

Article

Optimal Subsampling for Upper Expectation Parametric Regression

Zhaolei Liu

Zhongtai Securities Institute for Financial Studies, Shandong University, Jinan 250100, China;
201812065@mail.sdu.edu.cn

Abstract: In classic regression analysis, the error term of the model typically conforms to the requirement of being independent and identically distributed. However, in the realm of big data, it is exceedingly common for the error term to exhibit varying distributions due to discrepancies in data collection timing and sources. In this article, we expand upon and refine the upper expectation parameter regression model, introducing a novel upper expectation loss form, which handles distribution heterogeneity via group-specific μ_i , to ensure consistent and efficient parameter estimation. Furthermore, we establish the asymptotic properties of this estimation. To address the challenges posed by big data or the restrictions of privacy, we propose a method utilizing Poisson subsampling to devise a new loss function. Under certain assumptions, this method satisfies the condition of asymptotic normality. Additionally, based on the asymptotic properties, we determine the optimal sampling probability and introduce the optimal subsampling technique. Our sampling method surpasses uniform sampling, which reduces MSE by about 50% compared to uniform sampling, and is straightforward to implement in practical scenarios. Subsequent simulation experiments and real-world examples further demonstrate the effectiveness of our approach.

Keywords: upper expectation; parametric regression; optimal subsampling probability; distribution heterogeneity

MSC: 62J05; 62D10



Academic Editor: Xuejun Wang

Received: 7 March 2025

Revised: 26 March 2025

Accepted: 27 March 2025

Published: 30 March 2025

Citation: Liu, Z. Optimal Subsampling for Upper Expectation Parametric Regression. *Mathematics* **2025**, *13*, 1133. <https://doi.org/10.3390/math13071133>

Copyright: © 2025 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Big data usually refers to datasets that are large in different aspects: there can be many observations, many variables, or both. Big data presents new challenges for statistical methods and inference. Excessive data may result in data that cannot be stored and calculated in one machine. In addition, the data may also come from different sources and have inhomogeneities, which would cause failure of traditional statistical methods. Specifically, the question we are interested in is whether it is possible to extract a computationally feasible model that is suitable for data from different times or different sources, or more generally, data with different potential distributions. Our goal is to obtain a robust parameter estimate and a prediction based on the expected upper limit.

Almost all classical statistical models rely on various assumptions, with the most crucial being that the model in question possesses a probability distribution, whether known or unknown. Classical linear expectations and determinant statistics hinge on the certainty of this distribution or model. However, in cases of model heterogeneity, traditional statistical methods may become inapplicable. For instance, the classical maximum likelihood may not exist or be uniquely determined due to the absence of a definitive likelihood function. In addition, the classical least squares estimation is invalid because the parameters are

defined by linear expectations. Moreover, classic statistical models, such as linear regression models, may not be well defined, since their identifiability hinges on the average certainty, and the regression function becomes unidentifiable in its absence. Consequently, to attain the objective of statistical inference, it is imperative to devise a novel statistical framework and corresponding methodologies.

In the absence of certainty in distribution, the expectations derived are often nonlinear. Early studies on nonlinear expectations can be traced back to [1] in the realm of robust statistics and [2] in the field of imprecise probability. Over recent decades, theories and methods related to nonlinear expectations have seen substantial development and have gained recognition in application areas such as financial risk measurement and control. Ref. [3] presented a notable instance of nonlinear expectation within the context of backward stochastic differential equations, termed *g*-expectation. As an extension, ref. [4] introduced *g*-expectation and its related forms. Within the framework of nonlinear expectations, the most prevalent distribution is the so-called *G*-normal distribution, which was first put forth by [4]. Moreover, refs. [5,6] established the law of large numbers and the central limit theorem, serving as the theoretical cornerstone for nonlinear expectations.

Addressing statistical problems arising from distribution heterogeneity, ref. [7] examined *k*-diverse distributions and introduced the upper expectation regression model. Subsequently, ref. [8] advanced a mini-max risk and mini-mean risk regression approach within the context of distribution heterogeneity. Furthermore, ref. [9] introduced the notion of “maximin effects” along with a suitable estimator, evaluating its predictive accuracy from a theoretical perspective in mixture models with either known or unknown group structures. Then, ref. [10] focused on learning models that ensure uniform performance through distributionally robust optimization, incorporating considerations of the worst-case distribution and tail effects.

On the other hand, in the big data era, the rapid proliferation of data introduces fresh obstacles to numerous traditional statistical challenges. Foremost among these is the practical impossibility of utilizing standard computer data storage and analysis techniques. To address this issue, a multitude of statistical and computational methodologies have been devised thus far. The main strategies include subsampling, divide-and-conquer, and online update [11–14]. In this paper, we primarily consider the method of subsampling.

A central concept of the subsampling approach is to employ nonuniform sampling probabilities, ensuring that data points with higher information content are more likely to be selected. A notable method in this regard is the leverage-score-based subsampling introduced by [15]. Subsequently, ref. [16] suggested an information-driven optimal subsample selection technique specifically for linear models. This technique avoids random sampling and instead selects subsamples deterministically for statistical analysis. Additionally, ref. [17] derived the optimal Poisson subsampling probability for quasi-likelihood estimation and devised a distributed optimal subsampling strategy.

The main contribution of this paper is to improve and develop the upper expectation regression method within the framework of big data or under privacy constraints for models with distribution heterogeneity. This heterogeneity is common in practical applications due to factors such as differences in data sources and environments, which always lead to variations in data distributions, such as the model for the influencing factors of air quality in Section 5. Upper expectation regression differs from classical regression in that it tends to use larger values to predict the response variable and obtain mini-max prediction risk. Unlike the method proposed by [7], we address model heterogeneity by introducing group-specific μ_i values, which allows us to obtain a consistent estimator for β , thereby avoiding the potential inconsistency issue of beta found in the literature.

Another major contribution is to study the asymptotic theory of mini-max estimates for upper expectation regression under subsampling conditions. And then we provide a method to obtain the optimal subsampling probability based on the asymptotic theory. Furthermore, we employ an effective and robust estimation and prediction method, making sampling more stable and feasible. This is further supported by simulations and real data.

The rest of this paper is organized as follows. The second section briefly reviews the motivation, methods, and theoretical properties of the upper expectation regression. And we improve the upper expectation regression. The subsampling method and asymptotic theory are studied in the third section. Section 4 provides the selection method and specific implementation form of the optimal subsampling probability. In the fifth section, simulation and real data examples are given to prove the effectiveness and feasibility of the proposed method. The proof of the theorem is postponed to the Appendix A.

2. Upper Expectation Regression Model

2.1. Preliminary of Upper Expectation Model

We analyze the given linear regression model, which is expressed as follows:

$$Y = \beta' \mathbf{x} + \varepsilon, \quad (1)$$

where Y represents a scalar response variable, $\mathbf{x} = (x_1, \dots, x_p)'$ is a p -dimensional covariate vector, and $\beta = (\beta_1, \dots, \beta_p)'$ is a p -dimensional vector comprising unknown parameters. For simplicity, we need the independence assumption. In this way, we need that the conditional expectation of ε with given $\mathbf{x}$ is a constant independent of $\mathbf{x}$. That is,

$$E[\varepsilon|\mathbf{x}] = \mu,$$

where μ is a constant when $\mathbf{x}$ is given.

In the classic regression model, it is often assumed that the error ε is an independent and identically distributed random variable. However, in practical applications, due to the different time and sources of data collection, the assumption of independent and identically distributed for error terms may not always hold.

Firstly, we briefly review the k -sample upper expectation linear regression in [7]. The essential difference between this and the classic regression model is that the error term ε has distribution heterogeneity. The possible distribution of the error term forms a set

$$\mathcal{F} = \{F_1, \dots, F_k\},$$

where k is the number of different distributions and it is finite.

Under the framework of sublinear expectations, the distribution of ε can be defined as

$$\mathbb{E}[\varphi(\varepsilon)] = \sup_{F \in \mathcal{F}} E_F[\varphi(\varepsilon)], \varphi \in C_{l,Lip}.$$

Subsequently, we express the conditional expectations as follows:

$$\bar{\mu} = \mathbb{E}[\varepsilon|\mathbf{x}], \quad \underline{\mu} = -\mathbb{E}[-\varepsilon|\mathbf{x}].$$

Given these definitions, we introduce the concept of upper expectation regression:

$$\mathbb{E}[Y|\mathbf{x}] = \beta' \mathbf{x} + \bar{\mu}.$$

Let $\{(y_{ij}, x_{ij})\}_{j=1}^n$ be a sample in model (1), where $i = 1, \dots, k$, which means that the data are divided into k groups. We assume that samples in different groups have different

distributions and that samples in the same group have the same distribution. For simplicity, we assume that the number of samples in each group is equal, i.e., $n_1 = n_2 = \dots = n_k = n$. In actual situations, the number of samples in each group may be different, but when the number is not much different, we obtain almost the same theoretical results.

In order to achieve the upper expectation loss, we use the empirical version of it. Specifically, the empirical version of the upper expectation loss is

$$Q(\beta) = \max_{1 \leq i \leq k} \frac{1}{n} \sum_{j=1}^n [y_{ij} - \beta' x_{ij} - \bar{\mu}]^2.$$

By minimizing the upper expectation loss $Q(\beta)$, we can obtain the estimator of β , and we call it the mini-max estimator of β .

$$(\hat{\beta}_G, \hat{\mu}) = \arg \min_{\beta \in B, \bar{\mu} \in U} \max_{1 \leq i \leq k} \frac{1}{n} \sum_{j=1}^n [y_{ij} - \beta' x_{ij} - \bar{\mu}]^2.$$

We write $\sigma_i^2 = E[(\varepsilon_{ij} - \bar{\mu})^2]$ and $\sigma_{i_*}^2 = \max_{1 \leq i \leq k} \sigma_i^2$.

