Estimating Stochastic Linear Combination of Non-linear Regressions

Di Wang*, **Xiangyu Guo, Chaowen Guan, Shi Li** and **Jinhui Xu**
Department of Computer Science and Engineering
State University of New York at Buffalo
Buffalo, NY 14260

Abstract

In this paper we study the problem of estimating stochastic linear combination of non-linear regressions, which has a close connection with many machine learning and statistical models such as non-linear regressions, the Single Index, Multi-index, Varying Coefficient Index Models and Two-layer Neural Networks. Specifically, we first show that with some mild assumptions, if the variate vector x is multivariate Gaussian, then there is an algorithm whose output vectors have ℓ_2-norm estimation errors of $O(\sqrt{\frac{p}{n}})$ with high probability, where p is the dimension of x and n is the number of samples. Then we extend our result to the case where x is sub-Gaussian using the zero-bias transformation, which could be seen as a generalization of the classic Stein's lemma. We also show that with some additional assumptions there is an algorithm whose output vectors have ℓ_∞-norm estimation errors of $O(\frac{1}{\sqrt{p}} + \sqrt{\frac{p}{n}})$ with high probability. Finally, for both Gaussian and sub-Gaussian cases we propose a faster sub-sampling based algorithm and show that when the sub-sample sizes are large enough then the estimation errors will not be sacrificed by too much. Experiments for both cases support our theoretical results. To the best of our knowledge, this is the first work that studies and provides theoretical guarantees for the stochastic linear combination of non-linear regressions model.

Introduction

We study the problem of estimating *stochastic linear combination of non-linear regressions*. The model can be formally defined as follows.

Definition 1 (Stochastic Linear Combination of Non-linear Regressions)**.** Given variates $x \in \mathbb{R}^p$ and $z_1, \cdots, z_k \in \mathbb{R}$ such that $\mathbb{E}[x] = 0$ and z_i's for all $i \in [k]$ are i.i.d random variables independent of x with $\mathbb{E}[z_i] = 0$ and $\text{Var}(z_i) = 1$, the response y is given by

$$y = \sum_{i=1}^{k} z_i f_i(\langle \beta_i^*, x \rangle) + \epsilon, \tag{1}$$

where $\beta_1^*, \beta_2^*, \cdots, \beta_k^* \in \mathbb{R}^p$ are unknown parameters, f_i's for all $i \in [k]$ are known (but could be non-convex) *link functions*, and ϵ is some random noise (from an unknown distribution) satisfying $\mathbb{E}[\epsilon] = 0$ and is independent of x and z_i's.

The goal is to estimate the parameters β_j^* for all $j \in [k]$ from n observations $(x_1, y_1, \{z_{1,i}\}_{i=1}^k)$, $(x_2, y_2, \{z_{2,i}\}_{i=1}^k), \cdots, (x_n, y_n, \{z_{n,i}\}_{i=1}^k)$.

This model has a close connection with many models in Statistics, Machine Learning, Signal Processing and Information Theory: (1) when $k = 1$, the model is reduced to the non-linear regression estimation problem which has been studied in (Zhang, Yang, and Wang 2018; Beck and Eldar 2013; Yang et al. 2016; Cook and Lee 1999) and is related to compressed sensing and image recovery as well; (2) when $k = 1$ but the link function f_1 is unknown, it becomes the **Single Index Model**, which is one of the most fundamental models in statistics and has been studied for many years (Ichimura 1993; Horowitz 2009; Kakade et al. 2011; Yang, Balasubramanian, and Liu 2017; Radchenko 2015); (3) when $k \geq 1$, z_i's are deterministic but f_i's are unknown, this model will be a special case of the **Multi-index Model** which has been studied in (Li 1991; Li 1992; Li, Duan, and others 1989; Yang et al. 2017); (4) when $k \geq 1$, z_i's are stochastic but f_i's are unknown, it will be the **Varying Coefficient Index Model** which was introduced by (Ma and Song 2015) and has wide applications in economics and medical science (Fan and Zhang 2008); (5) when all f_i's are the same, the model can be viewed as a **Two-layer Neural Network** with k hidden nodes and random hidden-output layer weights.

To estimate the parameters in Model (1), the main challenge is that without the assumption that f_i's are convex or similarities between them, it is hard to establish an objective function that can be efficiently optimized using optimization methods such as (Stochastic) Gradient Descent. Recently, some works including (Yang, Balasubramanian, and Liu 2017; Na et al. 2018; Yang et al. 2017) studied and proposed efficient algorithms for the Single Index, Multi-index and Varying Coefficient Index models using Stein's Lemma. Their theoretical guarantees are measured in terms of $\|\beta_j - c\beta_j^*\|_2, j \in [k]$, where β_j is the estimator for β_j^* and c is a constant depending on many parameters in the models (such as f_i's, β_j^*'s and the distribution for x). However there is a common issue related to the constant c in these

*The first two authors contributed equally.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

results: They did not provide a method to compute or even estimate c. To address this issue, we measure the error in terms of $\beta_j - \beta_j^*$ for all $j \in [k]$; that is, we do not introduce the constant c. The key question the paper tries to answer is:

Is there an efficient method whose output vectors β_1, $\beta_2, \cdots, \beta_k$ have small errors compared to $\beta_1^*, \beta_2^*, \cdots, \beta_k^*$?

In this paper, we answer the question in the affirmative under some mild assumptions on the model. Specifically, our contributions can be summarized as follows.

1. We first consider the case where x multivariate Gaussian. In this case, we show that there is a special structure for each β_j^*, $j \in [k]$: $\beta_j^* = c_j \beta_j^{ols}$, where c_j is a constant depending on the link function f_j and x, and β_j^{ols} is the Ordinary Lest Square estimator w.r.t yz_j and x, *i.e.,* $\beta_j^{ols} = (\mathbb{E}[xx^T])^{-1}\mathbb{E}[z_j yx]$. Based on this key observation, we propose an algorithm which estimates c_j's and β_j^{ols}'s, and outputs $\{\beta_j\}_{j=1}^k$ satisfying $\|\beta_j - \beta_j^*\|_2 \leq O(\sqrt{\frac{p}{n}})$ for each $j \in [k]$ with high probability. Moreover, in order to make our algorithm faster, instead of using linear regression estimator to approximate β_j^{ols}, we use the sub-sampling covariance linear regression estimator (Dhillon et al. 2013). We show that if the sub-sample size is large enough, the error bound is almost the same as in the previous ones.
2. We then extend our result to the case when x is (bounded) sub-Gaussian. The challenge is that the result for the Gaussian case depends on some properties of Gaussian distribution which are not satisfied in the sub-Gaussian case. To overcome this, we use the zero-bias transformation (Goldstein, Reinert, and others 1997), which could be seen as a generalization of the Stein's lemma (Brillinger 1982). Particularly, we show that instead of the equality $\beta_j^* = c_j \beta_j^{ols}$, we have the ℓ_∞ norm estimation error $\|\beta_j^* - c_j \beta_j^{ols}\|_\infty \leq O(\frac{1}{\sqrt{p}})$ with some additional mild assumptions. Based on this and the same idea from the Gaussian case, we show that there exists an algorithm whose output vectors $\{\beta_j\}_{j=1}^k$ satisfy $\|\beta_j - \beta_j^*\|_\infty \leq O(\frac{1}{\sqrt{p}} + \sqrt{\frac{p}{n}})$ with high probability. Similarly, we also propose a sub-sampled version of our algorithm as in the Gaussian case.
3. At the end, we show the experimental results on both Gaussian and sub-Gaussian cases with single/mixed type of link functions, and these results support our theoretical results above.

