Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Individualized Group Learning

Chencheng Cai, Rong Chen & Min-ge Xie

To cite this article: Chencheng Cai, Rong Chen & Min-ge Xie (2023) Individualized Group Learning, Journal of the American Statistical Association, 118:541, 622-638, DOI: 10.1080/01621459.2021.1947306

To link to this article: https://doi.org/10.1080/01621459.2021.1947306

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 118, NO. 541, 622–638: Theory and Methods
https://doi.org/10.1080/01621459.2021.1947306

Individualized Group Learning

Chencheng Cai[a], Rong Chen[b], and Min-ge Xie[b]

[a]Department of Statistical Science, Fox School of Business, Temple University, Philadelphia, PA; [b]Department of Statistics, Rutgers University, Piscataway, NJ

ABSTRACT

Many massive data sets are assembled through collections of information of a large number of individuals in a population. The analysis of such data, especially in the aspect of individualized inferences and solutions, has the potential to create significant value for practical applications. Traditionally, inference for an individual in the dataset is either solely relying on the information of the individual or from summarizing the information about the whole population. However, with the availability of big data, we have the opportunity, as well as a unique challenge, to make a more effective individualized inference that takes into consideration of both the population information and the individual discrepancy. To deal with the possible heterogeneity within the population while providing effective and credible inferences for individuals in a dataset, this article develops a new approach called the individualized group learning (iGroup). The iGroup approach uses local nonparametric techniques to generate an individualized group by pooling other entities in the population which share similar characteristics with the target individual, even when individual estimates are biased due to limited number of observations. Three general cases of iGroup are discussed, and their asymptotic performances are investigated. Both theoretical results and empirical simulations reveal that, by applying iGroup, the performance of statistical inference on the individual level are ensured and can be substantially improved from inference based on either solely individual information or entire population information. The method has a broad range of applications. An example in financial statistics is presented.

ARTICLE HISTORY
Received April 2021
Accepted June 2021

KEYWORDS
Bayesian inference; Clustering; Fusion learning; Individualized Inference; Kernel smoothing; Nonparametric; Similarity meaures

1. Introduction

With the massive data sets readily available in the digital and information era, advanced statistical learning methodologies for analysis of big data are in high demand. Traditional statistical methods are often used to discover the general rule of the population. However, in many applications we are also interested in an individual entity for personalized solutions or products. For instance, in precision medicine, each patient has his/her own traits. Therefore, it is crucial and beneficial to make individualized treatments and prescribe personalized medicine (Wang et al. 2007; Qian and Murphy 2011; Zhao et al. 2012; Yang, Miescke, and McCullagh 2012; Collins and Varmus 2015; Liu and Meng 2016). In business, the so-called market of one strategy that makes a customer feel that he or she is exclusive or preferred by the firm, becomes popular for companies to design personalized products. Indeed, individualized learning and inference matters in many applications (Shen, Liu, and Xie 2020).

Since no two patients or two customers are exactly the same, heterogeneity often exists in a population. It poses a challenge to combine the data from different individuals, especially for making improved inferences in individualized learning. A class of conventional methods is to cluster/group individual entities into subgroups and, assuming homogeneity within each subgroup, then use the data in the same subgroup for statistical analysis (Binder 1978; Ng and Han 1994; Agrawal et al. 1998; Jain, Murty, and Flynn 1999; Liao 2005; Xu and Wunsch 2005; Gan, Ma, and Wu 2007; Jain 2010). The clustering and grouping in the conventional methods are typically performed in *a priori*. Such approaches have several disadvantages. First, the constitution of subgroups often depends on a predetermined total number of subgroups, which is a parameter that is either difficult or not reliable to choose in practice. Second, since analytic outcomes and inference (e.g., estimated parameters and testing) are the same for all individuals in the same subgroup, such a procedure potentially diminishes hidden local structures. More importantly, in many cases, there may not be clear-cut and well-divided subgroups in the population. In these situations, the conventional subgroup analysis may impose an artificial grouping structure to the population, which can potentially lead to large biases and invalid inference for many individuals. Another class of conventional methods is to assume mixture models, including classical hierarchical models and Bayesian nonparametric models (Duda and Hart 1973; Ferguson 1973; Antoniak 1974; Lo 1984; Lindsay 1995; Figueiredo and Jain 2000; Teh et al. 2005). Similar to the clustering method, the mixture models assume that the population contains several homogeneous subpopulations, but unlike clustering, there is no clear boundary between the subpopulations. However, inference on each individual is not the focus of such a procedure. It is often

CONTACT Rong Chen *rongchen@stat.rutgers.edu* Department of Statistics, Rutgers University, Piscataway, NJ 08901-8554.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

done as an afterthought, by estimating the mixture likelihood. Furthermore, a mixture model may not be able to explain the population heterogeneity when the assumed latent structure is invalid. In addition, when given an observation, it is usually difficult to tell which subpopulation it belongs to.

In this article, we propose a new method called individualized group learning, abbreviated as *iGroup*. Instead of grouping at the population level, the iGroup approach focuses on each individual and forms an individualized group for the target individual, by locating individuals that share similar characteristics of the target. It sidesteps aforementioned difficulties by forming an iGroup specifically for the target individual while ignoring other entities that have little in common with the target.

In this article, two sets of information are used in our proposed framework to define similarity and to form groups. One is individual level estimator $\hat{\theta}_k$, which is a direct estimation of θ_k, the parameter of interest, for each individual $k \in \{0, 1, \ldots, K\}$ in a parametric model with observation $\boldsymbol{x}_k$, without any grouping. The other is some additional information $\boldsymbol{z}_k$, which is not directly related to θ_k but can reveal similarity between the individuals as well as their parameters. Both $\hat{\theta}_k$ and $\boldsymbol{z}_k$ can provide useful information in identifying groups so that closeness in the space of $(\hat{\theta}_k, \boldsymbol{z}_k)$ implies closeness in the space of θ_k. Depending on the feasibility and availability of the two information sets, iGroup can be constructed based on three different information sets: $\{\hat{\theta}_k\}$, $\{\boldsymbol{z}_k\}$, $\{\hat{\theta}_k, \boldsymbol{z}_k\}$. They will be discussed in detail in later sections.