The following theorem gives the asymptotic normality of the mini-max estimator of β .

Theorem 1 ([7]). We postulate that $E[\mathbf{x}\mathbf{x}']$ is a positive definite matrix, with $\sigma_{i_*}^2$ exceeding σ_i^2 for all $i_* \neq i$. As n approaches infinity, it follows that

$$\sqrt{n}(\hat{\beta}_G - \beta) \xrightarrow{d} N(0, \sigma_{i_*}^2 (E[\mathbf{x}\mathbf{x}'])^{-1}),$$

where $\xrightarrow{d}$ denotes convergence in distribution, and $N(0, \sigma_{i_*}^2 (E[\mathbf{x}\mathbf{x}'])^{-1})$ represents the standard normal distribution.

However, the prediction based on $\hat{\mu}$ may not align with the upper expectation prediction, as $\hat{\mu}$ might not consistently estimate the upper expectation $\bar{\mu}$. Prior to concluding this section, we propose a two-stage estimation approach to develop a consistent estimator for $\bar{\mu}$ and subsequently formulate a prediction grounded in upper expectation. Utilizing the consistent estimator $\hat{\beta}_G$ derived previously, the second-stage estimator for $\bar{\mu}$ is defined as

$$\hat{\mu}_S = \max_{1 \leq i \leq k} \frac{1}{n} \sum_{j=1}^n (y_{ij} - x'_{ij} \hat{\beta}_G)$$

Let $c = 1 - E[\mathbf{x}'](E[\mathbf{x}\mathbf{x}'])^{-1}E[\mathbf{x}]$ and

$$\Omega^{-1}(\mathbf{x}) = (E[\mathbf{x}\mathbf{x}'])^{-1} + \frac{(E[\mathbf{x}\mathbf{x}'])^{-1}E[\mathbf{x}]E[\mathbf{x}'](E[\mathbf{x}\mathbf{x}'])^{-1}}{c}$$

We present the following theorem based on these definitions.

Theorem 2 ([7]). Given the conditions outlined in Theorem 1, if the sequences $\{\varepsilon_{ij}, j = 1, \dots, n\}$ and $\{\varepsilon_{sj}, j = 1, \dots, n\}$ are independent for $i \neq s$, then the second-stage estimator satisfies the asymptotic distribution:

$$\sqrt{n}(\hat{\mu}_S - \bar{\mu}) \xrightarrow{d} N(0, \sigma_*^2 + \sigma_i^2 E[\mathbf{x}'\Omega^{-1}(\mathbf{x})E[\mathbf{x}]]),$$

where $\bar{\mu}$ is equivalent to $\mu_{\bar{i}}$ for some index $\bar{i}$ within the set $\{1, \dots, k\}$. Note that $E[\mathbf{x}']E[\Omega^{-1}(\mathbf{x})]E[\mathbf{x}]$ has been simplified to $E[\mathbf{x}'\Omega^{-1}(\mathbf{x})E[\mathbf{x}]]$ for clarity.

The proof of Theorems 1 and 2 can be found in [7].

2.2. Improvement of the Method

We firstly consider the upper expectation loss given in the Section 2.1:

$$\mathbb{E}[(Y - \beta'x - \tilde{\mu})^2] = \max_{1 \leq i \leq k} E_i[Y_i - \beta'x_i - \tilde{\mu}]^2.$$

In different groups of data, the expectation of the error term in the right of the formula can be different. But in the expectation loss, each group uses the same $\tilde{\mu}$ value. By comparing the loss of each group, the selected data group with the largest expectation loss is likely to be the group for which the expectation of the error term is farthest from $\tilde{\mu}$. This is somewhat unreasonable to some extent.

To solve this problem, we propose a new upper expectation loss

$$\max_{1 \leq i \leq k} E_i[Y_i - \beta'x_i - \mu_i]^2, \quad (2)$$

where μ_i denotes the expectation of the error term of the group i , and E_i denotes the expectation under the distribution of the data in the group i .

Due to the value of μ_i being unknown, in order to proceed smoothly in the following estimation, we can make an a priori estimation of μ_i . That is, in each group, we would use simple least squares estimation to approximate the value of μ_i to be $\hat{\mu}_i$, which is

$$\hat{\mu}_i = \arg \min_{\mu_i \in U} \frac{1}{n} \sum_{j=1}^n [y_{ij} - \beta'x_{ij} - \mu_i]^2. \quad (3)$$

In this optimization equation, β is also used as a parameter. In other words, there is an estimate of the value of β , but we do not need an estimate of β here, so it is not listed here.

Then, we give a new empirical version of the upper expectation loss:

$$Q_1(\beta) = \max_{1 \leq i \leq k} \frac{1}{n} \sum_{j=1}^n [y_{ij} - \beta'x_{ij} - \hat{\mu}_i]^2. \quad (4)$$

By minimizing the upper expectation loss, we can obtain the estimator of β . We write it as $\hat{\beta}_N$.

For convenience, we use the same notation as in Section 2.1. In this situation, we write $\sigma_i^2 = E[(\varepsilon_{ij} - \mu_i)^2]$ and $\sigma_{i_*}^2 = \max_{1 \leq i \leq k} \sigma_i^2$. The σ_i^2 s that appear later are all defined here.

To better express our method, we summarize it in Algorithm 1.

Algorithm 1 The two-stage estimation process for μ_i and β ,

Require:

$\{(y_{ij}, x_{ij})\}_{j=1}^n$ be a sample in model (1), where $i = 1, \dots, k$;

Ensure:

the estimations $\hat{\mu}_i$ and $\hat{\beta}_N$;

1: **for** $i = 1$ to n **do**

2: Use (3) to solve for the parameter estimate $\hat{\mu}_i$;

3: **end for**

4: Via minimizing the empirical upper expectation loss (4), obtain the estimation $\hat{\beta}_N$;

5: **return** $\hat{\mu}_i$ and $\hat{\beta}_N$.

Before presenting the asymptotic results, we first make the following assumption about ε_{ij} primarily concerning data independence and distributional assumptions.

Assumption 1. There exists an index decomposition $I_i, i = 1, \dots, k$ such that when $(ij) \in I_i$, $\varepsilon_{i1}, \dots, \varepsilon_{in_i}$ are identically distributed with bounded variance, and

$$\sum_{j=1}^n E(\varepsilon_{ij} - \mu_i)^2 1\{|\varepsilon_{ij} - \mu_i| > \delta\} \rightarrow 0, \quad \text{every } \delta > 0.$$

The following theorem gives the asymptotic normality of the new mini-max estimator of β .

Theorem 3. Under Assumption 1 and further assuming $E[\mathbf{x}\mathbf{x}']$ is a positive definite matrix, $\sigma_{i_*}^2 > \sigma_i^2$ for all $i_* \neq i$. Then, when $n \rightarrow \infty$, we have

$$\sqrt{n}[\hat{\beta}_N - \beta] \xrightarrow{d} N\left(0, \sigma_{i_*}^2 (E[\mathbf{x}\mathbf{x}'])^{-1}\right),$$

where $\xrightarrow{d}$ stands for convergence in distribution, and $N\left(0, \sigma_{i_*}^2 (E[\mathbf{x}\mathbf{x}'])^{-1}\right)$ is a classic normal distribution.

The proof of the theorem is provided in the Appendix A.

3. General Poisson Subsampling

For large datasets, the increase in the amount of the data has caused great difficulties in calculations and storage. In order to solve this problem, we adopt the Poisson subsampling method. We first consider Poisson subsampling in general, that is, we firstly consider the method without specifying the probability of each sample being selected. Then, we provide a sampling method and give a specific implementation method in later chapters.

3.1. Poisson Subsampling Method

We consider the same datasets as those in Section 2.1. Then, let p_{ij} be probability to obtain the j th sample point in the i th group, where $i = 1, \dots, k$ and $j = 1, \dots, n$. Let S_i denote the set of observation values and sampling probabilities of the sampled subsample in the i th group. That is,

$$S_i = \{\delta_{ij}(x_{ij}, y_{ij}, p_{ij}), j = 1, \dots, n\},$$

where δ_{ij} is a random variable with the Bernoulli distribution. We write $\delta_{ij} \sim \text{Bernoulli}(p_{ij})$.

According to the introduction in the previous section, we know that $Q_1(\beta)$ is the empirical version of the upper expectation loss $\max_{1 \leq i \leq k} E_i[Y_i - \beta' \mathbf{x}_i - \hat{\mu}_i]^2$.

