ESSAYS IN ECONOMETRICS AND FINANCE

Xiaoying Lan

A dissertation

submitted to the Faculty of

the Department of Economics

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Boston College

Morrissey College of Arts and Sciences

Graduate School

July 2022

ESSAYS IN ECONOMETRICS AND FINANCE

Xiaoying Lan

Dissertation Committee:

Professor Shakeeb Khan (Co-Chair)

Professor Zhijie Xiao (Co-Chair)

Professor Arthur Lewbel

Binary choice models can be easily estimated (using, e.g. maximum likelihood estimation) when the distribution of the latent error is known, as in Logit or Probit. In contrast, most estimators with unknown error distribution (e.g., maximum score, maximum rank correlation, or Klein-Spady) are computationally difficult or numerically unstable, making estimation impractical with more than a few regressors.

The first chapter proposes an estimator that is convex at each iteration, and so is numerically well behaved even with many regressors and large sample sizes. The proposed estimator, which is root-n consistent and asymptotically normal, is based on batch gradient descent, while using a sieve to estimate the unknown error distribution function. Simulations show that the estimator has lower mean bias and root mean squared error than Klein-Spady estimator. It also requires less time to compute.

The second chapter discusses the same estimator in high dimensional setting. The estimator is consistent with rate lower than root-n when the number of regressors grows slower than the number of observations and asymptotic normal when the square of the number of regressors grows slower than the number of observations. Both theory and simulation show that higher learning rate is needed with higher number of regressors.

The third chapter provides an application of the proposed estimator to bankruptcy prediction. With more than 20 regressors, the proposed estimator performs better

than logistic regression in terms of Area Under the Receiver Operating Characteristics using firm data one year or two years prior to bankruptcy, but worse than logistic regression using firm data three years prior to bankruptcy.

Contents

List of Tables

List of Figures

Acknowledgement

I am grateful to my dissertation advisors, Professor Shakeeb Khan, Professor Zhijie Xiao and Professor Arthur Lewbel for their constant guidance, help and patience. Thank you to Professor Shakeeb Khan, for your endless support and exceptional words of wisdom. My research was greatly enriched and dramatically improved by your immense knowledge. Thank you to Professor Zhijie Xiao, for your unerring suggestions and ceaseless support from my graduate study in China to the completion of doctoral studies in Boston College. Thank you to Professor Arthur Lewbel, for your insightful comments and invaluable advice for my dissertation.

Thank you to my friends, Jiayang Hu, Xiang Zheng and Praveen Saini for your supportive friendship.

Thank you to Licheng Liu for your unwavering encouragement.

Thank you to my parents for your unconditional love. I love you always!

To my parents

Chapter 1

Estimation and Inference of Semiparametric Binary Choice Model

1.1 Introduction

Binary choice models are widely used in empirical economics. Examples are modeling the choice of who to vote for, or whether to buy a product or not. Logit and Probit models are commonly used in empirical applications, but they impose strong, rarely justified functional form restrictions on the model's error distribution.

To deal with this drawback, many methods have been developed to specify and estimate binary choice models that do not impose these error functional form restrictions. However, these alternative methods are not widely use in practice, because they tend to be computationally complex and numerically poorly behaved, or they require additional strong restrictions on the regressors. These problems have become more acute in recent years, as data sets and models have grown larger, with many regressors.

The goal of this paper is to provide a binary choice model estimator that doesn't impose a functional form on the errors, and doesn't require strong restrictions on the regressors, but is still computationally easy and numerically well behaved, even with big data.

The standard binary choice model says that y_i equals one if $x_i^T\beta^* > \epsilon_i$ and zero otherwise, where x_i is a p-vector of regressors, β^* is a p-vector of coefficients and the random variable ϵ is an unobserved latent error term. Formally, this model is $y_i = \mathbb{1}\{x_i^T\beta^* > \epsilon_i\}$, where $\mathbb{1}$ is the indicator function that equals one if its argument is true and zero otherwise. Our goal is estimation of the coefficients β^* from a set of independent, identically distributed (i.i.d.) observations (y_i, x_i^T).

This paper proposes an iterative estimator based on the batch gradient descent (BGD) algorithm, and its asymptotic properties, including convergence rate and limiting distribution, are derived.

Let g denotes the unknown cumulative distribution function of each ϵ. The standard BGD algorithm requires that the error distribution g be known. To allow for an unknown distribution, we use a sieve method to approximate this distribution. First we apply BGD algorithm to estimate β for a given choice of g. Then, using that estimate of β, we apply a sieve estimator, the Series Logit Estimator (SLE), to get an estimate of the function g. This procedure is then iterated many times until the estimates of β and g converge. Each step in this process is computationally easy and numerically very well behaved, because the underlying objective functions at each stage are convex.

The resulting estimator is shown to be $\sqrt{n}$-consistent and asymptotically normal, with a limiting distribution that can be calculated, to allow for inference.

The estimator described above is for a fixed number of parameters p. We next consider a high dimensional setting, where p goes to infinity as n goes to infinity. The resulting estimator is shown to have $\sqrt{p/n}$ consistency as $p/n \to 0$, and is asymptotically normal under the stronger condition that $p^2/n \to 0$. These results suggest that the proposed estimator can be used in big data settings with many covariates.

Some Monte Carlo analyses are performed to assess the finite sample properties of the proposed estimators, and to compare with other existing estimators in the literature.

1.2 Literature Review

1.2.1 Literature of Semiparametric Estimators

When the distribution of error is known, maximum likelihood estimation (MLE) is widely used, i.e, find the value of β that gives the highest value of a likelihood function. One of the advantages of MLE is that the estimator touches CramérRao lower bound, the lowest variance an unbiased estimator can get. When the distribution of error is unknown, MLE leads to an inconsistent estimator. White (1982) finds that MLE is consistent to a well defined limit, which may not be the true value under misspecification.

With some specific assumptions on error distributions one can get MLE that converges to $\gamma\beta$, where γ is unknown, non-zero scalar, i.e, the estimator is consistent up to a scaling factor. Ruud (1983) finds that MLE is still consistent if the expectation of each regressor conditional on $x_i^T\beta$ is linear in $x_i^T\beta$.

Semiparametric binary choice model was first introduced by Manski (1975). He proposes maximum score estimator, the basic assumption is $median(x_i|\epsilon_i) = 0$, which is much weaker than the assumption of Ruud (1983). The MS estimator is the following:

$$\hat{\beta}_{MS} = argmax_{||\beta||=1}\frac{1}{n}\sum_i^n y_i\mathbb{1}[x_i^T\beta^* > 0] + (1-y_i)\mathbb{1}[x_i^T\beta^* < 0]$$

The convergent rate of MS estimator is $\frac{1}{n^{1/3}}$, which is slower than $\frac{1}{\sqrt{n}}$. In addition, the limiting distribution is complex, which makes statistics inference difficult. To address this problem, Horowitz (1992) proposes the smoothed maximum score (SMS) estimator which has a better performance than MS estimator in terms of convergence rate and asymptotic variance. SMS estimator is the following:

$$\hat{\beta}_{SMS} = argmax_{||\beta||=1}\frac{1}{n}\sum_{i}^{n}(2y_i - 1)K(\frac{x_i^T\beta^*}{h_n})$$

where $K(.)$ is kernel function and h_n is a bandwidth parameter satisfying $h_n n \to \infty$. SMS estimator is at least $\frac{1}{n^{2/5}}$ and asymptotic normal. Both MS estimator and SMS estimator converge slower than $\frac{1}{\sqrt{n}}$.

Han (1987) proposes maximum rank correlation estimator. The basic assumption is that ϵ_i is independent of x_i, which means $\epsilon_i - \epsilon_j$ is independent of x_i and x_j. The MRC estimator is the following:

$$\hat{\beta}_{MRC} = argmax\sum_{i\neq j}\mathbb{1}[y_i > y_j]\mathbb{1}[x_i^T\beta^* > x_j^T\beta^*]$$

As proved by Sherman (1993), $\hat{\beta}_{MRC}$ converges to β with the convergence rate of $\frac{1}{\sqrt{n}}$.

Cosslett (1983) proposes an estimator based on MLE. Their method include two parts, first they approximate the distribution of error using basic distribution functions. Secondly, they estimate β^* via MLE and repeat the process until converge. Ichimura (1987) proposes semiparametric least square (SLS) estimator for single index model, where he uses kernel estimator to approximate the error distribution. It also has an asymptotic normal distribution with rate $\frac{1}{\sqrt{n}}$.

However, the above estimators involve finding the maximum of a non-concave maximum likelihood functions or other functions. This is computationally hard when we use methods like grid search to find the maximum. With more than 4 regressors it's almost impossible to implement those method in practice. Some methods help relieve the problem of computation. E.g, the objective function of MRC estimator is neither globally concave nor smooth, therefore traditional methods like NewtonRaphson algorithm and NelderMead algorithm can't be applied to it. Wang (2007) proposes iterative marginal optimization (IMO) to estimate MRC, which updates covariates one by one. IMO is stable since it guarantees monotonic increase of MRC objective function in each iteration, but it sill requires grid search in their algorithm, with $O(n^2logn)$ operations for each grid search step. The objective function of our estimator is globally convex. As a result, our estimator is computationally easy compared to the above estimators.

Some estimators are computationally easy, but the first stage suffers from the curse of dimensionality. Powell, Stock and Stoker (1989) proposes Weighted average derivative estimator for index model. However, it requires kernel estimation of joint density function of all regressors. As a result, it suffers from curse of dimensionality. Besides, it requires all regressors to be continuous. Ahn et al. (2018) propose a computationally easy estimator, they match observations with same expected value of $x_i^T\beta$ but different value of x_i. However, the first stage is still a kernel estimation of error distribution. In addition, the estimator is not robust to heteroscedasticity and discrete regressors.(see Khan and Tamer (2018)). We estimate the error distribution based on $x_i^T\beta$ rather than all the regressors. This makes our estimator free from the curse of dimensionality and we can get error distribution while estimating β^*.

Some estimators add more assumptions about error distribution to gain computational efficiency. Lewbel et al. (2012) finds that the estimator calculated by special regressors method is $\sqrt{n}$-consistent and asymptotically normal. However, it requires a very thick tailed regressor or bounded error support. Dominitz and Sherman (2005) proposes the iterative least square estimator (ILS) based on Klein and Spady (1993):

$$\begin{aligned}\hat{\beta}_k &= argmax_\beta \sum_{i=1}^{n}(\hat{y}_i(\hat{\beta}_{k-1}) - x_i^T\beta)^2 \\ &= \hat{\beta}_{k-1} - x_k'\hat{u}_k(\hat{\beta}_{k-1})\end{aligned}$$

ILS estimator is very easy to compute but requires error distribution to be log-concave, which excludes some common distributions like Cauchy distribution. Besides, one of the tuning parameters that controls the tail of error distribution is very sensitive to estimation. Our estimator does not put any shape restrictions on error distribution, which means that our method can be applied more widely than special regressor method and ILS.

In computer science literature, iterative algorithm is widely used to estimate β. Kalai and Sastry (2009) use monotonic regression to estimate error terms and an algorithm to update β. Their method is simple and fast in programming and achieve linear complexity, but they focus mainly on converge of error distribution and don't prove the consistency of β. Agarwal et al. (2013) propose an estimator based on Kalai and Sastry (2009). They proved consistency but the estimator require the underlying distribution function is known. Our estimator share the same objective function as Agarwal et al. (2013), which means that we obtain linear complexity while updating β. What's more, we prove our estimator is $\sqrt{n}$-consistent and asymptotically normal.

1.2.2 Literature of the Technical Part of the Estimator

The estimator is related to three kind of literature. The first kind talks about the convex objective function. Agarwal et al. (2013) propose the convex objective function to get β^*. We use the same function as theirs. The objective function is also implied or mentioned by Kalai and Sastry (2009) and Ravikumar, Wainwright and Yu (2008).

Secondly, our estimator is related to gradient descend method, which is widely used in machine learning literature (see Mustapha, Mohamed and Ali (2020),Ruder (2016)). One variation is stochastic gradient descent (SGD) algorithm, which updates β^* use one data point in each iteration, therefore the total number of iteration times is n. The SGD estimator is easy to compute since the algorithm of updating β^* is linear if the objective function is convex. It is one type of NewtonRaphson estimator and a special case of stochastic approximation method of Robbins and Monro (1951). SGD algorithm usually requires the learning parameter to shrink to 0 as iterate times goes to infinity. Polyak and Juditsky (1992) proposes SGD average algorithm which averages β^* across each iteration and gain efficiency than the basic SGD algorithm. See Kushner and Yin (2003) for more details about the variations of SGD algorithm. Another variation is batch gradient descent (BGD) algorithm, which use all data points in each iteration until converge. See Wilson and Martinez (2003) and Hinton, Srivastava and Swersky (2012) for more discussion on the differences and combinations of the two algorithm. Our algorithm uses BGD and then averages β^* across each iteration.

At last, our estimator uses method of sieve to estimate the unknown distribution. The method of sieve is first proposed by Grenander (1981). It uses a sequence of finite-dimensional spaces, which is called sieve, to approximate unknown infinite-dimensional space. The complexity of sieves should increase with the number of observations and the sieves should be dense in the unknown space. We use Series Logit Estimator(SLE), which is used in Hirano, Imbens and Ridder (2003) to estimate propensity score. It is a special case of sieve MLE proposed by Geman and Hwang (1982), they prove the consistency of sieve MLE with i.i.d data. As for dependent and heterogeneous data, White (1991) provides a more detailed analysis. Hirano, Imbens and Ridder (2003) use logistic model with power series called series Logit estimator (SLE). They only require some smoothness properties of the unknown distribution. Our estimator is similar to two-step sieve estimator. Two-step sieve estimator starts with unknown function nonparametrically and then estimates the parametric part with GMM or MLE and other methods. Under some regularity conditions, the parametric part of two-step sieve estimator can get $\sqrt{n}-$asymptotic normality, see Chen (2007), Chen, Linton and Van Keilegom (2003) for more discussion. As for the nonparametric part of sieve estimator, like Chen (2007) point out, we don't have a universal theory of a pointwise limiting distribution.

1.3 Model

In this section we show the algorithm, the assumptions and the theorem.

$$y_i = \mathbb{1}\{x_i^T\beta^* > \epsilon\} \tag{1.1}$$

x_i is a p-vector of regressors, β^* is a p-vector of coefficients $(\beta^{*1}, \beta^{*2}...\beta^{*p})$, $\mathbb{1}$ is an indicator function and ϵ is a random variable. We make one condition on the distribution of error that is l-Lipschitz condition: $0 \leq g(b) - g(a) \leq l * (b - a)$ for all $a \leq b$, where $g : \mathbb{R} \to \mathbb{R}$ is the cumulative distribution function (CDF) of ϵ. We want to estimate β^* from i.i.d. data points (y_i, x_i^T). We assume $\beta^{*1} = 1$, therefore we only update and estimate $p - 1$ coefficients. Here we assume all the regressors are not a constant.

Remark 1. *Here it means we don't estimate the location of error distribution because we want to compare estimator with known distribution and estimator with unknown distribution. The estimator of location coefficient may have different convergence rate with unknown distribution.*

1.3.1 Estimator with Known g

First we introduce the convex objective function proposed by Agarwal et al. (2013):

$$\zeta(\beta; (x_i, y_i)) = G(x_i^T \beta) - y x_i^T \beta \tag{1.2}$$

There exists a convex function G such that $G' = g$ if g is monotone increasing function and satisfies l-Lipschitz condition according to Lemma 1. Notice that the loss function is convex since G is convex.

Secondly, we introduce the batch gradient descent algorithm(BGD), which uses all the data points at each iteration. The gradient of $\zeta(.)$ is the following:

$$\nabla \zeta(\beta; (x_i, y_i)) = (g(x_i^T \beta) - y_i) x_i^T \tag{1.3}$$

We will use all the data points to calculate the average of gradient in each iteration :

$$\frac{1}{n} \sum_{i=1}^{n} \nabla \zeta(\beta; (x_i, y_i)) = \frac{1}{n} \sum_{i=1}^{n} (g(x_i^T \beta) - y_i) x_i^T \tag{1.4}$$

The following is BGD algorithm:

Algorithm 1 BGD algorithm: k denotes the iterate times. The total number of iteration is K. C_k is a fixed $p * p$ positive-denite matrix. γ_k is learning speed depended on k.

1: Guess $\hat{\beta}_0$.
2: Iterate $\hat{\beta}_k = \hat{\beta}_{k-1} - \gamma_k C_k(\frac{1}{n}\sum_{i=1}^n (g(x_i^T\beta) - y_i)x_i^T)$ until you get $\hat{\beta}_K$.

At last we get BGD average (BA) estimator $\hat{\beta}_{BA}$ by averaging $\hat{\beta}_k$ across different k and let $K = n$:

$$\hat{\beta}_{BA} = \frac{1}{n}\sum_{k=1}^{n}\hat{\beta}_k \tag{1.5}$$

Remark 2. *BGD estimator usually requires less iterate times than n. We follow the requirements for SGD by letting $K = n$ and averaging $\hat{\beta}_k$ across different k for two reasons. Firstly, it's easy to prove under such assumptions. Secondly, we follow the same assumptions here as the ones in next section so that we can compare the limiting distribution of $\hat{\beta}_{BA}$ with the limiting distribution of estimator with unknown distribution.*

We follow the assumptions by Toulis, Airoldi et al. (2017).

Assumption 1. $\{\gamma_k\} = \gamma_1 k^{-\gamma}$, *where $\gamma_1 > 1$ is the learning parameter, $\gamma \in (0.5, 1]$.*

Assumption 2. *function $g(.)$ satisfies l-Lipschitz conditions, i.e, $0 \leq g(b) - g(a) \leq l*(b-a)$ and g(.) is non-decreasing and differentiable almost surely.*

Assumption 3. *The matrix $\hat{I}_i(\beta) \equiv g'(x_i\beta)x_ix_i^T$ has nonvanishing trace, that is , there exists constant $b > 0$ such that trace($\hat{I}_i(\beta)$) $\geq b$ almost surely, for all β. The matrix $I(\beta^*) = E(\hat{I}_i(\beta^*))$, has minimum eigenvalue $\underline{\lambda}_f > 0$ and maximum eigenvalue $\overline{\lambda}^f < \infty$. Typical regularity conditions holds.(Lehmann and Casella (2006), Theorem 5.1,page 463).*

Assumption 4. *C_k is a fixed positive-definite matrix, such that $C_k = C + O(\gamma_n)$, where$||C|| = 1$, $C \succ 0$ and symmetric, and C commutes with $I(\beta)$. Every C_k has a greatest eigenvalue $\overline{\lambda}_c$ and smallest eigenvalue $\underline{\lambda}_c$.*

Assumption 5. $\frac{1}{n}\sum_{i}^{n} x_i x_i^T$ *converges to a symmetric positive-definite matrix.*

Remark 3. *Assumption 1 guarantees that* $\sum_i \gamma_i = \infty$ *and* $\sum_i \gamma_i^2 < \infty$ *as mentioned by Robbins and Monro (1951), which is a necessary condition for the converge of SGD estimator. Assumption 2 means that* $G(.)$ *is Lipschitz-continuous following the traditional optimization literature (see Nesterov (2003)). In assumption 3, the matrix* $I(\beta^*)$ *has minimum and maximum eigenvalue is equivalent to strong convexity condition. Assumption 5 guarantees the use of central limit theorem, which can be relaxed to allow non i.i.d. data.*

Theorem 1. *Under assumptions 1-5 and for* $k \leq n$*, use BGD algorithm 1 we get*

$$\mathbb{E}||\hat{\beta}_k - \beta^*||^2 \leq \frac{8\overline{\lambda}_c^2 \sigma_x^2 C_1(1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f)}{2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f} k^{-\gamma}$$
$$+ exp(-log(1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f)\phi(k))[||\beta_0 - \beta^*|| + (1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f)^{n_1} A]$$

with k *sufficiently large, where* $A = 4\overline{\lambda}_c^2 \sum_i \gamma_i^2 < \infty$ *and* $\phi(k) = k^{1-\gamma}$ *if* $\gamma \in (0.5, 1]$ *and* $\phi(k) = logk$ *if* $\gamma = 1$*.* C_1 *and* n_1 *are some constants.*

When $K = n$ and $\gamma = 1$, $\hat{\beta}_K$ is consistent to β^* at the rate of $\frac{1}{\sqrt{n}}$.

$\hat{\beta}_k$ exhibits the same convergency rate as traditional stochastic gradient descent estimator, which uses single data point once per iteration. $\hat{\beta}_k$ is robust to initial condition since the g is bounded, see Moulines and Bach (2011) for bounded gradient discussion. For unbounded g function, there will be extra term that grows exponentially with the value of initial point.

Theorem 2. *Under assumptions 1-5 and for* $\gamma \in (0.5, 1)$*, we get*

$$(i) \quad \sqrt{n}(\hat{\beta}_{BA} - \beta^*) \to N(0, \Sigma_2^{-1}\Sigma_1\Sigma_2^{-1})$$

where $\Sigma_1 = \mathbb{E}g(x_i^T\beta^*)(1 - g(x_i^T\beta^*))x_i x_i^T$ *and* $\Sigma_2 = \mathbb{E}g'(x_i^T\beta^*)x_i x_i^T$*.*

$$(ii) \quad \hat{\Sigma}_2^{-1}\hat{\Sigma}_1\hat{\Sigma}_2^{-1} \to \Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}$$

where $\hat{\Sigma}_1 = \frac{1}{n}\sum_i^n g(x_i^T\hat{\beta}_{BA})(1 - g(x_i^T\hat{\beta}_{BA}))x_i x_i^T$ *and* $\hat{\Sigma}_2 = \frac{1}{n}\sum_i^n g'(x_i^T\hat{\beta}_{BA})x_i x_i^T$*.*

With the assumption of the model that g is bounded, we can get the similar result in Toulis, Airoldi et al. (2017) that $\hat{\beta}_{BA}$ is asymptotically normal distributed.

Like Logit and Probit and any other generalized linear model, the sample partial effect and average partial effect of regressors converge to the partial effect and average partial effect respectively with the rate of $\sqrt{n}$. This rate of partial effect will be different for estimator with unknown g in the next section.