To the best of our knowledge, this is the first paper studying and providing the estimation error bound for Model (1) in both Gaussian and sub-Gaussian cases.

Due to the space limit, omitted proofs and the background are included in the full version of the paper. The source code of experiments can be found at github.com/anonymizepaper/SLSE.

Related Work

As we mentioned above, there is no previous work on Model (1) with guarantees on the ℓ_2 or ℓ_∞ norm of the errors $\beta_j - \beta_j^*$. Hence, below we compare with the results which are close to ours.

When the link functions f_j's are unknown, Model (1) is just the Varying Coefficient Index Model. (Na et al. 2018) provided the first efficient algorithm for this model. Although they considered the high dimensional sparse case, their method requires the underlying distribution of x to be known, an unrealistic assumption for most applications. Moreover, their estimation errors are measured by the differences between β_j's and $c\beta_j^*$'s for an unknown c, while in our results we have fixed $c = 1$.

When the link functions f_j's are all the same, then our model can be reduced to the two-layer neural network with random hidden-output layer weights. Previous work on the convergence results all focused on the gradient descent type of methods such as those in (Zhang et al. 2019; Zou et al. 2018; Nitanda and Suzuki 2019). However, our method is based on Stein's lemma and its generalization. Compared with the gradient descent type methods, our algorithm is non-interactive (that is, we do not need to update estimators in each iteration) and parameter-free (that is, we don not need to tune the step-size). Moreover, our method can be extended to the case where the link functions f_j's are different.

Our method is motivated by Stein's lemma (Brillinger 1982) and its generalization, the zero-bias transformation. Several previous studies have used Stein's Lemma in various machine learning problems. For example, (Erdogdu, Dicker, and Bayati 2016; Erdogdu 2016) used it to accelerate some optimization procedures, (Liu and Wang 2016) applied it to Bayesian inference and (Yang, Balasubramanian, and Liu 2017; Yang et al. 2016; Na et al. 2018; Wei, Yang, and Wang 2019) used it and its generalizations in the Single Index, Multi-index, Varying Coefficient Index and Generative models, respectively. The zero-bias transformation has also been used in (Erdogdu, Bayati, and Dicker 2019) for estimating the Generalized Linear Model. However, due to the difference between the models, these algorithms cannot be applied to our problem.

Preliminaries

In this section, we review some necessary definitions and lemmas.

Definition 2 (Sub-Gaussian)**.** For a given constant κ, a random variable $x \in \mathbb{R}$ is said to be sub-Gaussian if it satisfies $\sup_{m \geq 1} \frac{1}{\sqrt{m}} \mathbb{E}[|x|^m]^{\frac{1}{m}} \leq \kappa$. The smallest such κ is the sub-Gaussian norm of x and it is denoted by $\|x\|_{\psi_2}$.

Similarly, a random vector $x \in \mathbb{R}^p$ is called a sub-Gaussian vector if there exists a constant κ such that $\sup_{v \in S^{p-1}} \|\langle x, v\rangle\|_{\psi_2} \leq \kappa$, where S^{p-1} is the set of all p-dimensional unit vector.

In order to extend our results to the sub-Gaussian case, we will use the zero-bias transformation which is proposed by (Goldstein, Reinert, and others 1997). It is a generalization of the classic Stein's lemma in (Brillinger 1982).

Definition 3. Let z be a random variable with mean 0 and variance σ^2. Then there exists a random variable z^* such that for all differentiable functions f we have $\mathbb{E}[zf(z)] = \sigma^2 \mathbb{E}[f'(z^*)]$. The distribution of z^* is said to be the z-zero-bias distribution.

The standard Gaussian distribution is the unique distribution whose zero-bias distribution is itself. This is just the basic **Stein's lemma**.

Lemma 1. (Dhillon et al. 2013) Assume that $\mathbb{E}[x] = 0, \mathbb{E}[x_i x_i^T] = \Sigma \in \mathbb{R}^{p\times p}$, and $\Sigma^{-\frac{1}{2}}x$ and y are sub-Gaussian with norms κ_x and γ respectively. If $n \geq \Omega(\gamma\kappa_x p)$, then with probability at least $1 - 3\exp(-p)$ we have

$$\|\Sigma^{\frac{1}{2}}(\tilde{\beta}^{ols} - \beta^{ols})\|_2 \leq C_1\kappa_x\gamma\sqrt{\frac{p}{n}}, \tag{2}$$

where $\beta^{ols} = \Sigma^{-1}\mathbb{E}[yx]$ is the OLS estimator w.r.t y and x, $\tilde{\beta}^{ols} = (X^TX)^{-1}X^TY$ is the empirical one, and $C_1 > 0$ is some universal constant.

Lemma 2. (Erdogdu, Bayati, and Dicker 2019) Let $B^\delta(\tilde{\beta})$ denote the ball centered around $\tilde{\beta}$ with radius δ. For $i = 1, 2, \cdots, n$, let $x_i \in \mathbb{R}^p$ be i.i.d random vectors with a covariance matrix Σ. Given a function g that is uniformly bounded by L and G-Lipschitz, with probability at least $1 - \exp(-p)$ we have

$$\sup_{\beta\in B^\delta(\tilde{\beta})}\left|\frac{1}{n}\sum_{i=1}^n g(\langle x_i,\beta\rangle) - \mathbb{E}[g(\langle x,\beta\rangle)]\right| \leq 2(G(\|\tilde{\beta}\|_2+\delta)\|\Sigma\|_2 + L)\sqrt{\frac{p}{n}}.$$

Assumption 1. We assume that for each $j \in [k]$, the random variable yz_j is sub-Gaussian with its sub-Gaussian norm $\|yz_j\|_{\psi_2} = \gamma$.

Note that this assumption holds if y is bounded and z_j is sub-Gaussian or z_j is bounded and y is sub-Gaussian.