Due to the large amount of data, a natural idea is to perform statistical analysis on the sampled subsets. Specifically, by using the obtained sample set, we define a new weighted upper expectation loss as

$$\max_{1 \leq i \leq k} E_i \left[\frac{\delta}{p} (Y_i - \beta' \mathbf{x}_i - \hat{\mu}_i)^2 \right].$$

Because of the independence of δ and (x, Y) , we know the weighted upper expectation loss is equal to the upper expectation loss. That is,

$$\begin{aligned} \mathbb{E} \left[\frac{\delta}{p} (Y - \beta' \mathbf{x} - \hat{\mu})^2 \right] &= \max_{1 \leq i \leq k} E_i \left[\frac{\delta}{p} (Y - \beta' \mathbf{x} - \hat{\mu}_i)^2 \right] = \max_{1 \leq i \leq k} E_i \left[\frac{\delta}{p} \right] E_i \left[(Y - \beta' \mathbf{x} - \hat{\mu}_i)^2 \right] \\ &= \max_{1 \leq i \leq k} E_i \left[(Y - \beta' \mathbf{x} - \hat{\mu}_i)^2 \right] = \mathbb{E} \left[(Y - \beta' \mathbf{x} - \hat{\mu})^2 \right]. \end{aligned}$$

Then, we define the empirical version of the weighted upper expectation loss as

$$Q_1^*(\beta) = \max_{1 \leq i \leq k} \frac{1}{n} \sum_{j \in S_i} \frac{1}{p_{ij}} [y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2 = \max_{1 \leq i \leq k} \frac{1}{n} \sum_{j=1}^n \frac{\delta_{ij}}{p_{ij}} [y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2. \quad (5)$$

We can solve the parameter β by minimizing the weighted estimation function. It means that

$$\tilde{\beta} = \arg \min_{\beta \in B} Q_1^*(\beta).$$

From simple calculations, we obtain

$$Q_1^*(\beta) = \max_{1 \leq i \leq k} \frac{1}{n} \sum_{j \in S_i} \frac{1}{p_{ij}} [y_{ij}^2 + \beta' x_{ij} x'_{ij} \beta + \hat{\mu}_i^2 - 2\beta' y_{ij} x_{ij} - 2y_{ij} \hat{\mu}_i + 2\beta' x_{ij} \hat{\mu}_i].$$

One advantage of Poisson subsampling is that the probability p_{ij} only depends on the dataset $\{(x_{ij}, y_{ij})\}_{j=1}^n$. Therefore, the probability p_{ij} can be generated block by block, without using all the data at once. In addition, according to the above formula, we only need to calculate the sum of y_{ij}^2 , $x_{ij} x'_{ij}$, $y_{ij} x_{ij}$, y_{ij} , and x_{ij} in each group. These statistics can be sent to the central machine for calculation, without transmitting the original data of each group. Thereby, it can reduce the time and cost of communication.

3.2. Theoretical Properties

In order to establish asymptotic results of $\tilde{\beta}$, we need the following assumptions.

Assumption 2. The regression parameter lies in the l_1 ball $\Lambda = \{\beta \in \mathbb{R}^p : \|\beta\|_1 \leq C\}$, where C is a constant. This means that $\tilde{\beta}$ and $\hat{\beta}_N$ are the inner points of Λ .

Assumption 3. We assume

$$E_i(\|\mathbf{x}_i\|^3) < \infty$$

for all $i \in \{1, \dots, k\}$.

Assumption 4. For every $i \in \{1, \dots, k\}$, we have

$$E_i[(y_i - \beta' \mathbf{x}_i - \hat{\mu}_i)^3] < \infty.$$

The size of the subsample r^* is a random variable, and we have $E(r^*) = \sum_{i=1}^n p_i$. We use r to denote $E(r^*)$. In this article, it is obvious to assume that $r < n$.

Assumption 5. We assume that $\max_{j=1, \dots, n} (np_j)^{-1} = O_P(r^{-1})$.

Assumption 2 is that the neighborhood of the estimated values $\tilde{\beta}$ and $\hat{\beta}_N$ have reasonable properties. Assumption 3 gives several moment conditions for variables. Assumption 4 requires the moment condition of the loss function. Assumptions 3 and 4 are used in the proof of Theorems 4 and 5 as the key moment conditions in the Appendix A. And Assumption 5 puts a limit on the sampling probability of each point.

First, we give the property of $Q_1^*(\beta)$.

Theorem 4. If the Assumptions 2–5 hold, then we have

$$Q_1^*(\beta) = Q_1(\beta) + o_p(1).$$

The proof of Theorem 4 is in the Appendix A. From Theorem 4, we can see that the $Q_1^*(\beta)$ we proposed is reasonable.

Then, we give the asymptotic normality of $\tilde{\beta}$.

Theorem 5. *If Assumptions 2–5 hold, then when $n \rightarrow \infty$, $r \rightarrow \infty$, we have*

$$V^{-1/2}(\tilde{\beta} - \hat{\beta}_N) \xrightarrow{d} N(0, I).$$

where $V = \Sigma_{i_*}^{-1} V_c \Sigma_{i_*}^{-1}$, and

$$V_c = \frac{1}{n^2} \sum_{j=1}^n \frac{\{y_{i_*j} - \hat{\beta}'_N x_{i_*j} - \hat{\mu}_{i_*}\}^2 x_{i_*j} x'_{i_*j}}{p_{i_*j}} - \frac{1}{n^2} \sum_{j=1}^n \{y_{i_*j} - \hat{\beta}'_N x_{i_*j} - \hat{\mu}_{i_*}\}^2 x_{i_*j} x'_{i_*j},$$

$$\Sigma_i = n^{-1} \sum_{j=1}^n x_{ij} x'_{ij}.$$

The proof of Theorem 5 is in the Appendix A.

4. Optimal Subsampling Strategies

In this section, we derive the best subsampling probability to obtain a better estimate of $\hat{\beta}_N$. After the theoretical derivation, we give a selection method for actual calculations. We mainly draw on the method of [17].

4.1. Theoretical Method

We would find the optimal subsampling probability according to the result of Theorem 5. That is, we minimize the asymptotic mean square error of $\tilde{\beta}$ approaching $\hat{\beta}_N$. This is equivalent to minimizing $\text{tr}(V)$. This method is called A-optimality in the language of optimal design (See [18]).

Method 1 (A-optimality). *For ease of presentation, define the statistics in the i th group:*

$$d_{ij}^V = |y_{ij} - \hat{\beta}'_N x_{ij} - \hat{\mu}_i| \left\| \Sigma_i^{-1} x_{ij} \right\|_2, j = 1, \dots, n.$$

Let $d_{i(1)}^V \leq d_{i(2)}^V \leq \dots \leq d_{i(n)}^V$ denote the order statistics of $\{d_{ij}^V\}_{j=1}^n$. By minimizing the asymptotic mean square error, i.e., minimizing $\text{tr}(V)$, we obtain the sampling probability of the i th group p_{ij} as

$$p_{i(j)}^V = (r-s) \frac{d_{i(j)}^V}{\sum_{k=1}^{n-s} (d_{i(k)}^V)}, 1 \leq j \leq n-s, \quad (6)$$

and we have

$$p_{i(j)}^V = 1, n-s+1 \leq j \leq n, \quad (7)$$

where

$$s = \min \left\{ t \mid 0 \leq t \leq r, \quad (r-t)d_{i(n-t)}^V \leq \sum_{j=1}^{n-t} d_{i(j)}^V \right\}. \quad (8)$$

This means that s satisfies

$$(r-s+1)d_{i(n-s+1)}^V \geq \sum_{j=1}^{n-s+1} d_{i(j)}^V, \quad (r-s)d_{i(n-s)}^V < \sum_{j=1}^{n-s} d_{i(j)}^V.$$

$\{p_{ij}^V\}_{j=1}^n$ can be obtained directly by using the data of each group of covariates and response variables.

We know that the range of sampling probability values is $[0, 1]$. Therefore, the sampling probability calculated based on (6) should be less than or equal to 1. As (j) increases, the sampling probability gets closer to 1. To ensure that the sampling probability remains less than or equal to 1, we introduce a threshold s that satisfies (8). When (j) exceeds this threshold, we directly set the sampling probability to 1.

However, when calculating $\{p_{ij}^V\}_{j=1}^n$, we need to calculate $\|\Sigma_i^{-1}x_{ij}\|_2$ for $j = 1, \dots, n$. It takes $O(nd^2)$ time. To further reduce the calculation time, a simple and natural way is to use the matrix V_c directly, and thus, we do not need to calculate $\|\Sigma_i^{-1}x_{ij}\|_2$. In this way, we could change the method by minimizing $tr(V_c)$ to calculate the optimal subsampling probability. This method is called the linear optimality criterion, or L-optimality for short (see [18]).

Below, we describe in detail the process of obtaining optimal subsampling probability by minimizing $tr(V_c)$.

Method 2 (L-optimality). For ease of presentation, define the statistics in the i th group:

$$d_{ij}^{Vc} = |y_{ij} - \hat{\beta}'_N x_{ij} - \hat{\mu}_i| \|x_{ij}\|_2, j = 1, \dots, n.$$

Let $d_{i(1)}^{Vc} \leq d_{i(2)}^{Vc} \leq \dots \leq d_{i(n)}^{Vc}$ denote the order statistics of $\{d_{ij}^{Vc}\}_{j=1}^n$. By minimizing the asymptotic mean square error, i.e., minimizing $tr(V_c)$, we obtain the sampling probability of the i th group p_{ij} as

$$p_{i(j)}^{Vc} = (r-s) \frac{d_{i(j)}^{Vc}}{\sum_{k=1}^{n-s} d_{i(k)}^{Vc}}, 1 \leq j \leq n-s, \quad (9)$$

and we have

$$p_{i(j)}^{Vc} = 1, n-s+1 \leq j \leq n, \quad (10)$$

where

$$s = \min \left\{ t \mid 0 \leq t \leq r, \quad (r-t)d_{i(n-t)}^{Vc} \leq \sum_{j=1}^{n-t} d_{i(j)}^{Vc} \right\}.$$

This means that s satisfies

$$(r-s+1)d_{i(n-s+1)}^{Vc} \geq \sum_{j=1}^{n-s+1} d_{i(j)}^{Vc}, \quad (r-s)d_{i(n-s)}^{Vc} < \sum_{j=1}^{n-s} d_{i(j)}^{Vc}.$$

4.2. Robust Implementation

In the above calculation, there is $\frac{\delta_{ij}}{p_{ij}}$ as the weight in the estimation function. Therefore, we consider the data points which satisfy $y_{ij} - \hat{\beta}'_N x_{ij} - \hat{\mu}_i \approx 0$. In this case, the probability of these data being sampled is quite small, but they may also be sampled. If these data points are sampled, the estimation equation may be sensitive to the data points. This may cause the failure of Methods 1 and 2 in practical applications. In order to make the estimation more stable, we adopt a more robust sampling method to implement Methods 1 and 2 in the previous section.