1.3.2 Estimator with Unknown g

When g is unknown, we use sieve method to get the feasible estimator. Kalai and Sastry (2009) use isotonic regression rather than sieve methods to approximate the error function. They don't provide asymptotic properties for their estimator. In addition, sieve methods exhibit better performance than simple isotonic regression. The following is the k_{th} updating for β

$$\tilde{\beta}_k = \tilde{\beta}_{k-1} - \gamma_k C_k \frac{1}{n} \sum_{i=1}^{n} \nabla \tilde{\zeta}_{k-1}(\tilde{\beta}_{k-1}; (x_i, y_i)) \tag{1.6}$$

where $\tilde{\zeta}_{k-1}(\tilde{\beta}_{k-1}; (x_i, y_i))$ is the estimation for $\zeta(\tilde{\beta}_{k-1}; (x_i, y_i))$ using series logistic estimator(SLE) by Hirano, Imbens and Ridder (2003). Denote $R^q(x_i^T\beta)$ as a q-vector of orthogonal basic functions for $x_i^T\beta$ with $\mathbb{E}R^q(x_i^T\beta)R^q(x_i^T\beta)^T = I_q$ conditional on β, where I_q is a $q * q$ identity matrix. One easy way to build $R^q(x_i^T\beta)$ is by power series. Denote $r^q(x_i^T\beta) = (1, (x_i^T\beta), (x_i^T\beta)^2 ... (x_i^T\beta)^{p-1})^T$, then $R^q(x_i^T\beta) = (\mathbb{E}R^q(x_i^T\beta)R^q(x_i^T\beta)^T)^{-\frac{1}{2}} r^q(x_i^T\beta)$. Newey (1994, 1997) proves that $sup||R^q(x_i^T\beta)|| \leq Cq$ for some constant C for orthonormal polynomials. We use $R^q(x_i^T\beta)$ to approximate $g(x_i^T\beta)$. Denote $L(.)$ as $\frac{exp(.)}{(1+exp(.))}$. Then SLE for $R^q(x_i^T\beta)$ is $L(R^q(x_i^T\beta)^T\hat{\pi}_q^k)$ with

$$\hat{\pi}_q^k = arg \max_{\pi} \sum_{i}^{n} (y_i log L(R^q(x_i^T\beta)^T\pi) + (1 - y_i) log(1 - L(R^q(x_i^T\beta)^T\pi))) \tag{1.7}$$

The advantage of SLE is that the objective function above is globally concave so that we can use optimization algorithm like SGD, BGD or the simplex search method of Lagarias et al. (1998) to get $\hat{\pi}_q^k$. Then we get approximation for gradient:

$$\nabla\tilde{\zeta}_{k-1}(\tilde{\beta}_{k-1};(x_i,y_i)) = (L(R^q(x_i^T\tilde{\beta}_{k-1})^T\hat{\pi}_q^{k-1}) - y_i)x_i \tag{1.8}$$

Algorithm 2 Sieve BGD algorithm

1: Guess β^* and $g(.)$ as $\tilde{\beta}_0$ and $g_0(.)$.
2: Update $\tilde{\beta}_1$ using equation $\tilde{\beta}_1 = \tilde{\beta}_0 - \gamma_1 C_1 \frac{1}{n}\sum_{i=1}^{n}(g_0(x_i^T\tilde{\beta}_0) - y_i)x_i$.
3: Calculate $R^q(x_i^T\tilde{\beta}_1)$ and update $g_1(.) = L(R^q(x_i^T\beta_1)^T\hat{\pi}_q^1)$ using equation 1.7.
4: For $k \geq 2$, update $\tilde{\beta}_k$ using equation 1.6 and 1.8.
5: For $k \geq 2$, Calculate $R^q(x_i^T\tilde{\beta}_k)$ and update $g_k(.) = L(R^q(x_i^T\beta_k)^T\hat{\pi}_q^k)$ using equation 1.7.
6: Repeat step 4 and 5 until $\tilde{\beta}_K$.

Remark 4. *We update all β in equation 1.8, which means we update p coefficients. We will standardize it in the last step. In the end, we will estimate $p-1$ coefficient.*

At last we get sieve BGD average (SBA) estimator $\tilde{\beta}_{SBA}$ by averaging $\tilde{\beta}_k$ across different k and let $K = n$:

$$\tilde{\beta}_{SBA} = \frac{1}{n}\sum_{k=1}^{n}\frac{\tilde{\beta}_k}{\tilde{\beta}_k^1} \tag{1.9}$$

Where $\tilde{\beta}_k^1$ is the first component of $\tilde{\beta}_k$.

The assumptions below are following Hirano, Imbens and Ridder (2003)

Assumption 6. *the support* $\mathbf{X}$ *of* X *is a compact subset of* $\mathbb{R}^r$.

Assumption 7. *g is s times continuously differentiable, with $s \geq 5$.*

Assumption 8. *g is bounded away from zero and one on* $\mathbf{X}$.

Assumption 9. *the density of X is bounded away from zero on* $\mathbf{X}$.

Assumption 10. $q \to \infty$ *as* $n \to \infty$ *and* $q^5/n \to 0$.

Remark 5. *Assumption 6 can be relaxed to allow* $\mathbf{X}$ *to be* $\mathbb{R}^r$ *with some tail restriction on the density of X. X with normal distribution works well in simulation.*

Theorem 3. *Under assumptions 1-10 and using sieve BGD algorithm 2 we get*

$$\mathbb{E}||\frac{\tilde{\beta}_k}{\tilde{\beta}_k^1} - \beta^*||^2 \leq \frac{8\overline{\lambda}_c^2\sigma_x^2 C_2(1+2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f1})}{2\gamma_1\underline{\lambda}_c\underline{\lambda}_f}k^{-\gamma}$$
$$+ exp(-log(1+2\gamma_1\underline{\lambda}_c\underline{\lambda}_f)\phi(k))[||\beta_0 - \beta^*|| + (1+2\gamma_1\underline{\lambda}_c\underline{\lambda}_f)^{n_2}A]$$

with k sufficiently large, where $A = 4\overline{\lambda}_c^2\sum_i \gamma_i^2 < \infty$ and $\phi(k) = k^{1-\gamma}$ if $\gamma \in (0.5, 1]$ and $\phi(k) = logk$ if $\gamma = 1$. C_2 and n_2 are some constants.

The result is similar to theorem 1. $\frac{\tilde{\beta}_K}{\tilde{\beta}_K^1}$ is consistent to β^* with rate $\frac{1}{\sqrt{n}}$ if $K = n$ and $\gamma = 1$.

Not surprisingly we get the similar convergence rate as the $\hat{\beta}_{BA}$ in the previous section since the data is averaging and the final estimator is averaging across different iterations. Polyak and Juditsky (1992) suggested that averaging SGD estimator is $\sqrt{n}$ consistent.

Theorem 4. *Under assumptions 1-10, assume and $\gamma \in (0.5, 1)$, we get*

$$(i) \quad \sqrt{n}(\tilde{\beta}_{SBA} - \beta^*) \to N(0, \Sigma_{22}^{-1}\Sigma_1\Sigma_{22}^{-1})$$

where $\Sigma_1 = \mathbb{E}g(x_i^T\beta^*)(1 - g(x_i^T\beta^*))x_i x_i^T$ and $\Sigma_{22} = \mathbb{E}(g'(x_i^T\beta^*)x_i x_i^T - f(x_i^T\beta^*))$, where $f(x_i^T\beta^*)) = \lim_{q\to\infty} x_i R^q(x_i^T\beta^*)^T\mathbb{E}R^q(x_j^T\beta^*)g'(x_j^T\beta^*)x_j^T$ and $R^q(x_i^T\beta^*)$ is orthogonal polynomial function of $x_i^T\beta^*$.

$$(ii) \quad \tilde{\Sigma}_{22}^{-1}\tilde{\Sigma}_1\tilde{\Sigma}_{22}^{-1} \to \Sigma_{22}^{-1}\Sigma_1\Sigma_{22}^{-1}$$

where $\tilde{\Sigma}_1 = \frac{1}{n}\sum_i^n g_n(x_i^T\tilde{\beta}_{SBA})(1-g_n(x_i^T\tilde{\beta}_{SBA}))x_i x_i^T$, $\tilde{\Sigma}_{22} = \frac{1}{n}\sum_i^n(g_n'(x_i^T\tilde{\beta}_{SBA})x_i x_i^T - \tilde{f}(x_i^T\tilde{\beta}_{SBA}))$ and $\tilde{f}(x_i^T\tilde{\beta}_{SBA})) = x_i R^q(x_i^T\tilde{\beta}_{SBA})^T(\frac{1}{n}\sum_j^n R^q(x_j^T\tilde{\beta}_{SBA})g'(x_j^T\tilde{\beta}_{SBA})x_j^T)$.

Variance of $\tilde{\beta}_{SBA}$ differs from variance of $\hat{\beta}_{BA}$ in that it has an extra term $f(x_k^T\beta^*)$ in Σ_{22} compared with Σ_2. This extra term stands for the variance of estimated error function.

Sample partial effect of regressors converges to the expectation of partial effect with known g and slower rate that $\sqrt{n}$ since the lower convergence rate of the approximated error function to true error function. However, Sample average partial effect of regressors converges to the expectation of average partial effect with known g and the rate that $\sqrt{n}$. The difference between sample average partial effect and the expectation of average partial effect with known g is consist two parts. The first part is the difference between sample average partial effect and the expectation of sample average partial effect with unknown g, which equals $O(\frac{1}{\sqrt{n}})$. The second part is the difference between the expectation of sample average partial effect with unknown g and the expectation of sample average partial effect with known g, which also equals $O(\frac{1}{\sqrt{n}})$ with details in the above theorem.

1.4 Simulation

This section presents the result of Monte Carlo experiments. In the following simulation, we use the binary choice model:

$$y_i = \mathbb{1}\{x_i^T \beta > \epsilon\}$$

In this subsection we present simulation results when p is small. First for different initial points, $\tilde{\beta}_{SBA}$ always converges to the neighborhood of the same point. Secondly, we present that ur estimator always converges to the neighborhood of the same point with different γ. At last we compare computation time, mean bias and root mean squared error of our estimator with the estimator (ILS) proposed by Dominitz and Sherman (2005). Through out this subsection, x_i and β is a vector of length 9, the true value of β is $\{1, 1, 2, 4, 5, -1, -2, -4, -5\}$. The regressors are independent of each other. The first regressor equals 1 across i and we do not estimate β_1. We standardized $\hat{\beta}$ by dividing β by $\hat{\beta}_2$. $q = 3$, which means we use 1, z, z^2 and z^3 to estimate the underlying distribution.

We use different initial points, see table 1.1. $\gamma = 0.6$ and ϵ follows standard normal distribution. The number of observation is 5000. Table 1.2 shows that our

Table 1.1: Initial point

β^*	initial 1	initial 2	initial 3	initial 4	initial 5
1	1	1	100	1	100
1	0	1	0	1	100
2	0	2	0	-100	-100
4	0	4	0	100	100
5	0	5	0	100	100
-1	0	-1	0	1	100
-2	0	-2	0	1	100
-4	0	-4	0	-100	-100
-5	0	-5	0	1	-100

estimator converges to the neibourhood of the true point even if the initial point is far away like initial point 5.

Table 1.2: Result for different initial points

β^*	initial 1	initial 2	initial 3	initial 4	initial 5
2	2.06613	2.06621	2.06613	2.06532	2.06038
4	4.37991	4.37958	4.3991	4.38118	4.37315
5	5.35199	5.35179	5.35199	5.35326	5.34327
-1	-1.14102	-1.1408	-1.14102	-1.14101	-1.13701
-2	-2.12314	-2.21338	-2.12314	-2.12282	-2.11651
-4	-4.32496	-4.3246	-4.32496	-4.32598	-4.31775
-5	-5.37754	-5.37742	-5.37754	-5.37807	-5.36886

Then we test the sensitivity of different value of γ on the convergence of our estimator. we use the initial point 1 and 5. Table 1.3 and table 1.4 show that if the initial point is close to true value, our estimator is not sensitive to γ. However, if the initial point is far away like initial point 5, we should choose $\gamma \leq 0.9$. In theorem 4 , $\gamma \in (0.5, 1)$. So if we choose γ close to 1, the estimator is not performed well. The estimator still works well if we choose γ close to 0.5.

Table 1.3: Result for different γ with initial point 1

beta	0.55	0.6	0.7	0.8	0.9	0.99
2	2.06613	2.06613	2.06611	2.06609	2.06603	2.06583
4	4.37992	4.37991	4.37987	4.3798	4.37962	4.37898
5	5.35201	5.35199	5.35195	5.35186	5.35166	5.35092
-1	-1.14102	-1.14102	-1.14101	-1.14098	-1.14091	-1.14066
-2	-2.12315	-2.12314	-2.12311	-2.12306	-2.12294	-2.12248
-4	-4.32498	-4.32496	-4.32492	-4.32484	-4.32467	-4.324
-5	-5.37757	-5.37754	-5.37749	-5.3774	-5.37716	-5.37631

Table 1.4: Result for different γ with initial point 5

β^*	0.55	0.6	0.7	0.8	0.9	0.99
2	2.06117	2.06038	2.05776	2.0512	2.02429	1.93057
4	4.37413	4.37315	4.36991	4.36173	4.32815	4.21098
5	5.3445	5.34327	5.33918	5.32901	5.28767	5.14088
-1	-1.13752	-1.13701	-1.13536	-1.13137	-1.11561	-1.05874
-2	-2.11731	-2.11651	-2.11389	-2.1075	-2.0819	-1.98916
-4	-4.31877	-4.31775	-4.31433	-4.30555	-4.26894	-4.14277
-5	-5.37016	-5.36886	-5.3646	-5.35372	-5.30896	-5.15212

At last, we calculate the computation time, mean bias and root mean squared error. ϵ follows either standard normal distribution or Cauchy distribution with location equivalent to 0 and scale equivalent to 1. The number of observation is 5000 or 10000. We calculate the average time of each experiment, mean bias and root mean square error with 500 experiments.

MRC estimator and MS estimator are not feasible in the binary choice model with more than 4 estimators. We compare our estimator with Dominitz and Sherman (2005), they use iterative least square estimator(ILS) with kernel estimation of the distribution of error, which is similar to our estimator. One major problem is that there are 3 tuning parameters in the process. It's hard to adjust the tuning parameters to calculate the estimator.

Table 1.5: Computation time(second)

	Our estimator		ILS	
Sample size	Normal error	Cauchy error	Normal error	Cauchy error
5000	349.896	201.324	758.784	746.196

We can see from Table 1.5 that our estimator spends much less time than ILS. For the sample size of 5000 and normal distribution, the time spent by our estimator is around 6 minutes, which is reasonable and feasible for empirical studies. For Cauchy distribution, our estimator spend less than 4 minutes.

Table 1.6 shows mean bias and rmse of $\tilde{\beta}_{SBA}$ with error being normal distribution and Cauchy distribution. The mean bias is very small. Both mean bias and root mean squared error (rmse) decrease with size. The bias and rmse are larger under Cauchy distribution.

Table 1.6: Mean bias and rmse

	Normal distribution				Cauchy distribution			
	5000		10000		5000		10000	
$\tilde{\beta}_{SBA}$	bias	rmse	bias	rmse	bias	rmse	bias	rmse
2	-0.00098	0.12496	0.00038	0.08679	0.03378	0.2381	0.00958	0.17107
4	0.00868	0.22910	-0.00002	0.15255	0.07183	0.44649	0.02003	0.31867
5	0.01452	0.28469	0.00013	0.18786	0.08584	0.54389	0.02014	0.38005
-1	-0.00238	0.07796	0.00056	0.05196	-0.01910	0.15741	-0.00075	0.098
-2	-0.00463	0.12256	0.00066	0.08249	-0.03391	0.25052	-0.00667	0.16069
-4	-0.01337	0.22625	-0.00515	0.15335	-0.07077	0.44018	-0.01366	0.29937
-5	-0.01290	0.28205	-0.00420	0.19361	-0.07397	0.53917	-0.02374	0.36411

Table 1.7: Discrete regressors

	5000		10000	
$\tilde{\beta}_{SBA}$	bias	rmse	bias	rmse
2	0.05802055	0.32051897	0.04287214	0.23202075
4	0.0191393	0.57339057	-0.0191574	0.40693718
5	0.05646336	0.70977338	-0.0083537	0.50601113
-1	-0.0325505	0.20010574	-0.0151618	0.13791517
-2	-0.0573911	0.32089004	-0.0378252	0.22680123
-4	-0.0181102	0.5577853	0.02030826	0.41024683
-5	-0.0532119	0.69020763	0.00231334	0.51130765

Table 1.7 show the mean bias and rmse of $\tilde{\beta}_{SBA}$ when all regressors are discrete with value 0 and 1. The error term is normal Cauchy distributed. The mean bias is small. However, rmse is relatively large compare the result with the result with the continuous regressors.

Table 1.8: Normal distribution comparison

	$\tilde{\beta}_{SBA}$		ILS	
	5000		5000	
β^*	bias	rmse	bias	rmse
2	-0.00098	0.12496	-0.11740	0.18355
4	0.00868	0.22910	-0.23175	0.34909
5	0.01452	0.28469	-0.29818	0.43953
-1	-0.00238	0.07796	0.05668	0.11038
-2	-0.00463	0.12256	0.11885	0.18408
-4	-0.01337	0.22625	0.23428	0.35091
-5	-0.01290	0.28205	0.29641	0.43390

Table 1.8 and Table 1.9 are the mean bias and Root mean square error of our estimator and ILS estimator. The bias and rmse of ILS estimator is high because it's hard to adjust the tuning parameters. We can see from table 1.9

Table 1.9: Cauchy distribution comparison

	$\tilde{\beta}_{SBA}$		ILS	
	5000		5000	
β^*	bias	rmse	bias	rmse
2	0.03378	0.23810	-0.34290	0.37925
4	0.07183	0.44649	-0.68575	0.74352
5	0.08585	0.54390	-0.84691	0.91930
-1	-0.01910	0.15741	0.16329	0.19464
-2	-0.03392	0.25052	0.34542	0.37953
-4	-0.07078	0.44019	0.68166	0.73546
-5	-0.07397	0.53917	0.85384	0.92192

that our estimator has less bias than ILS estimator. The bias is even larger if we use Cauchy distribution in table 3 because cauchy distribution is not log-concave which violates the assumption of Dominitz and Sherman (2005).

We don't compare our estimator with other estimators mentioned in the literature review section because most of them suffer from the curse of dimensionality which requires more data or from the optimization problem with non global convex objective functions.

1.5 Conclusion

In this chapter a new estimator is proposed in binary choice model with a semi-parametric setting. If the distribution of error term is unknown, many estimators suffer from curse of dimensionality or optimization problem of non-globally convex objective function.

Our estimator overcome those problems by minimizing a globally convex objective function using single index and approximating the distribution of error term by sieve estimation.

The estimator is calculated through iterations. Firstly, guess β and g as initial value. Secondly, update β according to g from last step by Batch Gradient Descent estimation. Thirdly, update g according to β from last step by Series Logit Estimation.

The estimator is $\sqrt{n}$ consistent and asymptotic normal. With Batch Gradient

Descent estimation, it's easy to compute the estimator and also the variance. We can do inference with the calculated variance. At last, continue previous two steps until satisfaction.

Simulations show that the estimator is computationally easy and performs better than the estimator proposed by Dominitz and Sherman (2005). Other estimators are computationally hard or need more observations.

In the next two chapter, we will develop our estimator into high dimension and apply it to bankruptcy prediction.

1.6 Appendix

Lemma 1. *Suppose $g : R \to R$ is a non-decreasing function, then there exists a convex function $G : R \to R$ such that $G' = g$.*

Proof. Define $G(x) = \int_d^x g(t)dt$, where d is a constant. Then $G(x)$ is convex since $G'(x) = g(x) \geq 0$. □

Lemma 2. *Suppose X is a $v*1$ vector of random variables $X_1, X_2...X_v$ on product probability space $(\Omega, \mathcal{F}, P)$. P is the product of measures $P_1, P_2...P_v$. The domain of at least one of random variables is $\mathbb{R}$ and the measure of it is continuous. $\mathbb{E}(X^TX)$ is positive definite matrix. g(.) is a non-negative continuous function on $\mathbb{R}$. $\mathbb{E}g(X^T\beta) > 0$ for constant vector β with length v. Then $\mathbb{E}g(X^T\beta)(X^TX)$ is positive definite matrix.*

Proof. . We know $\mathbb{E}(X^TX)$ and $\mathbb{E}g(X^T\beta)(X^TX)$ are semi-positive definite matrix. If $\det\mathbb{E}(X^TX) = 0$ if and only if there is linear relation between $X_1, X_2...X_v$, then there is no linear relation between $g(X^T\beta)X_1, g(X^T\beta)X_2...g(X^T\beta)X_v$ and we finish the proof.The sufficiency is obvious and we only prove the necessity. There exists a linear relation among columns of $\mathbb{E}(X^TX)$ since $\det\mathbb{E}(X^TX) = 0$. Denote $\mathbb{E}(X^TX)$ as $[A_1, A_2...A_v]$. Suppose $A_1 = a_2 * A_2 + a_3 * A_3 + ... + a_v * A_v$, where $a_1, a_2...a_v$ are constant, and at least one of them is not zero.By changing the second column into $a_2 * A_2 + a_3 * A_3 + ... + a_v * A_v$, we get a new matrix denoted as $[B_1, B_2...B_v]^2$, By changing the second rows into $a_2 * B_2 + a_3 * B_3 + ... + a_v * B_v$ we get the new matrix, and the first $2*2$ elements are the following:

$$\begin{bmatrix} \mathbb{E}(X_1^2) & \mathbb{E}(X_1(a_2X_2 + a_3X_3 + ... + a_vX_v)) \\ \mathbb{E}(X_1(a_2X_2 + a_3X_3 + ... + a_vX_v)) & \mathbb{E}(a_2X_2 + a_3X_3 + ... + a_vX_v)^2 \end{bmatrix}$$

Then the determinant of the above matrix is 0, then by Hölder's inequality, $X_1 = a_2 * X_2 + a_3 * X_3 + ... + a_v * X_v$. □

Theorem 1. *Under assumptions 1-5 and for $k \leq n$, use BGD algorithm 1 we get*

$$\mathbb{E}||\hat{\beta}_k - \beta^*||^2 \leq \frac{8\overline{\lambda}_c^2 \sigma_x^2 C_1 (1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f)}{2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f} k^{-\gamma}$$
$$+ exp(-log(1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f)\phi(k))[||\beta_0 - \beta^*|| + (1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f)^{n_1} A]$$

with k sufficiently large, where $A = 4\overline{\lambda}_c^2 \sum_i \gamma_i^2 < \infty$ and $\phi(k) = k^{1-\gamma}$ if $\gamma \in (0.5, 1]$ and $\phi(k) = logk$ if $\gamma = 1$. C_1 and n_1 are some constants.