Assumption 2. We assume that there exist constants $G, L > 0$ such that for each $j \in [k]$, f_j' is G-Lipschitz and bounded by L. Also for $j \in [k]$, we let $\mathbb{E}[f_j'(\langle x, \beta_j^*\rangle)] \neq 0$.

Notations For a positive semi-definite matrix $M \in \mathbb{R}^{p\times p}$, we define the M-norm for a vector w as $\|w\|_M^2 = w^TMw$. Also we will denote $B_M^\delta(\tilde{\beta})$ as the ball around $\tilde{\beta}$ with radius δ under M-norm, *i.e.,* $B_M^\delta(\tilde{\beta}) = \{\beta : \|M^{\frac{1}{2}}(\beta - \tilde{\beta})\|_2 \leq \delta\}$. $\lambda_{\min}(M)$ is the minimal singular value of the matrix M. For a semi positive definite matrix $M \in \mathbb{R}^{p\times p}$, let its SVD be $M = U^T\Sigma U$, where $\Sigma = \text{diag}(\lambda_1, \cdots, \lambda_p)$, then $M^{\frac{1}{2}}$ is defined as $M^{\frac{1}{2}} = U^T\Sigma^{\frac{1}{2}}U$ with $\Sigma^{\frac{1}{2}} = \text{diag}(\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_p})$.

Gaussian Case

In this section we consider the case where x is sampled from some multivariate Gaussian distribution, then we will extend our idea to the sub-Gaussian distribution case in next section.

Our algorithm is based on the following key observation using some properties of the multivariate Gaussian distribution.

Theorem 1. Consider Model (1) in Definition 1 under Assumptions 1 and 2. Moreover, assume that the observations $\{x_i\}_{i=1}^n$ are i.i.d sampled from $\mathcal{N}(0, \Sigma)$. Then each $\beta_j^*, j \in [k]$ can be written as

$$\beta_j^* = c_j \times \beta_j^{ols}, \tag{3}$$

where $\beta_j^{ols} = \Sigma^{-1}\mathbb{E}[z_j yx]$ and c_j is the root of the function $l_j(c) - 1$ where

$$l_j(c) = c\mathbb{E}[f_j'(\langle x, \beta_j^{ols}\rangle c)]. \tag{4}$$

From Theorem 1 we can see that, in order to estimate β_j^*, it is sufficient to estimate the terms $\beta_j^{ols} = \Sigma^{-1}\mathbb{E}[z_j yx]$ and c_j. If we denote $z_j y$ as the response and x as the variate, then the term β_j^{ols} is just the **Ordinary Least Square (OLS)** estimator. Thus we can use its empirical form $\tilde{\beta}_j^{ols} = (\sum_{i=1}^n x_i^T x_i)^{-1}\sum_{i=1}^n z_{i,j}y_i x_i = (X^TX)^{-1}X^TY_j$ as an estimator, where $X = [x_1^T; x_2^T; \cdots; x_n^T] \in \mathbb{R}^{n\times d}$ is the data matrix and $Y_j = [z_{1,j}y_1, \cdots, z_{n,j}y_n]^T$ is the corresponding response vector.

After getting the estimator of β_j^{ols}, denoted by $\tilde{\beta}_j^{ols}$, we use it to approximate c_j. That is we find the root $\hat{c}_j$ of the empirical version of $l_j(c) - 1$, *i.e.,* $\hat{l}_j(c) - 1$, where

$$\hat{l}_j(c) = \frac{c}{n}\sum_{i=1}^n [f_j'(\langle x_i, \tilde{\beta}_j^{ols}\rangle c)].$$

Note that there are numerous methods available to find a root of a function, such as Newton's root-finding method with quadratic convergence and Halley's method with cubic convergence. We also note that this step only cost $O(n)$ per-iteration. After that, we could estimate each β_j^* by $\hat{\beta}_j^{nlr} = \hat{c}_j\tilde{\beta}_j^{ols}$. In total, we have Algorithm 1.

Algorithm 1 SLS: Scaled Least Squared Estimators

Input: Data $\{(x_i, y_i, \{z_{i,j}\}_{j=1}^k)\}_{i=1}^n$, link functions $\{f_j\}_{j\in[k]}$.

1: **Option 1:** Let $X = [x_1, x_2, \cdots, x_n]^T \in \mathbb{R}^{n\times p}$ and compute the $\hat{\Sigma}^{-1} = (X^TX)^{-1}$.
2: **Option 2:** Construct a sub-sampling based OLS estimator, that is let $S \subset [n]$ be a random subset and take $\hat{\Sigma}_S^{-1} = \frac{|S|}{n}(X_S^TX_S)^{-1}$, where $X_S \in \mathbb{R}^{|S|\times p}$ is the data matrix constrained on indices of S.
3: **for** $j = 1, 2\cdots, k$ **do**
4: Let $Y_i = [z_{1,j}y_1, \cdots, z_{n,j}y_n]^T \in \mathbb{R}^n$. For **Option 1**, Compute the ordinary least squares estimator $\tilde{\beta}_j^{ols} = (\Sigma)^{-1}X^TY_j$. For **Option 2**, take $\tilde{\beta}_j^{ols} = (\hat{\Sigma}_S)^{-1}X^TY_j$.
5: Denote $\tilde{y}_j = X\tilde{\beta}_j^{ols}$. Then use the Newton's root-finding method to the function $\frac{c}{n}\sum_{i=1}^n[f_j'(\tilde{y}_{j,i}c)] - 1$, denote the root as $\hat{c}_j$:
6: **for** $t = 1, 2, \cdots$ until convergence **do**
7: $c = c - \frac{c\frac{1}{n}\sum_{i=1}^n f_j'(c\tilde{y}_{j,i}) - 1}{\frac{1}{n}\sum_{i=1}^n\{f_j'(c\tilde{y}_{j,i}) + c\tilde{y}_{j,i}f_j^{(2)}(c\tilde{y}_{j,i})\}}$.
8: $\hat{\beta}_j^{nlr} = \hat{c}_j \cdot \tilde{\beta}_j^{ols}$.
9: **return** $\left(\hat{\beta}_j^{nlr}\right)_{j\in[k]}$

The following theorem shows that the converge rate of the estimation error for each $\|\hat{\beta}_j^{nlr} - \beta_j^*\|_2$ is $O(\sqrt{\frac{p}{n}})$ under some additional mild assumptions on link functions $\{f_j\}_{j=1}^k$.