We use the following subsampling probability:

$$p_{ij}^{sos} = (1-\rho)p_{ij}^{os} + \rho rn^{-1}, \quad j = 1, \dots, n, \quad (11)$$

where p_{ij}^{os} is p_{ij}^V or p_{ij}^{Vc} , and $\rho \in (0, 1)$.

$\{p_{ij}^{sos}\}_{j=1}^n$ is the convex combination of $\{p_{ij}^{os}\}_{j=1}^n$ and the uniform subsampling probability rn^{-1} . It inherits the advantages of both. Here, ρ is a preference tuning parameter. A smaller ρ results in a more optimal sampling probability, while a larger ρ yields a more robust sampling probability. When ρ is close to 1, the estimation function is not sensitive to data points, so the estimation obtained becomes more stable. On the other hand, no matter what value ρ takes, the order of p_{ij}^{sos} is consistent with p_{ij}^{os} . Thus, the estimator can still retain the advantages of optimal subsampling.

In practical implementation, the parameters $\hat{\beta}_N$ in p_{ij}^{os} are unknown. Therefore, we need to estimate the values of $\hat{\beta}_N$ to continue the following calculations.

In order to estimate $\hat{\beta}_N$, we use uniform sampling to obtain an a priori estimate. In detail, we first take a uniform subsample in each group for which the subsample size is r_0 . Then, we substitute the obtained subsamples into the estimation function to calculate the pilot estimator of $\hat{\beta}_N$. In addition, we also use the $\hat{\mu}_i$ in the calculation of the pilot estimator. We write the pilot estimator as $\tilde{\beta}_0$.

In addition, the typical situation is that $r \ll n$ is required in our subsampling setting. In order to ensure that each $p_{ij}^{sos} \in [0, 1]$, we need take the inverses of $p_{ij}^{sos} \wedge 1$ as weights in the weighted upper expectation loss.

At the end of this section, we summarize the estimation methods under the optimal sampling probability as Algorithm 2, paving the way for the simulations and real-data applications in the next section.

Algorithm 2 The robust subsampling method for β .

Require:

$\{(y_{ij}, x_{ij})\}_{j=1}^n$ be a sample in model (1), where $i = 1, \dots, k$;

Ensure:

the estimations $\tilde{\beta}$;

1: **for** $i = 1$ to n **do**

2: Use (3) to solve for the parameter estimate $\hat{\mu}_i$;

3: **end for**

4: Use uniform sampling to get a priori estimate $\tilde{\beta}_0$ based on (5);

5: Use (6) and (7) or (9) and (10) to obtain the subsampling probability p_{ij}^V or p_{ij}^{Vc} ;

6: Obtain the robust subsampling probability p_{ij}^{sos} according to (11);

7: Compute the estimation $\tilde{\beta}$ according to (5) with the subsampling probability p_{ij}^{sos} ;

8: **return** $\tilde{\beta}$.

5. Simulation Study and Real Data Analysis

5.1. Simulation Study

In this part, we give several simulation examples. By comparing the performance of our proposed optimal subsampling methods and uniform sampling method, the rationality and effectiveness of our proposed method are further illustrated.

From the simulations given later, we discover the following conclusions. Whether it is the case where the mean value is certain or the mean value is uncertain, the optimal subsampling method we proposed has significantly better performance than the simple uniform subsampling method. And as the sampling ratio increases, the performance of the method becomes better.

5.1.1. Experiment 1

In the first simulation experiment, we consider the following linear model:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon.$$

During the simulation, we examine a scenario where the error term possesses a defined mean, but its variance is uncertain. For this setup, we select regression coefficients as $\beta_j = 1$ for $j = 1, 2, 3$. Additionally, the predictors X_j are independently and identically distributed as $N(0, 1)$ for $j = 1, 2, 3$. The error term ε is presumed to adhere to a normal distribution with a mean of zero and an unknown variance σ^2 , thereby forming a generalized regression model accommodating variance uncertainty. During the simulation process, we draw variance values σ_i^2 for i ranging from 1 to k from a uniform distribution spanning from $[0, 4]$. Following this, we generate error values ε_{ij} for j ranging from 1 to n , using normal distributions centered at zero with variances σ_i^2 .

We have three subsampling methods. The first method is the simple uniform sampling method. This method serves as a control to reflect the effectiveness and rationality of the method we proposed. The second method is the Method 1 we proposed in Section 4, which is called A-optimality. And the third method is the Method 2 in Section 4, which is called L-optimality. Let $\rho = 0.2$ be the preference tuning parameter. We chose $k = 10$ and $n = 10,000$, and we chose that the number of samples per group r was 100, 300, and 500. This implies that by adopting such a practice with varying sample sizes for subsampling, we can compare the impact of different sample sizes r in the method. We sampled the generated data 500 times and compared the performance of these three methods by calculating the deviation and mean square error of the parameter estimates.

The simulation results are reported in Tables 1–3. The bolded results in the tables correspond to the best-performing method.

In order to further illustrate the effect of our sampling method in the upper expected regression, we changed the experimental settings, which represents stronger heterogeneity. We considered increasing the number of groups k and appropriately reducing the number of samples in each group. We explored the effect of parameter estimation in the new setting. We chose $k = 50$ and $n = 2000$ as the new experiment settings. And the other experimental conditions remained unchanged.

Table 1. The mean and MSE of three methods when $r = 100$ in experiment 1, with $k = 10$ and $n = 10,000$.

Method	Statistics	β_1	β_2	β_3
uniform	mean	1.072	1.082	1.075
	MSE	7.25×10^{-2}	8.05×10^{-2}	7.13×10^{-2}
A-optimality	mean	1.050	1.052	1.044
	MSE	4.54×10^{-2}	4.08×10^{-2}	3.92×10^{-2}
L-optimality	mean	1.043	1.052	1.056
	MSE	4.57×10^{-2}	4.08×10^{-2}	4.18×10^{-2}

Table 2. The mean and MSE of three methods when $r = 300$ in experiment 1, with $k = 10$ and $n = 10,000$.

Method	Statistics	β_1	β_2	β_3
uniform	mean	1.045	1.037	1.033
	MSE	2.86×10^{-2}	2.54×10^{-2}	2.35×10^{-2}
A-optimality	mean	1.015	1.025	1.028
	MSE	1.51×10^{-2}	1.29×10^{-2}	1.46×10^{-2}
L-optimality	mean	1.015	1.034	1.023
	MSE	1.52×10^{-2}	1.53×10^{-2}	1.51×10^{-2}

Table 3. The mean and MSE of three methods when $r = 500$ in experiment 1, with $k = 10$ and $n = 10,000$.

Method	Statistics	β_1	β_2	β_3
uniform	mean	1.025	1.027	1.028
	MSE	1.50×10^{-2}	1.66×10^{-2}	1.43×10^{-2}
A-optimality	mean	1.014	1.028	1.026
	MSE	8.51×10^{-3}	9.68×10^{-3}	8.96×10^{-3}
L-optimality	mean	1.009	1.024	1.023
	MSE	8.67×10^{-3}	9.19×10^{-3}	8.59×10^{-3}

The simulation results are reported in Tables 4–6. Based on the entire simulation experiment, we draw the following conclusions:

1. By comparing the results of the three methods, it can be observed that our proposed A-optimality and L-optimality approaches significantly outperformed the uniform sampling method in terms of mean squared error. On the other hand, there was little difference in the effectiveness of the mean estimates obtained by the three methods. This demonstrates the effectiveness and stability of our methods, particularly for the method of A-optimality.
2. By comparing Tables 1–3 or Tables 4–6, it can be observed that as the number of sampling instances r increased, the estimation performance of all three methods improved significantly, both in terms of the mean and mean squared error. This underscores the importance of the number of sampling instances.
3. By comparing Tables 1–3 and Tables 4–6, it can be observed that in cases of high heterogeneity, where k is large and n is small, the estimation performance of all three methods deteriorates. However, as the number of sampling instances r increases, this discrepancy diminishes.

Table 4. The mean and MSE of three methods when $r = 100$ in experiment 1, with $k = 50$ and $n = 2000$.

Method	Statistics	β_1	β_2	β_3
uniform	mean	1.146	1.034	1.112
	MSE	1.54×10^{-1}	1.40×10^{-1}	1.46×10^{-1}
A-optimality	mean	1.089	1.030	1.104
	MSE	8.02×10^{-2}	7.87×10^{-2}	9.23×10^{-2}
L-optimality	mean	1.091	1.009	1.104
	MSE	8.23×10^{-2}	7.99×10^{-2}	9.34×10^{-2}

Table 5. The mean and MSE of three methods when $r = 300$ in experiment 1, with $k = 50$ and $n = 2000$.

Method	Statistics	β_1	β_2	β_3
uniform	mean	1.061	1.008	1.064
	MSE	5.18×10^{-2}	4.14×10^{-2}	4.65×10^{-2}
A-optimality	mean	1.040	0.971	1.028
	MSE	2.66×10^{-2}	2.29×10^{-2}	2.06×10^{-2}
L-optimality	mean	1.038	0.965	1.039
	MSE	2.51×10^{-2}	2.23×10^{-2}	2.43×10^{-2}

Table 6. The mean and MSE of three methods when $r = 500$ in experiment 1, with $k = 50$ and $n = 2000$.