To ease our notation, from now on, let us say our goal is to provide an estimation on θ_0 for the individual 0. The estimator is constructed with a specified loss function L, the observations $(\boldsymbol{x}_0, \boldsymbol{z}_0)$ on individual 0 and all other available observations $\mathcal{D}_x = \{\boldsymbol{x}_k\}_{k=1}^K$ and $\mathcal{D}_z = \{\boldsymbol{z}_k\}_{k=1}^K$. By focusing on individualized local structures, the proposed iGroup learning is robust and effective for handling heterogeneity arising from diverse sources in big data, and it is ideally suited for specific objective-oriented applications in an individualized inference. Additionally, in terms of computation, by ignoring a large number of irrelevant entities and zooming directly to the relevant individuals, the iGroup learning is parallel in nature and can scale up better for big data. In this article, we investigate the validity and theoretical property of iGroup learning and provide simulation studies and applications to demonstrate the grouping effectiveness of the proposed methodology.

As all individualized inference is based on borrowing the information from "other" individuals with similar features or characteristics as the individual under study, it inevitably resembles nonparametric kernel smoothing methods and k-nearest neighbor methods in many ways. In a way, the regular nonparametric smoothing methods can be viewed as a special case of individualized inference. In this article, we focus on a more general class of individualized inference problems that are not covered by the existing standard kernel smoothing methods, even though our approach remains under the principle of finding "similar" individuals; hence, resembles nonparametric kernel smoothing methods. Major differences, including problem setting, objectives, error in features, aggregation of different source of information, and theoretical foundation, will be pointed out throughout the article and will be summarized in the Section 6.

Another closely related to the recent development is the individualized fusion learning (*i*Fusion) approach proposed in Shen, Liu, and Xie (2020). The *i*Fusion approach is developed under the asymptotic settings that $n_k \to \infty$, $n_k / \sum_{i=1}^K n_k = O(1)$ and K is large but finite, where $n_k = |\boldsymbol{x}_k|$ is the effective sample size for individual k. The requirement that $n_k \to \infty$ ensures the individual studies are not biased, which permits Shen, Liu, and Xie (2020) to directly extend the standard theory in the kernel smoothing literature to demonstrate that the *i*Fusion approach is effective with good theoretical properties (including consistency, oracle efficiency, and asymptotic normality) under their assumed setting. Furthermore, the target neighbor, referred to as *clique* in the *i*Fusion approach, is defined only through the parameter space using $\{\hat{\theta}_k\}$'s. The *i*Group approach in this article, however, focuses on a different setting where each individual has only a limited number of observations with $n_k = O(1)$ and infinite numbers of individuals are available as $K \to \infty$, under which *i*-Fusion is not applicable. A key development of the proposed iGroup method is that we need to make the efforts to develop new theories to overcome the biases from individual estimates, a task that is not covered by the standard kernel smoothing methods and *i*Fusion. Furthermore, in addition to borrow information through $\{\hat{\theta}_k\}$, we also investigate how we can effectively borrow strength from other individuals when the information sets $\{\boldsymbol{z}_k\}$ and $\{\hat{\theta}_k, \boldsymbol{z}_k\}$ are available.

The proposed iGroup methodology has a wide range of applications. In general, the proposed method can be useful in situations where there are two sources of information: some limited amount of data on the individual level and some extra features that provide similarity measures among individuals. For example, the situation arises in evaluation of risk scores of companies with their own financial data with additional company features for borrowing information from similar companies; in prediction sales volume of many products with short historical time series with additional product features for similarity measures; in assessment of treatment effects in health research with temporal measurements and individual characteristics of patients; and many others. In this article, we demonstrate the application of iGroup method in financial risk management and compare iGroup to some existing approaches. The example is on improve the individualized inference in estimation of value at risk (VaR) (prediction of a small quantile value of future stock return distribution). It is a difficult problem due to the lack of observations in the tail of the distribution; hence, it is naturally beneficial to borrow information from other stocks with similar features or behavior of the specific stock (company) of interest. Features such as industry sectors and various financial characteristics can be used. In this article we use the linear relationship (represented by the coefficients of a linear model) between the stock returns of the company and the three Fama–French factors common to all stocks. The Fama–French Model is a popular model commonly used in finance and the coefficients reflects the riskiness, size and market perceived growth potential of the underlying company. Empirical study shows that, by using the estimated coefficients of the Fama–French model as the additional information $\boldsymbol{z}_k$ in the iGroup approach, we were able to obtain more accurate and robust estimator of VaR. For more detailed information about Fama–

French factors and our iGroup approach, see Section 5. The observed features and estimates with nonignorable errors are used to form individualized group to improve inference for the target individuals.

The rest of the article is arranged as below. In Section 2, we introduce the general framework of iGroup learning. Section 3 focuses on three different information sets with asymptotic analysis and theoretical results. Section 4 provides three simulated studies and Section 5 provides a real application on financial risk management. Section 6 concludes.

