}\right]^T = 0$$

As $(\partial\theta^*)^{-1} \neq 0, \partial\theta \neq 0,$ it follows that,

$$\sum_{i=1}^{m}\frac{\partial g\left(x_m,\hat{\theta}\right)}{\partial\theta} = u^*$$

Statistical insignificant estimators are, thus, harmful. Overfitting is a special case, instead of the worst case, of statistical insignificancy, because overfitting requires statistically insignificant estimators must appear in orthogonal terms. When overfitting happens, predictions are still unbiased; while in other cases, including statistically insignificant estimators makes predictions biased. Excluding statistical insignificancy from regression is important.

Corollary 3.3 From *Lemma 3.1*, *Lemma 3.2,* and *Central Limit Theorem (CLT)*, averages of decentralized parameters in *MVUE* are normally distributed:

$$\sqrt{m}\left(\hat{\theta}-\theta\right)\overset{d}{\rightarrow}N\left(0,\left(\mathrm{E}\left(\frac{\partial g(x_m,\theta)}{\partial\theta}\right)^{T}\left(\mathrm{E}(g(x_m,\theta)g(x_m,\theta)^{T})\right)^{-1}\mathrm{E}\left(\frac{\partial g(x_m,\theta)}{\partial\theta}\right)\right)^{-1}\right)$$

Corollary 3.4 Estimations, including *OLS, MM, MLE* can be considered as special cases of *MVUE*. When using least squares as criterion function, minimizing criterion function is equivalent to $\frac{\partial\left(g(x_m,\hat{\theta})^{T}g(x_m,\hat{\theta})\right)}{\partial\theta}=0,$ and thus equivalent to $\sum_{i=1}^{m}\frac{\partial g(x_m,\hat{\theta})}{\partial\theta}=0.$ Particularly, *MLE* (and *Quasi-MLE*) are equivalent to maximizing a function, with respects to estimators (thus the first derivative equals to 0), subject to a constraint, also known as score, requiring that sum of probability equals to 1. Equivalently, *MLE* can be understand as *MM* on score.

Corollary 3.5 (Generalized Cramer-Rao lower bound) If distribution or (and) variance of residuals is known, *MVUE* can be simplified. $\mathrm{E}(g(x_m,\theta)g(x_m,\theta)^{T})$ is estimator of variance of residuals. Initial version of *Cramer-Rao lower bound* aims to compare variance of estimators in *Quasi-MLE* and *MLE.* In *MLE*, as distribution is known, $\mathrm{E}(g(x_m,\theta)g(x_m,\theta)^{T})$ can be calculated more easily. In addition, if derivatives of fitting functions, for example in linear regressions, are simple, then *MVUE* can be simplified as well.

3.2 MVUE: Ensemble Algorithm is alternative to gradient descents

Ensemble Algorithm is a technique, that firstly run different regressions (called sub-regression) separately, and then take (weighted) averages of results. When using *Ensemble Algorithm*, we automatically assume that true functions of all sub-regressions are identical. Underlying logics of regressing panel data are related to *Ensemble Algorithm*.

Lemma 3.3 Consider a dataset D. If we directly perform a regression, minimum variance of unbiased estimators is denoted as var_{MVUE}. Now, if we separate dataset D into different parts (sizes of data in each part are different) and perform *Ensemble Algorithm*, minimum variances of estimators obtained from *Ensemble Algorithm* should be equals to var_{MVUE}.