Proof. We start from Eq. (3) and k is the iterative times,

$$\hat{\beta}_k - \beta^* = \hat{\beta}_{k-1} - \beta^* - \gamma_k C_k \frac{1}{n} \sum_i^n \nabla \zeta(\hat{\beta}_{k-1}; (x_i, y_i))$$

then,

$$\begin{aligned} ||\hat{\beta}_k - \beta^*||^2 =& ||\hat{\beta}_{k-1} - \beta^*||^2 \\ &- 2\gamma_k (\hat{\beta}_{k-1} - \beta^*)^T C_k \frac{1}{n} \sum_i^n \nabla \zeta(\hat{\beta}_{k-1}; (x_i, y_i)) \\ &+ \gamma_k^2 ||C_k \frac{1}{n} \sum_i^n \nabla \zeta(\hat{\beta}_{k-1}; (x_i, y_i))||^2 \end{aligned} \tag{1.10}$$

for the third term,

$$\begin{aligned} &\gamma_k^2 ||C_k \frac{1}{n} \sum_i^n \nabla \zeta(\hat{\beta}_{k-1}; (x_i, y_i))||^2 \\ &\leq 4\gamma_k^2 \overline{\lambda}_c^2 \sigma_x^2 \end{aligned}$$

its expectation is bounded as

$$\begin{aligned} &\mathbb{E}(\gamma_k^2 ||C_k \nabla \frac{1}{n} \sum_i^n \nabla \zeta(\hat{\beta}_{k-1}; (x_i, y_i))||^2) \\ &\leq 4\gamma_k^2 \overline{\lambda}_c^2 \sigma_x^2 \end{aligned}$$

for the second term,

$$
\begin{aligned}
&\mathbb{E}(-2\gamma_k(\hat{\beta}_{k-1}-\beta^*)^T C_k \frac{1}{n}\sum_i^n \nabla\zeta(\hat{\beta}_{k-1};(x_i,y_i)))\\
&= -2\gamma_k\mathbb{E}((\hat{\beta}_{k-1}-\beta^*)^T C_k \frac{1}{n}\sum_i^n \nabla\zeta(\hat{\beta}_{k-1};(x_i,y_i)))\\
&= -2\gamma_k\mathbb{E}((\hat{\beta}_{k-1}-\beta^*)^T C_k \mathbb{E}(\nabla\zeta(\hat{\beta}_{k-1};(x_i,y_i))|\hat{\beta}_{k-1})) + \gamma_k(\mathbb{E}||\hat{\beta}_{k-1}-\beta^*||^2)^{\frac{1}{2}} O(\frac{1}{\sqrt{n}})\\
&= -2\gamma_k\mathbb{E}((\hat{\beta}_{k-1}-\beta^*)^T C_k \mathbb{E}(\nabla\zeta(\hat{\beta}_{k-1};(x_i,y_i)) - \nabla\zeta(\beta^*;(x_i,y_i))|\hat{\beta}_{k-1}))\\
&+ \gamma_k(\mathbb{E}||\hat{\beta}_{k-1}-\beta^*||^2)^{\frac{1}{2}} O(\frac{1}{\sqrt{n}})\\
&\leq -2\gamma_k \underline{\lambda}_c \underline{\lambda}_f \mathbb{E}||\hat{\beta}_{k-1}-\beta^*||^2 + \gamma_k(\mathbb{E}||\hat{\beta}_{k-1}-\beta^*||^2)^{\frac{1}{2}} O(\frac{1}{\sqrt{n}})
\end{aligned}
$$

The last inequality comes from strong convexity by Assumption 3 and 2. $\mathbb{E}\nabla\zeta(\beta^*;(x_i,y_i)) = 0$ is implied by Eq.1.1

$$
\begin{aligned}
&g(x_i^T\beta^*) - \mathbb{E}(y_i|x_i) = 0\\
\implies &g(x_i^T\beta^*)x_i - \mathbb{E}(y_i|x_i)x_i = 0\\
\implies &\mathbb{E}(\nabla\zeta(\beta^*;(x_i,y_i)) = 0
\end{aligned}
$$

Then we can rewrite Eq. 1.10 as

$$
\begin{aligned}
\mathbb{E}||\hat{\beta}_k-\beta^*||^2 \leq &(1-2\gamma_k\underline{\lambda}_c\underline{\lambda}_f)\mathbb{E}||\hat{\beta}_{k-1}-\beta^*||^2 + \gamma_k(\mathbb{E}||\hat{\beta}_{k-1}-\beta^*||^2)^{\frac{1}{2}} O(\frac{1}{\sqrt{n}}) + 4\gamma_k^2\overline{\lambda}_c^2\sigma_x^2\\
&\frac{1}{(1+2\gamma_k\underline{\lambda}_c\underline{\lambda}_f)}\mathbb{E}||\hat{\beta}_{k-1}-\beta^*||^2 + \gamma_k(\mathbb{E}||\hat{\beta}_{k-1}-\beta^*||^2)^{\frac{1}{2}} O(\frac{1}{\sqrt{n}}) + 4\gamma_k^2\overline{\lambda}_c^2\sigma_x^2
\end{aligned}
$$

We know $\mathbb{E}||\hat{\beta}_k-\beta^*||^2$ converges to 0 with rate of at least $\frac{1}{n^{\frac{1}{4}}}$ when $\gamma = 1$ by calculating the upper bound of $(\mathbb{E}||\hat{\beta}_{k-1}-\beta^*||^2)^{\frac{1}{2}}$ as $\mathbb{E}||\hat{\beta}_{k-1}-\beta^*||^2+1$ and corollary 2.1 in Toulis, Airoldi et al. (2017) with $a_k = 4\gamma_k^2\overline{\lambda}_c^2\sigma_x^2$ and $b_k = 2\gamma_k\underline{\lambda}_c\underline{\lambda}_f$. However, with rate of less or equal to $\frac{1}{\sqrt{n}}$, we can rewrite the bound for $\mathbb{E}||\hat{\beta}_k-\beta^*||^2$ as

$$
\frac{1}{(1+2\gamma_k\underline{\lambda}_c\underline{\lambda}_f)}\mathbb{E}||\hat{\beta}_{k-1}-\beta^*||^2 + 4\gamma_k^2(\overline{\lambda}_c^2\sigma_x^2 + C_1)
$$

for some constant C_1. Then by corollary 2.1 in Toulis, Airoldi et al. (2017) with $a_k = 4\gamma_k^2(\overline{\lambda}_c^2\sigma_x^2 + C_1)$ and $b_k = 2\gamma_k\underline{\lambda}_c\underline{\lambda}_f$ we get

$$\mathbb{E}||\hat{\beta}_k - \beta^*||^2 \leq \frac{8\overline{\lambda}_c^2\sigma_x^2 C_1(1+2\gamma_1\underline{\lambda}_c\underline{\lambda}_f)}{2\gamma_1\underline{\lambda}_c\underline{\lambda}_f}k^{-\gamma}$$
$$+ exp(-log(1+2\gamma_1\underline{\lambda}_c\underline{\lambda}_f)\phi(k))[||\beta_0 - \beta^*|| + (1+2\gamma_1\underline{\lambda}_c\underline{\lambda}_f)^{n_1}A]$$

with k sufficiently large, where $A = 4\overline{\lambda}_c^2\sum_i \gamma_i^2 < \infty$ and $\phi(k) = k^{1-\gamma}$ if $\gamma \in (0.5, 1]$ and $\phi(k) = logk$ if $\gamma = 1$. C_1 and n_1 is some constant. □

Theorem 2. *Under assumptions 1-5 and for $\gamma \in (0.5, 1)$, we get*

$$(i) \quad \sqrt{n}(\hat{\beta}_{BA} - \beta^*) \to N(0, \Sigma_2^{-1}\Sigma_1\Sigma_2^{-1})$$

where $\Sigma_1 = \mathbb{E}g(x_i^T\beta^*)(1-g(x_i^T\beta^*))x_ix_i^T$ *and* $\Sigma_2 = \mathbb{E}g'(x_i^T\beta^*)x_ix_i^T$.

$$(ii) \quad \hat{\Sigma}_2^{-1}\hat{\Sigma}_1\hat{\Sigma}_2^{-1} \to \Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}$$

where $\hat{\Sigma}_1 = \frac{1}{n}\sum_i^n g(x_i^T\hat{\beta}_{BA})(1-g(x_i^T\hat{\beta}_{BA}))x_ix_i^T$ *and* $\hat{\Sigma}_2 = \frac{1}{n}\sum_i^n g'(x_i^T\hat{\beta}_{BA})x_ix_i^T$.

Proof. First, we write updating function in algorithm 1 as

$$\frac{1}{n}\sum_{i=1}^n \nabla\zeta_{k-1}(\hat{\beta}_{k-1}; (x_i, y_i)) = \frac{1}{\gamma_k}(\hat{\beta}_{k-1} - \hat{\beta}_k)$$

By calculating taylor expansion on $\frac{1}{n}\sum_{i=1}^n \nabla\zeta_{k-1}(\hat{\beta}_{k-1}; (x_i, y_i))$ we get

$$\frac{1}{n}\sum_{i=1}^n \nabla\zeta_{k-1}(\hat{\beta}_{k-1}; (x_i, y_i)) = \frac{1}{n}\sum_{i=1}^n \nabla\zeta_{k-1}(\beta^*; (x_i, y_i)) + \frac{1}{n}\sum_{i=1}^n \frac{\partial\nabla\zeta_{k-1}(\beta^*; (x_i, y_i))}{\partial\beta}(\hat{\beta}_{k-1} - \beta^*)$$

If we proove $\frac{1}{n}\sum_{k=1}^n \frac{1}{\gamma_k}(\tilde{\beta}_{k-1} - \tilde{\beta}_k) = o(1/\sqrt{n})$, then $\sqrt{n}(\hat{\beta}_{BA} - \beta^*)$ behaves like

$$(\frac{1}{n}\sum_{i=1}^n \frac{\partial\nabla\zeta_{k-1}(\beta^*; (x_i, y_i))}{\partial\beta})^{-1}\frac{1}{\sqrt{n}}(\sum_{i=1}^n \nabla\zeta_{k-1}(\beta^*; (x_i, y_i)) - \beta^*)$$

then,

$$\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}(\hat{\beta}_{k-1}-\hat{\beta}_k) \le \frac{1}{n}(-\frac{1}{\gamma_n}(\hat{\beta}_n-\beta^*)+\sum_{k=1}^{n-1}|(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}})(\hat{\beta}_k-\beta^*)|+\frac{1}{\gamma_1}(\hat{\beta}_0-\beta^*))$$
$$\le \frac{1}{n}(-\frac{1}{\gamma_n}(\hat{\beta}_n-\beta^*)+O(1)\sum_{k=1}^{n-1}\frac{1}{\sqrt{k}}+\frac{1}{\gamma_1}(\hat{\beta}_0-\beta^*))$$
$$= o(1/\sqrt{n})$$

This means $\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}(\hat{\beta}_{k-1}-\hat{\beta}_k)$ is negligible. Then we get

$$\frac{1}{\sqrt{n}}(\sum_{i=1}^{n}\nabla\zeta_{k-1}(\beta^*;(x_i,y_i))-\beta^*)\to N(0,\Sigma_1)$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\nabla\zeta_{k-1}(\beta^*;(x_i,y_i))}{\partial\beta}\xrightarrow{p}\Sigma_2$$

where $\Sigma_1=\mathbb{E}g(x_i^T\beta^*)(1-g(x_i^T\beta^*))x_ix_i^T$ and $\Sigma_2=\mathbb{E}g'(x_i^T\beta^*)x_ix_i^T$. □

Theorem 3. *Under assumptions 1-10 and using sieve BGD algorithm 2 we get*

$$\mathbb{E}||\frac{\tilde{\beta}_k}{\tilde{\beta}_k^1}-\beta^*||^2 \le \frac{8\overline{\lambda}_c^{-2}\sigma_x^2C_2(1+2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f1})}{2\gamma_1\underline{\lambda}_c\underline{\lambda}_f}k^{-\gamma}$$
$$+exp(-log(1+2\gamma_1\underline{\lambda}_c\underline{\lambda}_f)\phi(k))[||\beta_0-\beta^*||+(1+2\gamma_1\underline{\lambda}_c\underline{\lambda}_f)^{n_2}A]$$

with k sufficiently large, where $A=4\overline{\lambda}_c^{-2}\sum_i\gamma_i^2<\infty$ and $\phi(k)=k^{1-\gamma}$ if $\gamma\in(0.5,1]$ and $\phi(k)=logk$ if $\gamma=1$. C_2 and n_2 is some constants.

Proof. the following are notations and definitions from Hirano, Imbens and Ridder (2003) with some changes. we use matrix norm $||A||=\sqrt{tr(A'A)}$. Define

$$L_n(\pi)=\frac{1}{n}\sum_{i=1}^{n}(y_ilnL(R_q^{\tilde{\beta}}(x_i)'\pi)+(1-y_i)lnL(1-R_q^{\tilde{\beta}}(x_i)'\pi))$$

$R_q^{\tilde{\beta}}(x_i)\equiv R^q(x_i^T\tilde{\beta})$, $R_q^{\beta^*}(x)\equiv R^q(x^T\beta^*)$, $R^q(.)$ is the basis functions in Hirano, Imbens and Ridder (2003) with order q. $\mathbb{E}R_q^{\tilde{\beta}}(x_i)'R_q^{\tilde{\beta}}(x_i)=1$. Denote

$\iota(q) = sup_{x\in X}||R_q^{\tilde{\beta}}(x_i)||$, where $\iota(q) \leq Cq$ for some constant C. $L(.)$ is logistic distribution. Define

$$\hat{\pi}_q = \underset{\pi}{argmax} L_n(\pi)$$

then, we have

$$\begin{aligned}||\tilde{\beta}_k - \beta^*||^2 =& ||\tilde{\beta}_{k-1} - \beta^*||^2 - 2\gamma_k \frac{1}{n}\sum_{i=1}^{n}(\tilde{\beta}_{k-1} - \beta^*)^T C_k \nabla\tilde{\zeta}(\tilde{\beta}_{k-1};(x_i,y_i))\\ &+ \gamma_k^2||\frac{1}{n}\sum_{i=1}^{n} C_k\nabla\tilde{\zeta}(\tilde{\beta}_{k-1};(x_i,y_i))||^2\end{aligned}$$

where $\nabla\tilde{\zeta}(\beta_{k-1};(x_i,y_i) = (L(R_q^{\tilde{\beta}_{k-1}}(x_i)'\hat{\pi}_q) - y_i)x_i$.

for the second term, by maximize $L_n(\pi)$, we get

$$\frac{1}{n}\sum_{i=1}^{n}(L(R_q^{\tilde{\beta}_{k-1}}(x_i)'\hat{\pi}_q) - y_i)R_q^{\tilde{\beta}_{k-1}}(x_i) = 0. \quad (1.11)$$

We can approximate $L(R_q^{\tilde{\beta}_{k-1}}(x_k)'\hat{\pi}_q)$ and $g(x_k^T\beta^*)$ with $R_q^{\tilde{\beta}_{k-1}}(x_k)'\tilde{\pi}_q$ and $R_q^{\beta^*}(x_k)'\tilde{\pi}_q^*$, according to Lorentz (1986). Then equation becomes

$$\frac{1}{n}\sum_{i=1}^{n}(R_q^{\tilde{\beta}_{k-1}}(x_i)'\tilde{\pi}_q - y_i)R_q^{\tilde{\beta}_{k-1}}(x_i) = O(q^{-s}). \quad (1.12)$$

then we can get $\tilde{\pi}_q$

$$\tilde{\pi}_q = \frac{\frac{1}{n}\sum_{i=n}^{n} R_q^{\tilde{\beta}_{k-1}}(x_i)y_i}{\frac{1}{n}\sum_{i=n}^{n} R_q^{\tilde{\beta}_{k-1}}(x_i)'R_q^{\tilde{\beta}_{k-1}}(x_i)} + (\frac{1}{n}\sum_{i=n}^{n} R_q^{\tilde{\beta}_{k-1}}(x_i)'R_q^{\tilde{\beta}_{k-1}}(x_i))^{-1}O(q^{-s}). \quad (1.13)$$

Denote $\pi_q = \mathbb{E}(R_q^{\tilde{\beta}_{k-1}}(x_i)g(x_i^T\beta^*)|\tilde{\beta}_{k-1})$, then $\tilde{\pi}_q - \pi_q = O(\frac{1}{\sqrt{n}}) + O(\frac{q^{3/2-s}}{\sqrt{n}})$ and

$||\tilde{\pi}_q - \pi_q|| = O(\frac{q}{\sqrt{n}}) + O(\frac{q^{5/2-s}}{\sqrt{n}})$ then,

$$
\begin{aligned}
&\mathbb{E}(2\gamma_k \frac{1}{n}\sum_{i=1}^{n}(\tilde{\beta}_{k-1}-\beta^*)^T C_k \nabla\hat{\zeta}(\tilde{\beta}_{k-1};(x_i,y_i))) \\
\geq& 2\gamma_k \underline{\lambda}_c \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}(L(R_q^{\tilde{\beta}_{k-1}}(x_i)'\hat{\pi}_q) - y_i)(x_i^T\tilde{\beta}_{k-1} - x_i^T\beta^*) \\
=& 2\gamma_k \underline{\lambda}_c \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}(R_q^{\tilde{\beta}_{k-1}}(x_i)'\tilde{\pi}_q - y_i)(x_i^T\tilde{\beta}_{k-1} - x_i^T\beta^*) + \gamma_k O(q^{-s})(\mathbb{E}_{\beta_{k-1}}||\tilde{\beta}_{k-1}-\beta^*||^2)^{\frac{1}{2}} \\
\geq& 2\gamma_k \underline{\lambda}_c \mathbb{E}\mathbb{E}((R_q^{\tilde{\beta}_{k-1}}(x_i)'\tilde{\pi}_q - y_i)(x_i^T\tilde{\beta}_{k-1} - x_i^T\beta^*)|\tilde{\beta}_{k-1}) \\
&+ \gamma_k(O(\frac{q^2}{\sqrt{n}}) + O(\frac{q^{7/2-s}}{\sqrt{n}}))(\mathbb{E}||\beta_{k-1}-\beta^*||^2)^{\frac{1}{2}} + O(\frac{q^2}{\sqrt{n}})\mathbb{E}||\tilde{\beta}_{k-1}-\beta^*||^2 \\
\geq& 2\gamma_k \underline{\lambda}_c \mathbb{E}\mathbb{E}((R_q^{\tilde{\beta}_{k-1}}(x_i)'(\frac{\frac{1}{n}\sum_{i=n}^{n} R_q^{\tilde{\beta}_{k-1}}(x_i)y_i}{\frac{1}{n}\sum_{i=n}^{n} R_q^{\tilde{\beta}_{k-1}}(x_i)'R_q^{\tilde{\beta}_{k-1}}(x_i)} - \frac{\frac{1}{n}\sum_{i=n}^{n} R_q^{\beta^*}(x_i)y_i}{\frac{1}{n}\sum_{i=n}^{n} R_q^{\beta^*}(x_i)'R_q^{\beta^*}(x_i)}))(x_i^T\tilde{\beta}_{k-1} - x_i^T\beta^*)|\tilde{\beta}_{k-1}) \\
&+ 2\gamma_k \underline{\lambda}_c \mathbb{E}\mathbb{E}(R_q^{\tilde{\beta}_{k-1}}(x_i)'\frac{\frac{1}{n}\sum_{i=n}^{n} R_q^{\beta^*}(x_i)y_i}{\frac{1}{n}\sum_{i=n}^{n} R_q^{\beta^*}(x_i)'R_q^{\beta^*}(x_i)} - y_i)(x_i^T\tilde{\beta}_{k-1} - x_i^T\beta^*)|\tilde{\beta}_{k-1}) \\
&+ \gamma_k(O(\frac{q^{5/2}}{n}) + O(\frac{q^{4-s}}{n}))(\mathbb{E}||\tilde{\beta}_{k-1}-\beta^*||^2)^{\frac{1}{2}} + O(\frac{q^2}{\sqrt{n}})\mathbb{E}||\tilde{\beta}_{k-1}-\beta^*||^2
\end{aligned}
$$

The second inequality is coming from

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(R_q^{\tilde{\beta}_{k-1}}(x_i)'\tilde{\pi}_q - y_i)(x_i^T\tilde{\beta}_{k-1} - x_i^T\beta^*) - \mathbb{E}((R_q^{\tilde{\beta}_{k-1}}(x_i)'\tilde{\pi}_q - y_i)(x_i^T\tilde{\beta}_{k-1} - x_i^T\beta^*)|\tilde{\beta}_{k-1}) \\
=&\frac{1}{n}\sum_{i=1}^{n}(R_q^{\tilde{\beta}_{k-1}}(x_i)'(\tilde{\pi}_q - \pi_q))(x_i^T\tilde{\beta}_{k-1} - x_i^T\beta^*) + \frac{1}{n}\sum_{i=1}^{n}(R_q^{\tilde{\beta}_{k-1}}(x_i)'\pi_q - y_i)(x_i^T\tilde{\beta}_{k-1} - x_i^T\beta^*) \\
&+ \mathbb{E}((R_q^{\tilde{\beta}_{k-1}}(x_i)'(\tilde{\pi}_q - \pi_q))(x_i^T\tilde{\beta}_{k-1} - x_i^T\beta^*)|\tilde{\beta}_{k-1}) + \mathbb{E}((R_q^{\tilde{\beta}_{k-1}}(x_i)'\pi_q - y_i)(x_i^T\tilde{\beta}_{k-1} - x_i^T\beta^*)|\tilde{\beta}_{k-1}) \\
=&\frac{1}{n}\sum_{i=1}^{n}(R_q^{\tilde{\beta}_{k-1}}(x_i)'(\tilde{\pi}_q - \pi_q))(x_i^T\tilde{\beta}_{k-1} - x_i^T\beta^*) + \mathbb{E}((R_q^{\tilde{\beta}_{k-1}}(x_i)'(\tilde{\pi}_q - \pi_q))(x_i^T\tilde{\beta}_{k-1} - x_i^T\beta^*)|\tilde{\beta}_{k-1}) \\
+&\frac{1}{n}\sum_{i=1}^{n}(R_q^{\tilde{\beta}_{k-1}}(x_i)'\pi_q - y_i)(x_i^T\tilde{\beta}_{k-1} - x_i^T\beta^*) + \mathbb{E}((R_q^{\tilde{\beta}_{k-1}}(x_i)'\pi_q - y_i)(x_i^T\tilde{\beta}_{k-1} - x_i^T\beta^*)|\tilde{\beta}_{k-1}) \\
=&(O(\frac{q^{5/2}}{n}) + O(\frac{q^{4-s}}{n}))||\tilde{\beta}_{k-1}-\beta^*|| + O(\frac{q^2}{\sqrt{n}})||\tilde{\beta}_{k-1}-\beta^*||^2
\end{aligned}
$$

The proof is similar to the bound on (5) in the addendum of Hirano, Imbens and Ridder (2003).