Theorem 2. Consider Option 1 in Algorithm 1. Under the Assumptions 1, 2 and the assumptions in Theorem 1, for each $j \in [k]$ we define the function $\ell_j(c, \beta) = c\mathbb{E}[f_j'(\langle x, \beta\rangle c)]$ and its empirical counter part as

$$\hat{\ell}_j(c, \beta) = \frac{c}{n}\sum_{i=1}^{n} f_j'(\langle x_i, \beta\rangle c).$$

Assume that there exist some constants $\eta, \bar{c}_j$ such that $\ell_j(\bar{c}_j, \beta_j^{ols}) > 1 + \eta$. Then there exists $c_j > 0$ satisfying the equation $1 = \ell_j(c_j, \beta_j^{ols})$ for each $j \in [k]$.

Further, assume that n is sufficiently large:

$$n \geq \Omega(p\|\Sigma\|_2\|\beta_j^*\|_2^2)$$

Then, with probability at least $1 - k\exp(-p)$ there exist constants $\hat{c}_j \in (0, \bar{c}_j)$ satisfying the equations

$$1 = \frac{\hat{c}_j}{n}\sum_{i=1}^{n} f_j'(\langle x_i, \tilde{\beta}_j^{ols}\rangle \hat{c}_j).$$

Moreover, if for all $j \in [k]$ the derivative of $z \mapsto \ell_j(z, \beta_j^{ols})$ is bounded below in absolute value (*does not change sign*) by $M > 0$ in the interval $z \in [0, c_j]$. Then with probability at least $1 - 4k\exp(-p)$ the outputs $\{\hat{\beta}_j^{nlr}\}_{j=1}^k$ satisfy for each $j \in [k]$

$$\|\hat{\beta}_j^{nlr} - \beta_j^*\|_2 \leq O(\max\{1, \|\beta_j^*\|_2\}\lambda_{\min}^{-\frac{1}{2}}(\Sigma)\sqrt{\frac{p}{n}}), \quad (5)$$

where $G, L, \gamma, M, c_j, \eta$ are assumed to be $\Theta(1)$ and thus omitted in the Big-O and Ω notations (see Appendix for the explicit forms).

Note that in Theorem 2 the link functions f_j are not required to be convex. Hence this is quite useful in non-convex learning models.

Time Complexity Analysis Under Option 1 of Algorithm 1, we can see that the first step takes $O(np^2 + p^3)$ time, calculating $\tilde{\beta}_j^{ols}$ for all $j \in [k]$ takes $O(k(np + p^2))$ time and each iteration of finding $\hat{c}_j$ takes $O(n)$ time. Thus, if k ,the number of link functions f_j, is a constant, then the total time complexity is $O(np^2 + p^3 + nT)$, where T is the number of iterations for finding c_j.

However, the term np^2 is prohibitive in the large scale setting where n, p are huge (see (Wang and Xu 2018) for details). To further reduce the time complexity, we propose another estimator based on sub-sampling.

Note that the term $O(np^2)$ comes from calculating the empirical covariance matrix X^TX. Thus, to reduce the time complexity, instead of calculating the covariance via the whole dataset, we here use the sub-sampled covariance matrix. More precisely, we first randomly sample a set of indices $S \subset [n]$ whose size $|S|$ will be specified later. Then we calculate $\frac{|S|}{n}(X_S^TX_S)^{-1}$ to estimate $(X^TX)^{-1}$, where $X_S \in \mathbb{R}^{|S|\times p}$ is the data matrix constrained on indices of S. We can see that the time complexity in this case will only be $O(|S|p^2 + p^3)$. The following lemma, which is given by (Dhillon et al. 2013; Erdogdu, Dicker, and Bayati 2016) shows the convergence rate of the OLS estimator based on the sub-sampled covariance matrix. This is a generalization of Lemma 1.

Lemma 3. Under the same assumptions as in Lemma 1, if $|S| \geq \Omega(\gamma\kappa_x p)$, then with probability at least $1 - 3\exp(-p)$ the sub-sampled covariance OLS estimator $\tilde{\beta}^{ols} = \frac{|S|}{n}(X_S^TX_S)^{-1}X^TY$ satisfies

$$\|\tilde{\beta}^{ols} - \beta^{ols}\|_2 \leq C_2\kappa_x\gamma\sqrt{\frac{p}{|S|}}.$$

We have the following approximation error based the sub-sampled covariance OLS estimator:

Theorem 3. Under the same assumptions as in Theorem 2, in Algorithm 1 if we use Option 2 with $|S| \geq \Omega(\gamma\kappa_x p)$, then with probability at least $1 - 4k\exp(-p)$ the outputs $\{\hat{\beta}_j^{nlr}\}_{j=1}^k$ satisfy for each $j \in [k]$, $\|\hat{\beta}_j^{nlr} - \beta_j^*\|_2 \leq O(\max\{1, \|\beta_j^*\|_2\}\lambda_{\min}^{-\frac{1}{2}}(\Sigma)\sqrt{\frac{p}{|S|}})$.

Moreover, it is also possible to accelerate the algorithm using the sub-sampling method in the step 5 (finding the root) and we can see the estimation error will be the same as in Theorem 3 (by the proof of Theorem 3). Due to the space limit, we omit it here.

Extension to Sub-Gaussian Case

Note that Theorem 2 is only suitable for the case when x is Gaussian. This is due to the requirements on some properties of the Gaussian distribution in the proof of Theorem 1. In this section we will first extend Theorem 1 to the sub-Gaussian case.

Remember that the proof of Theorem 1 is based on the classic Stein's lemma. Thus, in order to generalize to sub-Gaussian case, we will use the zero-bias transformation in Definition 3 since it is a generalization of the Stein's lemma. With some additional assumptions, we can get the following theorem.

Theorem 4. Let $x_1, \cdots, x_n \in \mathbb{R}^p$ be i.i.d realizations of a random vector x which is sub-Gaussian with zero mean, whose covariance matrix Σ has $\Sigma^{\frac{1}{2}}$ being diagonally dominant [1], and whose distribution is supported on a ℓ_2-norm ball of radius r. Let $v = \Sigma^{-\frac{1}{2}}x$ be the whitened random vector of x with sub-Gaussian norm $\|v\|_{\psi_2} = \kappa_x$. If for all $j \in [k]$, each v has constant first and second conditional moments (*i.e.,* $\forall s \in [p]$ and $\tilde{\beta}_j = \Sigma^{\frac{1}{2}}\beta_j^*$, $\mathbb{E}[v_s|\sum_{t\neq s}\tilde{\beta}_j v_t]$ and $\mathbb{E}[v_s^2|\sum_{t\neq s}\tilde{\beta}_j v_t]$ are deterministic) and the link functions f_j' satisfy Assumption 2. Then for $c_j = \frac{1}{\mathbb{E}[f_j'(\langle x, \beta_j^*\rangle)]}$, the following holds for the model in (1) for all $j \in [k]$

$$\|\frac{1}{c_j}\cdot\beta_j^* - \beta_j^{ols}\|_\infty \leq 16Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|\beta_j^*\|_\infty^2}{\sqrt{p}}, \quad (6)$$

[1] A square matrix is said to be diagonally dominant if, for every row of the matrix, the magnitude of the diagonal entry in a row is larger than or equal to the sum of the magnitudes of all the other (non-diagonal) entries in that row.

where ρ_q for $q = \{2, \infty\}$ is the conditional number of Σ in ℓ_q norm and $\beta_j^{ols} = \Sigma^{-1}\mathbb{E}[xyz_j]$ is the OLS vector w.r.t yz_j and x.