Proof Consider dataset D containing m numbers of data. Randomly separate D into t parts, so that $D = [D_1, D_2 \dots D_t]$. Each part contains $m_1, m_2 \dots m_t$ numbers of data, $m = m_1 + m_2 + \cdots m_t$ In each sub-regression, when performing minimum variance unbiased estimation, by *Corollary 3.3*, estimators in each sub-regressions are normally distributed,

$$\sqrt{m_i}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \left(\mathrm{E}\left(\frac{\partial g(x_m, \theta)}{\partial \theta}\right)^T \left(\mathrm{E}(g(x_m, \theta) g(x_m, \theta)^T)\right)^{-1} \mathrm{E}\left(\frac{\partial g(x_m, \theta)}{\partial \theta}\right)\right)^{-1}\right)$$

$$where\ 1 \leq i \leq t$$

Consider taking a weighted average of estimators of estimators from different parts. Weights must be linear; otherwise, weighted average should be biased. Firstly, suppose that the weights are proportional to $m_1, m_2 \dots m_t$. Weighted averages of estimators follow normal distribution, whose variance is $\frac{1}{m}\left(\mathrm{E}\left(\frac{\partial g(x_m, \theta)}{\partial \theta}\right)^T \mathrm{E}(g(x_m, \theta) g(x_m, \theta)^T)^{-1} \mathrm{E}\left(\frac{\partial g(x_m, \theta)}{\partial \theta}\right)\right)^{-1}$, same as result of *GMM, 2SLS* or *MVUE*. (In GMM, weight matrix ${\phi_i}^{-1}$ is asymptotically proportional to $\frac{1}{m_i}$).

If weights are not proportional to m_i $(1 \leq i \leq t)$, normally distributed terms are simultaneously included. By *Lemma 3.2 and Corollary 3.2*, variances would increase, if terms with 0 expected value is included. Using proof by contradiction, when weights are proportional to size of data, m_i $(1 \leq i \leq t)$ variances of estimators obtained from *Ensemble Algorithm* are minimized, which are equal to var_{MVUE}. Thus, *Ensemble Algorithm* is alternative to directly performing regressions.

3.3 Theorem of Exact-Identification Estimation: eliminating optimization error

Theorem of Exact-Identification Estimation (EIE Estimation) If separate samples into equal parts and in each part fitting function is exact identified by data; then, averages of estimators from each part are normally distributed, variance of which are equals to var_{MVUE}.

Proof: This theorem can be considered as a special case of *Lemma 3.3*, where we separate dataset D into equal parts. Thus, no more proofs are required.

Aside from minimized estimation error, the biggest advantage of *EIE estimation* is to eliminate optimization error, by preventing using gradients. When size of data in each sub-regression is sufficiently small, regressions would converge to equations, as SSE in each sub-regression exactly equals to 0. However, solving equations need to neither calculate gradients nor distinguish global minimum from local minimums. Thus, optimization error is eliminated.

It is possible that, depending on properties of fitting functions, solving the equations prementioned is difficult. Even though, *EIE estimation* is still an instrumental supplement of *gradient descents.* Since value of global minimum of SSE in all sub-regressions are known to be 0, gradients can be designed not to stop until SSE exactly equal to 0. Hence, local minimums are no longer obstacles. Simultaneously, as we don’t only refer to gradients, vanishing gradient problem will not be troublesome.

Generally speaking, easiness to perform optimizations is affected by decision variables, objective function(s) and constraints. Changes in any of the three factors may significantly simplify optimization process. For example, in some cases, allocating constraints into decision variables can make constraints linear, and subsequently make optimization process simpler. In this case, *EIE Estimation* takes advantages of that SSE (global minimum) is non-negative, and optimization process can be significantly simplified.

In summary, *EIE estimation* is a special case, instead of antonym, of *gradient descents*. *EIE estimation* is in the interaction of *GMM, 2SLS* and *Ensemble Algorithm*. Properties of *EIE estimation* are similar to its counterparties, while *EIE estimation* is featured with minimized estimation and optimization error.

3.4 Statistical Inferences are alternative to regularizations: preventing overfittings

As prementioned in *Corollary 3.2*, overfitting is only a special case, instead of the worst case, of including statistically insignificant estimators. For example, in this regression:

$$Y = \alpha_1 X_1 + \alpha_2 {X_2}^{\beta} + \varepsilon$$

If, for example, α_1 is statistically insignificant, (overfitting), predictions are still unbiased. If β is statistically insignificant, predictions must be biased. Meanwhile, even though β is significant, if variance of β is large, predictions will also be impacted non-linearly.