2. General Framework

2.1. Problem Setup

Assume for each individual $k \in \{0, 1, 2, \ldots, K\}$, we observe $(\boldsymbol{x}_k, \boldsymbol{z}_k)$, where observations $\boldsymbol{x}_k$ and $\boldsymbol{z}_k$ differ in their utilities. Specifically, $\boldsymbol{x}_k$ is the observed data that is directly related to the parameter of interest θ_k at the individual level, with a known distribution $\boldsymbol{x}_k \sim p(\cdot|\theta_k)$. The exogenous variable $\boldsymbol{z}_k$ serves as a proxy that reveals the similarity among θ's in the population level. Specifically, we assume that $\boldsymbol{z}_k$ is related to an unknown parameter $\boldsymbol{\eta}_k$ through an unknown distribution $q(\cdot; \boldsymbol{\eta}_k)$, and the parameter θ is an unknown continuous function of $\boldsymbol{\eta}$, that is, $\theta = g(\boldsymbol{\eta})$, where the function $g(\cdot)$ is not necessarily an one-to-one mapping. The continuity of $g(\cdot)$ guarantees that closeness in $\boldsymbol{\eta}$ implies closeness in θ. The hierarchical structure and the relationship among the variables are demonstrated in Figure 1, where $\pi(\cdot)$ is an unknown (prior) population distribution of θ, which may be heterogeneous in nature. Although $\pi(\cdot)$ is unknown and unspecified, it appears in the calculations when we study the theoretical property of the proposed approach. The distribution $p(\cdot; \theta_k)$ is known except the parameter θ_k, but both the function $g(\cdot)$ and the distribution $q(\cdot; \cdot)$ are unknown. The role of the exogenous variable $\boldsymbol{z}_k$ will be discussed further in later sections. In some cases $\boldsymbol{z}_k$ may not be available. Without further clarification, all unconditioned expectations $\mathbb{E}[\cdot]$ are assumed to take over all random variables including θ_k, which follows the unknown prior $\pi(\cdot)$. Posterior expectations on θ conditioned on certain observed information are explicitly noted with π in the subscript such as $\mathbb{E}_\pi[\theta_0|\hat{\theta}_0]$.

In the VaR example in Section 5, we are interested in the 1% quantile value $q_{0.01}$ of the underlying return distribution of stock k, with $\boldsymbol{x}_k$ being the recent 100 day's (assumed to follow the same distribution but not necessarily independent). The parameter θ_k is $q_{0.01}$ of the stock k at time t. The parameter $\boldsymbol{\eta}_k$ and exogenous variable $\boldsymbol{z}_k$ are the underlying true and estimated Fama–French coefficients of the returns of the stock in the same period.

Denote by $C_0(\epsilon) = \{k|\tilde{d}(\theta_k, \theta_0) < \epsilon, k = 0, \ldots, K\}$ an ϵ-neighborhood (or a *clique*) of individual 0, where $\tilde{d}(\cdot, \cdot)$ is a distance/similarity measure and ϵ is the threshold value. Thus, the clique $C_0(\epsilon)$ is a set of indexes of individuals that are similar to individual 0. In our model development, we impose two regularity assumptions as below.

Assumption 1 (Dense Assumption). There exists a constant $d \geqslant 1$ such that for all $i = 1, \ldots, K$, $|C_0(\epsilon)| \asymp K\epsilon^d$ in probability when $K \to \infty, \epsilon \to 0$.

Assumption 2 (Smooth Parameter Assumption). There exists a positive constant κ, such that for all $\theta, \theta' \in \Omega_\theta$

$$\sup_{\boldsymbol{x}} |p(\boldsymbol{x}; \theta) - p(\boldsymbol{x}; \theta')| \leqslant \kappa \|\theta - \theta'\|,$$

where $\|\cdot\|$ is a metric on Ω_θ.

The dense assumption suggests that individual 0 of interest is not isolated from other individuals, that is, for arbitrarily small ϵ, there are a sufficiently large number of other individuals in its neighborhood as $K \to \infty$. The smooth parameter assumption guarantees that whenever θ and θ' are close, the distributions of $\boldsymbol{x}$ and $\boldsymbol{x}'$ induced from θ and θ', respectively, are close to each other. Under these two assumptions, it is beneficial to aggregate information from the neighborhood to estimate θ since one can always find sufficient number of similar individuals in the neighborhood of individual θ. A key consideration in this aggregation is the familiar bias-variance tradeoff—aggregation over a larger group increases the sample size thus reduces estimation variance, but it also brings bias.

The model setting shown in Figure 1 can be generalized as follows: Support $\boldsymbol{x}_k$ follows an underlying distribution $\sim p_k$ and $\theta_k = \theta(p_k)$ is a known functional of the underlying distribution p_k. Assumption 2 insures the smoothness of $\theta(\cdot)$ as a function of p_k. In addition, we also have $\theta_k = f(\boldsymbol{\eta}_k)$. Hence, for the estimation of θ_0 for individual 0, we can use two sources of information, $\boldsymbol{x}_0$ for p_0 and $\boldsymbol{z}_0$ for η_0. Moreover, because of the smoothness assumptions of the functions of $\theta(\cdot)$ and $f(\cdot)$, we can borrow information from other individuals with p_i similar to p_0 and $\boldsymbol{\eta}_k$ similar to $\boldsymbol{\eta}_0$, through the use of $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$. Note that the objective is not to estimate the function $f(\cdot)$ though some components of the proposed estimator is similar to a nonparametric estimation of it. The objective to use both sets of information from other individuals to obtain a more efficient estimator of θ_0. For clarity of presentation, we express our estimators and theoretical results under the setting shown in Figure 1, though they can be easily extended to cover the more general setting, with careful treatment of any latent parameters in a parametric formulation of p_k, or in a nonparametric formulation of p_k.

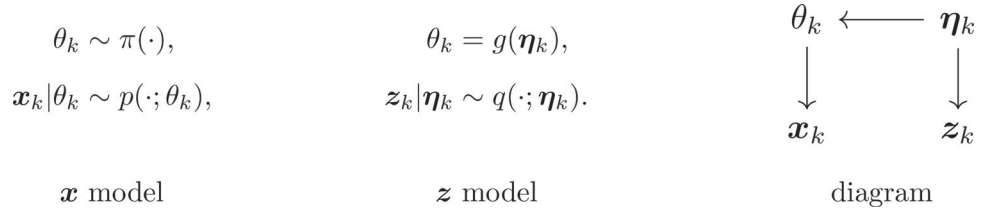

Figure 1. Hierarchical structure and parameter diagram.