Remark 1. Note that compared with the equality relationship between β_j^* and $c_j\beta^{ols}$ in Theorem 1, in Theorem 4 we only has the ℓ_∞ norm of their difference. Also, here we need more assumptions on the distribution of x, and these assumptions ensure that the estimation error decays in the rate of $O(\frac{1}{\sqrt{p}})$.

Theorem 4 indicates that we can use the same idea as in the Gaussian case to estimate each β_j^*. Note that the forms of c_j in Theorem 1 and 4 are different. In Theorem 1 each c_j is based on β_j^{ols}, while in Theorem 4 it is based on β_j^*. However, due to the closeness of β_j^* and β_j^{ols} in (6), we can still use $\frac{1}{\mathbb{E}[f_j'(\langle x_i, \beta_j^{ols}\rangle \tilde{c}_j)]}$ to approximate c_j, where $\tilde{c}_j$ is the root of $c\mathbb{E}[f_j'(\langle x_i, \beta_j^{ols}\rangle c)] - 1$. Because of this similarity, we can still use Algorithm 1 for the sub-Gaussian case under the assumptions in Theorem 4, and we can get the following estimation error:

Theorem 5. Consider Option 1 in Algorithm 1. Under Assumptions 1, 2 and the assumptions in Theorem 4, for each $j \in [k]$, if we define the function $\ell_j(c, \beta) = c\mathbb{E}[f_j'(\langle x, \beta\rangle c)]$ and its empirical counter part as

$$\hat{\ell}_j(c, \beta) = \frac{c}{n}\sum_{i=1}^{n} f_j'(\langle x_i, \beta\rangle c).$$

Assume that there exist some constants $\eta, \bar{c}_j$ such that $\ell_j(\bar{c}_j, \beta_j^{ols}) > 1 + \eta$. Then there exists $\tilde{c}_j > 0$ satisfying the equation $1 = \ell_j(\tilde{c}_j, \beta_j^{ols})$ for each $j \in [k]$.

Further, assume that n is sufficiently large:

$$n \geq \Omega(\|\Sigma\|_2 p^2 \rho_2 \rho_\infty^2 \|\beta_j^*\|_\infty^2 \max\{1, \|\beta_j^*\|_\infty^2\}).$$

Then, with probability at least $1 - k\exp(-p)$ there exist constants $\hat{c}_j \in (0, \bar{c}_j)$ satisfying the equations

$$1 = \frac{\hat{c}_j}{n}\sum_{i=1}^{n} f_j'(\langle x_i, \tilde{\beta}_j^{ols}\rangle \hat{c}_j).$$

Moreover, if for all $j \in [k]$, the derivative of $z \mapsto \ell_j(z, \beta_j^{ols})$ is bounded below in absolute value (*does not change sign*) by $M > 0$ in the interval $z \in [0, \max\{\bar{c}_j, c_j\}]$. Then with probability at least $1 - 4k\exp(-p)$ the outputs $\{\hat{\beta}_j^{nlr}\}_{j=1}^k$ satisfy for each $j \in [k]$

$$\|\hat{\beta}_j^{nlr} - \beta_j^*\|_\infty \leq O\big(\sqrt{\rho_2}\rho_\infty \lambda_{\min}^{-\frac{1}{2}}(\Sigma)\sqrt{\frac{p}{n}}\|\beta_j^*\|_\infty$$
$$\times \max\{1, \|\beta_j^*\|_\infty\} + \rho_2\rho_\infty^2 \frac{\max\{\|\beta_j^*\|_\infty^2, 1\}\|\beta_j^*\|_\infty^2}{\sqrt{p}}\big),$$

where $\eta, G, L, \gamma, M, \bar{c}_j, r, \kappa_x, c_j$ are assumed to be $\Theta(1)$ and thus omitted in the Big-O and Ω notations (see Appendix for the explicit forms).

Draft Proof. We first prove the following lemma (see Appendix for the proof):

Lemma 4. Under the assumptions in Theorem 5, with probability at least $1 - k\exp(-p)$ for all $j \in [k]$

$$|\hat{c}_j - \tilde{c}_j| \leq O(M^{-1}GL\bar{c}_j^2\kappa_x^2\gamma\|\Sigma\|_2^{1/2}\|\beta_j^{ols}\|_2\sqrt{\frac{p}{n}}).$$

For convenience we will assume G, L, γ, M, $\bar{c}_j$, r, κ_x, $c_j = \Theta(1)$.

We have for each $\hat{\beta}_j^{nlr}$,

$$\|\hat{\beta}_j^{nlr} - \beta_j^*\|_\infty \leq \|\hat{c}_j\tilde{\beta}_j^{ols} - \tilde{c}_j\beta_j^{ols}\|_\infty + \|\tilde{c}_j\beta_j^{ols} - \beta_j^*\|_\infty$$
$$\leq \|\hat{c}_j\tilde{\beta}_j^{ols} - \tilde{c}_j\beta_j^{ols}\|_\infty + \|\tilde{c}_j\beta_j^{ols} - c_j\beta_j^{ols}\|_\infty$$
$$+ \|c_j\beta_j^{ols} - \beta_j^*\|_\infty. \tag{7}$$

To bound the terms in (7) we first bound $|\tilde{c}_j - c_j|$. By definition we have $c_j\mathbb{E}[f_j'(\langle x, \beta_j^*\rangle)] = 1$ and $\tilde{c}_j\mathbb{E}[f_j'(\langle x, \beta_j^{ols}\rangle\tilde{c}_j)] = 1$, then we get

$$\begin{aligned}
&|\ell_j(\tilde{c}_j, \beta_j^{ols}) - \ell_j(c_j, \beta_j^{ols})| = |1 - \ell_j(c_j, \beta_j^{ols})| \\
&= |c_j\mathbb{E}[f_j'(\langle x, \beta_j^*\rangle)] - c_j\mathbb{E}[f_j'(\langle x, \beta_j^{ols}\rangle c_j)]| \\
&\leq G|c_j|\mathbb{E}[\langle x, \beta_j^* - c_j\beta_j^{ols}\rangle] \\
&\leq Gc_j\|\beta_j^* - c_j\beta_j^{ols}\|_\infty\mathbb{E}\|x\|_1 \\
&\leq Gc_j\kappa_x\|\beta_j^* - c_j\beta_j^{ols}\|_\infty,
\end{aligned}$$

where the last inequality is by the definition of the sub-Gaussian norm (see Definition 1). Thus, by the assumption of the bounded deviation of $\ell(c, \beta_j^{ols})$ on $\max\{\bar{c}_j, c_j\}$ we have

$$\begin{aligned}
M|\tilde{c}_j - c_j| &\leq |\ell_j(\tilde{c}_j, \beta_j^{ols}) - \ell_j(c_j, \beta_j^{ols})| \\
&\leq Gc_j\kappa_x\|\beta_j^* - c_j\beta_j^{ols}\|_\infty,
\end{aligned}$$

and further by Theorem 4 we have

$$|\tilde{c}_j - c_j| \leq O(\sqrt{\rho_2}\rho_\infty\frac{\|\beta_j^*\|_\infty^2}{\sqrt{p}}). \tag{8}$$