Statistical inference can be performed in all *MVUE* (including *EIE estimation*). It would be instrumental for handling statistical insignificancy and evaluating estimators. From *Corollary 3.3,* estimators are normally distributed. Thus, by performing *t-tests* (as variances are estimated), we may successfully exclude insignificant estimators—both in orthogonal terms and non-orthogonal terms.

Meanwhile, it is possible that biases, resulting from excluding statistically significant estimators, or resulting from early stop, are over-compensated by smaller variances. Even though, statistical inference is still an instrumental supplement. With the help of statistical inference, we can better measure the trade-off between biases and variances, trying to achieve a balance between costs and benefits:

$$\sum(y_i - \hat{y_i})^2 = bias^2 + variance + random\ terms$$

Shallow and deep NN: mitigated approximation error and interpretability

In chapter 2, a fundamental nonparametric regression is established, main weakness of which is convergence rate of approximation. In this chapter, we firstly discuss a special group of function approximations, *GAAM*. Wisely choosing basis functions in *GAAM* can alleviate approximation error. Then, we propose *Theorem of re-structure*, which proves that DNN is a special case of *GAAM* and thus can be interpretable. Adding more hidden layers can be proven to be equivalent to increasing amount of basis functions. Finally, we reinforce *EIE estimation*, in order to prevent multiple global minimums when estimating DNN.

4.1 Generalized Additive Approximation Model (GAAM) is interpretable

Definition 4.1 For a *Generalized Additive Model* $y = f_1(x) + f_2(x) + \cdots f_n(x) + \varepsilon$, if $f_1(x) + f_2(x) + \cdots + f_n(x)$ can be considered as a function approximation, then a *Generalized Additive Model* can be defined as *Generalized Additive Approximation Model*.

One-hidden-layer Neural Network is obviously a special case of *GAAM*. Just like *Decision Trees* has nothing to do with forestry, *Neural Network* should only be considered as one of function approximations.

Corollary 4.1 *GAAM* is a special case of *Generalized Additive Model*. Because *Generalized Additive Models* are all considered as interpretable, *GAAM* must interpretable. From another aspect, if connections among variables are as simple as linear, function approximations are interpretable.

Corollary 4.2 Sufficient but unnecessary condition for a function to be *GAAM* is that:

When $n \gg 0$, for arbitrary 1$\leq i \leq n$, basis functions $f_1(x), f_2(x), \dots f_n(x)$ can approximately be:

$$f_i(x) = \begin{cases} 1 & if\ x = x_i \\ 0 & otherwise \end{cases}$$

This special case of *GAAM* can be defined as *Spike-GAAM*.

Different from its perplexing definition, underlying logic of *Spike-GAAM* is a close relative of *Newton-Leibniz Theorem*. The only difference is that we only take the value of $f(x)$, instead of multiplying $f(x)$ by dx. Hence, outputs of *Spike-GAAM* functions are themselves, instead of integrals of functions.

A special case of *Spike-GAAM* is *Weierstrass First Theorem*. Basis functions of *Bernstein Polynomials* are identical to probability distribution function (pdf) of *binomial distribution,* $\binom{m}{l}x^l(1-x)^{m-l}$. When $m \gg 0$, by properties of binomial distribution, such function obviously satisfies that:

$$f(x) = \begin{cases} 1 & if\ x = x_0 \\ 0 & otherwise \end{cases}$$

Corollary 4.3 In *Spike-GAAM*, when $m \gg 0$, all basis functions are approximately mutually orthogonal. Thus, *Primary Component Analysis* has already been embedded.