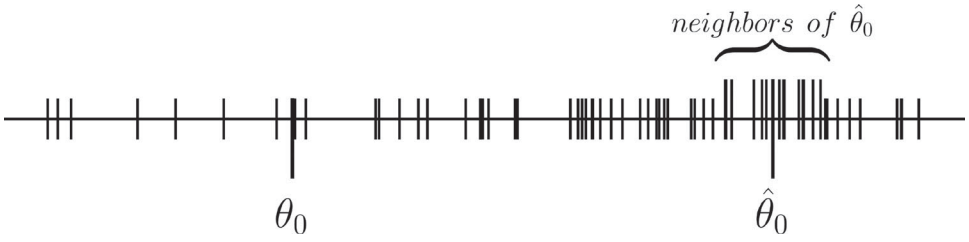

Figure 2. A one-dimension example in which $\hat{\theta}_0$ is away from θ_0. If one naively selects individuals according to $\hat{\theta}_0$ and $\hat{\theta}_k$ directly, individuals adjacent to $\hat{\theta}_0$, but not those close to θ_0, are often selected.

2.2. Aggregated Estimation in iGroup

There are two common methods to aggregate information by creating "pooled" estimators for θ_0. The first approach constructs a weighted estimator $\hat{\theta}_0^{(c)}(\boldsymbol{x}_0, \boldsymbol{z}_0, \mathcal{D}_x, \mathcal{D}_z)$ for the target individual 0, directly using the point estimators $\hat{\theta}_k$ of other individuals based on $\boldsymbol{x}_k$. The second approach aggregates objective functions $M_k(\theta) = M_k(\theta, \boldsymbol{x}_k)$ of other individuals, where the point estimator $\tilde{\theta}_0^{(c)}$ is obtained by optimizing an aggregated objective function. Specifically, these two methods can be formulated as

$$\begin{aligned}(\text{Aggregating estimators})\hat{\theta}_0^{(c)} &= \frac{\sum_{k=0}^{K}\hat{\theta}_k w(k;0)}{\sum_{k=0}^{K} w(k;0)} \\ &= \frac{\sum_{k=0}^{K}\theta(\boldsymbol{x}_k) w(k;0)}{\sum_{k=0}^{K} w(k;0)}, \qquad (1)\end{aligned}$$

$$(\text{Aggregating objective functions})\tilde{\theta}_0^{(c)} = \arg\min_{\theta}\sum_{k=0}^{K} M_k(\theta, \boldsymbol{x}_k) w(k;0), \qquad (2)$$

where $w(k;0)$ is the weight assigned to individual k when constructing iGroup estimator for individual 0. In Equation (1), we point out that $\hat{\theta}_k = \theta(\boldsymbol{x}_k)$ is a given estimator $\theta(\cdot)$ of the observation $\boldsymbol{x}_k$, though we will use $\hat{\theta}_k$ for simplicity. In practice, one can choose either $\hat{\theta}_0^{(c)}$ or $\tilde{\theta}_0^{(c)}$ based on the availability of the point estimator $\hat{\theta}_k$ and the objective function M_k. In fact, the aggregating estimator in Equation (1) is a special case of Equation (2) when one uses the squared loss function $M_k(\theta, \boldsymbol{x}_k) = (\theta(\boldsymbol{x}_k) - \theta)^2$. However, the analytical form of Equation (1) is more intuitive and easier to analyze. Other loss functions such as the log-likelihood functions or other functions that leads to M-estimators can be used in Equation (2). As to be shown in Section 3, both estimators have similar convergence rate, under certain conditions on the loss function used. Discussion on loss function is provided in Section 3.1.

The weight $w(k;0)$ is crucial for the aggregated estimators as it controls how much information is borrowed from other individuals. We propose to incorporate both individual level estimator $\hat{\theta}_k$ and exogenous observation $\boldsymbol{z}_k$ into the weight function as both can provide useful information of θ_0. Specifically, let

$$w(k;0) = w(\hat{\theta}_k, \boldsymbol{z}_k; \hat{\theta}_0, \boldsymbol{z}_0) = w_1(\boldsymbol{z}_k, \boldsymbol{z}_0) w_2(\hat{\theta}_k, \hat{\theta}_0 | \boldsymbol{z}_0, \boldsymbol{z}_k). \qquad (3)$$

The weight is decomposed into two parts. The first part $w_1(\boldsymbol{z}_k, \boldsymbol{z}_0)$ measures the similarity between $\boldsymbol{z}_k$ and $\boldsymbol{z}_0$, and can be a kernel function

$$w_1(\boldsymbol{z}_k, \boldsymbol{z}_0) = \mathcal{K}_1\left(\frac{\|\boldsymbol{z}_k - \boldsymbol{z}_0\|}{b_1}\right), \qquad (4)$$

When $\mathcal{K}_1$ has a finite support, the weight function has a hard grouping structure—individuals lying far enough from individual 0 are not considered at all. Otherwise, it has a soft grouping structure such that dissimilar individuals are assigned with nonzero but tiny weights.

The second part $w_2(\hat{\theta}_k, \hat{\theta}_0 | \boldsymbol{z}_0)$ measures the similarity between $\hat{\theta}$'s. Again, $\hat{\theta}_k = \theta(\boldsymbol{x}_k)$ is a function of $\boldsymbol{x}_k$. One can view $\hat{\theta}_k$ as a low-dimensional summary statistic of the high dimensional observation $\boldsymbol{x}_k$, and $w_2(\hat{\theta}_k, \hat{\theta}_0 | \boldsymbol{z}_0) = w_2(\theta(\boldsymbol{x}_k), \theta(\boldsymbol{x}_0) | \boldsymbol{z}_0)$ in fact measures the distance between $\boldsymbol{x}_k$ and $\boldsymbol{x}_0$, through the function $\theta(\cdot)$. However, unlike $w_1(\cdot)$, using a distance measure such as $\mathcal{K}_2(\|\hat{\theta}_k - \hat{\theta}_0\|/b_2)$ is not a good practice, since it is directly correlated with $\hat{\theta}_k$ in the weighted average operation in (1) and (2). It ignores the error in $\hat{\theta}_0$ and $\hat{\theta}_k$ and $\hat{\theta}_0$ may be biased. Note that when $K \to \infty$ and $b_2 \to 0$, the kernel concentrates on a smaller and smaller area adjacent to $\hat{\theta}_0$. In this area, aggregating individual $\hat{\theta}_k$ will not improve the estimation of θ_0. An example of one-dimensional case is shown in Figure 2. Vertical bars mark the locations of $\hat{\theta}_k$. When $\hat{\theta}_0$ is away from its target value θ_0, a small bandwidth b_2 tends to give large weights to individuals in a local region around $\hat{\theta}_0$. Aggregating these individual $\hat{\theta}_k$ in such a local region will not correct the bias $\hat{\theta}_0 - \theta_0$.