Method	Statistics	β_1	β_2	β_3
uniform	mean	1.052	0.981	1.040
	MSE	2.75×10^{-2}	2.61×10^{-2}	2.47×10^{-2}
A-optimality	mean	1.026	0.958	1.009
	MSE	1.36×10^{-2}	1.29×10^{-2}	1.00×10^{-2}
L-optimality	mean	1.022	0.964	1.012
	MSE	1.33×10^{-2}	1.47×10^{-2}	8.45×10^{-3}

5.1.2. Experiment 2

We reconsider the linear model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon,$$

which is the same in form as in Experiment 1. In this model, we consider the situation that the mean and variance of the error are both uncertain. This means that $\varepsilon \sim N(\mu, \sigma^2)$ have uncertain mean μ and variance σ^2 .

During the simulation, we also specified that X_j follows a normal distribution $N(0, 1)$ for $j = 1, 2, 3$. The error terms ε were generated as follows: First, the mean values μ_i and variance values σ_i^2 were drawn from uniform distributions $U(-1, 1)$ and $U(0, 1)$, respectively. Subsequently, the error values ε_{ij} for $j = 1, \dots, k$ were sampled from the normal distribution $N(\mu_i, \sigma_i^2)$. Additionally, we also set $k = 10$ and $n = 10,000$ or $k = 50$ and $n = 2000$, as well as r to be 100, 300, or 500.

The simulation results are reported in Tables 7–12. The simulation results are similar to those of Experiment 1, which are briefly summarized here:

1. In summary, by comparing the methods, our proposed A-optimality and L-optimality approaches outperformed the uniform sampling method in terms of mean squared error, while the mean estimates of all three methods showed little difference in effectiveness.
2. It can be observed that as the number of sampling instances r increased, the estimation performance of all three methods improved significantly.
3. Upon reviewing Tables 7–12, it becomes apparent that when heterogeneity is high, the estimation capabilities of all three methods suffer.

Table 7. The mean and MSE of three methods when $r = 100$ in experiment 2, with $k = 10$ and $n = 10,000$.

Method	Statistics	β_1	β_2	β_3
uniform	mean	1.086	1.063	1.072
	MSE	8.11×10^{-2}	6.92×10^{-2}	7.30×10^{-2}
A-optimality	mean	1.028	1.047	1.046
	MSE	4.23×10^{-2}	4.22×10^{-2}	4.57×10^{-2}
L-optimality	mean	1.042	1.057	1.038
	MSE	4.68×10^{-2}	4.61×10^{-2}	3.88×10^{-2}

Table 8. The mean and MSE of three methods when $r = 300$ in experiment 2, with $k = 10$ and $n = 10,000$.

Method	Statistics	β_1	β_2	β_3
uniform	mean	1.028	1.036	1.038
	MSE	2.61×10^{-2}	2.48×10^{-2}	2.53×10^{-2}
A-optimality	mean	1.008	1.017	1.030
	MSE	1.44×10^{-2}	1.44×10^{-2}	1.37×10^{-2}
L-optimality	mean	1.015	1.021	1.028
	MSE	1.40×10^{-2}	1.59×10^{-2}	1.64×10^{-2}

Table 9. The mean and MSE of three methods when $r = 500$ in experiment 2, with $k = 10$ and $n = 10,000$.

Method	Statistics	β_1	β_2	β_3
uniform	mean	1.010	1.028	1.036
	MSE	1.58×10^{-2}	1.60×10^{-2}	1.60×10^{-2}
A-optimality	mean	1.001	1.014	1.026
	MSE	8.39×10^{-3}	8.38×10^{-3}	9.72×10^{-3}
L-optimality	mean	1.009	1.010	1.018
	MSE	8.86×10^{-3}	9.90×10^{-3}	7.79×10^{-3}

Table 10. The mean and MSE of three methods when $r = 100$ in experiment 2, with $k = 50$ and $n = 2000$.

Method	Statistics	β_1	β_2	β_3
uniform	mean	1.125	1.055	1.116
	MSE	1.50×10^{-3}	1.37×10^{-1}	1.53×10^{-1}
A-optimality	mean	1.063	1.011	1.088
	MSE	7.69×10^{-2}	8.01×10^{-2}	7.81×10^{-2}
L-optimality	mean	1.085	1.009	1.082
	MSE	8.80×10^{-2}	7.02×10^{-2}	8.23×10^{-2}

Table 11. The mean and MSE of three methods when $r = 300$ in experiment 2, with $k = 50$ and $n = 2000$.

Method	Statistics	β_1	β_2	β_3
uniform	mean	1.068	1.000	1.057
	MSE	5.44×10^{-2}	4.16×10^{-2}	4.65×10^{-2}
A-optimality	mean	1.032	0.975	1.024
	MSE	2.65×10^{-2}	2.54×10^{-2}	2.23×10^{-2}
L-optimality	mean	1.028	0.973	1.035
	MSE	2.89×10^{-2}	2.33×10^{-2}	2.31×10^{-2}

Table 12. The mean and MSE of three methods when $r = 500$ in experiment 2, with $k = 50$ and $n = 2000$.

Method	Statistics	β_1	β_2	β_3
uniform	mean	1.035	0.974	1.028
	MSE	2.98×10^{-2}	2.34×10^{-2}	2.67×10^{-2}
A-optimality	mean	1.013	0.961	1.016
	MSE	1.34×10^{-2}	1.42×10^{-2}	9.99×10^{-3}
L-optimality	mean	1.018	0.962	1.011
	MSE	1.46×10^{-2}	1.61×10^{-2}	9.50×10^{-3}

5.2. Real Data Analysis

An increasing number of urban areas are grappling with persistent air pollution, primarily due to fine particulate matter, especially PM2.5. PM2.5 comprises particles suspended in the air with aerodynamic diameters smaller than 2.5 micrometers. These particles are recognized for their impact on visibility, human health, and climate patterns. Epidemiological studies indicate that exposure to PM2.5 can lead to respiratory ailments, severe cardiovascular diseases, and potentially fatal outcomes.

There are many reasons for PM2.5 pollution, including SO₂ concentration, O₃ concentration, temperature, wind speed, etc. The main consideration in this article is the impact of these four factors on PM2.5 concentration. The dataset used in this article was taken from the UCI machine learning knowledge base. Specifically, we obtained the PM2.5 concentration, SO₂ concentration, O₃ concentration, temperature, and wind speed of the ancient city of Beijing from 1 March 2013 to 29 February 2016. Due to the large time span involved in the model, model heterogeneity may arise. Therefore, we considered grouping the data based on time for processing. Then, we considered the linear influence of the covariate on the response variable, and we tested the effect of the subsampling method proposed in this article.

We first centralized the 25,217 data collected over three years, with the PM2.5 concentration after centralization as the response variable, and the central SO₂ concentration, O₃ concentration, temperature, and wind speed as four covariates. These covariates all have an impact on PM2.5 concentrations. For instance, SO₂ can undergo chemical reactions to transform into components of PM2.5, while wind speed affects the concentration of PM2.5 through dispersion and dilution effects. Next, we divided the data into 12 groups. The grouping method was grouped by time. Every three months was a group. This yielded about 2100 data points in each group. The number of data in each group was not necessarily equal.

For the linear regression problem, we considered two different situations. The first case is that the model has only variance uncertainty, that is, the mean value of the error term in each group is equal. The second case is that the model has both variance uncertainty and mean uncertainty. The two sampling methods we proposed performed well in both cases, and they were significantly better than ordinary uniform sampling methods.

Case 1 (Mean certainty model)

We assumed that the model only has variance heterogeneity. In other words, the errors from different group have the same mean. Consider the following model:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon.$$

Our purpose is to estimate the parameters β_1 , β_2 , β_3 , and β_4 , which are the same in every group. Use the subsampling methods we proposed above, we obtained the following

table that describes the mean prediction error and computational time of each method under different settings. The sample size r in the subsampling of each group was set to 50.

Case 2 (Mean–variance heterogeneity model)

Then, we considered a model with not only mean heterogeneity but also variance heterogeneity in the errors of different groups. Consider the following model:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon.$$

In this model, $\varepsilon_i \sim N(\mu_i, \sigma_i^2)$. We also set the sample size in the subsampling of each group to 50. We have the following table that describes the mean prediction error and computational time of each method under different settings.

The results are presented in Tables 13 and 14. Through the actual data analysis of the two cases, we obtained the following conclusions:

1. First, in terms of computational time, it can be observed that uniform sampling was significantly the fastest, while A-optimality was the slowest. This is because the sampling probability for uniform sampling is easily obtained, whereas A-optimality requires the calculation of the inverse of the covariance matrix, which is more complex than L-optimality.
2. Based on the mean prediction error metric, we found that the A-optimality and L-optimality methods significantly outperformed the uniform sampling method, with the A-optimality method being the best.
3. By comparing Tables 13 and 14, it can be observed that the prediction performance of the mean–variance heterogeneous model was significantly better than that of the variance heterogeneous model, while their computational times were comparable. This suggests that the mean–variance model is more suitable for the real data example.

Table 13. The computational time and MPE of three methods in mean certainty model when $r = 50$.

Method	Uniform	A-Optimality	L-Optimality
computational time	0.67 s	4.05 s	3.13 s
MPE	0.737	0.630	0.632

Table 14. The computational and MPE of three methods in mean–variance heterogeneity model when $r = 50$.

Method	Uniform	A-Optimality	L-Optimality
computational time	0.68 s	4.03 s	3.10 s
MPE	0.676	0.513	0.550

6. Conclusions

In the background with heterogeneous distributions, the task of statistical modeling and inference remains a significant challenge. In this paper, we improved the k -sample upper expectation regression model in the context of distribution heterogeneity by employing group-specific μ_i values to address the impact of heterogeneity and obtaining consistent estimates. Additionally, we delved into optimal subsampling techniques specifically designed for big data scenarios, overcoming the challenges posed by large datasets and privacy protection through a robust optimal sampling probability. We analyzed the theoretical properties of our method and conducted comprehensive numerical experiments on both simulated and actual datasets to assess its practical effectiveness. Both theoretical analysis and numerical results confirm the validity of our method when dealing with large

datasets. Furthermore, real data analysis demonstrates that our method can be applied to various similar fields, such as finance and environmental monitoring.