4.2 Selecting proper basis function can decrease approximation error

Extension 4.1 When $m \gg 0$, linear combination of sigmoid functions approximately satisfies:

$$\left(1+e^{mx-\frac{1}{m}}\right)^{-1} - \left(1+e^{mx+\frac{1}{m}}\right)^{-1} = \begin{cases} a & if\ x = 0 \\ 0 & otherwise \end{cases}, where\ a\ is\ constant$$

Barron (1996) shown that, when using sigmoid as basis functions, convergence rate of approximation is dependent on reciprocal of number of parameters, m^{-1}, and independent from number of variables, n. Thus, substituting polynomial basis functions in *Bernstein Polynomials* with sigmoid basis functions can mitigate approximation error in multivariable regressions.

By and large, *GAAM* (or one-hidden-layer NN) with sigmoid basis function can be unbiased, consistent, efficient, and interpretable. All goals of nonparametric regression have been achieved at this stage. Deep NN, as what will be showed later, is optional to make improvements.

4.3 Universal Approximation Theorem is a special case of Spike-GAAM

Extension 4.2 Linear combinations of “squash” functions, including sigmoid, when $m \gg 0$, can also approximately satisfies that:

$$f(x) = \begin{cases} a & if\ x = x_0 \\ 0 & otherwise \end{cases}$$

Hence, *Universal Approximation Theorem,* is neither universal nor the answer for everything. This theorem should be considered as comparable with *Weierstrass First Theorem* and *Newton-Leibniz Theorem;* and be considered as a special case of *Spike-GAAM.*

4.4 DNN is as asymptotically interpretable as Decision Trees

Basis functions pervasively used can be classified into three families: polynomial family (*Weierstrass First* and *Second Theorem*); exponential family (*sigmoid, SoftMax, tanh*); piecewise family (*ReLU, MAXOUT, (linear)Taylor's expansion*).

These three families, however, are not distinctly difference with each other: they are all subsets of *Spike-GAAM.* When $m \gg 0$, all basis functions from these three families satisfy that:

$$f(x) = \begin{cases} a & if\ x = x_0 \\ 0 & otherwise \end{cases}$$

Lemma 4.1 *One-hidden-layer Neural Network* with *ReLU* is equivalent to *Decision Tree*.

Proof This statement can be easily proved by definitions. *Decision Tree* is a metaphor, whose correct name should be *"Binary Regression"*, where variables can be binary (dummy). Basis functions, *ReLU*, are also binary. In addition, connections among variables are additive in *Binary Regression*, as well as in *One-hidden-layer Neural Network.* Hence, these two models are equal.

Lemma 4.2 *Multiple-hidden-layer Neural Network* with *ReLU* can be decomposed into a *one-hidden-layer Neural Network* with more basis functions.

Proof Since *one-hidden-layer NN* with *ReLU* is equivalent to *Decision Tree,* it can be written as:

```
$if\ x > x_0$
  $then\ y = y_0$
  $else\ y = 0$
```

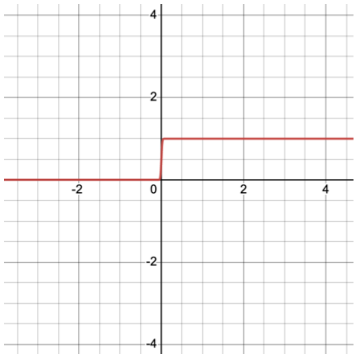
Subsequently, *two-hidden-layer NN* with *ReLU* can be written as:

```
$if\ x > x_0$
  $then\ y = y_0$
    $if\ y > y_1$
    $then\ y = y_2$
  $else\ y = y_3$
```

In summary, *ReLU* of *ReLU* is still *ReLU*.