We propose the following weight function that considers the distribution $p(\hat{\theta}|\theta)$ instead of the point estimator $\hat{\theta}$. Specifically, let

$$w_2(\hat{\theta}_k, \hat{\theta}_0 | \boldsymbol{z}_0, \boldsymbol{z}_k) = \frac{\int p(\hat{\theta}_k|\theta) p(\hat{\theta}_0|\theta) p(\theta|\boldsymbol{z}_0) d\theta}{p(\hat{\theta}_k|\boldsymbol{z}_k) p(\hat{\theta}_0|\boldsymbol{z}_0)}. \qquad (5)$$

Notice that, the posterior distribution of θ_0, given $(\hat{\theta}_0, \boldsymbol{z}_0)$, is

$$\begin{aligned}p(\theta_0|\hat{\theta}_0, \boldsymbol{z}_0) &= p(\theta_0, \hat{\theta}_0|\boldsymbol{z}_0)/p(\hat{\theta}_0|\boldsymbol{z}_0) \\ &= p(\hat{\theta}_0|\theta_0) p(\theta_0|\boldsymbol{z}_0)/p(\hat{\theta}_0|\boldsymbol{z}_0).\end{aligned}$$

If $\theta_k \equiv \theta_0$ (hence $\hat{\theta}_k$ provides useful information about θ_0), then the predictive distribution of $\hat{\theta}_k$, given $(\hat{\theta}_0, \boldsymbol{z}_0)$, is

$$\begin{aligned}p(\hat{\theta}_k|\hat{\theta}_0, \boldsymbol{z}_0) &= \int p(\hat{\theta}_k|\theta) p(\theta|\hat{\theta}_0, \boldsymbol{z}_0) d\theta \\ &= \frac{\int p(\hat{\theta}_k|\theta) p(\hat{\theta}_0|\theta) p(\theta|\boldsymbol{z}_0) d\theta}{p(\hat{\theta}_0|\boldsymbol{z}_0)}.\end{aligned}$$

Thus, the weight function $w_2(\hat{\theta}_k, \hat{\theta}_0 | \boldsymbol{z}_0, \boldsymbol{z}_k)$ in Equation (5) is the Radon–Nikodym derivative between the predictive distribution $p(\hat{\theta}_k|\hat{\theta}_0, \boldsymbol{z}_0)$ and the sampling distribution $p(\hat{\theta}_k|\boldsymbol{z}_k)$. As a result, for any measurable function $h(\cdot)$, we have

$$\mathbb{E}_{p(\hat{\theta}_k|\boldsymbol{z}_k)}[h(\hat{\theta}_k) w_2(\hat{\theta}_k, \hat{\theta}_0 | \boldsymbol{z}_0, \boldsymbol{z}_k)] = \mathbb{E}_{p(\hat{\theta}_k|\hat{\theta}_0, \boldsymbol{z}_0)}[h(\hat{\theta}_k)].$$

That is, the weighted expectation of $h(\hat{\theta}_k)$ under the sampling distribution $p(\hat{\theta}_k|\boldsymbol{z}_k)$ equals to its expectation under the predictive distribution $p(\hat{\theta}_k|\hat{\theta}_0, \boldsymbol{z}_0)$ if $\theta_k = \theta_0$. This property brings invariance under different sampling distributions. More importantly, it shows that the weighted averages, such as Equations (1) and (2), estimates the expectations under the predictive distribution. This gives the iGroup estimators promising asymptotic properties as we will discuss later in Section 3.

The shape (thin or flat) of the weight $w_2(\cdot)$ as a function of $\hat{\theta}_k$ does not change with the number of individuals K. However, the shape is influenced by the variation (accuracy) of $\hat{\theta}$. The larger the variance of $\hat{\theta}$ is, the flatter the weight function tends to be. If $\hat{\theta}_k$ is estimated without any measurement error, the weight $w_2(\hat{\theta}_k, \hat{\theta}_0|\boldsymbol{z}_0, \boldsymbol{z}_k)$ is proportional to the indicator function $I_{\{\hat{\theta}_k=\hat{\theta}_0\}}$. It reduces to the case in which the individual estimator $\hat{\theta}_0$ or the individual objective function $M_0(\theta)$ is used without grouping.

2.3. Evaluating the Weight Functions

The weight function $w_1(\boldsymbol{z}_k, \boldsymbol{z}_0)$ in Equation (4) can be directly evaluated. Similar to a bandwidth selection problem for kernel smoothing, one can choose the bandwidth b_1 for $w_1(\boldsymbol{z}_k, \boldsymbol{z}_0)$ in Equation (4) by either using the plug-in method (Chiu 1991) or through cross-validation procedure. The plug-in bandwidth is proportional to $K^{-\frac{1}{d+4}}$ (see Section 3). Also, the leave-one-out cross-validation process gives an empirical optimal bandwidth, as discussed in Section 3.6.

The evaluation of the weight function $w_2(\hat{\theta}_k, \hat{\theta}_0|\boldsymbol{z}_0, \boldsymbol{z}_k)$ in Equation (5) is more complicated, since the conditional probability $p(\hat{\theta}|\boldsymbol{z})$ and the integral $\int p(\hat{\theta}_0|\theta)p(\hat{\theta}_k|\theta)p(\theta|\boldsymbol{z}_0)d\theta$ are unknown as the relationship between θ and $\boldsymbol{z}$ is not explicit. We propose an approximation method to evaluate $w_2(\hat{\theta}_k, \hat{\theta}_0|\boldsymbol{z}_0, \boldsymbol{z}_k)$ below.