We rewrite the last inequality as:

$$2\gamma_k\underline{\lambda}_c\mathbb{E}\mathbb{E}((R_q^{\tilde{\beta}_{k-1}}(x_i)'(\frac{\frac{1}{n}\sum_{i=n}^{n}R_q^{\tilde{\beta}_{k-1}}(x_i)y_i}{\frac{1}{n}\sum_{i=n}^{n}R_q^{\tilde{\beta}_{k-1}}(x_i)'R_q^{\tilde{\beta}_{k-1}}(x_i)}-\frac{\frac{1}{n}\sum_{i=n}^{n}R_q^{\beta^*}(x_i)y_i}{\frac{1}{n}\sum_{i=n}^{n}R_q^{\beta^*}(x_i)'R_q^{\beta^*}(x_i)}))(x_i^T\tilde{\beta}_{k-1}-x_i^T\beta^*)|\tilde{\beta}_{k-1})$$

$$+2\gamma_k\underline{\lambda}_c\mathbb{E}\mathbb{E}(R_q^{\tilde{\beta}_{k-1}}(x_i)'\frac{\frac{1}{n}\sum_{i=n}^{n}R_q^{\beta^*}(x_i)y_i}{\frac{1}{n}\sum_{i=n}^{n}R_q^{\beta^*}(x_i)'R_q^{\beta^*}(x_i)}-y_i)(x_i^T\tilde{\beta}_{k-1}-x_i^T\beta^*)|\tilde{\beta}_{k-1})$$

$$+\gamma_k(O(\frac{q^{5/2}}{n})+O(\frac{q^{4-s}}{n}))(\mathbb{E}||\tilde{\beta}_{k-1}-\beta^*||^2)^{\frac{1}{2}}+O(\frac{q^2}{\sqrt{n}})\mathbb{E}||\tilde{\beta}_{k-1}-\beta^*||^2$$

$$\geq 2\gamma_k\underline{\lambda}_c\mathbb{E}\mathbb{E}((R_q^{\tilde{\beta}_{k-1}}(x_i)'\frac{\frac{1}{n}\sum_{i=n}^{n}R_q^{\tilde{\beta}_{k-1}}(x_i)(g(x_i^T\beta^*)-g(x_i^T\tilde{\beta}_{k-1}))}{\frac{1}{n}\sum_{i=n}^{n}R_q^{\tilde{\beta}_{k-1}}(x_i)'R_q^{\tilde{\beta}_{k-1}}(x_i)}(x_i^T\tilde{\beta}_{k-1}-x_i^T\beta^*)|\tilde{\beta}_{k-1})$$

$$+2\gamma_k\underline{\lambda}_c\mathbb{E}\mathbb{E}(g(x_i^T\tilde{\beta}_{k-1})-g(x_i^T\beta^*)(x_i^T\tilde{\beta}_{k-1}-x_i^T\beta^*)|\tilde{\beta}_{k-1})$$

$$+\gamma_k(O(\frac{q^{5/2}}{n})+O(\frac{q^{4-s}}{n})+O(q^{2-s}))(\mathbb{E}||\tilde{\beta}_{k-1}-\beta^*||^2)^{\frac{1}{2}}+O(\frac{q^2}{\sqrt{n}})\mathbb{E}||\tilde{\beta}_{k-1}-\beta^*||^2$$

$$\geq 2\gamma_k\underline{\lambda}_c\mathbb{E}\mathbb{E}((R_q^{\tilde{\beta}_{k-1}}(x_i)'\mathbb{E}R_q^{\tilde{\beta}_{k-1}}(g(x_i^T\beta^*)-g(x_i^T\tilde{\beta}_{k-1}))(x_i^T\tilde{\beta}_{k-1}-x_i^T\beta^*)|\tilde{\beta}_{k-1})$$

$$+2\gamma_k\underline{\lambda}_c\mathbb{E}\mathbb{E}(g(x_i^T\tilde{\beta}_{k-1})-g(x_i^T\beta^*)(x_i^T\tilde{\beta}_{k-1}-x_i^T\beta^*)|\tilde{\beta}_{k-1})$$

$$+\gamma_k(O(\frac{q^{5/2}}{n})+O(\frac{q^{4-s}}{n})+O(q^{2-s}))(\mathbb{E}||\tilde{\beta}_{k-1}-\beta^*||^2)^{\frac{1}{2}}$$

$$+O(\frac{q^{5/2}}{\sqrt{n}})\mathbb{E}||\tilde{\beta}_{k-1}-\beta^*||^2$$

$$\geq 2\gamma_k\underline{\lambda}_c\mathbb{E}\mathbb{E}((R_q^{\tilde{\beta}_{k-1}}(x_i)'\mathbb{E}R_q^{\tilde{\beta}_{k-1}}g(x_i^T\beta^*)-g(x_i^T\beta^*))(x_i^T\tilde{\beta}_{k-1}-x_i^T\beta^*)|\tilde{\beta}_{k-1})$$

$$+\gamma_k(O(\frac{q^{5/2}}{n})+O(\frac{q^{4-s}}{n})+O(q^{2-s}))(\mathbb{E}||\tilde{\beta}_{k-1}-\beta^*||^2)^{\frac{1}{2}}$$

$$+O(\frac{q^{5/2}}{\sqrt{n}})\mathbb{E}||\tilde{\beta}_{k-1}-\beta^*||^2$$

If we take the derivative of $\mathbb{E}\mathbb{E}((R_q^{\tilde{\beta}_{k-1}}(x_i)'\mathbb{E}R_q^{\tilde{\beta}_{k-1}}g(x_i^T\beta^*)-g(x_i^T\beta^*))(x_i^T\tilde{\beta}_{k-1}-x_i^T\beta^*)$ w.r.t. $\tilde{\beta}_{k-1}$ and take value at β^*, we get

$$\mathbb{E}(g'(x_i^T\beta^*)x_ix_i^T-x_iR^q(x_i^T\beta^*)^T\mathbb{E}R^q(x_j^T\beta^*)g'(x_j^T\beta^*)x_j^T)$$

The matrix becomes singular when n goes to infinity. So we must normalized one of β^* in the beginning or at the end of updating process. Denote the minimum eigenvalue of the matrix as λ_{f1}.

By requiring $s \geq 5$ and we consider $q = n^d$, $d < 1/5$. the bound become

$$O(\frac{1}{\sqrt{n}})(\mathbb{E}||\tilde{\beta}_{k-1} - \beta^*||^2)^{\frac{1}{2}} + o(1)\mathbb{E}||\tilde{\beta}_{k-1} - \beta^*||^2$$

for the third term,

$$\begin{aligned} &\gamma_k^2 \mathbb{E}||\frac{1}{n}\sum_{i=1}^{n} C_k \nabla\tilde{\zeta}(\tilde{\beta}_{k-1}; (x_i, y_i))||^2 \\ \leq &4\gamma_k^2 \overline{\lambda}_c^2 \sigma_x^2 \end{aligned}$$

Then,

$$\begin{aligned} \mathbb{E}||\tilde{\beta}_k - \beta^*||^2 \leq (1 - 2\gamma_k \underline{\lambda}_c \underline{\lambda}_{f1} + \gamma_k o(1))\mathbb{E}||\tilde{\beta}_{k-1} - \beta^*||^2 \\ + \gamma_k(O(\sqrt{1/n}))(\mathbb{E}||\tilde{\beta}_{k-1} - \beta^*||^2)^{\frac{1}{2}} + 4\gamma_k^2 \overline{\lambda}_c^2 \sigma_x^2 \end{aligned}$$

then, if n is sufficiently large,

$$\begin{aligned} \mathbb{E}||\tilde{\beta}_k - \beta^*||^2 \leq &(1 - 2\gamma_k \underline{\lambda}_c \underline{\lambda}_{f1})\mathbb{E}||\tilde{\beta}_{k-1} - \beta^*||^2 \\ &+ \gamma_k(O(\sqrt{1/n}))(\mathbb{E}||\tilde{\beta}_{k-1} - \beta^*||^2)^{\frac{1}{2}} + 4\gamma_k^2 \overline{\lambda}_c^2 \sigma_x^2 \\ \leq &\frac{1}{1 + 2\gamma_k \underline{\lambda}_c \underline{\lambda}_{f1}} \mathbb{E}||\tilde{\beta}_{k-1} - \beta^*||^2 \\ &+ \gamma_k(O(\sqrt{1/n}))(\mathbb{E}||\tilde{\beta}_{k-1} - \beta^*||^2)^{\frac{1}{2}} + 4\gamma_k^2 \overline{\lambda}_c^2 \sigma_x^2 \end{aligned}$$

By the same argument as the proof of theorem 1, we get

$$\begin{aligned} \mathbb{E}||\tilde{\beta}_k - \beta^*||^2 \leq \frac{8\overline{\lambda}_c^2 \sigma_x^2 C_2 (1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f1})}{2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f} k^{-\gamma} \\ + exp(-log(1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f)\phi(k))[||\tilde{\beta}_0 - \beta^*|| + (1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f)^{n_2} A] \end{aligned}$$

with k sufficiently large, where $A = 4\overline{\lambda}_c^2 \sum_i \gamma_i^2 < \infty$ and $\phi(k) = k^{1-\gamma}$ if $\gamma \in (0.5, 1]$ and $\phi(k) = logk$ if $\gamma = 1$. C_2 and n_2 is some constants. □

Theorem 4. *Under assumptions 1-10, assume and $\gamma \in (0.5, 1)$, we get*

$$(i) \quad \sqrt{n}(\tilde{\beta}_{SBA} - \beta^*) \rightarrow N(0, \Sigma_{22}^{-1}\Sigma_1\Sigma_{22}^{-1})$$

where $\Sigma_1 = \mathbb{E}g(x_i^T\beta^*)(1 - g(x_i^T\beta^*))x_ix_i^T$ *and* $\Sigma_{22} = \mathbb{E}(g'(x_i^T\beta^*)x_ix_i^T - f(x_i^T\beta^*))$, *where* $f(x_i^T\beta^*)) = \lim_{q\to\infty} x_iR^q(x_i^T\beta^*)^T\mathbb{E}R^q(x_j^T\beta^*)g'(x_j^T\beta^*)x_j^T$ *and* $R^q(x_i^T\beta^*)$ *is orthogonal polynomial function of* $x_i^T\beta^*$.

$$(ii) \quad \tilde{\Sigma}_{22}^{-1}\tilde{\Sigma}_1\tilde{\Sigma}_{22}^{-1} \rightarrow \Sigma_{22}^{-1}\Sigma_1\Sigma_{22}^{-1}$$

where $\tilde{\Sigma}_1 = \frac{1}{n}\sum_i^n g_n(x_i^T\tilde{\beta}_{SBA})(1-g_n(x_i^T\tilde{\beta}_{SBA}))x_ix_i^T$, $\tilde{\Sigma}_{22} = \frac{1}{n}\sum_i^n(g_n'(x_i^T\tilde{\beta}_{SBA})x_ix_i^T - \tilde{f}(x_i^T\tilde{\beta}_{SBA}))$ *and* $\tilde{f}(x_i^T\tilde{\beta}_{SBA})) = x_iR^q(x_i^T\tilde{\beta}_{SBA})^T(\frac{1}{n}\sum_j^n R^q(x_j^T\tilde{\beta}_{SBA})g'(x_j^T\tilde{\beta}_{SBA})x_j^T)$ *and* $g_n(.)$, $g_n'(.)$ *are approximated functions for* $g(.)$, $g'(.)$, *respectively.*

Proof. First, we write equation updating function as

$$\frac{1}{n}\sum_{i=1}^n \nabla\tilde{\zeta}_{k-1}(\tilde{\beta}_{k-1};(x_i,y_i)) = \frac{1}{\gamma_k}(\tilde{\beta}_{k-1} - \tilde{\beta}_k).$$

By taylor expansion on $\frac{1}{n}\sum_{i=1}^n \nabla\tilde{\zeta}_{k-1}(\tilde{\beta}_{k-1};(x_i,y_i))$ we get

$$\frac{1}{n}\sum_{i=1}^n \nabla\tilde{\zeta}_{k-1}(\tilde{\beta}_{k-1};(x_i,y_i)) = \frac{1}{n}\sum_{i=1}^n \nabla\tilde{\zeta}_{k-1}(\beta^*;(x_i,y_i)) + \frac{1}{n}\sum_{i=1}^n \frac{\partial\nabla\tilde{\zeta}_{k-1}(\beta^*;(x_i,y_i))}{\partial\beta}(\tilde{\beta}_{k-1} - \beta^*)$$

We know that $\frac{1}{n}\sum_{i=1}^n \nabla\tilde{\zeta}_{k-1}(\beta^*;(x_i,y_i)) - \frac{1}{n}\sum_{i=1}^n \nabla\zeta(\beta^*;(x_i,y_i))$ is negeligible from the similar argument in theorem 3, then if we prove $\frac{1}{n}\sum_{k=1}^n \frac{1}{\gamma_k}(\tilde{\beta}_{k-1} - \tilde{\beta}_k)$ is negligible $o(1/\sqrt{n})$ and

$$\frac{1}{n}\sum_{i=1}^n \frac{\partial\nabla\tilde{\zeta}_{k-1}(\beta^*;(x_i,y_i))}{\partial\beta} \xrightarrow{p} (\frac{1}{n}\sum_{i=1}^n \frac{\partial\nabla\zeta(\beta^*;(x_i,y_i))}{\partial\beta} + \lim_{q\to\infty} x_iR^q(x_i^T\beta^*)^T\mathbb{E}R^q(x_j^T\beta^*)g'(x_j^T\beta^*)x_j^T)$$

is negligible $o(1/\sqrt{n})$ then $\frac{1}{n}\sum_{k=1}^{n}(\tilde{\beta}_k - \beta^*)$ behaves like

$$(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \nabla\zeta(\beta^*;(x_i,y_i))}{\partial\beta} + \lim_{q\to\infty} x_i R^q(x_i^T\beta^*)^T \mathbb{E}R^q(x_j^T\beta^*)g'(x_j^T\beta^*)x_j^T)^{-1} * (\frac{1}{n}\sum_{i=1}^{n}\nabla\zeta(\beta^*;(x_i,y_i))) \\ \to N(0, \Sigma_{22}^{-1}\Sigma_1(\Sigma_{22}^{-1})^T)$$

where $\Sigma_{22} = \mathbb{E}(g'(x_i^T\beta^*)x_i x_i^T - f(x_i^T\beta^*))$ and $f(x_i^T\beta^*)) = \lim_{q\to\infty} x_i R^q(x_i^T\beta^*)^T \mathbb{E}R^q(x_j^T\beta^*)g'(x_j^T\beta^*)x_j^T$.

At last, $\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}(\tilde{\beta}_{k-1} - \tilde{\beta}_k)$ should be $o(1/\sqrt{n})$, which means negligible.

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}(\tilde{\beta}_{k-1} - \tilde{\beta}_k) &\le \frac{1}{n}(-\frac{1}{\gamma_n}(\tilde{\beta}_n - \beta^*) + \sum_{k=1}^{n-1}|(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}})(\tilde{\beta}_k - \beta^*)| + \frac{1}{\gamma_1}(\tilde{\beta}_0 - \beta^*)) \\
&\le \frac{1}{n}(-\frac{1}{\gamma_n}(\tilde{\beta}_n - \beta^*) + C\sum_{k=1}^{n-1}\frac{1}{\sqrt{k}} + \frac{1}{\gamma_1}(\tilde{\beta}_0 - \beta^*)) \\
&= o(1/\sqrt{n})
\end{aligned}$$

□

Chapter 2

Estimation and Inference of High Dimensional Semiparametric Binary Choice Model

2.1 Introduction

In the previous chapter we propose a novel estimator for binary choice model in semiparametric setting which exhibits root-n consistency and asymptotic normality. In this new chapter, the same estimator is discussed under the high dimensional assumption of regressors. High dimension means the number of regressors goes to infinity as the number of observations goes to infinity, i.e., p goes to infinity as n goes to infinity. The following are three types of high dimensional cases:

- $p/n \to 0$ or more restrictive condition $p^2/n \to 0$ and $p^3/n \to 0$.
- p grows as fast as or faster than n.
- Lasso or ridge restriction.

The first case requires the magnitude of p should be less than the one of n. For linear regression, the requirements are mainly discussed in Portnoy (1984) and Portnoy (1985). For maximum likelihood estimation, see Sur and Candès

(2019) for more information. More detailed discussions of the assumptions and requirements are in literature subsection.

The second case needs central limit theorem in high dimensions, which is mainly discussed in Chernozhukov, Chetverikov and Kato (2017). They showed that the approximation error of the probability that $\frac{1}{\sqrt{n}}\sum_i e_i$ belongs to a hyper-rectangle is close to 0 even if p is much greater that n, where e_i is independent vectors.

The third case is widely discussed after lasso regression proposed by Tibshirani (1996) and ridge regression by Hoerl and Kennard (1970). There are two types of lasso regressions: lasso estimator which runs regression plus ℓ_1 restriction and post lasso regressor which apply original regression to the model selected by first step lass regression. Zhao and Yu (2006) argue the lasso estimator select consistent true model under the irrepresentable condition even if p grows much faster than n. Zhang and Huang (2008) stated under a sparse Riesz condition the lasso estimator is also consistent and select a right dimensional model. For post lasso estimator in linear regression, Belloni and Chernozhukov (2011) showed it converges at least as fast as lass estimator with less convergence.

The new chapter provides asymptotic properties for the same estimators $\hat{\beta}_{BA}$ and $\tilde{\beta}_{SBA}$ with known g and unknown g respectively. Both $\hat{\beta}_{BA}$ and $\tilde{\beta}_{SBA}$ are $\sqrt{\frac{p}{n}}$ consistent. Asymptotic normality of linear combination of $\hat{\beta}_{BA}$ and $\tilde{\beta}_{SBA}$ are also provided. Simulation shows that we need comparatively larger γ to make sure the estimator satisfies the condition $\frac{p^2}{n^{2\gamma-1}} \to 0$ in theorem 8.

2.2 Literature Review

Our estimator can be extended to high dimensional case. High dimension means the number of regressors goes to infinity as the number of observations goes to infinity. The following are three examples mentioned by Fan, Lv and Qi (2011) :

- Augment standard Vector autoregression (VAR) models by Bernanke, Boivin and Eliasz (2005).

- Spatial regression using home-price data (Fan and Lv (2010)).

- Volatility matrix estimation in finance.

The theorem under high dimensional setting is different from traditional theorem under fixed number of regressors (see Portnoy (1984, 1985, 1988); Fan, Liao and Yao (2015); Chernozhukov, Hansen and Spindler (2015); Chernozhukov, Chetverikov and Kato (2017); Belloni et al. (2017)). Huber studied asymptotic properties of M-estimator in his influential paper Huber (1973). Yohai and Maronna (1979) obtain similar result as Huber (1973). Portnoy (1984) finds the smooth M-estimator for linear regression model is consistent under the assumption $\frac{plog(p)}{n} \rightarrow 0$. Fan et al. (2020) get consistency under the "exponential moment condition" by Spokoiny (2012, 2013). We get the same consistent rate $\sqrt{\frac{p}{n}}$ as Yohai and Maronna (1979), Portnoy (1984) and Fan et al. (2020) by adding assumption that $var(x_i^T \beta^*)$ is bounded (see assumption 11).

For asymptotic normality, we need stronger condition $\frac{p^2}{n} \rightarrow 0$ (see assumption 12). Portnoy (1985) and Mammen (1989, 1993) obtain normality even if $\frac{p^2}{n}$ is large in linear regression model. As for MLE of generalized linear model, Portnoy (1988) shows the assumption $\frac{p^2}{n} \rightarrow 0$ is the minimum requirement for the validity of asymptotic normality, which echoes our assumption needed for normality since we use MLE when apply SLE to update error distribution. Sur and Candès (2019) consider logistic regression in high dimension. They find an area where MLE exists below a nonlinear line of $\gamma - \kappa$ map where κ is dimensionality and γ is signal strength. They also provide 'average' behavior of the MLE, i.e, the true parameters are centered around a multiple of true parameter and the asymptotic variance of the MLE are also centered. They provide the limiting distribution of the MLE when the true parameters are 0. We follow the same assumption that $var(x_i^T \beta^*)$ is bounded by Sur and Candès (2019).

In high dimensional setting, our major competitor is the rank estimator proposed by Fan et al. (2020). They generalize Han's MRC estimator and obtaid consistency if the condition $p/n \to 0$ is satisfied. Under a stronger condition that $p^2/n \to 0$ and the condition $log(n/p^2)p^{3/2}/n^{1/4} \to 0$, they find asymptotic normality of the estimator. However, they use the algorithm by Wang (2007), which still suffers from the computational problem. Our estimator has the same convergence rate and also gains asymptotic normality. The biggest advantage of the estimator compared with ranked estimator is that it is computationally easier because of the globally convexity and smoothness of objective function.

2.3 Model

In this section, we provides addition assumptions and theorem in high dimensional settings. The same estimators are used in this chapter and the following theorem show asymptotic properties under known g and unknown g. $\hat{\beta}_{BA}$ is used with known g and $\tilde{\beta}_{SBA}$ is used with unknown g. Below is the additional assumptions:

Assumption 11. *$var(x_i^T\beta^*)$ is bounded.*

Assumption 12. *$p \to \infty$ as $n \to \infty$ and $p/n \to 0$.*

Assumption 13. *$p \to \infty$ as $n \to \infty$ and $p^2/n \to 0$.*

Assumption 11 and 12 are necessary conditions for consistency and Assumption 12 is a necessary condition to normality. Although assumption 11 is not important for consistency of linear regression (see Portnoy (1984)), it is important to get consistency for MLE since the estimator needs existence of MLE (Sur and Candès (2019)). It is also important to sieve estimation since the approximation requires compact domain.

2.3.1 Estimator with Known g

Theorem 5. *Under assumption 1-5 and 11-12 and for $k \leq n$, use BGD algorithm 1 ,we get*

$$\mathbb{E}||\hat{\beta}_k - \beta^*||^2 \leq \frac{8\overline{\lambda}_c^2 \sigma_x^2 C_3(1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f)}{2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f} p k^{-\gamma}$$
$$+ exp(-log(1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f)\phi(k))[||\beta_0 - \beta^*|| + (1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f)^{n_3} A]$$

with k sufficiently large, where $A = 4\overline{\lambda}_c^2 \sum_i \gamma_i^2 < \infty$ and $\phi(k) = k^{1-\gamma}$ if $\gamma \in (0.5, 1]$ and $\phi(k) = logk$ if $\gamma = 1$. C_3 and n_3 are some constants.

$\hat{\beta}_k$ is consistent to β^* with rate $\sqrt{\frac{p}{n}}$ if $K = n$ and $\gamma = 1$.

Not surprisingly, the result is similar to the one with p fixed. The only difference is that the expectation of the norm of $\hat{\beta}_k - \beta^*$ is increasing with p.

Theorem 6. *Under assumption 1-5 and 11-13 and for any $\varsigma \in \mathbb{R}^p$ with $||\varsigma|| = 1$, choose $\gamma \in (0.5, 1)$ so that $\frac{p^2}{n^{2\gamma-1}} \to 0$ and we get*

$$(i) \quad \sqrt{n}\frac{\varsigma'(\hat{\beta}_{BA} - \beta^*)}{(\varsigma'\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}\varsigma)^{\frac{1}{2}}} \to N(0,1)$$

where $\Sigma_1 = \mathbb{E}g(x_i^T\beta^*)(1 - g(x_i^T\beta^*))x_i x_i^T$ and $\Sigma_2 = \mathbb{E}g'(x_i^T\beta^*)x_i x_i^T$.

$$(ii) \quad \varsigma'\hat{\Sigma}_2^{-1}\hat{\Sigma}_1\hat{\Sigma}_2^{-1}\varsigma \to \varsigma'\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}\varsigma$$

where $\hat{\Sigma}_1 = \frac{1}{n}\sum_i^n g(x_i^T\hat{\beta}_{BA})(1 - g(x_i^T\hat{\beta}_{BA}))x_i x_i^T$ and $\hat{\Sigma}_2 = \frac{1}{n}\sum_i^n g'(x_i^T\hat{\beta}_{BA})x_i x_i^T$.

This result is similar to Portnoy (1985).

2.3.2 Estimator with Unknown g

Theorem 7. *Under assumption 1-12 and for $k \leq n$, use sieve BGD algorithm 2 ,we get*

$$\mathbb{E}||\frac{\tilde{\beta}_k}{\tilde{\beta}_k^1} - \beta^*||^2 \leq \frac{8\overline{\lambda}_c^2\sigma_x^2 C_4(1+2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f1})}{2\gamma_1\underline{\lambda}_c\underline{\lambda}_f} pk^{-\gamma}$$
$$+ exp(-log(1+2\gamma_1\underline{\lambda}_c\underline{\lambda}_f)\phi(k))[||\beta_0 - \beta^*|| + (1+2\gamma_1\underline{\lambda}_c\underline{\lambda}_f)^{n_4}A]$$

with k sufficiently large, where $A = 4\overline{\lambda}_c^2\sum_i \gamma_i^2 < \infty$ and $\phi(k) = k^{1-\gamma}$ if $\gamma \in (0.5, 1]$ and $\phi(k) = logk$ if $\gamma = 1$. C_4 and n_4 are some constants.

$\frac{\tilde{\beta}_K}{\tilde{\beta}_K^1}$ is consistent to β^* with rate $\sqrt{\frac{p}{n}}$ if $K = n$ and $\gamma = 1$.