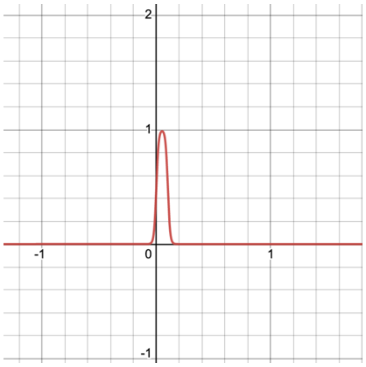
Theorem 4.1 (Theorem of Restructure) When $m \gg 0$, *ANN*, including *DNN*, with *squash* basis functions, can be asymptotically equivalent to *Decision Trees*. Thus, *DNN* can be transformed into *one-hidden-layer NN* with more basis functions, which is underlying reason for lower approximation error.

Proof *Squash* basis functions, defined with "S" shapes, can asymptotically be considered as linear combination of two *ReLU*. For example, for $y = \frac{e^{100x}}{1+e^{100x}}$:



Identical to *Extension 4.2*, two squash functions can be considered as a *spike*. For example, for

$y = \frac{e^{100x}}{1+e^{100x}} - \frac{e^{(100x-10)}}{1+e^{(100x-10)}}$:



From *Lemma 4.2*, When $m \gg 0$, *ANN* and *DNN* with *squash* basis functions can be considered as asymptotically equivalent to *Decision Trees*.

Corollary 4.4 In *Theorem 4.1*, asymptotical equivalence is proved. In another word, *squashed* basis function is not equivalent to combinations of *ReLU*. Hence, *squashed* basis function is differentiable and thus *EIE estimation* can be deployed.

Extension 4.3 Linear combination of *ReLU* can be estimated in another way, other than using *gradients* or *EIE estimation*. *Local regression* is also aimed to decompose true function into a linear combination of lines. Underlying logic of *local regression* is that, within small changes of input x, by *Taylor's Expansion*, changes in output y is approximately linear. Hence, forms of fitting functions (combination of lines) become similar with forms of true functions. Such process is similar, when using *ANN* with *MAXOUT*. Hence, linear combination of *ReLU* (or $MAXOUT$) can be estimated and interpreted similar as *local regression.*

Extension 4.4 Performing nonlinear transformation before function approximation, other than *squashed* functions, may also work. Recall *Weierstrass Second Theorem*, trigonometric transformations are performed before function approximations. Since nonlinear transformation, if uniformly continuous, can be approximated by linear combinations of *squashed* functions. Notice that convolution is typically one of *smooth* nonlinear transformations. Hence, it may be possible that *Convolutional NN* can be interpreted in similar way.

Extension 4.5 *Recurrent NN* can be interpreted similarly, under the background of stochastic processes. In parametric Econometrics, recurrency is typically used to capture persistency. For example, *GARCH* is recurrent, where high volatility in the past (inputs) is likely, but not necessarily, to bring high current volatility (output).

Corollary 4.6 Full-connected DNN, after re-structure, is a special case of one-hidden-layer NN (and *GAAM*). According to *corollary 4.1*, full-connected DNN is interpretable. From another perspective, full connected *DNN* is a special case of *Decision Tree*, thus, it is interpretable.

Thus, *DNN* is optional in many cases in econometrics. If preferred, we may also add more basis functions manually in *one-hidden-layer NN*. As what will be discussed in the following contents, one of advantages of *one-hidden-layer NN* is easiness to be estimated.

4.5 Re-structured EIE: prevent multiple global minimums in DNN

For example, consider that:

$$\begin{aligned}\big(F(x)\big) &= (a_0 + a_1x^1 + a_2x^2 + \cdots + a_mx^m) + (a_0 + a_1x^1 + a_2x^2 + \cdots + a_mx^m)^2 \\ &\quad + (a_0 + a_1x^1 + a_2x^2 + \cdots + a_mx^m)^3 \ldots + (a_0 + a_1x^1 + a_2x^2 + \cdots + a_mx^m)^m \\ &= b_0 + b_1x^1 + b_2x^2 + \cdots + b_{m^2}x^{m^2}\end{aligned}$$

Obviously, parameters in new basis functions are different from those in initial basis functions. In *DNN, SSE*, with respect to parameters before re-structuring, may have multiple global minimums, regardless of using *EIE* or *gradient descents*. Directly estimating *one-hidden-layer NN,* $F\big(F(x)\big) = b_0 + b_1x^1 + b_2x^2 + \cdots + b_{m^2}x^{m^2}$, can prevent multiple global minimums.