Denote the estimator of θ_k and the observed exogenous variable $\boldsymbol{z}_k$ as the tuple $(\hat{\theta}_k, \boldsymbol{z}_k), k = 0, \ldots, K$. To calculate the weight in Equation (5), we treat them as $K+1$ samples from the joint distribution of $(\hat{\theta}, \boldsymbol{z})$. We use the kernel method to estimate the conditional probability $p(\hat{\theta}|\boldsymbol{z})$ nonparametrically by

$$\hat{p}(\hat{\theta}|\boldsymbol{z}) = \frac{\sum_{j=0}^{K} \mathcal{K}_1\left(\frac{\|\boldsymbol{z}-\boldsymbol{z}_j\|}{b_1}\right)\mathcal{K}_2\left(\frac{\|\hat{\theta}-\hat{\theta}_j\|}{b_2}\right)}{\sum_{j=0}^{K} \mathcal{K}_1\left(\frac{\|\boldsymbol{z}-\boldsymbol{z}_j\|}{b_1}\right)},$$

where $\mathcal{K}_1, \mathcal{K}_2$ are two kernel functions with b_1, b_2 as the corresponding bandwidths. To estimate the integral in (5), we use the interpretation discussed above that it is the conditional distribution $p(\hat{\theta}_k|\hat{\theta}_0, \boldsymbol{z}_0)$ given $\theta_k = \theta_0$. Hence, we need samples from the joint distribution of $(\hat{\theta}, \hat{\theta}', \boldsymbol{z})$ observed from the same individual with parameter θ. However, this is infeasible because in our problem setting, no two individual share the same true parameter θ and for each individual only one $\hat{\theta}$ is observed. To generate samples from such a distribution, we consider a bootstrap method. Denote $\hat{\theta}_k^{(1)}$ and $\hat{\theta}_k^{(2)}$ as the two bootstrap estimators for θ_k, obtained by resampling $\boldsymbol{x}_k$ with replacement (not applicable when $\boldsymbol{x}_k$ has few observations). Then $(\hat{\theta}_k^{(1)}, \hat{\theta}_k^{(2)}, \boldsymbol{z}_k), k = 0, \ldots, K$ is an approximate sample of $(\hat{\theta}, \hat{\theta}', \boldsymbol{z})$, guaranteeing $\hat{\theta}_k^{(1)}, \hat{\theta}_k^{(2)}, \boldsymbol{z}_k$ are generated from the same individual k. Therefore, the integral can be estimated by

$$\int p(\hat{\theta}_0|\theta)p(\hat{\theta}_k|\theta)p(\theta|\boldsymbol{z}_0)d\theta \approx \frac{\sum_{j=0}^{K} \mathcal{K}_1\left(\frac{\|\boldsymbol{z}_0-\boldsymbol{z}_j\|}{b_1}\right)\mathcal{K}_2\left(\frac{\|\hat{\theta}_0-\hat{\theta}_j^{(1)}\|}{b_2}\right)\mathcal{K}_3\left(\frac{\|\hat{\theta}_k-\hat{\theta}_j^{(2)}\|}{b_3}\right)}{\sum_{j=0}^{K} \mathcal{K}_1\left(\frac{\|\boldsymbol{z}_0-\boldsymbol{z}_j\|}{b_1}\right)}, \quad (6)$$

where $\mathcal{K}_1, \mathcal{K}_2$, and $\mathcal{K}_3$ are three kernel functions with b_1, b_2, and b_3 as the corresponding bandwidths. The bandwidths can be selected by either minimizing asymptotic mean integrated squared error (AMISE) or a rule-of-thumb bandwidth estimator. This estimation of the integral is an approximation that requires K to be sufficiently large.

3. Theoretical Results

In this section, we consider several model settings for which we apply the proposed iGroup method and discuss their corresponding theoretical properties, especially in terms of their asymptotic performance. In particular, we first define a target estimator Θ_0 that minimizes the Bayes risk, and then investigate the asymptotic performance of iGroup estimators in Equations (1) and (2) in approximating the target estimator Θ_0. We also quantify the bias and variance of iGroup estimators as well as the target estimator Θ_0 in term of estimating θ_0. Throughout this article, we consider the asymptotic framework that the number of individuals K goes to infinity, while the number of observations for each individual n is fixed and finite.

3.1. Risk Decomposition and the Target Estimator

We are interested in making inference about individual 0, with given data information $\mathcal{D}_x, \mathcal{D}_z$ that may include the observations $\boldsymbol{x}_0$ and $\boldsymbol{z}_0$ plus information from other relevant individuals. Let $\delta_0(\mathcal{D}_x, \mathcal{D}_z)$ be a point estimator for θ_0, which is constructed with information sets $\mathcal{D}_x$ and $\mathcal{D}_z$. The iGroup estimator $\hat{\theta}_0^{(c)}$ in (1) is such an estimator. Similarly, $\delta_0(\mathcal{D}_x)$ and $\delta_0(\mathcal{D}_z)$ are point estimators constructed solely based on either $\mathcal{D}_x$ or $\mathcal{D}_z$. Under squared loss, the overall risk of δ_0 in estimating θ_0 can be decomposed into two nonnegative parts: the expected squared error of δ_0 in estimating the corresponding posterior mean and the overall risk of the posterior mean itself, as shown in Proposition 1.

Proposition 1. Suppose θ_0 has a prior distribution $\pi(\cdot)$. Under squared loss, we have the following overall risk decomposition:

$$\begin{aligned}
\mathbb{E}[(\delta_0(\mathcal{D}_x, \mathcal{D}_z) - \theta_0)^2] &= \mathbb{E}[(\delta_0(\mathcal{D}_x, \mathcal{D}_z) - \mathbb{E}_\pi[\theta_0|\boldsymbol{x}_0, \boldsymbol{z}_0])^2] \\
&\quad + \mathbb{E}[(\mathbb{E}_\pi[\theta_0|\boldsymbol{x}_0, \boldsymbol{z}_0] - \theta_0)^2], \\
\mathbb{E}[(\delta_0(\mathcal{D}_x) - \theta_0)^2] &= \mathbb{E}[(\delta_0(\mathcal{D}_x) - \mathbb{E}_\pi[\theta_0|\boldsymbol{x}_0])^2] \\
&\quad + \mathbb{E}[(\mathbb{E}_\pi[\theta_0|\boldsymbol{x}_0] - \theta_0)^2], \\
\mathbb{E}[(\delta_0(\mathcal{D}_z) - \theta_0)^2] &= \mathbb{E}[(\delta_0(\mathcal{D}_z) - \mathbb{E}_\pi[\theta_0|\boldsymbol{z}_0])^2] \\
&\quad + \mathbb{E}[(\mathbb{E}_\pi[\theta_0|\boldsymbol{z}_0] - \theta_0)^2],
\end{aligned}$$