Theorem 8. *Under assumption 1-13 and for any $\varsigma \in \mathbb{R}^p$ with $||\varsigma|| = 1$, choose $\gamma \in (0.5, 1)$ so that $\frac{p^2}{n^{2\gamma-1}} \to 0$ and we get*

$$(i) \quad \sqrt{n}\frac{\varsigma'(\tilde{\beta}_{SBA} - \beta^*)}{(\varsigma'\Sigma_{22}^{-1}\Sigma_1\Sigma_{22}^{-1}\varsigma)^{\frac{1}{2}}} \to N(0,1)$$

where $\Sigma_1 = \mathbb{E}g(x_i^T\beta^*)(1-g(x_i^T\beta^*))x_ix_i^T$ and $\Sigma_{22} = \mathbb{E}(g'(x_i^T\beta^*)x_ix_i^T - f(x_i^T\beta^*))$, where $f(x_i^T\beta^*)) = \lim_{q\to\infty} x_iR^q(x_i^T\beta^*)^T\mathbb{E}R^q(x_j^T\beta^*)g'(x_j^T\beta^*)x_j^T$ and $R^q(x_i^T\beta^*)$ is orthogonal polynomial function of $x_i^T\beta^*$.

$$(ii) \quad \varsigma'\tilde{\Sigma}_{22}^{-1}\tilde{\Sigma}_1\tilde{\Sigma}_{22}^{-1}\varsigma \to \varsigma'\Sigma_{22}^{-1}\Sigma_1\Sigma_{22}^{-1}\varsigma$$

where $\tilde{\Sigma}_1 = \frac{1}{n}\sum_i^n g_n(x_i^T\tilde{\beta}_{SBA})(1-g_n(x_i^T\tilde{\beta}_{SBA}))x_ix_i^T$, $\tilde{\Sigma}_{22} = \frac{1}{n}\sum_i^n(g_n'(x_i^T\tilde{\beta}_{SBA})x_ix_i^T - \tilde{f}(x_i^T\tilde{\beta}_{SBA}))$ and $\tilde{f}(x_i^T\tilde{\beta}_{SBA})) = x_iR^q(x_i^T\tilde{\beta}_{SBA})^T(\frac{1}{n}\sum_j^n R^q(x_j^T\tilde{\beta}_{SBA})g'(x_j^T\tilde{\beta}_{SBA})x_j^T)$.

Remark 6. *For large p, we need relative large γ so that condition $\frac{p^2}{n^{2\gamma-1}} \to 0$ is satisfied.*

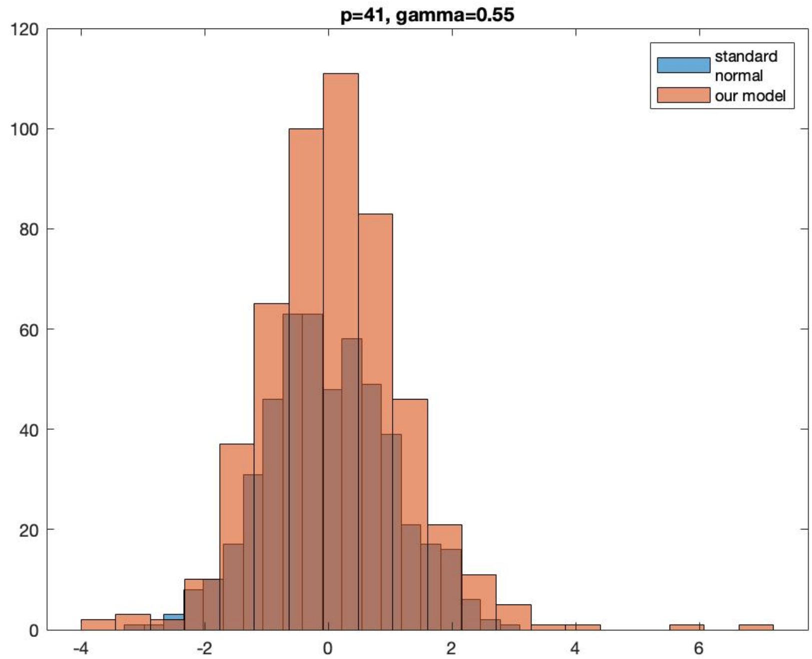

Figure 2.1: High dimension (a)

2.4 Simulation

We set $p = 41$ and run our model at $\gamma = 0.7$ and $\gamma = 0.9$. We set $\varsigma = (1, 1, 1, 1, 1...)^T$, then calculate $\sqrt{n}\frac{\varsigma'(\bar{\beta}_K - \beta^*)}{(\varsigma'\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}\varsigma)^{\frac{1}{2}}}$. We compare our result to standard normal distribution for different γ.

Figure 2.4 to Figure 2.4 show that as p is large, we need comparatively larger γ to make sure the estimator satisfies the condition $\frac{p^2}{n^{2\gamma-1}} \to 0$ in theorem 8 . when γ equals 0.75 or less, we need $\frac{p^4}{n} \to 0$, which is not possible when $p = 41$ and $n = 5000$. In Figure 2.4, the simulation is far from normal distribution. From Figure 2.4 to Figure 2.4, the simulation are closer to standard normal distribution as γ close to 1. In Figure 2.4, the histogram of our estimator is close to standard normal when γ is close to 1.

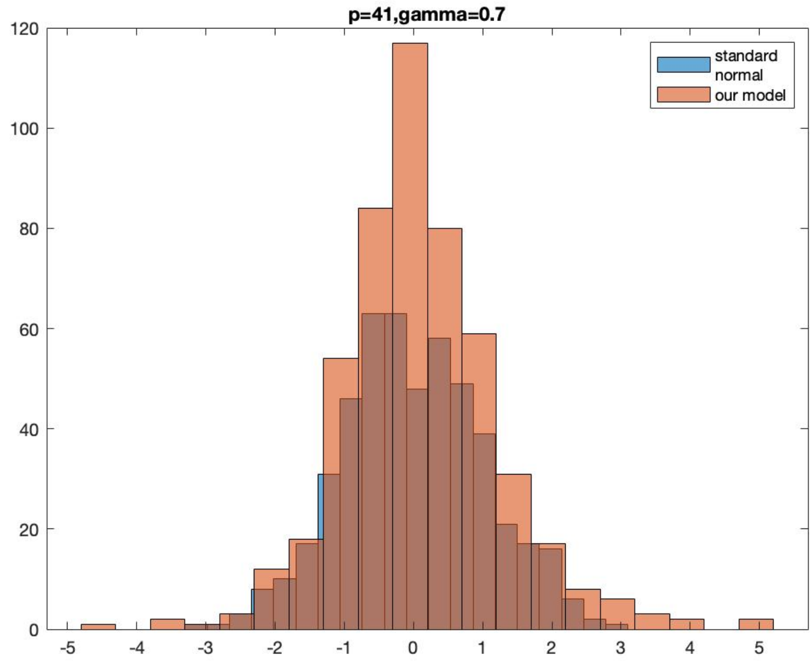

Figure 2.2: High dimension (b)

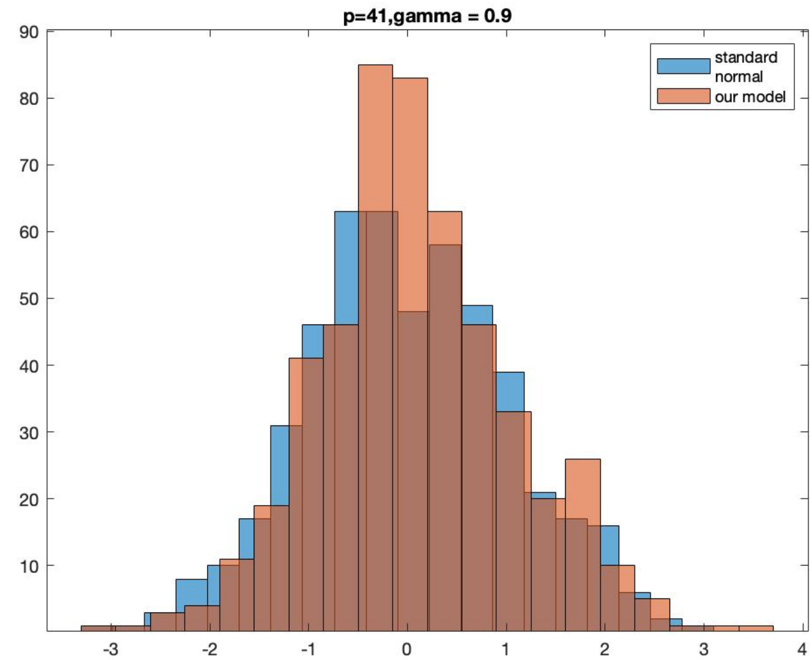

Figure 2.3: High dimension (c)

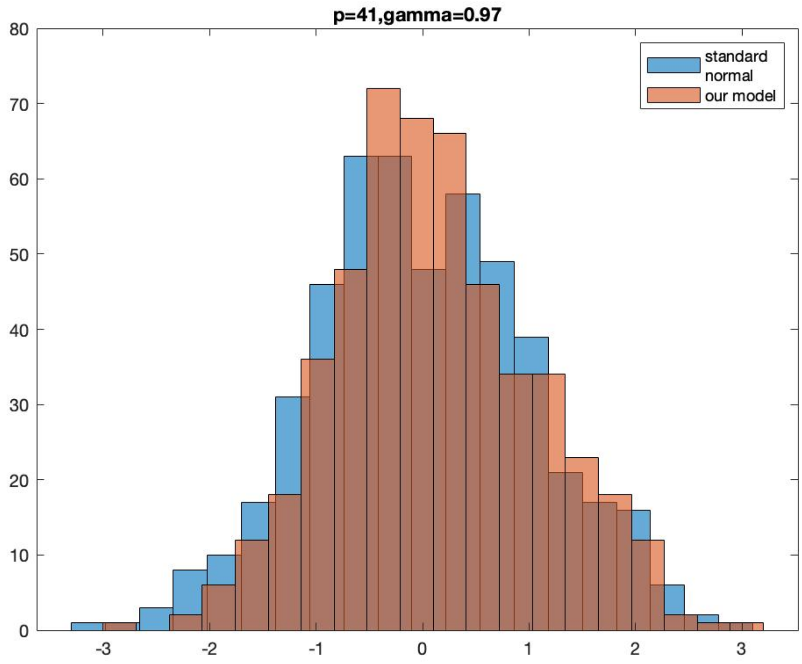

Figure 2.4: High dimension (d)

2.5 Conclusion

Our model is extended to high dimension in this chapter. Firstly, we restrict the high dimension to the case that $\frac{p}{n} \to 0$. Secondly, we calculate the asymptotic normality of linear combination of $\hat{\beta}_{BA}$ and $\tilde{\beta}_{SBA}$ as Portnoy (1984) and Portnoy (1985). Our assumptions are similar to Sur and Candès (2019), i.e., $var(x_i^T\beta^*)$ is bounded.

$\hat{\beta}_{BA}$ and $\tilde{\beta}_{SBA}$ are $\sqrt{\frac{p}{n}}$ consistent with $\frac{p}{n} \to 0$. $\hat{\beta}_{BA}$ and $\tilde{\beta}_{SBA}$ are also asymptotic normal with linear combination and $\frac{p^2}{n} \to 0$. Simulation shows that higher γ is needed with higher p.

This paper can be improved in three ways. Firstly, we will develop the model to allow the number of regressors exceed the number of observations. Recent papers are considering ultra-high dimensional data, see Belloni and Chernozhukov (2011), Chernozhukov, Hansen and Spindler (2015).

Secondly, we will introduce data selection methods like lasso to solve ultra dimensional problem and overcome overfitting. By introducing some bias, methods like lasso will select the accurate variables, see Zhao and Yu (2006) and Zhang and Huang (2008) for more discussion.

At last, panel data will be considered with dynamic version of our model in the future.

2.6 Appendix

Theorem 5. *Under assumption 1-5 and 11-12 and for $k \leq n$, use BGD algorithm 1 ,we get*

$$\mathbb{E}||\hat{\beta}_k - \beta^*||^2 \leq \frac{8\overline{\lambda}_c^{2}\sigma_x^2 C_3(1+2\gamma_1\underline{\lambda}_c\underline{\lambda}_f)}{2\gamma_1\underline{\lambda}_c\underline{\lambda}_f} pk^{-\gamma}$$
$$+ exp(-log(1+2\gamma_1\underline{\lambda}_c\underline{\lambda}_f)\phi(k))[||\beta_0 - \beta^*|| + (1+2\gamma_1\underline{\lambda}_c\underline{\lambda}_f)^{n_3}A]$$

with k sufficiently large, where $A = 4\overline{\lambda}_c^{2}\sum_i \gamma_i^2 < \infty$ and $\phi(k) = k^{1-\gamma}$ if $\gamma \in (0.5, 1]$ and $\phi(k) = logk$ if $\gamma = 1$. C_3 and n_3 are some constants.

$\hat{\beta}_k$ is consistent to β^* with rate $\sqrt{\frac{p}{n}}$ if $K = n$ and $\gamma = 1$.

Proof. With assumption 12, we only have two changes here. The first one is

$$\mathbb{E}(-2\gamma_k(\hat{\beta}_{k-1} - \beta^*)^T C_k \frac{1}{n}\sum_i^n \nabla\zeta(\hat{\beta}_{k-1}; (x_i, y_i)))$$
$$= -2\gamma_k \mathbb{E}((\hat{\beta}_{k-1} - \beta^*)^T C_k \frac{1}{n}\sum_i^n \nabla\zeta(\hat{\beta}_{k-1}; (x_i, y_i)))$$
$$= -2\gamma_k \mathbb{E}((\hat{\beta}_{k-1} - \beta^*)^T C_k \mathbb{E}(\nabla\zeta(\hat{\beta}_{k-1}; (x_i, y_i))|\hat{\beta}_{k-1})) + \gamma_k(\mathbb{E}||\hat{\beta}_{k-1} - \beta^*||^2)^{\frac{1}{2}} O(\sqrt{\frac{p}{n}})$$
$$= -2\gamma_k \mathbb{E}((\hat{\beta}_{k-1} - \beta^*)^T C_k \mathbb{E}(\nabla\zeta(\hat{\beta}_{k-1}; (x_i, y_i)) - \nabla\zeta(\beta^*; (x_i, y_i))|\hat{\beta}_{k-1}))$$
$$+ \gamma_k(\mathbb{E}||\hat{\beta}_{k-1} - \beta^*||^2)^{\frac{1}{2}} O(\sqrt{\frac{p}{n}})$$
$$\leq -2\gamma_k \underline{\lambda}_c\underline{\lambda}_f \mathbb{E}||\hat{\beta}_{k-1} - \beta^*||^2 + \gamma_k(\mathbb{E}||\hat{\beta}_{k-1} - \beta^*||^2)^{\frac{1}{2}} O(\sqrt{\frac{p}{n}})$$

The second one is

$$\gamma_k^2 \mathbb{E}||\frac{1}{n}\sum_{i=1}^n C_k \nabla\zeta(\hat{\beta}_{k-1}; (x_i, y_i))||^2$$
$$\leq 4p\gamma_k^2\overline{\lambda}_c^{2}\sigma_x^2$$

Then,

$$\mathbb{E}||\hat{\beta}_k - \beta^*||^2 \leq (1 - 2\gamma_k \underline{\lambda}_c \underline{\lambda}_{f1} + \gamma_k o(1))\mathbb{E}||\hat{\beta}_{k-1} - \beta^*||^2$$
$$+\gamma_k(O(\sqrt{p/n}))(\mathbb{E}||\hat{\beta}_{k-1} - \beta^*||^2)^{\frac{1}{2}} + 4p\gamma_k^2\overline{\lambda}_c^2\sigma_x^2$$

then, if n is sufficiently large,

$$\begin{aligned}\mathbb{E}||\hat{\beta}_k - \beta^*||^2 \leq& (1 - 2\gamma_k \underline{\lambda}_c \underline{\lambda}_{f1})\mathbb{E}||\hat{\beta}_{k-1} - \beta^*||^2 \\ &+ \gamma_k(O(\sqrt{p/n}))(\mathbb{E}||\hat{\beta}_{k-1} - \beta^*||^2)^{\frac{1}{2}} + 4p\gamma_k^2\overline{\lambda}_c^2\sigma_x^2 \\ \leq& \frac{1}{1 + 2\gamma_k \underline{\lambda}_c \underline{\lambda}_{f1}}\mathbb{E}||\hat{\beta}_{k-1} - \beta^*||^2 \\ &+ \gamma_k(O(\sqrt{p/n}))(\mathbb{E}||\hat{\beta}_{k-1} - \beta^*||^2)^{\frac{1}{2}} + 4p\gamma_k^2\overline{\lambda}_c^2\sigma_x^2\end{aligned}$$

By corollary 2.1 in Toulis, Airoldi et al. (2017) and for $k \leq n$, use BGD algorithm 1 ,we get

$$\begin{aligned}\mathbb{E}||\hat{\beta}_k - \beta^*||^2 \leq& \frac{8\overline{\lambda}_c^2\sigma_x^2 C_3(1 + 2\gamma_1\underline{\lambda}_c\underline{\lambda}_f)}{2\gamma_1\underline{\lambda}_c\underline{\lambda}_f} pk^{-\gamma} \\ &+ exp(-log(1 + 2\gamma_1\underline{\lambda}_c\underline{\lambda}_f)\phi(k))[||\beta_0 - \beta^*|| + (1 + 2\gamma_1\underline{\lambda}_c\underline{\lambda}_f)^{n_3}A]\end{aligned}$$

with k sufficiently large, where $A = 4\overline{\lambda}_c^2\sum_i \gamma_i^2 < \infty$ and $\phi(k) = k^{1-\gamma}$ if $\gamma \in (0.5, 1]$ and $\phi(k) = logk$ if $\gamma = 1$. C_3 and n_3 are some constants.

□

Theorem 6. *Under assumption 1-5 and 11-13 and for any $\varsigma \in \mathbb{R}^p$ with $||\varsigma|| = 1$, choose $\gamma \in (0.5, 1)$ so that $\frac{p^2}{n^{2\gamma-1}} \to 0$ and we get*

$$(i) \quad \sqrt{n}\frac{\varsigma'(\hat{\beta}_{BA} - \beta^*)}{(\varsigma'\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}\varsigma)^{\frac{1}{2}}} \to N(0, 1)$$

where $\Sigma_1 = \mathbb{E}g(x_i^T\beta^*)(1 - g(x_i^T\beta^*))x_i x_i^T$ and $\Sigma_2 = \mathbb{E}g'(x_i^T\beta^*)x_i x_i^T$.

$$(ii) \quad \varsigma'\hat{\Sigma}_2^{-1}\hat{\Sigma}_1\hat{\Sigma}_2^{-1}\varsigma \to \varsigma'\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}\varsigma$$

where $\hat{\Sigma}_1 = \frac{1}{n}\sum_i^n g(x_i^T\hat{\beta}_{BA})(1-g(x_i^T\hat{\beta}_{BA}))x_ix_i^T$ *and* $\hat{\Sigma}_2 = \frac{1}{n}\sum_i^n g'(x_i^T\hat{\beta}_{BA})x_ix_i^T$.

Proof. There are two differences compare to the proof when p is fixed. The first it the following:

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}\varsigma'(\hat{\beta}_{k-1}-\hat{\beta}_k) &\leq \frac{1}{n}(-\frac{1}{\gamma_n}\varsigma'(\hat{\beta}_n-\beta^*)+\sum_{k=1}^{n-1}|(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}})\varsigma'(\hat{\beta}_k-\beta^*)|+\frac{1}{\gamma_1}\varsigma'(\hat{\beta}_0-\beta^*)) \\ &< \frac{1}{n}(-\frac{1}{\gamma_n}(\hat{\beta}_n-\beta^*)+\sum_{k=1}^{n-1}|(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}}||\varsigma'||||\hat{\beta}_k-\beta^*||+\frac{1}{\gamma_1}\varsigma'(\hat{\beta}_0-\beta^*)) \\ &< \frac{1}{n}(-\frac{1}{\gamma_n}(\hat{\beta}_n-\beta^*)+\sum_{k=1}^{n-1}|(k-(k-1))|C\sqrt{\frac{p}{k}}+\frac{1}{\gamma_1}\varsigma'(\hat{\beta}_0-\beta^*)) \\ &= o(\sqrt{\frac{p}{n}})\end{aligned}$$

this means $\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}(\hat{\beta}_{k-1}-\hat{\beta}_k)$ is negligible.

The second difference is the following:

The second-order term of Taylor expansion of $\nabla\zeta_{k-1}(\beta^*;(x_i,y_i))$ is

$$\frac{\partial^2\nabla\zeta_{k-1}(\hat{\beta}_k^*;(x_i,y_i))}{\partial\beta^2}$$

where $\hat{\beta}_k^* = \psi\hat{\beta}_k+(1-\psi)\beta^*$ and $\psi\in[0,1]$. $\frac{\partial^2\nabla\zeta_{k-1}(\hat{\beta}_k^*;(x_i,y_i))}{\partial\beta^2}$ is bounded since $\hat{\beta}_K = \beta^*+o(1)$ and Σ_2 has bounded derivatives. Then the second-order term of Taylor expansion of $\frac{1}{n^2}\sum_{k=1}^{n}\sum_{i=1}^{n}\varsigma'\nabla\zeta_{k-1}(\beta^*;(x_i,y_i))$ is bounded by $C\frac{1}{n}\sum_{k=1}^{n}||\mathbb{E}\varsigma'x_k||\frac{p}{k^\gamma}\leq C\frac{1}{n}\sum_{k=1}^{n}\frac{p^{\frac{3}{2}}}{k^\gamma}$, which is $o(\sqrt{\frac{p}{n}})$ if $\frac{p^2}{n^{2\gamma-1}}\to 0$.

then $\frac{1}{n}\sum_{k=1}^{n}(\hat{\beta}_k-\beta^*)$ behaves like

$$(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\nabla\zeta(\beta^*;(x_i,y_i))}{\partial\beta})^{-1}(\frac{1}{n}\sum_{i=1}^{n}\nabla\zeta(\beta^*;(x_i,y_i)))$$

then for any $\varsigma\in\mathbb{R}^p$ we get $\sqrt{n}\frac{\varsigma'(\hat{\beta}_{BA}-\beta^*)}{(\varsigma'\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}\varsigma)^{\frac{1}{2}}}\to N(0,1)$.

□

Theorem 7. *Under assumption 1-12 and for $k\leq n$, use sieve BGD algorithm 2*

,we get

$$\mathbb{E}||\frac{\tilde{\beta}_k}{\tilde{\beta}_k^1} - \beta^*||^2 \leq \frac{8\overline{\lambda}_c^2 \sigma_x^2 C_4 (1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f1})}{2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f} pk^{-\gamma}$$
$$+ exp(-log(1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f)\phi(k))[||\beta_0 - \beta^*|| + (1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f)^{n_4} A]$$

with k sufficiently large, where $A = 4\overline{\lambda}_c^2 \sum_i \gamma_i^2 < \infty$ and $\phi(k) = k^{1-\gamma}$ if $\gamma \in (0.5, 1]$ and $\phi(k) = logk$ if $\gamma = 1$. C_4 and n_4 are some constants.