For the second term in (7), by (8) we have

$$\|\tilde{c}_j\beta_j^{ols} - c_j\beta_j^{ols}\|_\infty \leq O(\sqrt{\rho_2}\rho_\infty\frac{\|\beta_j^{ols}\|_\infty\|\beta_j^*\|_\infty^2}{\sqrt{p}}), \tag{9}$$

where the last inequality is due to Lemma 4, Lemma 1 and the assumption of n.

For the first term in (7) we have

$$\begin{aligned}
&\|\hat{c}_j\tilde{\beta}_j^{ols} - \tilde{c}_j\beta_j^{ols}\|_\infty \\
&\leq \hat{c}_j\|\tilde{\beta}_j^{ols} - \beta_j^{ols}\|_\infty + |\hat{c}_j - \tilde{c}_j|\|\beta_j^{ols}\|_\infty \\
&\leq O\big((\tilde{c}_j + \|\Sigma\|_2^{1/2}\|\beta_j^{ols}\|_2\sqrt{\frac{p}{n}}) \times \lambda_{\min}(\Sigma)^{-\frac{1}{2}}\sqrt{\frac{p}{n}} \\
&+ \|\Sigma\|_2^{1/2}\|\beta_j^{ols}\|_2\sqrt{\frac{p}{n}}\|\beta_j^{ols}\|_\infty\big) \qquad (10) \\
&= O\big(\lambda_{\min}^{-\frac{1}{2}}(\Sigma)\sqrt{\frac{p}{n}}\max\{1, \|\beta_j^{ols}\|_\infty\}\big). \qquad (11)
\end{aligned}$$

By Theorem 4, we see that the third term of (7) is bounded as the following

$$\|c_j\beta_j^{ols} - \beta_j^*\|_\infty \leq O(\sqrt{\rho_2}\rho_\infty\frac{\|\beta_j^*\|_\infty^2}{\sqrt{p}}). \tag{12}$$

(9), (11) and (12) together give

$$\|\hat{\beta}_j^{nlr} - \beta_j^*\|_\infty \leq O\big(\sqrt{\rho_2}\rho_\infty \frac{\|\beta_j^{ols}\|_\infty \|\beta_j^*\|_\infty^2}{\sqrt{p}} + \lambda_{\min}^{-\frac{1}{2}}(\Sigma)\sqrt{\frac{p}{n}}\max\{1, \|\beta_j^{ols}\|_\infty\} + \sqrt{\rho_2}\rho_\infty \frac{\|\beta_j^*\|_\infty^2}{\sqrt{p}}\big). \tag{13}$$

By Theorem 4, we have

$$\|\beta_j^{ols}\|_\infty \leq O(\|\beta_j^*\|_\infty + \sqrt{\rho_2}\rho_\infty \frac{\|\beta_j^*\|_\infty^2}{\sqrt{p}}).$$

Plugging it into (13) completes the proof. □

Remark 2. Compared with the converge rate in the ℓ_2-norm of $O(\sqrt{\frac{p}{n}})$ in Theorem 2, Theorem 5 shows that for the sub-Gaussian case, the converge rate of the estimation error is $O(\frac{1}{\sqrt{p}} + \sqrt{\frac{p}{n}})$ in the ℓ_∞-norm (if we omit other terms). This is due to the estimation error in Theorem 4. Moreover, compared with the assumptions of link functions in Theorem 2, there are additional assumptions in Theorem 5.

In order to reduce the time complexity and make the algorithm faster, we can also use the sub-sampled covariance OLS estimator. This is the same as that in the Gaussian case.

Theorem 6. Under the same assumptions as in Theorem 5,if we use Option 2 in Algorithm 1, then with probability at least $1 - 4k\exp(-p)$, the outputs $\{\hat{\beta}_j^{nlr}\}_{j=1}^k$ satisfy for each $j \in [k]$

$$\|\hat{\beta}_j^{nlr} - \beta_j^*\|_\infty \leq O\big(\sqrt{\rho_2}\rho_\infty \lambda_{\min}^{-\frac{1}{2}}(\Sigma)\sqrt{\frac{p}{|S|}}\|\beta_j^*\|_\infty \times \max\{1, \|\beta_j^*\|_\infty\} + \rho_2\rho_\infty^2 \frac{\max\{\|\beta_j^*\|_\infty^2, 1\}\|\beta_j^*\|_\infty^2}{\sqrt{p}}\big).$$

Experiments

We conduct experiments on three types of link functions: polynomial, sigmoid, and logistic function, as well as an arbitrary mix of them. Formally, the polynomial link functions include $f(x) = x, x^3, x^5$; the sigmoid link function is defined as $f(x) = (1 + e^{-x})^{-1}$; the logistic link function refers to $f(x) = \log(1 + e^{-x})$. Due to the statistical setting we focused on in the paper, we will perform our algorithm on the synthetic data, and the same experimental setting has been used in the previous work such as (Na et al. 2018; Yang et al. 2017; Erdogdu, Dicker, and Bayati 2016).

Experimental setting. We sample all coefficient $z_{i,j}$ and noise ϵ i.i.d. from standard Gaussian distribution $N(0, 1)$ across each experiment. Each β_j^* is generated by sampling from $N(1, 16\mathbb{I}_d)$. We consider two distributions for generating x: Gaussian and Uniform distribution (corresponds to thr sub-Gaussian case). To satisfy the requirement of Theorem 6, in both cases the standard variance is scaled by $1/p$ and this is also used in (Erdogdu, Dicker, and Bayati 2016), where p is the data dimension. Thus, in the Gaussian case, $x \sim N(0, \frac{1}{p}\mathbb{I}_p)$, while in the sub-Gaussian case x is sampled from a uniform distribution, *i.e.,* $x \sim U([-1/p, 1/p])^p$. Finally, given the list $F = [f_1, \ldots, f_k]$ of link functions, the response y is computed via (1). It is notable that the experimental results with different number of link functions k are incomparable since when k is changed Model (1) will also be changed.