where $\mathbb{E}_\pi[\theta_0|\boldsymbol{x}_0,\boldsymbol{z}_0]$, $\mathbb{E}_\pi[\theta_0|\boldsymbol{x}_0]$ and $\mathbb{E}_\pi[\theta_0|\boldsymbol{z}_0]$ are the posterior means under prior $\pi(\cdot)$ and observations $(\boldsymbol{x}_0,\boldsymbol{z}_0)$, $\boldsymbol{x}_0$ and $\boldsymbol{z}_0$ correspondingly.

The proof is given in the supplemental material.

Proposition 1 reveals that the overall risk is minimized by setting δ_0 to the corresponding posterior mean under the prior $\pi(\cdot)$, which is the population-level (unknown) distribution for θ_0. Throughout this article, we call the estimator that minimizes the overall risk the *target estimator*. More specifically, under squared loss and different information sets, we denote the target estimators with

$$\Theta_0(\boldsymbol{x}_0;\ell_2)=\mathbb{E}_\pi[\theta_0|\boldsymbol{x}_0],\quad \Theta_0(\boldsymbol{z}_0;\ell_2)=\mathbb{E}_\pi[\theta_0|\boldsymbol{z}_0] \text{ and } \Theta_0(\boldsymbol{x}_0,\boldsymbol{z}_0;\ell_2)=\mathbb{E}_\pi[\theta_0|\boldsymbol{x}_0,\boldsymbol{z}_0]. \tag{7}$$

Here, ℓ_2 refers to the squared loss. For the ease of presentation, we also use a simple notation Θ_0 to represent one of the Bayes estimators in Equation (7) when its meaning is apparent.

Similarly, for a general loss function $L(\hat\theta,\theta)$, we define the target estimator as the Bayes estimator that minimizes the expected loss, given the available observation on individual 0 and the prior $\pi(\cdot)$ such that

$$\begin{aligned}\Theta_0(\boldsymbol{x}_0;L)&=\arg\min_\delta \mathbb{E}_\pi[L(\delta,\theta_0)|\boldsymbol{x}_0],\\ \Theta_0(\boldsymbol{z}_0;L)&=\arg\min_\delta \mathbb{E}_\pi[L(\delta,\theta_0)|\boldsymbol{z}_0],\\ \Theta_0(\boldsymbol{x}_0,\boldsymbol{z}_0;L)&=\arg\min_\delta \mathbb{E}_\pi[L(\delta,\theta_0)|\boldsymbol{x}_0,\boldsymbol{z}_0].\end{aligned} \tag{8}$$

A similar risk decomposition is demonstrated in Proposition 2. Again, for the ease of notation, we simply use Θ_0 to represent one of the Bayes estimators in Equation (8) when its meaning is apparent.

Proposition 2. Suppose θ_0 has a prior distribution $\pi(\cdot)$ and $L(\hat\theta,\theta)$ is a loss function, which is the second-order partially differentiable with respect to $\hat\theta$ such that $L'(\hat\theta,\theta)=\partial L/\partial\hat\theta$ and $L''(\hat\theta,\theta)=\partial^2 L/\partial\hat\theta^2$. Then for estimator δ_0 constructed based on information set $\mathcal{D}_x$, $\mathcal{D}_z$ or $(\mathcal{D}_x,\mathcal{D}_z)$, we have

$$\mathbb{E}[L(\delta_0,\theta_0)]=\frac{1}{2}\mathbb{E}[L''(\Theta_0,\theta_0)(\delta_0-\Theta_0)^2]+\mathbb{E}[L(\delta_0,\theta_0)]+o(\mathbb{E}[(\delta_0-\Theta_0)^2]),$$

where Θ_0 is the corresponding Bayes estimator based on the same information set as δ_0.

The proof is given in the supplemental material.

The target estimator Θ_0 as a function of $\boldsymbol{x}_0$ and $\boldsymbol{z}_0$ is not directly available, because neither the population distribution $\pi(\theta_0)$ nor the likelihood function $p(\boldsymbol{z}_0|\theta_0)$ is explicitly known or assumed. The iGroup estimator $\hat\theta_0^{(c)}$ in (1) constructed based on observed finite sample $\mathcal{D}_x,\mathcal{D}_z$ is desired to approach the target estimator Θ_0 when more and more similar individuals contribute to the estimator $\hat\theta_0^{(c)}$. See Diaconis and Freedman (1986) for discussions of target point estimators and target parameters in the Bayesian literature.

3.2. Case 1: With Exogenous Variable z Only

In the cases when the individual-level estimator $\hat\theta_k$ is not reliable to construct the individual groups, iGroup may be constructed with the exogenous variable $\boldsymbol{z}$ only. In this case, the corresponding target estimator is defined as follows:

$$\Theta_0(\boldsymbol{z}_0;\ell_2)=\mathbb{E}_\pi[\theta_0|\boldsymbol{z}_0], \tag{9}$$

where $p(\theta_0|\boldsymbol{z}_0)\propto p(\boldsymbol{z}_0|\theta_0)\pi(\theta_0)$. Although $\boldsymbol{x}_0$ is not used for grouping and thus does not appear in Equation (9), the data $\mathcal{D}_x$ are used in iGroup estimators in Equations (1) and (2).