However, there are still some shortcomings in our method that require subsequent work to address. First, the k -sample assumption necessitates prior knowledge of k -sample groups, which means we must assume that sample clusters with different distributions are already known, which are often derived from experience or other methods. This is a rather stringent requirement. Future work could consider developing methods that can accommodate complete distribution heterogeneity in situations where the grouping is unknown, for example, first by data-driven clustering and then by applying upper expectation regression. Additionally, when k is large, it often implies a smaller sample size in each group, resulting in a high degree of heterogeneity. In such cases, both computational complexity and estimation effectiveness face significant challenges. Overcoming these issues is also a direction for future work. Lastly, the robust optimal sampling probability involves a tune parameter ρ , and selecting the optimal ρ to balance robustness and efficiency of the sampling probability is also a research-worthy problem.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the author.

Acknowledgments: The author would like to express his sincere gratitude to the anonymous referees for their valuable comments and suggestions, which significantly improved the quality of this paper.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A

First, we present two lemmas that play crucial roles in subsequent proofs.

Lemma A1. *Given Assumptions 2–5, as both n and r tend to infinity, for any s_r that converges to zero in probability, we have*

$$\frac{1}{n} \sum_{j=1}^n \frac{\delta_{i_*j}}{p_{i_*j}} x_{i_*j}^T x_{i_*j} = \frac{1}{n} \sum_{j=1}^n x_{i_*j}^T x_{i_*j} + o_p(1).$$

Proof. Let $\mathcal{F} = \{x_{ij}, y_{ij}\}_{i=1, j=1}^{k, n}$. Direct calculation shows that conditionally on $\mathcal{F}$.

$$E \left\{ \frac{1}{n} \sum_{j=1}^n \frac{\delta_{i_*j}}{p_{i_*j}} x_{i_*j}^T x_{i_*j} \mid \mathcal{F} \right\} = \frac{1}{n} \sum_{j=1}^n x_{i_*j}^T x_{i_*j},$$

and

$$\begin{aligned}
& E \left\{ \frac{1}{n} \sum_{j=1}^n \frac{\delta_{i_*j}}{p_{i_*j}} x_{i_*j}^T x_{i_*j} - \frac{1}{n} \sum_{j=1}^n x_{i_*j}^T x_{i_*j} \mid \mathcal{F} \right\}^2 \\
&= \sum_{i=1}^n \frac{p_{i_*i}(1-p_{i_*i})}{p_{i_*i}^2} \left\{ \frac{x_{i_*i}^T x_{i_*i}}{n} \right\}^2 \\
&= \sum_{i=1}^n \frac{1}{p_{i_*i}} \left\{ \frac{x_{i_*i}^T x_{i_*i}}{n} \right\}^2 - \sum_{i=1}^n \left\{ \frac{x_{i_*i}^T x_{i_*i}}{n} \right\}^2 \\
&\leq \sum_{i=1}^n \frac{1}{p_{i_*i}} \left\{ \frac{x_{i_*i}^T x_{i_*i}}{n} \right\}^2 \\
&\leq \left(\max_{i=1, \dots, n} \frac{1}{np_{i_*i}} \right) \sum_{i=1}^n \frac{(x_{i_*i}^T x_{i_*i})^2}{n} \\
&\leq \left(\max_{i=1, \dots, n} \frac{1}{np_{i_*i}} \right) \sum_{i=1}^n \frac{\|x_{i_*i}\|^4}{n} \\
&= O_p(r^{-1}) \rightarrow 0.
\end{aligned}$$

Then, we obtain the conclusion

$$\frac{1}{n} \sum_{j=1}^n \frac{\delta_{i_*j}}{p_{i_*j}} x_{i_*j}^T x_{i_*j} = \frac{1}{n} \sum_{j=1}^n x_{i_*j}^T x_{i_*j} + o_p(1).$$

by using the Chebyshev's inequality. $\square$

Lemma A2. Under Assumptions 2–5, as $r \rightarrow \infty$ and $n \rightarrow \infty$,

$$R^*(\hat{\beta}_N) \rightarrow N(0, V_c),$$

in distribution, where

$$V_c = \frac{4}{n^2} \sum_{j=1}^n \frac{\{y_{i_*j} - \hat{\beta}'_N x_{i_*j} - \hat{\mu}_{i_*}\}^2 x_{i_*j} x'_{i_*j}}{p_{i_*j}} - \frac{4}{n^2} \sum_{j=1}^n \{y_{i_*j} - \hat{\beta}'_N x_{i_*j} - \hat{\mu}_{i_*}\}^2 x_{i_*j} x'_{i_*j}.$$

Proof. To prove the conclusion in the lemma, we use the Lindeberg–Feller central limit theorem. Thus, we only need to verify the conditions of the Lindeberg–Feller central limit theorem.

Let

$$r_j = \frac{-2}{n} \frac{\delta_{i_*j}}{p_{i_*j}} [y_{i_*j} - \hat{\beta}'_N x_{i_*j} - \hat{\mu}_{i_*}] x'_{i_*j}.$$

It is obvious that

$$R^*(\hat{\beta}_N) = \sum_{j=1}^n r_j.$$

By direct calculation, we can obtain

$$\begin{aligned}
E[r_j] &= \frac{-2}{n} E \left[\frac{\delta_{i_*j}}{p_{i_*j}} [y_{i_*j} - \hat{\beta}'_N x_{i_*j} - \hat{\mu}_{i_*}] x'_{i_*j} \right] \\
&= \frac{-2}{n} E \left[\frac{\delta_{i_*j}}{p_{i_*j}} \right] E \left[[y_{i_*j} - \hat{\beta}'_N x_{i_*j} - \hat{\mu}_{i_*}] x'_{i_*j} \right] \\
&= 0,
\end{aligned}$$

$$\begin{aligned}
\sum_{j=1}^n \text{var}(r_j) &= \frac{4}{n} \frac{\{y_{i_*j} - \hat{\beta}'_N x_{i_*j} - \hat{\mu}_{i_*}\}^2 x_{i_*j} x'_{i_*j}}{p_{i_*j}^2} \text{var}(\delta_{i_*j}) \\
&= \frac{4}{n} \frac{\{y_{i_*j} - \hat{\beta}'_N x_{i_*j} - \hat{\mu}_{i_*}\}^2 x_{i_*j} x'_{i_*j}}{p_{i_*j}^2} p_{i_*j} (1 - p_{i_*j}) \\
&= V_c.
\end{aligned}$$

And for any $\varepsilon > 0$,

$$\begin{aligned}
\sum_{j=1}^n E\left\{\|r_j\|^2 \mathbf{1}_{\|r_j\| > \varepsilon}\right\} &\leq \frac{1}{\varepsilon} \sum_{j=1}^n E\left\{\|r_j\|^3\right\} \\
&= \frac{1}{n^3} \frac{1}{\varepsilon} \sum_{j=1}^n \frac{\|y_{i_*j} - \hat{\beta}'_N x_{i_*j} - \hat{\mu}_{i_*}\|^3 \|\mathbf{x}_{i_*j}\|^3}{p_{i_*j}^2} \\
&\leq \frac{1}{\varepsilon} \left\{ \max_{j=1, \dots, n} \frac{1}{(np_{i_*j})^2} \right\} \sum_{j=1}^n \frac{\|y_{i_*j} - \hat{\beta}'_N x_{i_*j} - \hat{\mu}_{i_*}\|^3 \|\mathbf{x}_{i_*j}\|^3}{n}.
\end{aligned} \tag{A1}$$

According to the Assumptions 3 and 4, we know that

$$\sum_{j=1}^n \frac{\|y_{i_*j} - \hat{\beta}'_N x_{i_*j} - \hat{\mu}_{i_*}\|^3 \|\mathbf{x}_{i_*j}\|^3}{n} < \infty.$$

Combining Assumption 5, we can obtain a bound for (A1), namely,

$$\begin{aligned}
&\sum_{j=1}^n E\left\{\|r_j\|^2 \mathbf{1}_{\|r_j\| > \varepsilon}\right\} \\
&\leq \frac{1}{\varepsilon} O_p(r^{-2}) \rightarrow 0.
\end{aligned}$$

After verifying the conditions, according to the Lindeberg–Feller central limit theorem in [19], we obtain the conclusion of this lemma. $\square$

Proof of Theorem 3. Since the data in each group are independent and identically distributed, we have

$$\frac{1}{n} \sum_{j=1}^n (y_{ij} - \beta' x_{ij} - \hat{\mu}_i)^2 = \frac{1}{n} \sum_{j=1}^n (\varepsilon_{ij} - \hat{\mu}_i)^2 = \sigma_i^2 + \delta_n,$$

where δ_n is of order $O_p(1/n)$. Note that $\sigma_{i_*}^2 > \sigma_i^2$ for all $i \neq i_*$.