These experiments are divided into two parts, examining how the sample size n and the size of the sub-sample set S affect the algorithm performance. In the first part we vary n from 100 000 to 500 000 with fixed $p = 20$ and $|S| = n$, while in the second part we vary $|S|$ from $0.01n$ to n, with fixed $n = 500\,000$ and $p = 20$. In each part we test the algorithm against various data distribution/link function combinations. For each experiment, in order to support our theoretical analysis, we will use the (maximal) relative error as the measurement, that is when x is Gaussian we use $\max_{i\in[k]} \frac{\|\beta_i - \beta_i^*\|_2}{\|\beta_i^*\|_2}$ and when x is sub-Gaussian we will use $\max_{i\in[k]} \frac{\|\beta_i - \beta_i^*\|_\infty}{\|\beta_i^*\|_\infty}$. For each experiment we repeat 20 times and take the average as the final output.

Experiment results. Each of Figure 1-3 illustrates the result for a single type of link function. We can see that the relative error decreases steadily as the sample size n grows which is due to the $O(\frac{1}{\sqrt{n}})$ converge rate as our theorem states. Also, we can see that the size of S doesn't affect the final relative error much if $|S|$ is large enough, *i.e., in all cases, choosing large enough $|S| \geq 0.2n$ is sufficient to achieve a relative error roughly the same as when $|S| = n$,* which also has been mentioned in our theorems theoretically.

We further investigate the case when F contains different types of link functions. In Figure 4a, we let F contain polynomials with different degrees (x, x^3, x^5), and there are roughly $\frac{k}{3}$ functions for each degree. Similarly, in Figure 4b we also mix polynomial links with the other two types of links, i.e., logistic link and log-exponential link, and there are roughly $\frac{k}{3}$ functions for each type of link function. Our algorithm achieves similar performance as in the previous settings.

Conclusion

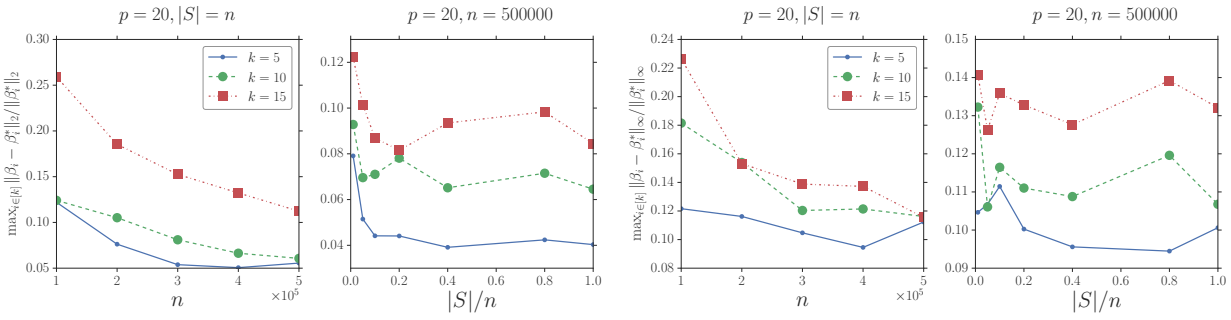
We studied a new model called *stochastic linear combination of non-linear regressions* and provided the first estimation error bounds for both Gaussian and bounded sub-Gaussian cases. Our algorithm is based on Stein's lemma and its generalization, the zero-bias transformation. Moreover, we used the sub-sampling of the covariance matrix to accelerate our algorithm. Finally, we conducted experiments whose results support our theoretical analysis.

Ackonwledgements

The research of the first and last authors was supported in part by NSF through grants CCF-1716400 and IIS-1910492.

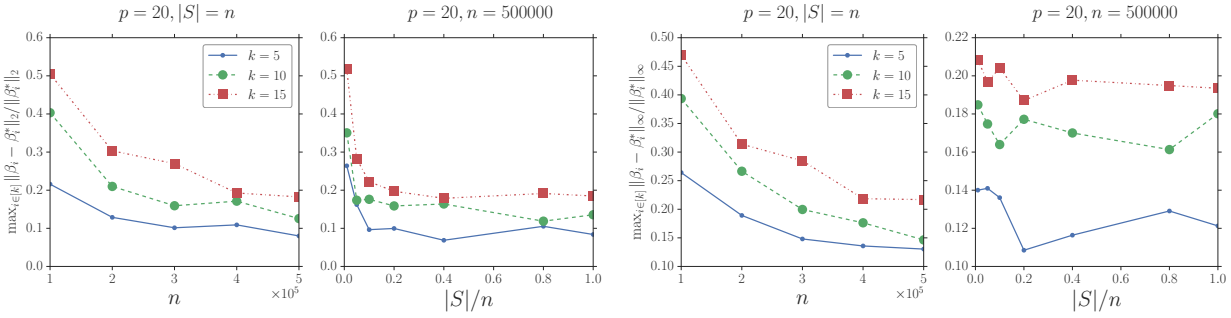
References

[Beck and Eldar 2013] Beck, A., and Eldar, Y. C. 2013. Sparse signal recovery from nonlinear measurements. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5464–5468. IEEE.