Proof. with assumption 12, we only have two changes here. The first one is

$$\begin{aligned}
&\mathbb{E}(2\gamma_k \frac{1}{n} \sum_{i=1}^n (\tilde{\beta}_{k-1} - \beta^*)^T C_k \nabla \tilde{\zeta}(\tilde{\beta}_{k-1}; (x_i, y_i))) \\
\geq& 2\gamma_k \underline{\lambda}_c \mathbb{E} \frac{1}{n} \sum_{i=1}^n (L(R_q^{\tilde{\beta}_{k-1}}(x_i)' \tilde{\pi}_q) - y_i)(x_i^T \tilde{\beta}_{k-1} - x_i^T \beta^*) \\
\geq& 2\gamma_k \underline{\lambda}_c \mathbb{E}\mathbb{E}((R_q^{\tilde{\beta}_{k-1}}(x_i)' \mathbb{E} R_q^{\tilde{\beta}_{k-1}} g(x_i^T \beta^*) - g(x_i^T \beta^*))(x_i^T \tilde{\beta}_{k-1} - x_i^T \beta^*) | \tilde{\beta}_{k-1}) \\
&+ \gamma_k (O(\frac{\sqrt{p} q^{5/2}}{n}) + O(\frac{\sqrt{p} q^{4-s}}{n}) + O(\sqrt{p} q^{2-s}))(\mathbb{E}||\tilde{\beta}_{k-1} - \beta^*||^2)^{\frac{1}{2}} \\
&+ O(\frac{q^{5/2}}{\sqrt{n}}) \mathbb{E}||\tilde{\beta}_{k-1} - \beta^*||^2
\end{aligned}$$

The second one is

$$\begin{aligned}
&\gamma_k^2 \mathbb{E}||\frac{1}{n} \sum_{i=1}^n C_k \nabla \tilde{\zeta}(\tilde{\beta}_{k-1}; (x_i, y_i))||^2 \\
\leq& 4p \gamma_k^2 \overline{\lambda}_c^2 \sigma_x^2
\end{aligned}$$

Then,

$$\mathbb{E}||\tilde{\beta}_k - \beta^*||^2 \leq (1 - 2\gamma_k \underline{\lambda}_c \underline{\lambda}_{f1} + \gamma_k o(1)) \mathbb{E}||\tilde{\beta}_{k-1} - \beta^*||^2$$
$$+ \gamma_k (O(\sqrt{p/n}))(\mathbb{E}||\tilde{\beta}_{k-1} - \beta^*||^2)^{\frac{1}{2}} + 4p \gamma_k^2 \overline{\lambda}_c^2 \sigma_x^2$$

then, if n is sufficiently large,

$$\begin{aligned}
\mathbb{E}||\tilde{\beta}_k - \beta^*||^2 \leq &(1 - 2\gamma_k \underline{\lambda}_c \underline{\lambda}_{f1})\mathbb{E}||\tilde{\beta}_{k-1} - \beta^*||^2 \\
&+ \gamma_k (O(\sqrt{p/n}))(\mathbb{E}||\tilde{\beta}_{k-1} - \beta^*||^2)^{\frac{1}{2}} + 4p\gamma_k^2 \overline{\lambda}_c^2 \sigma_x^2 \\
\leq &\frac{1}{1 + 2\gamma_k \underline{\lambda}_c \underline{\lambda}_{f1}} \mathbb{E}||\tilde{\beta}_{k-1} - \beta^*||^2 \\
&+ \gamma_k (O(\sqrt{p/n}))(\mathbb{E}||\tilde{\beta}_{k-1} - \beta^*||^2)^{\frac{1}{2}} + 4p\gamma_k^2 \overline{\lambda}_c^2 \sigma_x^2
\end{aligned}$$

By corollary 2.1 in Toulis, Airoldi et al. (2017) and for $k \leq n$, we get

$$\begin{aligned}
\mathbb{E}||\frac{\tilde{\beta}_k}{\tilde{\beta}_k^1} - \beta^*||^2 \leq &\frac{8\overline{\lambda}_c^2 \sigma_x^2 C_4 (1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f1})}{2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f} k^{-\gamma} \\
&+ exp(-log(1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f)\phi(k))[||\beta_0 - \beta^*|| + (1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_f)^{n_4} A]
\end{aligned}$$

with k sufficiently large, where $A = 4\overline{\lambda}_c^2 \sum_i \gamma_i^2 < \infty$ and $\phi(k) = k^{1-\gamma}$ if $\gamma \in (0.5, 1]$ and $\phi(k) = logk$ if $\gamma = 1$. C_4 and n_4 are some constants.

□

Theorem 8. *Under assumption 1-13 and for any $\varsigma \in \mathbb{R}^p$ with $||\varsigma|| = 1$, choose $\gamma \in (0.5, 1)$ so that $\frac{p^2}{n^{2\gamma-1}} \to 0$ and we get*

$$(i) \quad \sqrt{n} \frac{\varsigma'(\tilde{\beta}_{SBA} - \beta^*)}{(\varsigma' \Sigma_{22}^{-1} \Sigma_1 \Sigma_{22}^{-1} \varsigma)^{\frac{1}{2}}} \to N(0, 1)$$

where $\Sigma_1 = \mathbb{E}g(x_i^T \beta^*)(1 - g(x_i^T \beta^*))x_i x_i^T$ and $\Sigma_{22} = \mathbb{E}(g'(x_i^T \beta^*)x_i x_i^T - f(x_i^T \beta^*))$, where $f(x_i^T \beta^*)) = \lim_{q \to \infty} x_i R^q (x_i^T \beta^*)^T \mathbb{E} R^q (x_j^T \beta^*) g'(x_j^T \beta^*) x_j^T$ and $R^q(x_i^T \beta^*)$ is orthogonal polynomial function of $x_i^T \beta^*$.

$$(ii) \quad \varsigma' \tilde{\Sigma}_{22}^{-1} \tilde{\Sigma}_1 \tilde{\Sigma}_{22}^{-1} \varsigma \to \varsigma' \Sigma_{22}^{-1} \Sigma_1 \Sigma_{22}^{-1} \varsigma$$

where $\tilde{\Sigma}_1 = \frac{1}{n} \sum_i^n g_n(x_i^T \tilde{\beta}_{SBA})(1 - g_n(x_i^T \tilde{\beta}_{SBA}))x_i x_i^T$, $\tilde{\Sigma}_{22} = \frac{1}{n} \sum_i^n (g_n'(x_i^T \tilde{\beta}_{SBA})x_i x_i^T - \tilde{f}(x_i^T \tilde{\beta}_{SBA}))$ and $\tilde{f}(x_i^T \tilde{\beta}_{SBA})) = x_i R^q (x_i^T \tilde{\beta}_{SBA})^T (\frac{1}{n} \sum_j^n R^q (x_j^T \tilde{\beta}_{SBA}) g'(x_j^T \tilde{\beta}_{SBA}) x_j^T)$.

Proof. There are two differences compare to the proof when p is fixed. The first it the following:

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}\varsigma'(\tilde\beta_{k-1}-\tilde\beta_k) &\le \frac{1}{n}(-\frac{1}{\gamma_n}\varsigma'(\tilde\beta_n-\beta^*)+\sum_{k=1}^{n-1}|(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}})\varsigma'(\tilde\beta_k-\beta^*)|+\frac{1}{\gamma_1}\varsigma'(\tilde\beta_0-\beta^*))\\ &< \frac{1}{n}(-\frac{1}{\gamma_n}(\tilde\beta_n-\beta^*)+\sum_{k=1}^{n-1}|(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}}||\varsigma'||||\tilde\beta_k-\beta^*||+\frac{1}{\gamma_1}\varsigma'(\tilde\beta_0-\beta^*))\\ &< \frac{1}{n}(-\frac{1}{\gamma_n}(\tilde\beta_n-\beta^*)+\sum_{k=1}^{n-1}|(k-(k-1))|C\sqrt{\frac{p}{k}}+\frac{1}{\gamma_1}\varsigma'(\tilde\beta_0-\beta^*))\\ &= o(\sqrt{\frac{p}{n}})\end{aligned}$$

this means $\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}(\tilde\beta_{k-1}-\tilde\beta_k)$ is negligible.The second difference is the following:

The second-order term of Taylor expansion of $\nabla\tilde\zeta_{k-1}(\beta^*;(x_i,y_i))$ is

$$\frac{\partial^2\nabla\tilde\zeta_{k-1}(\tilde\beta_k^*;(x_i,y_i))}{\partial\beta^2}$$

where $\tilde\beta_k^*=\psi\tilde\beta_k+(1-\psi)\beta^*$ and $\psi\in[0,1]$. $\frac{\partial^2\nabla\tilde\zeta_{k-1}(\tilde\beta_k^*;(x_i,y_i))}{\partial\beta^2}$ is bounded since $\tilde\beta_K=\beta^*+o(1)$ and Σ_{22} has bounded derivatives. Then the second-order term of Taylor expansion

$$\frac{1}{n^2}\sum_{k=1}^{n}\sum_{i=1}^{n}\varsigma'\nabla\tilde\zeta_{k-1}(\beta^*;(x_i,y_i))$$

is bounded by $C\frac{1}{n}\sum_{k=1}^{n}||\mathbb{E}\varsigma' x_k||\frac{p}{k^\gamma}\le C\frac{1}{n}\sum_{k=1}^{n}\frac{p^{\frac{3}{2}}}{k^\gamma}$, which is $o(\sqrt{\frac{p}{n}})$ if $\frac{p^2}{n^{2\gamma-1}}\to 0$.

then $\frac{1}{n}\sum_{k=1}^{n}(\tilde\beta_k-\beta^*)$ behaves like

$$(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\nabla\zeta(\beta^*;(x_i,y_i))}{\partial\beta}+\lim_{q\to\infty}x_iR^q(x_i^T\beta^*)^T\mathbb{E}R^q(x_j^T\beta^*)g'(x_j^T\beta^*)x_j^T)^{-1}*(\frac{1}{n}\sum_{i=1}^{n}\nabla\zeta(\beta^*;(x_i,y_i)))$$

then for any $\varsigma\in\mathbb{R}^p$ we get

$$\sqrt{n}\frac{\varsigma'(\tilde\beta_{SBA}-\beta^*)}{(\varsigma'\Sigma_{22}^{-1}\Sigma_1\Sigma_{22}^{-1}\varsigma)^{\frac{1}{2}}}\to N(0,1)$$

where $\Sigma_1=\mathbb{E}g(x_i^T\beta^*)(1-g(x_i^T\beta^*))x_ix_i^T$ and $\Sigma_{22}=\mathbb{E}(g'(x_i^T\beta^*)x_ix_i^T-f(x_i^T\beta^*))$, where $f(x_i^T\beta^*)) = \lim_{q\to\infty} x_i R^q(x_i^T\beta^*)^T \mathbb{E} R^q(x_j^T\beta^*) g'(x_j^T\beta^*) x_j^T$ and $R^q(x_i^T\beta^*)$ is orthogonal polynomial function of $x_i^T\beta^*$.

□

Chapter 3

Application: Prediction of Bankruptcy Failure

3.1 Introduction

We apply our method to the prediction of bankruptcy failure using financial data. The issue is widely discussed in finance and accounting literature, see Belloni et al. (2017) for detailed history from 1930s. Early research focuses on univariate analysis. In 1968, Altman started the first multivariate study. Altman (1968) uses multivariate discriminant analysis (MDA) to analyze bankruptcy prediction, which is an extension of discriminant analysis. Since then, many methods have emerged. The following are the main methods with representative articles:

- Multivariate discriminant analysis(MDA): Altman (1968) , Deakin (1972), Grover (2003).
- Logit/Probit: Ohlson (1980), Mensah (1983), Gaeremynck and Willekens (2003).
- Mixed Logit: Jones and Hensher (2004).
- Time-series cum sums: Kahya and Theodossiou (1999).
- Proportional hazards: Shumway (2001).

- SVM: Barboza, Kimura and Altman (2017), Shin, Lee and Kim (2005).
- Neural network: Odom and Sharda (1990), Leshno and Spector (1996), Mai et al. (2019) .

Each of those methods will be discussed in literature review section.

We first use 5 variables with 680 bankruptcy firms from 1969 to 2010 and non-bankruptcy firms from 2009 to run our model, Probit and Logit model. Our model has different interpretation of the effect of variables. The effect of current ratio is significant in all of the models. Then 22 variables are used with data one year, two years and three years prior to bankruptcy. Our model performs better than Logit in terms of ROC curve with one year and two years data but worse with three years data.

3.2 Literature Review

Altman (1968) suggests using multivariate discriminant analysis (MDA) to analyze bankruptcy prediction. Similar to Logit and Probit analysis, MDA is a statistical method used to classify data. It maximizes distance between bankruptcy firms and non-bankruptcy firms and minimizes the within group variance. Like principal component analysis, it tries to find the classification that best explains the data. However, MDA relies on the parametric assumption that the distribution should be normal distribution, which is not reasonable and justified. The following variables are used in the paper:

Table 3.1: Factor name in Altman (1968)

	Factor name
X_1	Working Capital/Total Assets
X_2	Retained Earning/Total Assets
X_3	Earnings Before Interest and Taxes /Total Assets
X_4	Market value Equity/Book Value of Total Debt
X_5	Net Sales/Total Assets

The specific linear combination of the variables above is Altman's famous Z-score: $0.122X_1 + 0.014X_2 + 0.033X_3 + 0.006X_4 + 0.999X_5$. It is widely used in determining whether a firm is approaching to bankruptcy, which combines profitability, liquidity, leverage and firm activity together. The higher the above variables, the higher probability that a firm will survive. Altman selected 33 manufacturing firms that filed bankruptcy during 1946-1965 and 33 manufacturing non-bankruptcy firms using stratified sampling by industry and size.

Sine then, MDA has been commonly used until recently. Deakin (1972) uses 14 independent variables to predict. He divides them into 4 groups: non-liquid asset group, liquid asset to total asset group, liquid asset to current debt group and liquid asset to turnover group. He also tests the data through 1 year to 5 year before firms went bankruptcy. The result showes that debt ratio and the ratio of current assets to total assets are negative to the survival probability, while ratios like current ratio exhibits ambiguous effect in different year prior to bankruptcy. Zordan (1998) uses as many as 30 variables to predict. More recently, MDA is used with other methods like logit and neural network to compare the performance of those methods, see Lee and Choi (2013), Chung, Tan and Holdsworth (2008), Abdullah et al. (2008).

Logit and Probit are recently most widely used methods in classification. MDA requires the distribution of independent variable to be normal distribution, which are not reasonable and not even feasible for some discrete variables. Instead, Logit and Probit only require specific distribution of error terms and put no restriction on independent variables. The advantages of incorporating both continuous and discrete variables and less restrictions on independent variables make those two methods popular in bankruptcy analysis. Martin (1977) and Hanweck et al. (1977) implement Logit and Probit analysis respectively. Ohlson (1980) uses more than 100 bankruptcy firms and more than 2000 non-bankruptcy firms, which is substantially more than previous studies. The paper finds size of the firm also plays a significant role in determine the probability of bankruptcy. However, the effect of the ratio linked to current liquidity still exhibits less significance than other variables. Mensah (1983) uses more than 30 variables and compared the result of

MDA to Logit. It showed that there is no significant advantage of one method over another in all situations. Meyer and Pifer (1970) applies linear regression to 18 variables. The method does not impose assumption on error term. However, it suffers from the problem that some predictive probability may be greater than 1 or less than 0, which is not interpretable.

MDA or Probit and Logit deal with cross-sectional data. There are three ways to transform panel data into cross-sectional data. Firstly, pair each bankruptcy firm with non-bankruptcy firm by year, industry and size. Secondly, for all non-bankruptcy firms now, randomly pick one year data. Thirdly, use all the data from all the year to form a repeated cross-sectional data. First two methods decrease the number of observations while the last method ignore time varying effect.

Multiperiod Logit model or hazard model consider survival rate for each period. Shumway (2001) proves that Multiperiod Logit model is equivalent to discrete time hazard model. He argues that the static Logit or Probit is inconsistent if the true model is time varying and hazard model is consistent. Finally he presents the comparison between hazard model to MDA by Altman (1968) and Logit by Zmijewski (1984). In comparison between MDA and hazard model, they have different interpretations of the effect of retained earnings and sales. Hazard model also outperforms MDA in terms of out-of-sample prediction. In comparison between Logit and hazard, they show different significance level of some variables. However, Logit model show better performance in terms of out-of-sample prediction. This means changing from static model to dynamic model not necessarily increase prediction. It only changes the significance of some effects. He also finds that the prediction increases if one model include market variables like volatility of stock price. Campbell, Hilscher and Szilagyi (2008) uses the same dynamic Logit and confirmed the prediction power of market variables.

Time series CUSUM test is another version of time varying model. It cumulates and detects the long term effect of firm's bad performance. Kahya and Theodossiou (1999) find CUSUM test outperformance LDA (Linear Discriminant

Analysis) and logit in terms of expected loss. See Theodossiou (1993) for more technical details.

Most recently, machine learning methods gain popularity dealing with prediction problems like supportive vector machine (SVM), random forrest and neural network. The best advantage of machine learning methods is the prediction accuracy over other statistical methods. However, they lack interpretability of the effect of independent variables, which performs like a 'black box'.

SVM creates a hyperplane, linear or nonlinear, separates each class with largest margin between classes. The hyperplane is determined by the points that lie nearest to it. Barboza, Kimura and Altman (2017) test the out-of-sample performances of different machine learning methods like SVM, bagging, boosting and random forrest compared with traditional statistical methods like Logit and MDA. On average, those machine learning models outperform traditional methods by 10% accuracy and bagging, boosting and random forrest showed highest prediction power. Shin, Lee and Kim (2005) find that SVM performs even better than back-propagation neural network (BPN) with small samples.

Neural network uses several hidden layers to increase complexity of models. Logit model is a one layer neural network. Odom and Sharda (1990) show that neural network outperforms MDA in terms of out-of-sample prediction. Mai et al. (2019) applies deep learning method, which provided higher prediction power by including textual disclosures.

In the previous two chapters a new estimator without assuming specific distribution of error term. It shows the same interpretability as traditional MDA, Logit and Probit model and also add some complexities to allow variations of distribution of error term. Compared with most machine learning methods, semi-parametric model has the advantage of interpretability. We compare the new estimator with Logit and Probit in this chapter to see whether they have different the effects of variables and different prediction power.

Some models focus on specific industry:

- Banks: Espahbodi (1991).
- Small & mid-size firms: Laitinen (1991)
- Manufacturing firms: Altman (1968).
- Retail firms: Sharma and Mahajan (1980)
- Internet firms: Wang (2004).

We have data for all industries except for banks because they have much different financial structures than other industries.

3.3 Data

MDA and Logit/Probit are used in cross-sectional data analysis. Because the number of bankruptcy firms is small in a single year, multiple years data of bankruptcy firms is used in those models. Mixed Logit, Time-series cum sums and proportional hazards take time into account, which may exhibit better performances. Recently machine learning methods like neural network have been successful in the accuracy of prediction. However, those methods are harder to interpreted than traditional econometric methods. We compare our methods with Probit/Logit model in this chapter.

3.3.1 Data Description for 5 Variables

The number of factors studied varies from 1 to 57. Altman (1968) uses 5 factors. We first use 5 mostly used factors list in Belloni et al. (2017) (See table 3.2). We get the data from Compustat. Bankruptcy firms are those who filed chapter 11 or delisted from Compustat due to bankruptcy. The total number of bankruptcy firms is 680 from 1969 to 2010. The total number of non-bankruptcy firms is 2766 in 2009. Dependent variable equal 0 if the firm went bankruptcy. For bankruptcy firm, we use the data one or two year prior to bankruptcy.

Table 3.2: Factor name (5 variables)

	Factor name
reat	Retained earnings/Total assets
ebitat	Earnings before interest and taxes /Total assets
cr	Current Ratio
wcat	Working capital /Total assets
niat	Net income/Total assets

Table 3.3: Year 1 correlation matrix (5 variables)

Variables	(1) niat	(2) cr	(3) wcat	(4) ebitat	(5) reat
(1) niat	1				
(2) cr	0.006	1			
(3) wcat	0.684	0.003	1		
(4) ebitat	0.905	0.007	0.416	1	
(5) reat	0.496	0.005	0.586	0.448	1

Table 3.4: Year 2 correlation matrix (5 variables)

Variables	(1) niat	(2) cr	(3) wcat	(4) ebitat	(5) reat
(1) niat	1				
(2) cr	0.006	1			
(3) wcat	0.684	0.003	1		
(4) ebitat	0.905	0.007	0.416	1	
(5) reat	0.496	0.005	0.586	0.448	1

We have two regressions with the same dependent variables. They differ in independent variables:

1. Year 1: For bankruptcy firms, we use the data one year prior to bankruptcy. For non-bankruptcy firms, we use the data in 2009.

2. Year 2: For bankruptcy firms, we use the data two year prior to bankruptcy. For non-bankruptcy firms, we use the data in 2009.

The reason why we use bankruptcy firms across different year is the number of bankruptcy firms in a single year is small. The correlation matrix of year 1 data and year 2 data are almost the same because they only differ in bankruptcy firms. We compare regression result of our model with the result of Probit and Logit. We also compare prediction accuracy for out-of-sample bankruptcy firms from 2010 to 2019.

3.3.2 Data Description for 22 Variables

We get the data from Compustat. We exclude Banks. Bankruptcy firms are those who filed chapter 11 or delisted from Compustat due to bankruptcy. The total number of bankruptcy firms is 588 from 1969 to 2010. The total number of non-bankruptcy firms is 2766 in 2009. Dependent variable equal 0 if the firm went bankruptcy. We choose 22 variables listing in Belloni et al. (2017). For bankruptcy firm, we use the data one, two year or three year prior to bankruptcy.

1. Year 1: For bankruptcy firm, we use the data one year prior to bankruptcy. For non-bankruptcy firms, we use all the data from 1969 to 2009.

2. Year 2: For bankruptcy firm, we use the data 2 years prior to bankruptcy. For non-bankruptcy firms, we use all the data from 1969 to 2009.

3. Year 3: For bankruptcy firm, we use the data 3 years prior to bankruptcy. For non-bankruptcy firms, we use all the data from 1969 to 2009.

Different from previous section, we use all the data from 1969 to 2009 to get training data. We also compare prediction accuracy for out-of-sample bankruptcy firms from 2010 to 2019.

Table 3.5: Number of firms

	total firms	bankrupcy firms	non-bankruptcy firms
1 year ahead	51865	539	51326
2 year ahead	51849	523	51326
3 year ahead	51326	493	51326

Data description and correlation matrix for each year are presented in Appendix.

3.4 Results (5 variables)

Table 3.6: Year 1 result

	Our model	Probit	Logit
ebitat	-0.0737	-0.1061	-0.1828
	(0.1049)	(0.1932)	(0.1284)
cr	1.7084***	1.3397***	2.5221***
	(0.5081)	(0.2472)	(0.4557)
wcat	-0.0633	-0.1593	0.3003
	(0.3941)	(0.7202)	(0.2094)
niat	0.1157	0.1420	0.1989
	(0.1420)	(0.2639)	(0.1459)

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

Table 3.7: Year 1 average partial effect

	Our model	Probit	Logit
ebitat	-0.04321	-0.03556	-0.10289
cr	1.09879	0.44918	1.41958
wcat	-0.03704	-0.05341	0.16900
niat	0.06790	0.04760	0.11196

we can see from table 3.6 and 3.8 that the coefficient of current ratio is significant in all three models. Zmijewski (1984) finds the coefficient of Return on assets (niat) is positive and significant, which is not significant in three models here. This may due to the small year range (1972-1978 in their paper). As for the value of maximum likelihood function, our model reaches almost the same value as Logit in year 1 regression and has a bigger value than the value of Logit in year 2 regression. This means our model performs as good as Logit model in year 1 regression and are even better in year 2 regression. Probit model performs badly in both regressions.