The above two results lead to

$$\max_{1 \leq i \leq k} \frac{1}{n} \sum_{j=1}^n (\varepsilon_{ij} - \hat{\mu}_i)^2 = \sigma_{i_*}^2 + O_p(1/n).$$

Consequently, when n is large enough,

$$\max_{1 \leq i \leq k} \frac{1}{n} \sum_{j=1}^n (\varepsilon_{ij} - \hat{\mu}_i)^2 = \frac{1}{n} \sum_{j=1}^n (\varepsilon_{i_*j} - \hat{\mu}_{i_*})^2 + O_p(1/n).$$

Then,

$$\max_{1 \leq i \leq m} \frac{1}{n} \sum_{j=1}^n [y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2 = \frac{1}{n} \sum_{j=1}^n [y_{i_*j} - \beta' x_{i_*j} - \hat{\mu}_{i_*}]^2 + O_p(1/n).$$

We denote the true values of β and $\hat{\mu}_{i_*}$ as β^0 and $\mu_{i_*}^0$, respectively. And let β and $\hat{\mu}_{i_*}$ satisfy $\|\beta - \beta^0\| = O(1/\sqrt{n})$ and $\|\hat{\mu}_{i_*} - \mu_{i_*}^0\| = O(1/\sqrt{n})$. Then,

$$\begin{aligned} & \max_{1 \leq i \leq k} \frac{1}{n} \sum_{j=1}^n [y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2 \\ &= \max_{1 \leq i \leq k} \frac{1}{n} \sum_{j=1}^n \left[\varepsilon_{ij} - \mu_i^0 - (\beta - \beta^0)' x_{ij} - (\hat{\mu}_i - \mu_i^0) \right]^2 \\ &= \frac{1}{n} \sum_{j=1}^n \left\{ (\varepsilon_{i_*j} - \mu_{i_*}^0)^2 - 2 \left[(\beta - \beta^0)' x_{i_*j} + (\hat{\mu}_{i_*} - \mu_{i_*}^0) \right] (\varepsilon_{i_*j} - \mu_{i_*}^0) \right. \\ & \quad \left. + \left[(\beta - \beta^0)' x_{i_*j} + (\hat{\mu}_{i_*} - \mu_{i_*}^0) \right]^2 \right\} + O_p(1/n) \\ &= \frac{1}{n} \sum_{j=1}^n (\varepsilon_{i_*j} - \mu_{i_*}^0)^2 - \frac{2}{n} \sum_{j=1}^n \left[(\beta - \beta^0)' x_{i_*j} + (\hat{\mu}_{i_*} - \mu_{i_*}^0) \right] (\varepsilon_{i_*j} - \mu_{i_*}^0) \\ & \quad + \frac{1}{n} \sum_{j=1}^n \left[(\beta - \beta^0)' x_{i_*j} + (\hat{\mu}_{i_*} - \mu_{i_*}^0) \right]^2 + O_p(1/n). \end{aligned}$$

Note that the differences $\varepsilon_{i_*j} - \mu_{i_*}^0$ are independently and identically distributed with a mean of zero and a variance of $\sigma_{i_*}^2$. Additionally, it is evident that both $\frac{1}{n} \sum_{j=1}^n \left[(\beta - \beta^0)' x_{i_*j} + (\hat{\mu}_{i_*} - \mu_{i_*}^0) \right] (\varepsilon_{i_*j} - \mu_{i_*}^0)$ and $\frac{1}{n} \sum_{j=1}^n \left[(\beta - \beta^0)' x_{i_*j} + (\hat{\mu}_{i_*} - \mu_{i_*}^0) \right]^2$ converge in probability to terms of order $O_p(1)$.

On the other hand, we know that ε_{i_*j} and δ_n are independent of β . Thus, according to the method of solving $\hat{\beta}_N$, we obtain that minimizing $\max_{1 \leq i \leq m} \frac{1}{n} \sum_{j=1}^n [y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2$ is equivalent to minimizing

$$\sum_{j=1}^n \left\{ -2 \left[(\beta - \beta^0)' x_{i_*j} + (\hat{\mu}_{i_*} - \mu_{i_*}^0) \right] (\varepsilon_{i_*j} - \mu_{i_*}^0) + \left[(\beta - \beta^0)' x_{i_*j} + (\hat{\mu}_{i_*} - \mu_{i_*}^0) \right]^2 \right\}.$$

We rewrite this objective function as

$$Z_n(\gamma) = \sum_{j=1}^n \left[-\frac{2(\varepsilon_{i_*j} - \mu_{i_*}^0)}{\sqrt{n}} x_{i_*j}' \gamma + \frac{1}{n} \gamma' [x_{i_*j} x_{i_*j}'] \gamma \right],$$

where $Z_n(\gamma)$ is obviously convex and minimized at $\hat{\gamma}_n = \sqrt{n}(\hat{\beta}_N - \beta^0)'$. According to Assumption 1 and the Lindeberg–Feller central limit theorem in [19], we obtain

$$Z_n(\gamma) \xrightarrow{d} Z_0(\gamma) = -2W' \gamma + \gamma' E(\mathbf{x}\mathbf{x}') \gamma,$$

where $W \sim N(0, \sigma_{i_*}^2 E[\mathbf{x}\mathbf{x}'])$. Through minimizing $Z_n(\gamma)$ and simple deformation, we can obtain the conclusion of the theorem.

$$\sqrt{n}[\hat{\beta}_N - \beta] \xrightarrow{d} N(0, \sigma_{i_*}^2 (E[\mathbf{x}\mathbf{x}'])^{-1}).$$

□

Proof of Theorem 4. Let $\mathcal{F} = \{x_{ij}, y_{ij}\}_{i=1, j=1}^{k, n}$. Direct calculation shows that conditionally on $\mathcal{F}$.

$$E_i \left\{ \frac{1}{n} \sum_{j=1}^n \frac{\delta_{ij}}{p_{ij}} [y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2 \mid \mathcal{F} \right\} = \frac{1}{n} \sum_{j=1}^n [y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2,$$

and

$$\begin{aligned} & E_i \left\{ \frac{1}{n} \sum_{j=1}^n \frac{\delta_{ij}}{p_{ij}} [y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2 - \frac{1}{n} \sum_{j=1}^n [y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2 \mid \mathcal{F} \right\}^2 \\ &= \sum_{i=1}^n \frac{p_{ij}(1-p_{ij})}{p_{ij}^2} \left\{ \frac{[y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2}{n} \right\}^2 \\ &= \sum_{i=1}^n \frac{1}{p_{ij}} \left\{ \frac{[y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2}{n} \right\}^2 - \sum_{i=1}^n \left\{ \frac{[y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2}{n} \right\}^2 \\ &\leq \sum_{i=1}^n \frac{1}{p_{ij}} \left\{ \frac{[y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2}{n} \right\}^2 \\ &\leq \left(\max_{i=1, \dots, n} \frac{1}{np_{ij}} \right) \sum_{i=1}^n \frac{([y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2)^2}{n} \\ &= O_p(r^{-1}) \rightarrow 0. \end{aligned}$$

Then, we obtain the conclusion

$$\frac{1}{n} \sum_{j=1}^n \frac{\delta_{ij}}{p_{ij}} [y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2 = \frac{1}{n} \sum_{j=1}^n [y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2 + o_p(1),$$

by using Chebyshev's inequality.

Let i traverse 1 to k and take the maximum value, and we obtain

$$Q_1^*(\beta) = Q_1(\beta) + o_p(1).$$

□

Proof of Theorem 5. We divided the data into k different groups, and the data in each group have the same distribution. So, we have

$$\frac{1}{n} \sum_{j=1}^n (\varepsilon_{ij} - \hat{\mu}_i)^2 = \sigma_i^2 + \delta_n,$$

where δ_n is bounded by $O_p(1/n)$. It should be noted that $\sigma_{i_*}^2$ exceeds σ_i^2 for any i other than i_* . These two aforementioned findings imply that

$$\max_{1 \leq i \leq k} \frac{1}{n} \sum_{j=1}^n (\varepsilon_{ij} - \hat{\mu}_i)^2 = \sigma_{i_*}^2 + O_p(1/n).$$

Consequently, when n is large enough,

$$\max_{1 \leq i \leq k} \frac{1}{n} \sum_{j=1}^n (\varepsilon_{ij} - \hat{\mu}_i)^2 = \frac{1}{n} \sum_{j=1}^n (\varepsilon_{i_*j} - \hat{\mu}_{i_*})^2 + O_p(1/n).$$

Then,

$$\max_{1 \leq i \leq m} \frac{1}{n} \sum_{j=1}^n [y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2 = \frac{1}{n} \sum_{j=1}^n [y_{i_*j} - \beta' x_{i_*j} - \hat{\mu}_{i_*}]^2 + O_p(1/n).$$

Using the results obtained above in the sampling estimation function $Q_1^*(\beta)$, we obtain

$$Q_1^*(\beta) = \max_{1 \leq i \leq k} \frac{1}{n} \sum_{j=1}^n \frac{\delta_{ij}}{p_{ij}} [y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2 = \frac{1}{n} \sum_{j=1}^n \frac{\delta_{i_*j}}{p_{i_*j}} [y_{i_*j} - \beta' x_{i_*j} - \hat{\mu}_{i_*}]^2 + O_p(1/r).$$

Similarly, we have

$$Q_1(\beta) = \max_{1 \leq i \leq k} \frac{1}{n} \sum_{j=1}^n [y_{ij} - \beta' x_{ij} - \hat{\mu}_i]^2 = \frac{1}{n} \sum_{j=1}^n [y_{i_*j} - \beta' x_{i_*j} - \hat{\mu}_{i_*}]^2 + O_p(1/n).$$

Therefore, $Q_1^*(\beta)$ and $Q_1(\beta)$ are continuous and differentiable near the true value β_0 of β . We take the derivative of $Q_1^*(\beta)$ and $Q_1(\beta)$ at β_0 , respectively, and then we obtain $R^*(\beta)$ and $R(\beta)$.

$$R^*(\beta) = \frac{dQ_1^*(\beta)}{d\beta} = \frac{-2}{n} \sum_{j=1}^n \frac{\delta_{i_*j}}{p_{i_*j}} [y_{i_*j} - \beta' x_{i_*j} - \hat{\mu}_{i_*}] x'_{i_*j},$$

$$R(\beta) = \frac{dQ_1(\beta)}{d\beta} = \frac{-2}{n} \sum_{j=1}^n [y_{i_*j} - \beta' x_{i_*j} - \hat{\mu}_{i_*}] x'_{i_*j}.$$