In this example, $F\big(F(x)\big) = b_0 + b_1x^1 + b_2x^2 + \cdots + b_{m^2}x^{m^2}$ is approximate to the true functions, where parameters are exactly $(b_0, b_1, b_2, \ldots b_{m^2})$. Equivalently, by *Corollary 3.1*, a set of parameters, $(b_0, b_1, b_2, \ldots b_{m^2})$, minimizes SSE to the global minimum point, u. However, $b_0, b_1, b_2, \ldots b_{m^2}$, can be considered as functions, with respect with $a_0, a_1, a_2, \ldots a_m$. The function set:

$$\begin{bmatrix} b_0 \\ b_1 \\ \cdots \\ b_{m^2} \end{bmatrix} = \begin{bmatrix} f_0(a_0, a_1, a_2, \ldots a_m) \\ f_1(a_0, a_1, a_2, \ldots a_m) \\ \cdots \\ f_{m^2}(a_0, a_1, a_2, \ldots a_m) \end{bmatrix}$$

is not guaranteed to have a unique solution. Hence, when selecting proper basis functions, it is possible that only one set of parameters in *one-hidden-layer NN* can successfully minimize SSE to u.

4.6 Revisit GAAM: nonlinear connections between basis function may not be necessary

We can understand how different function approximations are related from another perspective. Consider a *one-hidden-layer NN*, with sigmoid basis functions.

Ignoring approximation errors, true functions can thus be regarded as a linear combination of sigmoid functions. Because sigmoid is uniformly continuous; subsequently, it can be approximate to algebraic polynomials. In this case, sigmoid can be considered as “betweenness”. Likewise, true functions can be decomposed into linear or non-linear combinations of different “betweenness”. Each “betweenness”, if uniformly continuous, can be approximated by linear combinations of polynomial basis function.

Equivalently, using different ways to decompose true functions may be possible to improve approximation accuracy. However, such improvements could be trivial.

Corollary 4.6 Uniformly continuous true functions can be “decomposed” into either identical basis functions or other different functions, connections among which can be either linear or nonlinear. It is possible that approximation error and be further diminished when connections are nonlinear; however, considering interpretability, it may not be necessary to try other connections.

Corollary 4.7 Considering *Theorem 4.1*, uniformly continuous true functions can be transformed into either identical basis functions or other different functions, into either single-hidden-layer or multiple-hidden-layers.

Conclusion

As we don't assume forms of true functions in nonparametric regressions, they are more flexible than parametric counterparties. Albeit flexibility, nonparametric regressions, including Neural Networks, are established on strict assumptions. It should neither be abused and misused, nor be considered as answer for everything.

Since connections among basis functions are pre-established, when parameters are known (or estimated), function approximations are at least implicitly known. Complexity, of both connections among basis functions and calculations within basis functions, affect easiness to interpret function approximations.

Estimation, optimization, and approximation error are main barriers in nonparametric regressions. Estimating process of nonparametric and parametric regressions are comparable. With the help of *MVUE standard*, *EIE estimation* and statistical inferences, estimation and optimization error can be proven to be minimized. Approximation error can also be alleviated, by choosing suitable basis functions and by adding more basis functions.

Weierstrass First Theorem is the root of other complex function approximations, including *deep Neural Networks*. Same as *one-hidden-layer NN*, *deep Neural Networks* are interpretable and equivalent to a linear combination of basis functions. Adding more hidden layers equals to increasing numbers of basis functions; however, we may also add basis functions manually in *shallow Neural Networks*. We may interchangeably use *shallow Neural Networks* and *deep Neural Networks*.