Table 3.8: Year 2 result

	Our model	Probit	Logit
ebitat	0.0335	0.0644	0.0363
	(0.0574)	(0.1003)	(0.0547)
cr	0.1013**	0.0670***	0.0864***
	(0.0418)	(0.0224)	(0.0307)
wcat	0.0361	-0.0126	0.0647
	(0.3407)	(0.3119)	(0.1816)
niat	-0.0331	-0.0766	-0.0308
	(0.0827)	(0.1458)	(0.0784)

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

Table 3.9: Year 2 average partial effect

	Our model	Probit	Logit
ebitat	0.07318	0.07062	0.07474
cr	0.18294	0.07340	0.17783
wcat	-0.12196	-0.01382	0.13329
niat	0.06098	-0.08398	-0.63328

As for the prediction, We calculate predicted probability for survive for bankruptcy firms and draw histogram. From figure 3.3 and 3.6, Logit performs better than our model in year 1 regression. Our model performs better in year 2 regression than Logit and porbit.

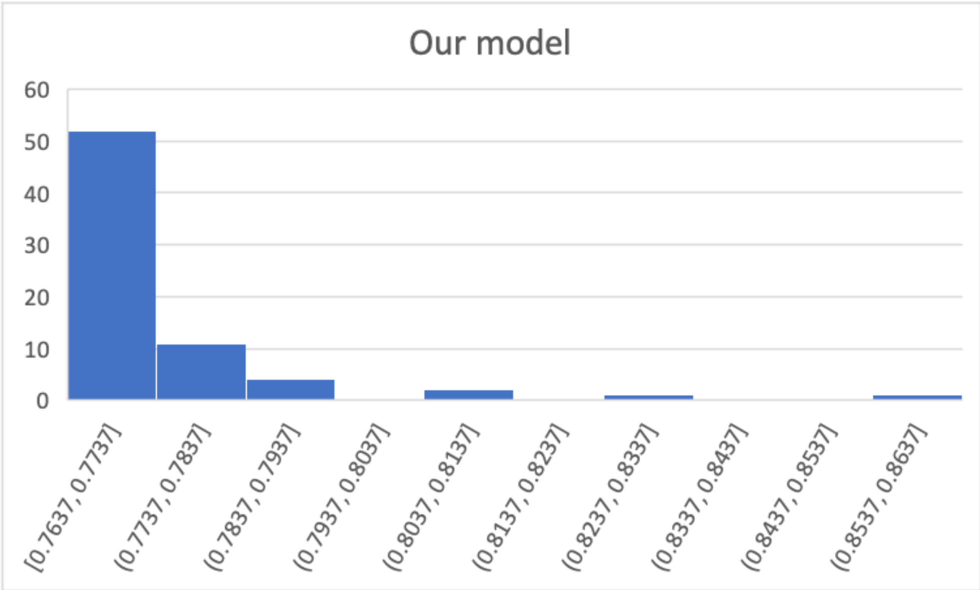

Figure 3.1: Year 1 prediction SBA

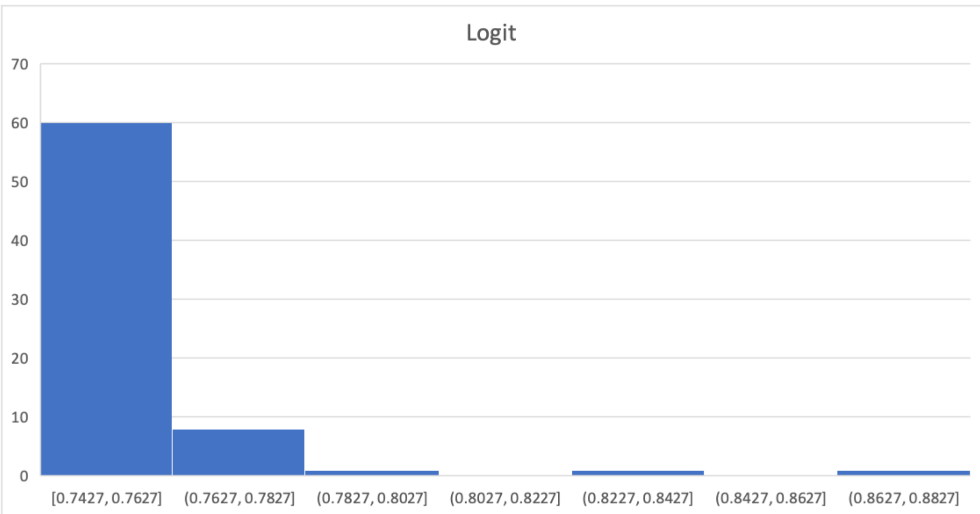

Figure 3.2: Year 1 prediction Logit

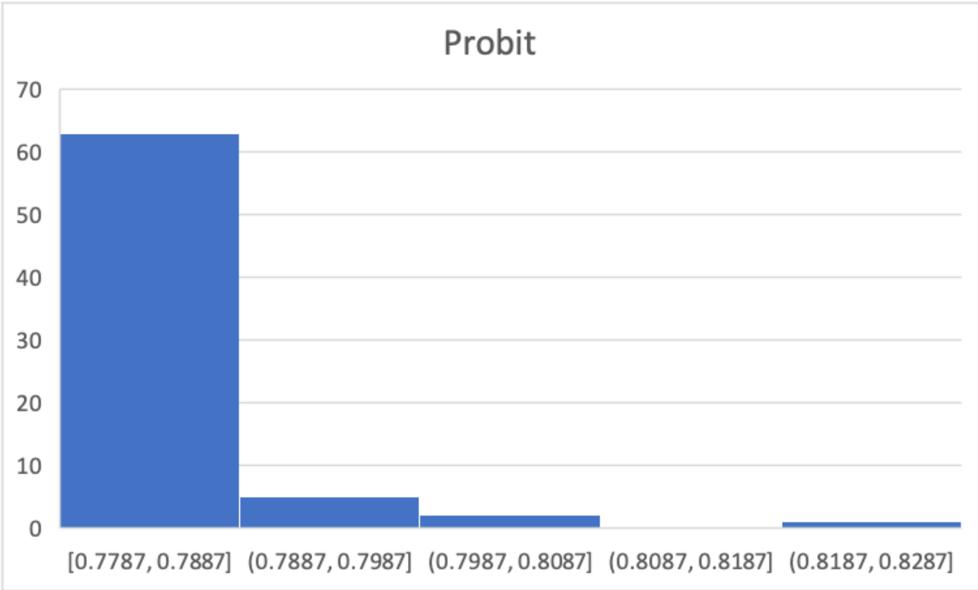

Figure 3.3: Year 1 prediction Probit

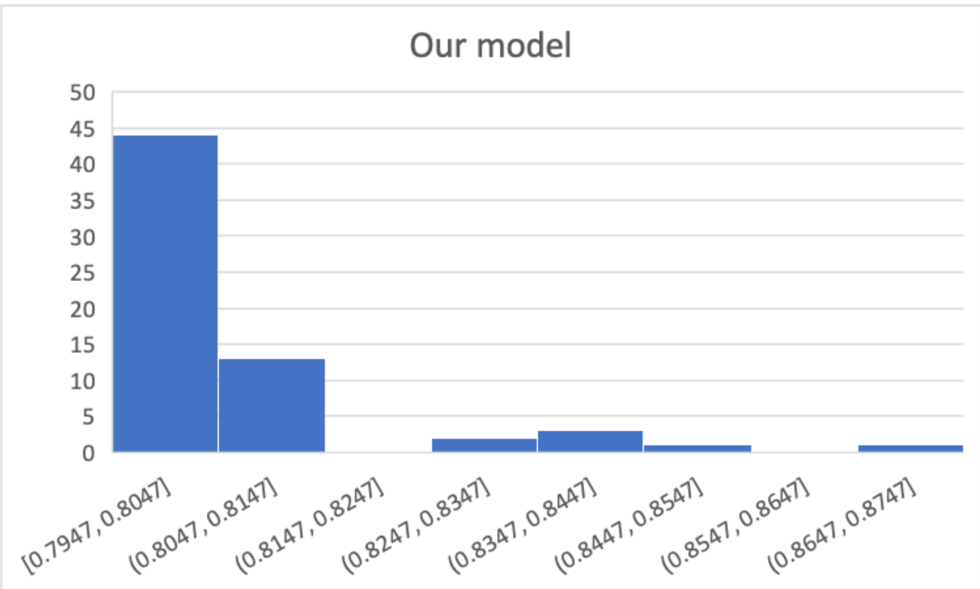

Figure 3.4: Year 1 prediction SBA

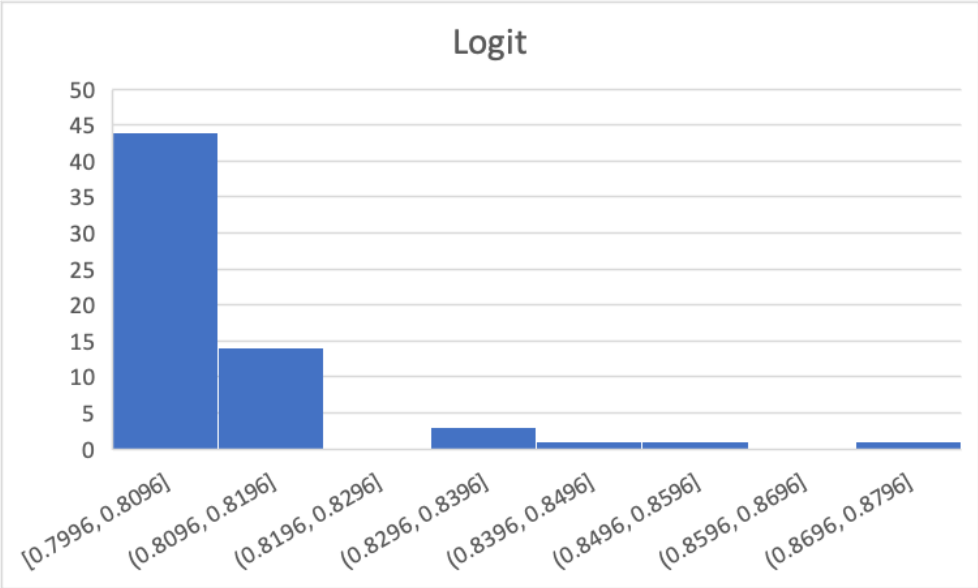

Figure 3.5: Year 1 prediction Logit

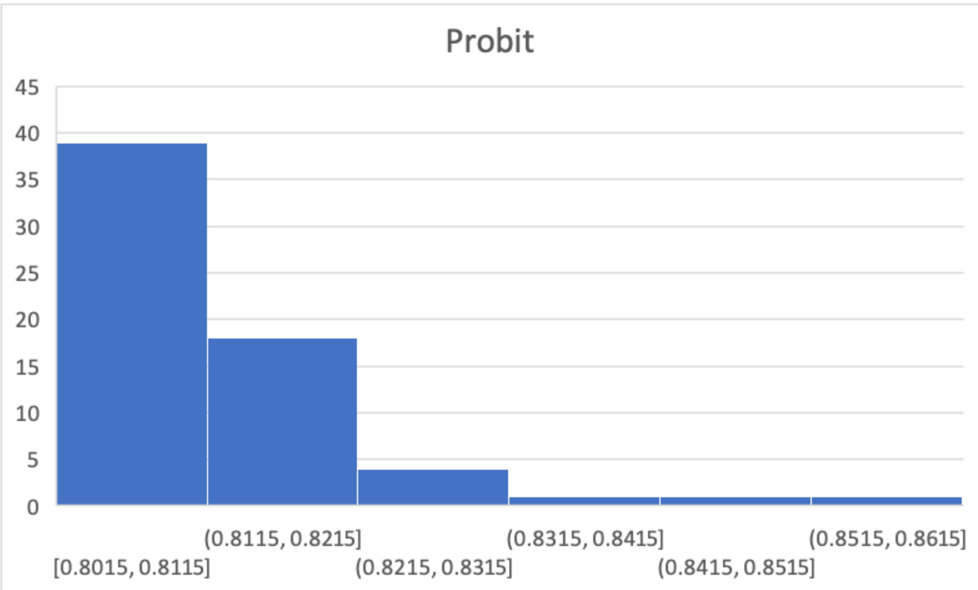

Figure 3.6: Year 1 prediction Probit

3.5 Results (22 variables)

We calculate the ROC curve and AUC for each regression with 2010-2019 data.We set $\gamma = 0.9$ because we have higher number of observations compared with the model in previous section. Changing value of γ has little effect on the result of prediction.

First we look at the ROC curve. Our model performs overwhelmingly better than Probit model since our model is above the curve of Probit almost everywhere with data 1 year prior to bankruptcy. However, with data 2 years prior to bankruptcy, our model is above Probit when threshold is lower while our model is lower than Probit when threshold is higher. With data 3 years prior to bankruptcy, Probit does better than our model.

Table 3.10: AUC

	Our Model	Logit Model
Year 1	0.806	0.807
Year 2	0.758	0.724
Year 3	0.709	0.753

As for AUC, our model performs better than Logit for data one year and two years prior to bankruptcy. However, the Logit is better in terms of data 3 years prior to bankruptcy. This may be the fact that the model is overfitting. If some irrelevant variables are included, overfitting problem will be more severe in our model than Logit because we use sieve methods to approximate the error function, which may increase more variance. This suggests model with lasso to select variables may exhibit better our-of-sample performance.

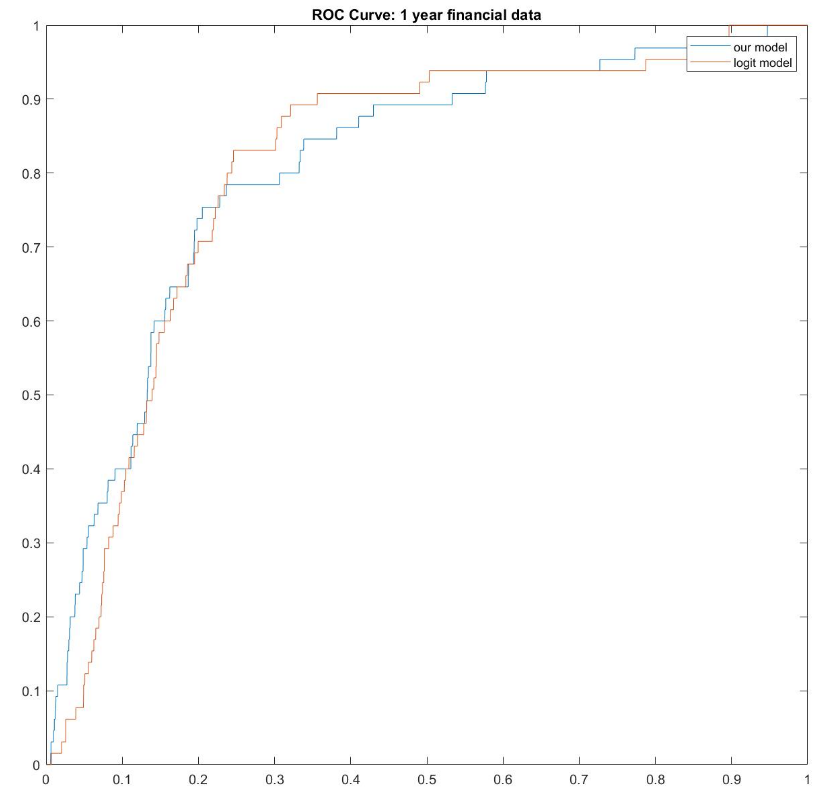

Figure 3.7: Year 1 ROC

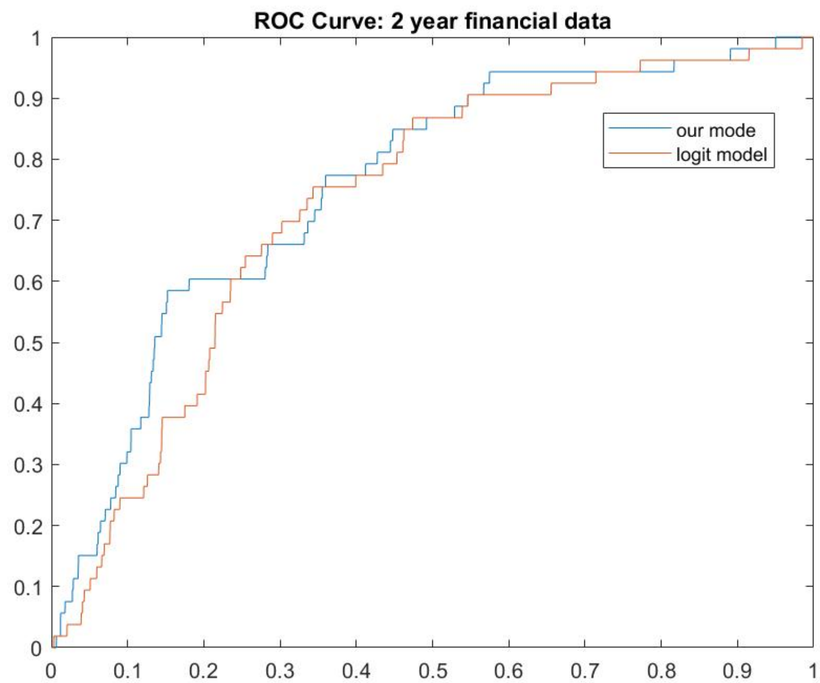

Figure 3.8: Year 2 ROC

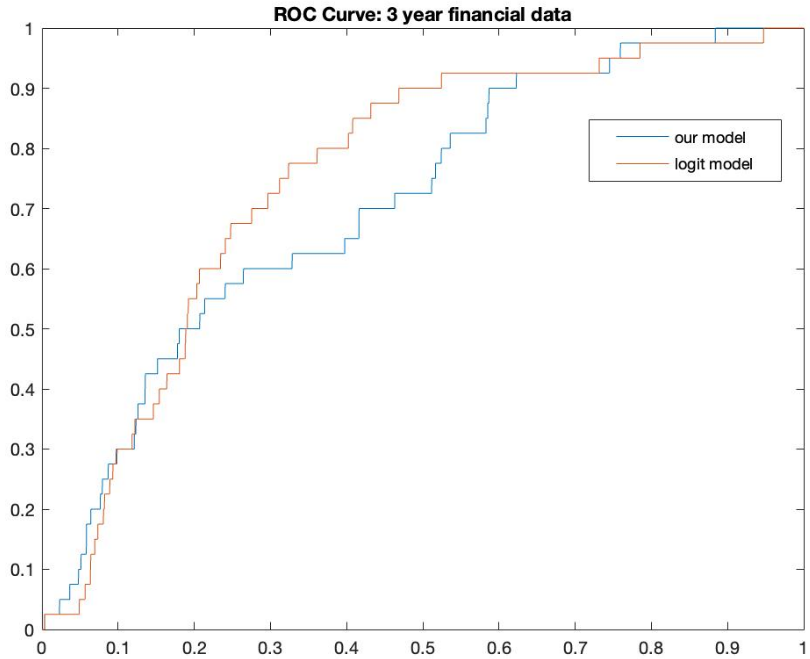

Figure 3.9: Year 3 ROC

3.6 Conclusion

In this chapter we apply our estimator to bankruptcy prediction and compare to Logit and Probit model. Bankruptcy prediction is widely discussed in finance literature. MDA, Logit/Probit, hazard model, SVM and neural network are among those methods used to predict bankruptcy.

We first compare our model with Probit and Logit model using 5 variables with 680 bankruptcy firms from 1969 to 2010 and non-bankruptcy firms from 2009. Data of bankruptcy firms is collected one year or two year prior to firm filing. for bankruptcy. Our model shows different effect for current ratio, which is significant in all the models. As for the prediction, Logit performs better than Probit and our model with year one data. However, our model performs better than Probit and Logit with year two data.

Then we use 22 variables with data one year, two years and three years prior to bankruptcy. Our model has higher AUC than Logit model with data one year, two years prior to bankruptcy. However, Logit model performs better with data three years prior to bankruptcy. This suggests our model is overfitting. Our model may show better result with lasso to select relevant variables.

3.7 Appendix

Table 3.11: Factor Name (22 variables)

	Factor Name
ni_ta	current assets/total assets
cr	net income/total assets
wc_ta	current ratio
re_ta	working capital/total assets
ebit_ta	retained earnings/ total assets
sale_ta	ebit/total assets
qr	sales/total assets
ca_ta	quick ratio
ni_nw	net income/net worth
tl_ta	total liability/total assets
cash_ta	cash/total assets
qa_ta	quick assets/total assets
ca_sale	current assets/sales
invt_sale	inventory/sales
oi_ta	operating income/total assets
ni_sale	net income/sales
ltd_ta	long term debt/total assets
tl_nw	total liability/net worth
wc_sale	working capital/sales
nw_tl	net worth/total liability
log_ta	log_total assets
wc_nw	working capital/net worth

Table 3.12: Year 1 correlation matrix (22 variables)

Variables	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14	-15	-16	-17	-18	-19	-20	-21	-22
(1) ni_ta	1																					
(2) cr	-0.401	1																				
(3) wc_ta	0.576	-0.617	1																			
(4) re_ta	0.568	-0.486	0.828	1																		
(5) ebit_ta	0.956	-0.278	0.43	0.477	1																	
(6) sale_ta	-0.192	0.151	-0.263	-0.249	-0.148	1																
(7) qr	0.001	-0.001	0.001	0.001	0.001	-0.001	1															
(8) ca_ta	-0.028	-0.009	-0.025	-0.036	-0.026	0.214	0.02	1														
(9) ni_nw	-0.002	0.001	0	-0.001	-0.002	0.003	0	-0.005	1													
(10) tl_ta	-0.58	0.617	-0.999	-0.829	-0.43	0.265	-0.001	0.028	0	1												
(11) cash_ta	-0.048	0.035	-0.076	-0.092	-0.039	0.014	0.029	0.424	-0.024	0.078	1											
(12) qa_ta	-0.044	0.003	-0.041	-0.058	-0.041	0.133	0.029	0.812	-0.01	0.044	0.625	1										
(13) ca_sale	-0.001	-0.001	0	0	-0.001	-0.011	0.001	0.021	0	0	0.057	0.037	1									
(14) invt_sale	-0.001	0	0	0	-0.002	-0.006	0	0.01	0	0	0.004	-0.003	0.113	1								
(15) oi_ta	0.944	-0.277	0.428	0.473	0.994	-0.09	0.001	-0.025	-0.002	-0.428	-0.035	-0.04	-0.001	-0.002	1							
(16) ni_sale	0.106	-0.026	0.022	0.037	0.108	0.018	0.001	-0.014	-0.003	-0.022	-0.062	-0.031	-0.297	-0.279	0.11	1						
(17) ltd_ta	-0.041	0.005	-0.011	-0.032	-0.033	0.013	-0.004	-0.049	0.001	0.034	0.005	-0.029	-0.002	-0.001	-0.033	-0.015	1					
(18) tl_nw	0	0	0	0.001	0	-0.002	0.001	0.001	-0.561	0	0.01	0.005	0	0	0	0	-0.002	1				
(19) wc_sale	0.064	-0.083	0.057	0.074	0.058	-0.005	0.002	0.019	0	-0.057	0.037	0.028	0.933	-0.056	0.059	-0.103	-0.012	0	1			
(20) nw_tl	0.018	-0.018	0.016	0.021	0.016	-0.065	0.02	0.112	0.001	-0.017	0.2	0.169	0.059	0	0.015	-0.029	-0.073	-0.003	0.066	1		
(21) log_ta	0.089	-0.073	0.068	0.1	0.083	-0.142	0.01	-0.394	0.002	-0.071	-0.257	-0.341	-0.017	-0.016	0.081	0.059	-0.039	0.001	0.015	-0.108	1	
(22) wc_nw	0	0	0	0	0	0.001	0.001	0.016	-0.725	0	0.023	0.018	0	0	-0.001	0	-0.001	0.66	0	0	-0.003	1