We obtain the estimator of $\tilde{\beta}$ by minimizing $Q_1^*(\beta)$. We can also obtain $\tilde{\beta}$ by solving $R^*(\beta)$ equal to 0.

$$R^*(\tilde{\beta}) = \frac{-2}{n} \sum_{j=1}^n \frac{\delta_{i_*j}}{p_{i_*j}} [y_{i_*j} - \tilde{\beta}' x_{i_*j} - \hat{\mu}_{i_*}] x'_{i_*j} = 0.$$

We write

$$R^*(\hat{\beta}_N) = \frac{-2}{n} \sum_{j=1}^n \frac{\delta_{i_*j}}{p_{i_*j}} [y_{i_*j} - \hat{\beta}_N' x_{i_*j} - \hat{\mu}_{i_*}] x'_{i_*j}.$$

By subtracting $R^*(\tilde{\beta})$ and $R^*(\hat{\beta}_N)$, we obtain

$$R^*(\tilde{\beta}) = R^*(\hat{\beta}_N) + \frac{2}{n} \sum_{j=1}^n \frac{\delta_{i_*j}}{p_{i_*j}} x_{i_*j}^T x_{i_*j} (\tilde{\beta} - \hat{\beta}_N).$$

By using Lemma A1, we have

$$\frac{1}{n} \sum_{j=1}^n \frac{\delta_{i_*j}}{p_{i_*j}} x_{i_*j}^T x_{i_*j} = \frac{1}{n} \sum_{j=1}^n x_{i_*j}^T x_{i_*j} + o_p(1).$$

Then, we can obtain the expression of $\tilde{\beta} - \hat{\beta}_N$ as follows.

$$(\tilde{\beta} - \hat{\beta}_N) = -\Sigma_{i_*}^{-1} R^*(\hat{\beta}_N) + o_p(\|\tilde{\beta} - \hat{\beta}_N\|).$$

According to Lemma A2, we know that $R^*(\hat{\beta}_N)$ obeys normal distribution. That is,

$$R^*(\hat{\beta}_N) \sim N(0, V_c).$$

After a simple calculation, we obtain

$$V^{-1/2}(\tilde{\beta} - \hat{\beta}_N) \xrightarrow{d} N(0, I).$$

This is the result of Theorem 5. $\square$

Proof of Methods 1 and 2. We employ d_{ij} to represent d_{ij}^V for simplicity. Generally, we assume that all d_{ij} values are positive. In cases where some d_{ij} values equal zero, we can assign zero subsampling probabilities to the corresponding pairs and focus on the subsampling probabilities among the remaining pairs.

To achieve the minimum asymptotic mean square error, specifically $tr(V)$ as stated in Theorem 5, we can solve the following optimization problem to derive the desired method:

$$\begin{aligned} \min \quad & \tilde{H} = \sum_{i=1}^n \text{tr} \left[\frac{1}{p_{ij}} \{y_{ij} - \hat{\beta}'_N \mathbf{x}_{ij} - \hat{\mu}_i\}^2 \|\Sigma_i^{-1} \mathbf{x}_{ij}\|_2^2 \right] \\ \text{s.t.} \quad & \sum_{i=1}^n p_{ij} = r, \quad 0 \leq p_{ij} \leq 1 \quad \text{for } i = 1, \dots, n. \end{aligned}$$

Without loss of generality, we further assume that $d_{i1} \leq d_{i2} \leq \dots \leq d_{in}$.

From the Cauchy–Schwarz inequality,

$$\begin{aligned} \tilde{H} &= \sum_{i=1}^n \left[\frac{1}{p_{ij}} \{y_{ij} - \hat{\beta}'_N \mathbf{x}_{ij} - \hat{\mu}_i\}^2 \|\Sigma_i^{-1} \mathbf{x}_{ij}\|_2^2 \right] \\ &= \frac{1}{r} \left(\sum_{j=1}^n p_{ij} \right) \left(\sum_{i=1}^n [p_{ij}^{-1} d_{ij}^2] \right) \geq \frac{1}{r} \left[\sum_{i=1}^n d_{ij} \right]^2, \end{aligned}$$

where the equality in it holds if and only if $p_{ij} \propto d_{ij}$.

When $p_{ij} = rd_{ij} / (\sum_{j=1}^n d_{ij})$ satisfies that $p_{ij} \leq 1$ for all $j = 1, \dots, n$, these p_i s give the optimal solution.

Otherwise, we can easily see that $p_{in} = 1$ when $rd_{in} / (\sum_{j=1}^n d_{ij}) > 1$. Thus, the original problem turns into solving the following optimization problem:

$$\begin{aligned} \min \quad & \tilde{H} = \sum_{i=1}^{n-1} \text{tr} \left[\frac{1}{p_{ij}} \{y_{ij} - \hat{\beta}'_N \mathbf{x}_{ij} - \hat{\mu}_i\}^2 \|\Sigma_i^{-1} \mathbf{x}_{ij}\|_2^2 \right] \\ \text{s.t.} \quad & \sum_{i=1}^{n-1} p_{ij} = r - 1, \quad 0 \leq p_{ij} \leq 1 \quad \text{for } i = 1, \dots, n-1. \end{aligned}$$

This is a recursion problem. We repeat the method given above until all $p_{ij} \leq 1$, where $j \leq s$ and

$$s = \min \left\{ t \mid 0 \leq t \leq r, \quad (r-t)d_{i(n-t)} \leq \sum_{j=1}^{n-t} d_{i(j)} \right\}.$$

Then, we obtain the optimal solution

$$p_{ij} = (r-s) \frac{d_{ij}}{\sum_{k=1}^{n-s} d_{ik}}, \quad 1 \leq j \leq n-s,$$

and

$$p_{ij} = 1, \quad n-s+1 \leq j \leq n.$$

The method that we obtain for Method 2 is similar for Method 1, so we omit the details. $\square$

References

1. Huber, P.J. *Robust Statistics*; John Wiley & Sons: Hoboken, NJ, USA, 1981.
2. Walley, P. *Statistical Reasoning with Imprecise Probabilities*; Chapman and Hall: London, UK; New York, NY, USA, 1991.

3. Peng, S. Backward SDE and related g-expectations. In *Backward Stochastic Differential Equations*, Pitman Research Notes in Math. Series, No. 364; El Karoui, M., Ed.; Longman: Harlow, UK, 1997; pp. 141–159.
4. Peng, S. G-Expectation, G-Brownian Motion and Related Stochastic Calculus of Itô's type. In *Stochastic Analysis and Applications: The Abel Symposium 2005*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 541–567.
5. Peng, S. Multi-dimensional G-Brownian motion and related stochastic calculus under G-expectation. *Stoch. Process. Their Appl.* **2008**, *118*, 2223–2253. [[CrossRef](#)]
6. Peng, S. Survey on normal distributions, central limit theorem, Brownian motion and the related stochastic calculus under sublinear expectations. *Sci. China Ser. A Math.* **2009**, *52*, 1391–1411. [[CrossRef](#)]
7. Lin, L.; Shi, Y.; Wang, X.; Yang, S. k-sample upper expectation linear regression-Modeling, identifiability, estimation and prediction. *J. Stat. Plan. Infer.* **2016**, *170*, 15–26. [[CrossRef](#)]
8. Lin, L.; Liu, Y.; Lin, C. Mini-max-risk and mini-mean-risk inferences for a partially piecewise regression. *Statistics* **2017**, *51*, 745–765.
9. Meinshausen, N.; Bühlmann, P. Maximin effects in inhomogeneous large-scale data. *Ann. Stat.* **2015**, *43*, 1801–1830. [[CrossRef](#)]
10. Duchi, J.C.; Namkoong, H. Learning Models with Uniform Performance via Distributionally Robust Optimization. *Ann. Stat.* **2020**, *49*, 1378–1406.
11. Schifano, E.D.; Wu, J.; Wang, C.; Yan, J.; Chen, M.H. Online updating of statistical inference in the big data setting. *Technometrics* **2016**, *58*, 393–403. [[CrossRef](#)] [[PubMed](#)]
12. Wang, H.; Zhu, R.; Ma, P. Optimal subsampling for large sample logistic regression. *J. Am. Stat. Assoc.* **2018**, *113*, 829–844. [[CrossRef](#)] [[PubMed](#)]
13. Wang, H.; Ma, Y. Optimal subsampling for quantile regression in big data. *Biometrika* **2021**, *108*, 99–112. [[CrossRef](#)]
14. Xu, G.; Shang, Z.; Cheng, G. Distributed Generalized Cross-Validation for Divide-and-Conquer Kernel Ridge Regression and Its Asymptotic Optimality. *J. Comput. Graph. Stat.* **2019**, *28*, 891–908.
15. Ma, P.; Mahoney, M.W.; Yu, B. A statistical perspective on algorithmic leveraging. *J. Mach. Learn. Res.* **2015**, *16*, 861–919.
16. Wang, H.Y.; Yang, M.; Stufken, J. Information-based optimal subdata selection for big data linear regression. *J. Am. Stat. Assoc.* **2019**, *114*, 393–405. [[CrossRef](#)]
17. Yu, J.; Wang, H.; Ai, M.; Zhang, H. Optimal Distributed Subsampling for Maximum Quasi-Likelihood Estimators with Massive Data. *J. Am. Stat. Assoc.* **2020**, *117*, 265–276. [[CrossRef](#)]
18. Pukelsheim, F. *Optimal Design of Experiments (Classics in Applied Mathematics, 50)*; Society for Industrial and Applied Mathematics, University City Science Center: Philadelphia, PA, USA, 2006.
19. Van der Vaart, A.W. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 2000; Volume 3.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.