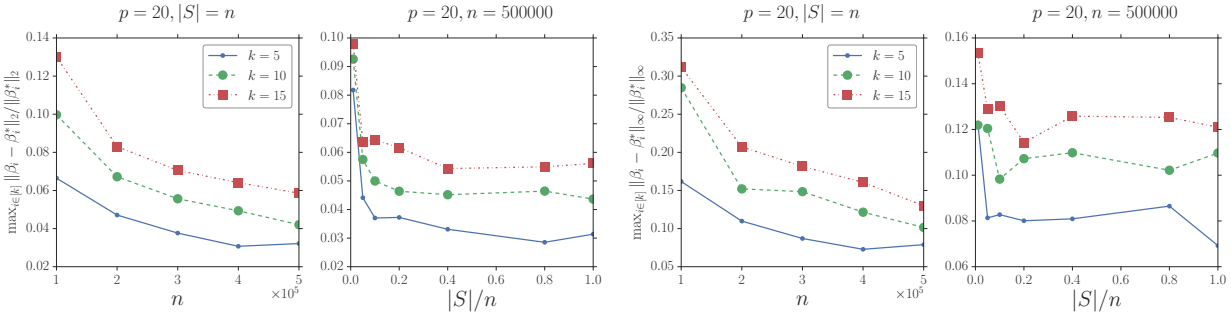
(a) Gaussian data (b) sub-Gaussian data

Figure 1: Single type of link function $f(x) = x^3$



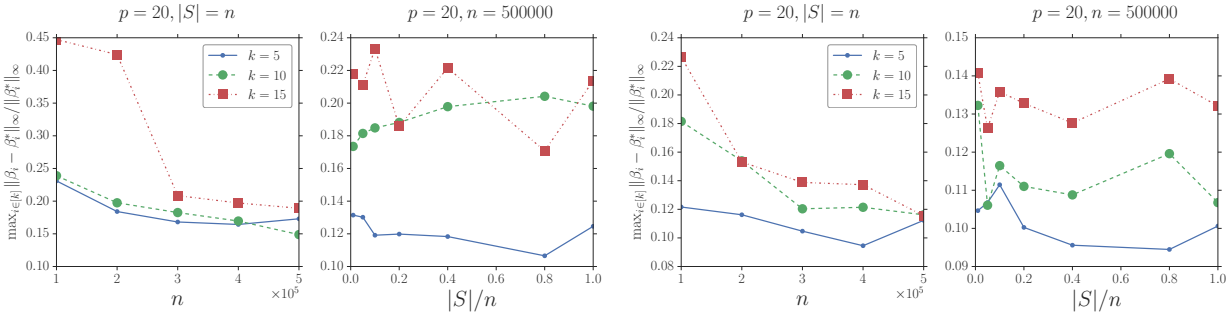
(a) Gaussian data (b) sub-Gaussian data

Figure 2: Single type of link function $f(x) = (1 + e^{-x})^{-1}$



(a) Gaussian data (b) sub-Gaussian data

Figure 3: Single type of link function $f(x) = \log(1 + e^{-x})$



(a) $f(x) = x, x^3, x^5$, sub-Gaussian data (b) $f(x) = x^3, (1 + e^{-x})^{-1}, \log(1 + e^{-x})$, sub-Gaussian data

Figure 4: Mixed different type of link functions

[Brillinger 1982] Brillinger, D. R. 1982. A generalized linear model with" gaussian" regressor variables. *A Festschrift For Erich L. Lehmann* 97.

[Cook and Lee 1999] Cook, R. D., and Lee, H. 1999. Dimension reduction in binary response regression. *Journal of the American Statistical Association* 94(448):1187–1200.

[Dhillon et al. 2013] Dhillon, P.; Lu, Y.; Foster, D. P.; and Ungar, L. 2013. New subsampling algorithms for fast least squares regression. In *Advances in neural information processing systems*, 360–368.

[Erdogdu, Bayati, and Dicker 2019] Erdogdu, M. A.; Bayati, M.; and Dicker, L. H. 2019. Scalable approximations for generalized linear problems. *The Journal of Machine Learning Research* 20(1):231–275.

[Erdogdu, Dicker, and Bayati 2016] Erdogdu, M. A.; Dicker, L. H.; and Bayati, M. 2016. Scaled least squares estimator for glms in large-scale problems. In *Advances in Neural Information Processing Systems*, 3324–3332.

[Erdogdu 2016] Erdogdu, M. A. 2016. Newton-stein method: an optimization method for glms via stein's lemma. *The Journal of Machine Learning Research* 17(1):7565–7616.

[Fan and Zhang 2008] Fan, J., and Zhang, W. 2008. Statistical methods with varying coefficient models. *Statistics and its Interface* 1(1):179.

[Goldstein, Reinert, and others 1997] Goldstein, L.; Reinert, G.; et al. 1997. Stein's method and the zero bias transformation with application to simple random sampling. *The Annals of Applied Probability* 7(4):935–952.

[Horowitz 2009] Horowitz, J. L. 2009. *Semiparametric and nonparametric methods in econometrics*, volume 12. Springer.

[Ichimura 1993] Ichimura, H. 1993. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics* 58(1-2):71–120.

[Kakade et al. 2011] Kakade, S. M.; Kanade, V.; Shamir, O.; and Kalai, A. 2011. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, 927–935.

[Li, Duan, and others 1989] Li, K.-C.; Duan, N.; et al. 1989. Regression analysis under link violation. *The Annals of Statistics* 17(3):1009–1052.

[Li 1991] Li, K.-C. 1991. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* 86(414):316–327.

[Li 1992] Li, K.-C. 1992. On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association* 87(420):1025–1039.

[Liu and Wang 2016] Liu, Q., and Wang, D. 2016. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, 2378–2386.

[Ma and Song 2015] Ma, S., and Song, P. X.-K. 2015. Varying index coefficient models. *Journal of the American Statistical Association* 110(509):341–356.

[Na et al. 2018] Na, S.; Yang, Z.; Wang, Z.; and Kolar, M. 2018. High-dimensional index volatility models via stein's identity. *arXiv preprint arXiv:1811.10790*.

[Nitanda and Suzuki 2019] Nitanda, A., and Suzuki, T. 2019. Refined generalization analysis of gradient descent for over-parameterized two-layer neural networks with smooth activations on classification problems. *arXiv preprint arXiv:1905.09870*.

[Radchenko 2015] Radchenko, P. 2015. High dimensional single index models. *Journal of Multivariate Analysis* 139:266–282.

[Wang and Xu 2018] Wang, D., and Xu, J. 2018. Large scale constrained linear regression revisited: Faster algorithms via preconditioning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[Wei, Yang, and Wang 2019] Wei, X.; Yang, Z.; and Wang, Z. 2019. On the statistical rate of nonlinear recovery in generative models with heavy-tailed data. In *International Conference on Machine Learning*, 6697–6706.

[Yang, Balasubramanian, and Liu 2017] Yang, Z.; Balasubramanian, K.; and Liu, H. 2017. High-dimensional non-gaussian single index models via thresholded score function estimation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3851–3860. JMLR. org.

[Yang et al. 2016] Yang, Z.; Wang, Z.; Liu, H.; Eldar, Y.; and Zhang, T. 2016. Sparse nonlinear regression: parameter estimation under nonconvexity. In *International Conference on Machine Learning*, 2472–2481.

[Yang et al. 2017] Yang, Z.; Balasubramanian, K.; Wang, Z.; and Liu, H. 2017. Learning non-gaussian multi-index model via second-order stein's method. *Advances in Neural Information Processing Systems* 30:6097–6106.

[Zhang et al. 2019] Zhang, X.; Yu, Y.; Wang, L.; and Gu, Q. 2019. Learning one-hidden-layer relu networks via gradient descent. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1524–1534.

[Zhang, Yang, and Wang 2018] Zhang, K.; Yang, Z.; and Wang, Z. 2018. Nonlinear structured signal estimation in high dimensions via iterative hard thresholding. In *International Conference on Artificial Intelligence and Statistics*, 258–268.

[Zou et al. 2018] Zou, D.; Cao, Y.; Zhou, D.; and Gu, Q. 2018. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*.