Table 3.13: Year 2 correlation matrix (22 variables)

Variables	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14	-15	-16	-17	-18	-19	-20	-21	-22
(1) ni_ta	1																					
(2) cr	-0.401	1																				
(3) wc_ta	0.576	-0.617	1																			
(4) re_ta	0.568	-0.486	0.828	1																		
(5) ebit_ta	0.956	-0.278	0.43	0.477	1																	
(6) sale_ta	-0.192	0.151	-0.263	-0.249	-0.148	1																
(7) qr	0.001	-0.001	0.001	0.001	0.001	-0.001	1															
(8) ca_ta	-0.028	-0.009	-0.025	-0.036	-0.026	0.214	0.02	1														
(9) ni_nw	-0.002	0.001	0	-0.001	-0.002	0.003	0	-0.005	1													
(10) tl_ta	-0.58	0.617	-0.999	-0.829	-0.43	0.265	-0.001	0.028	0	1												
(11) cash_ta	-0.048	0.035	-0.076	-0.092	-0.039	0.014	0.029	0.424	-0.024	0.078	1											
(12) qa_ta	-0.044	0.003	-0.041	-0.058	-0.041	0.133	0.029	0.812	-0.01	0.044	0.625	1										
(13) ca_sale	-0.001	-0.001	0	0	-0.001	-0.011	0.001	0.021	0	0	0.057	0.037	1									
(14) invt_sale	-0.001	0	0	0	-0.002	-0.006	0	0.01	0	0	0.004	-0.003	0.113	1								
(15) oi_ta	0.944	-0.277	0.428	0.473	0.994	-0.09	0.001	-0.025	-0.002	-0.428	-0.035	-0.04	-0.001	-0.002	1							
(16) ni_sale	0.106	-0.026	0.022	0.037	0.108	0.018	0.001	-0.014	-0.003	-0.022	-0.062	-0.031	-0.297	-0.279	0.11	1						
(17) ltd_ta	-0.041	0.005	-0.011	-0.032	-0.033	0.013	-0.004	-0.049	0.001	0.034	0.005	-0.029	-0.002	-0.001	-0.033	-0.015	1					
(18) tl_nw	0	0	0	0.001	0	-0.002	0.001	0.001	-0.561	0	0.01	0.005	0	0	0	0	-0.002	1				
(19) wc_sale	0.064	-0.083	0.057	0.074	0.058	-0.005	0.002	0.019	0	-0.057	0.037	0.028	0.933	-0.056	0.059	-0.103	-0.012	0	1			
(20) nw_tl	0.018	-0.018	0.016	0.021	0.016	-0.065	0.02	0.112	0.001	-0.017	0.2	0.169	0.059	0	0.015	-0.029	-0.073	-0.003	0.066	1		
(21) log_ta	0.089	-0.073	0.068	0.1	0.083	-0.142	0.01	-0.394	0.002	-0.071	-0.257	-0.341	-0.017	-0.016	0.081	0.059	-0.039	0.001	0.015	-0.108	1	
(22) wc_nw	0	0	0	0	0	0.001	0.001	0.016	-0.725	0	0.023	0.018	0	0	-0.001	0	-0.001	0.66	0	0	-0.003	1

Table 3.14: Year 3 correlation matrix (22 variables)

Variables	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14	-15	-16	-17	-18	-19	-20	-21	-22
(1) ni_ta	1																					
(2) cr	-0.401	1																				
(3) wc_ta	0.576	-0.617	1																			
(4) re_ta	0.568	-0.486	0.828	1																		
(5) ebit_ta	0.956	-0.278	0.43	0.477	1																	
(6) sale_ta	-0.192	0.151	-0.263	-0.249	-0.148	1																
(7) qr	0.001	-0.001	0.001	0.001	0.001	-0.001	1															
(8) ca_ta	-0.028	-0.009	-0.025	-0.036	-0.026	0.214	0.02	1														
(9) ni_nw	-0.002	0.001	0	-0.001	-0.002	0.003	0	-0.005	1													
(10) tl_ta	-0.58	0.617	-0.999	-0.829	-0.43	0.265	-0.001	0.028	0	1												
(11) cash_ta	-0.048	0.035	-0.076	-0.092	-0.039	0.014	0.029	0.424	-0.024	0.078	1											
(12) qa_ta	-0.044	0.003	-0.041	-0.058	-0.041	0.133	0.029	0.812	-0.01	0.044	0.625	1										
(13) ca_sale	-0.001	-0.001	0	0	-0.001	-0.011	0.001	0.021	0	0	0.057	0.037	1									
(14) invt_sale	-0.001	0	0	0	-0.002	-0.006	0	0.01	0	0	0.004	-0.003	0.113	1								
(15) oi_ta	0.944	-0.277	0.428	0.473	0.994	-0.09	0.001	-0.025	-0.002	-0.428	-0.035	-0.04	-0.001	-0.002	1							
(16) ni_sale	0.106	-0.026	0.022	0.037	0.108	0.018	0.001	-0.014	-0.003	-0.022	-0.062	-0.031	-0.297	-0.279	0.11	1						
(17) ltd_ta	-0.041	0.005	-0.011	-0.032	-0.033	0.013	-0.004	-0.049	0.001	0.034	0.005	-0.029	-0.002	-0.001	-0.033	-0.015	1					
(18) tl_nw	0	0	0	0.001	0	-0.002	0.001	0.001	-0.561	0	0.01	0.005	0	0	0	0	-0.002	1				
(19) wc_sale	0.064	-0.083	0.057	0.074	0.058	-0.005	0.002	0.019	0	-0.057	0.037	0.028	0.933	-0.056	0.059	-0.103	-0.012	0	1			
(20) nw_tl	0.018	-0.018	0.016	0.021	0.016	-0.065	0.02	0.112	0.001	-0.017	0.2	0.169	0.059	0	0.015	-0.029	-0.073	-0.003	0.066	1		
(21) log_ta	0.089	-0.073	0.068	0.1	0.083	-0.142	0.01	-0.394	0.002	-0.071	-0.257	-0.341	-0.017	-0.016	0.081	0.059	-0.039	0.001	0.015	-0.108	1	
(22) wc_nw	0	0	0	0	0	0.001	0.001	0.016	-0.725	0	0.023	0.018	0	0	-0.001	0	-0.001	0.66	0	0	-0.003	1

Bibliography

Abdullah, Nur Adiana Hiau, Abd Halim, Hamilton Ahmad and Rohani Md Rus. 2008. "Predicting corporate failure of Malaysias listed companies: Comparing multiple discriminant analysis, logistic regression and the hazard model." *International research journal of finance and economics* 15(2008):201–217.

Agarwal, Alekh, Sham M Kakade, Nikos Karampatziakis, Le Song and Gregory Valiant. 2013. "Least squares revisited: Scalable approaches for multi-class prediction." *arXiv preprint arXiv:1310.1949* .

Ahn, Hyungtaik, Hidehiko Ichimura, James L Powell and Paul A Ruud. 2018. "Simple estimators for invertible index models." *Journal of Business & Economic Statistics* 36(1):1–10.

Altman, Edward I. 1968. "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy." *The journal of finance* 23(4):589–609.

Barboza, Flavio, Herbert Kimura and Edward Altman. 2017. "Machine learning models and bankruptcy prediction." *Expert Systems with Applications* 83:405–417.

Belloni, Alexandre and Victor Chernozhukov. 2011. High dimensional sparse econometric models: An introduction. In *Inverse Problems and High-Dimensional Estimation.* Springer pp. 121–156.

Belloni, Alexandre, Victor Chernozhukov, Iván Fernández-Val and Christian

Hansen. 2017. "Program evaluation and causal inference with high-dimensional data." *Econometrica* 85(1):233–298.

Bellovary, Jodi L, Don E Giacomino and Michael D Akers. 2007. "A review of bankruptcy prediction studies: 1930 to present." *Journal of Financial education* pp. 1–42.

Bernanke, Ben S, Jean Boivin and Piotr Eliasz. 2005. "Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach." *The Quarterly journal of economics* 120(1):387–422.

Bharath, Sreedhar T and Tyler Shumway. 2008. "Forecasting default with the Merton distance to default model." *The Review of Financial Studies* 21(3):1339–1369.

Campbell, John Y, Jens Hilscher and Jan Szilagyi. 2008. "In search of distress risk." *The Journal of Finance* 63(6):2899–2939.

Chen, Xiaohong. 2007. "Large sample sieve estimation of semi-nonparametric models." *Handbook of econometrics* 6:5549–5632.

Chen, Xiaohong, Oliver Linton and Ingrid Van Keilegom. 2003. "Estimation of semiparametric models when the criterion function is not smooth." *Econometrica* 71(5):1591–1608.

Chernozhukov, Victor, Christian Hansen and Martin Spindler. 2015. "Valid post-selection and post-regularization inference: An elementary, general approach." *Annu. Rev. Econ.* 7(1):649–688.

Chernozhukov, Victor, Denis Chetverikov and Kengo Kato. 2017. "Central limit theorems and bootstrap in high dimensions." *The Annals of Probability* 45(4):2309–2352.

Chung, Kim Choy, Shin Shin Tan and David K Holdsworth. 2008. "Insolvency prediction model using multivariate discriminant analysis and artificial neural network for the finance industry in New Zealand." *International journal of business and management* 39(1):19–28.

Cosslett, Stephen R. 1983. "Distribution-free maximum likelihood estimator of the binary choice model." *Econometrica: Journal of the Econometric Society* pp. 765–782.

Cox, David R. 1972. "Regression models and life-tables." *Journal of the Royal Statistical Society: Series B (Methodological)* 34(2):187–202.

Deakin, Edward B. 1972. "A discriminant analysis of predictors of business failure." *Journal of accounting research* pp. 167–179.

Dominitz, Jeff and Robert P Sherman. 2005. "Some convergence theory for iterative estimation procedures with an application to semiparametric estimation." *Econometric Theory* 21(4):838–863.

Duchi, John, Elad Hazan and Yoram Singer. 2011. "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of machine learning research* 12(7).

Duffie, Darrell, Leandro Saita and Ke Wang. 2007. "Multi-period corporate default prediction with stochastic covariates." *Journal of financial economics* 83(3):635–665.

Efron, Bradley. 1975. "The efficiency of logistic regression compared to normal discriminant analysis." *Journal of the American Statistical Association* 70(352):892–898.

Espahbodi, Pouran. 1991. "Identification of problem banks and binary choice models." *Journal of Banking & Finance* 15(1):53–71.

Fabian, Vaclav et al. 1968. "On asymptotic normality in stochastic approximation." *The Annals of Mathematical Statistics* 39(4):1327–1332.

Fan, Jianqing and Jinchi Lv. 2010. "A selective overview of variable selection in high dimensional feature space." *Statistica Sinica* 20(1):101.

Fan, Jianqing, Jinchi Lv and Lei Qi. 2011. "Sparse high-dimensional models in economics." *Annu. Rev. Econ.* 3(1):291–317.

Fan, Jianqing, Yuan Liao and Jiawei Yao. 2015. "Power enhancement in high-dimensional cross-sectional tests." *Econometrica* 83(4):1497–1541.

Fan, Yanqin, Fang Han, Wei Li and Xiao-Hua Zhou. 2020. "On rank estimators in increasing dimensions." *Journal of Econometrics* 214(2):379–412.

Gaeremynck, Ann and Marleen Willekens. 2003. "The endogenous relationship between audit-report type and business termination: Evidence on private firms in a non-litigious environment." *Accounting and Business Research* 33(1):65–79.

Geman, Stuart and Chii-Ruey Hwang. 1982. "Nonparametric maximum likelihood estimation by the method of sieves." *The Annals of Statistics* pp. 401–414.

George, Abraham P and Warren B Powell. 2006. "Adaptive stepsizes for recursive estimation with applications in approximate dynamic programming." *Machine learning* 65(1):167–198.

Grenander, Ulf. 1981. Abstract inference. Technical report.

Grover, Jeffrey S. 2003. *Validation of a cash flow model: A non-bankruptcy approach.* Nova Southeastern University.

Han, Aaron K. 1987. "Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator." *Journal of Econometrics* 35(2-3):303–316.

Hanweck, Gerald A et al. 1977. Predicting bank failure. Technical report Board of Governors of the Federal Reserve System (US).

Hinton, Geoffrey, Nitish Srivastava and Kevin Swersky. 2012. "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent." *Cited on* 14(8):2.

Hirano, Keisuke, Guido W Imbens and Geert Ridder. 2003. "Efficient estimation of average treatment effects using the estimated propensity score." *Econometrica* 71(4):1161–1189.

Hoerl, Arthur E and Robert W Kennard. 1970. "Ridge regression: Biased estimation for nonorthogonal problems." *Technometrics* 12(1):55–67.

Horowitz, Joel L. 1992. "A smoothed maximum score estimator for the binary response model." *Econometrica: journal of the Econometric Society* pp. 505–531.

Huber, Peter J. 1973. "Robust regression: asymptotics, conjectures and Monte Carlo." *The annals of statistics* pp. 799–821.

Ichimura, Hidehiko. 1987. Estimation of single index models PhD thesis Massachusetts Institute of Technology.

Jain, Prateek, Ambuj Tewari and Purushottam Kar. 2014. "On iterative hard thresholding methods for high-dimensional m-estimation." *Advances in neural information processing systems* 27.

Jones, Stewart and David A Hensher. 2004. "Predicting firm financial distress: A mixed logit model." *The accounting review* 79(4):1011–1038.

Jurečková, Jana, Pranab Kumar Sen and Jan Picek. 2012. *Methodology in robust and nonparametric statistics*. CRC Press.

Kahya, Emel and Panayiotis Theodossiou. 1999. "Predicting corporate financial distress: A time-series CUSUM methodology." *Review of Quantitative Finance and Accounting* 13(4):323–345.

Kakade, Sham M, Varun Kanade, Ohad Shamir and Adam Kalai. 2011. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*. pp. 927–935.

Kalai, Adam Tauman and Ravi Sastry. 2009. The Isotron Algorithm: High-Dimensional Isotonic Regression. In *COLT*. Citeseer.

Khan, Sami and E Tamer. 2018. "Discussion of Simple Estimators for Invertible Index Models by H. Ahn, H. Ichimura, J. Powell, and P. Ruud." *Journal of Business & Economic Statistics* 36(1):11–15.

Khan, Shakeeb, Fu Ouyang, Elie Tamer et al. 2019. Inference on Semiparametric Multinomial Response Models. Technical report Boston College Department of Economics.

Klein, Roger W and Richard H Spady. 1993. "An efficient semiparametric estimator for binary response models." *Econometrica: Journal of the Econometric Society* pp. 387–421.

Komarova, Tatiana. 2013. "Binary choice models with discrete regressors: Identification and misspecification." *Journal of Econometrics* 177(1):14–33.

Kushner, Harold and G George Yin. 2003. *Stochastic approximation and recursive algorithms and applications*. Vol. 35 Springer Science & Business Media.

Lagarias, Jeffrey C, James A Reeds, Margaret H Wright and Paul E Wright. 1998. "Convergence properties of the Nelder–Mead simplex method in low dimensions." *SIAM Journal on optimization* 9(1):112–147.

Laitinen, Erkki K. 1991. "Financial ratios and different failure processes." *Journal of Business Finance & Accounting* 18(5):649–673.

Lee, Sangjae and Wu Sung Choi. 2013. "A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis." *Expert Systems with Applications* 40(8):2941–2946.

Lehmann, Erich L and George Casella. 2006. *Theory of point estimation.* Springer Science & Business Media.

Leshno, Moshe and Yishay Spector. 1996. “Neural network prediction analysis: The bankruptcy case.” *Neurocomputing* 10(2):125–147.

Lewbel, Arthur et al. 2012. *An overview of the special regressor method.* Boston College, Department of Economics.

Lo, Andrew W. 1986. “Logit versus discriminant analysis: A specification test and application to corporate bankruptcies.” *Journal of econometrics* 31(2):151–178.

Lorentz, G. 1986. “Approximation of functions. The second edition ed.” *American Mathematical Society, Rhode Island* .

Mai, Feng, Shaonan Tian, Chihoon Lee and Ling Ma. 2019. “Deep learning models for bankruptcy prediction using textual disclosures.” *European journal of operational research* 274(2):743–758.

Mammen, Enno. 1989. “Asymptotics with increasing dimension for robust regression with applications to the bootstrap.” *The annals of statistics* pp. 382–400.

Mammen, Enno. 1993. “Bootstrap and wild bootstrap for high dimensional linear models.” *The annals of statistics* pp. 255–285.

Manski, Charles F. 1975. “Maximum score estimation of the stochastic utility model of choice.” *Journal of econometrics* 3(3):205–228.

Manski, Charles F. 1984. Semiparametric Analysisi Of Discrete Response: Asymptotic Properties Of The Maximum Score Estimator. Technical report.

Martin, Daniel. 1977. “Early warning of bank failure: A logit regression approach.” *Journal of banking & finance* 1(3):249–276.

Martin, R and C Masreliez. 1975. “Robust estimation via stochastic approximation.” *IEEE Transactions on information Theory* 21(3):263–271.

Mensah, Yaw M. 1983. "The differential bankruptcy predictive ability of specific price level adjustments: some empirical evidence." *Accounting Review* pp. 228–246.

Meyer, Paul A and Howard W Pifer. 1970. "Prediction of bank failures." *The journal of finance* 25(4):853–868.

Moulines, Eric and Francis Bach. 2011. "Non-asymptotic analysis of stochastic approximation algorithms for machine learning." *Advances in neural information processing systems* 24.

Mustapha, Aatila, Lachgar Mohamed and Kartit Ali. 2020. An overview of gradient descent algorithm optimization in machine learning: application in the ophthalmology field. In *International Conference on Smart Applications and Data Analysis*. Springer pp. 349–359.

Nesterov, Yurii. 2003. *Introductory lectures on convex optimization: A basic course*. Vol. 87 Springer Science & Business Media.

Newey, Whitney K. 1994. "The asymptotic variance of semiparametric estimators." *Econometrica: Journal of the Econometric Society* pp. 1349–1382.

Newey, Whitney K. 1997. "Convergence rates and asymptotic normality for series estimators." *Journal of econometrics* 79(1):147–168.

Odom, Marcus D and Ramesh Sharda. 1990. A neural network model for bankruptcy prediction. In *1990 IJCNN International Joint Conference on neural networks*. IEEE pp. 163–168.

Ohlson, James A. 1980. "Financial ratios and the probabilistic prediction of bankruptcy." *Journal of accounting research* pp. 109–131.

Pagan, Adrian and Aman Ullah. 1999. *Nonparametric econometrics*. Cambridge university press.

Polyak, Boris T and Anatoli B Juditsky. 1992. "Acceleration of stochastic approximation by averaging." *SIAM journal on control and optimization* 30(4):838–855.

Portnoy, Stephen. 1984. "Asymptotic behavior of M-estimators of p regression parameters when p 2/n is large. I. Consistency." *The Annals of Statistics* pp. 1298–1309.

Portnoy, Stephen. 1985. "Asymptotic behavior of M estimators of p regression parameters when p2/n is large; II. Normal approximation." *The Annals of Statistics* pp. 1403–1417.

Portnoy, Stephen. 1988. "Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity." *The Annals of Statistics* pp. 356–366.

Powell, James L, James H Stock and Thomas M Stoker. 1989. "Semiparametric estimation of index coefficients." *Econometrica: Journal of the Econometric Society* pp. 1403–1430.

Press, S James and Sandra Wilson. 1978. "Choosing between logistic regression and discriminant analysis." *Journal of the American Statistical Association* 73(364):699–705.

Ravikumar, Pradeep, Martin Wainwright and Bin Yu. 2008. Single index convex experts: Efficient estimation via adapted bregman losses. In *Snowbird learning workshop*. Citeseer.

Robbins, Herbert and Sutton Monro. 1951. "A stochastic approximation method." *The annals of mathematical statistics* pp. 400–407.

Ruder, Sebastian. 2016. "An overview of gradient descent optimization algorithms." *arXiv preprint arXiv:1609.04747* .

Ruud, Paul A. 1983. "Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models." *Econometrica: Journal of the Econometric Society* pp. 225–228.

Ruud, Paul A. 1986. "Consistent estimation of limited dependent variable models despite misspecification of distribution." *Journal of Econometrics* 32(1):157–187.

Sharma, Subhash and Vijay Mahajan. 1980. "Early warning indicators of business failure." *Journal of marketing* 44(4):80–89.

Sherman, Robert P. 1993. "The limiting distribution of the maximum rank correlation estimator." *Econometrica: Journal of the Econometric Society* pp. 123–137.

Shin, Kyung-Shik, Taik Soo Lee and Hyun-jung Kim. 2005. "An application of support vector machines in bankruptcy prediction model." *Expert systems with applications* 28(1):127–135.

Shumway, Tyler. 2001. "Forecasting bankruptcy more accurately: A simple hazard model." *The journal of business* 74(1):101–124.

Spokoiny, Vladimir. 2012. "Parametric estimation. Finite sample theory." *The Annals of Statistics* 40(6):2877–2909.

Spokoiny, Vladimir. 2013. "Bernstein-von Mises Theorem for growing parameter dimension." *arXiv preprint arXiv:1302.3430* .

Sur, Pragya and Emmanuel J Candès. 2019. "A modern maximum-likelihood theory for high-dimensional logistic regression." *Proceedings of the National Academy of Sciences* 116(29):14516–14525.

Theodossiou, Panayiotis T. 1993. "Predicting shifts in the mean of a multivariate time series process: an application in predicting business failures." *Journal of the American Statistical Association* 88(422):441–449.

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1):267–288.

Toulis, Panos, Edoardo M Airoldi et al. 2017. "Asymptotic and finite-sample properties of estimators based on stochastic gradients." *The Annals of Statistics* 45(4):1694–1727.

Wang, Bin. 2004. *Strategy changes and internet firm survival.* University of Minnesota.

Wang, Hansheng. 2007. "A note on iterative marginal optimization: a simple algorithm for maximum rank correlation estimation." *Computational statistics & data analysis* 51(6):2803–2812.

White, Halbert. 1982. "Maximum likelihood estimation of misspecified models." *Econometrica: Journal of the econometric society* pp. 1–25.

White, Halbert. 1991. "Some Results for Sieve Estimation With Dependent Observa-tions." *Nonparametric and Semiparametric Methods in Econometrics and Statistics* .

Wilson, D Randall and Tony R Martinez. 2003. "The general inefficiency of batch training for gradient descent learning." *Neural networks* 16(10):1429–1451.

Yohai, Victor J and Ricardo A Maronna. 1979. "Asymptotic behavior of M-estimators for the linear model." *The Annals of Statistics* pp. 258–268.

Zhang, Cun-Hui and Jian Huang. 2008. "The sparsity and bias of the lasso selection in high-dimensional linear regression." *The Annals of Statistics* 36(4):1567–1594.

Zhao, Peng and Bin Yu. 2006. "On model selection consistency of Lasso." *The Journal of Machine Learning Research* 7:2541–2563.

Zmijewski, Mark E. 1984. "Methodological issues related to the estimation of financial distress prediction models." *Journal of Accounting research* pp. 59–82.

Zordan, Anthony Joseph. 1998. *Cash flow ratios as predictors of business failure.* Nova Southeastern University.