Recall that the relationship between θ_k and $\boldsymbol{\eta}_k$ is given by a deterministic relationship

$$\theta_k=g(\boldsymbol{\eta}_k),\quad \text{for } k=0,1,\ldots,K, \tag{10}$$

where $g(\cdot)$ is an unknown continuous function. Furthermore, $\boldsymbol{z}_k$ is a noisy observation of $\boldsymbol{\eta}_k$. Since $\boldsymbol{\eta}$ is a conceptual parameter, we may simply assume that

$$\boldsymbol{z}_k=\boldsymbol{\eta}_k+\epsilon_k,\quad \text{for } k=0,\ldots,K,$$

where the error satisfies $\mathbb{E}(\epsilon_k)=0$, $\text{var}(\epsilon_k)=\sigma_z^2\boldsymbol{\Sigma}_z$ with $\|\boldsymbol{\Sigma}_z\|=1$.

Suppose $\hat\theta_k$ is an unbiased estimator of θ_k. Then, the aggregated estimator in Equation (1) with $w(0,k)=w_1(\boldsymbol{z}_0,\boldsymbol{z}_k)$ becomes

$$\hat\theta_0^{(c)}=\frac{\sum_{k=0}^K \mathcal{K}\left(\frac{\|\boldsymbol{z}_k-\boldsymbol{z}_0\|}{b}\right)\hat\theta_k}{\sum_{k=0}^K \mathcal{K}\left(\frac{\|\boldsymbol{z}_k-\boldsymbol{z}_0\|}{b}\right)}. \tag{11}$$

In this specific case, the iGroup estimator is exactly a standard kernel smoothing estimator, except that $\boldsymbol{z}_k-\boldsymbol{z}_0$ is a noisy version of the ideal distance measure $\boldsymbol{\eta}_k-\boldsymbol{\eta}_0$. As a result, the target estimator here is Θ_0 in Equation (9) which may not be identical to θ_0. The discussion of the variance and bias of Θ_0 with respect to θ_0 is provided below in Theorem 3.

The boundary and asymptotic conditions/assumptions on the weight function $\mathcal{K}$ and the bandwidth b are summarized in Assumption 3.

Assumption 3 (Boundary and asymptotic conditions). The kernel function $\mathcal{K}(\cdot)$ satisfies

$$\mathcal{K}\geqslant 0,\quad \int|\mathcal{K}(u)|du<\infty,\quad \lim_{|u|\to\infty}u\mathcal{K}(u)\to 0.$$

And, in addition, when $K\to\infty$, b satisfies $b\to 0$, $b^dK\to\infty$.

Theorem 1. Under the conditions in Assumptions 1–3, we have

$$\hat\theta_0^{(c)}\longrightarrow\Theta_0(\boldsymbol{z}_0;\ell_2)\quad \text{in probability.}$$

The optimal choice of the bandwidth is $\hat b\asymp K^{-1/(d+4)}$ such that the optimal MSE is $\mathbb{E}[(\hat\theta_0^{(c)}-\Theta_0)^2]\asymp K^{-4/(d+4)}$.

Theorem 1 follows immediately from consistency theorem on a standard multivariate kernel smoothing estimator (Wasserman 2010). When the number of individuals K goes to infinity, the bias of $\hat{\theta}_0^{(c)}$ with bandwidth b is of order b^2 and the variance is of order $(b^d K)^{-1}$, where d is the dimension of $\boldsymbol{z}$ as defined in Assumption 1. In such case, the asymptotic optimal choice of bandwidth that minimizes the mean squared error, $b^4 + (b^d K)^{-1}$, is of order $K^{-1/(d+4)}$, same as a d-dimensional kernel smoothing problem.

Another way of combining individuals is aggregating the objective functions as shown in Equation (2). A combined estimator with respect to kernel $\mathcal{K}(\cdot)$ is defined by

$$\tilde{\theta}_0^{(c)} = \arg\min_\theta \sum_{k=0}^{K} \mathcal{K}\left(\frac{\|\boldsymbol{z}_k - \boldsymbol{z}_0\|}{b}\right) M_k(\theta).$$

The estimator is consistent and has a similar asymptotic performance to a d-dimensional kernel smoothing estimator as stated in Theorem 2. This approach is useful especially when $\hat{\theta}_k$ is not available, such as in the cases that the number of observations for each individual is less than the number of parameters.

Theorem 2. Suppose the conditions in Assumption 3 hold and in addition,

1. $M_k(\theta)$ is convex and second order partial differentiable with respect to θ,
2. for any given θ, $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{z}}[\frac{\partial M_{\boldsymbol{x}}(\theta)}{\partial \theta}]$ as a function of $\boldsymbol{z}$ is continuous,
3. $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{z}_0}[M_{\boldsymbol{x}}(\theta)]$ has a unique minimum at $\theta = \Theta_0(\boldsymbol{z}_0; \ell_2)$.

Then

$$\tilde{\theta}_0^{(c)} \longrightarrow \Theta_0 \quad \text{in probability.}$$

The optimal choice of bandwidth b is $\hat{b} \asymp K^{-1/(d+4)}$ and the optimized mean squared error is $\mathbb{E}[(\tilde{\theta}_0^{(c)} - \Theta_0)^2] \asymp K^{-4/(d+4)}$.

The proof is given in the supplemental material.

The above theorems suggest that the individualized combined estimator by aggregating either individual estimators $\hat{\theta}_k$ or objective functions $M_k(\theta)$ would result in an improvement in mean squared error and it shares a similar asymptotic performance as a d-dimensional kernel smoothing estimator.

When $\sigma_z = 0$, $\Theta_0(\boldsymbol{z}_0; \ell_2) = \mathbb{E}_\pi[\theta_0|\boldsymbol{z}_0] \equiv \theta_0$. Hence, estimating Θ_0 becomes estimating the unknown function $g(\cdot)$ evaluated at $\boldsymbol{z}_0$. When $\sigma_z > 0$, Θ_0 and θ_0 are in general different. Let B_0 and V_0 be the bias and variance of the target estimator $\Theta_0(\boldsymbol{z}_0; \ell_2)$ in estimating θ_0 such that

$$B_0(\theta_0) := \mathbb{E}_{\theta_0}[\Theta_0(\boldsymbol{z}_0; \ell_2)] - \theta_0, \quad V_0(\theta_0) = \mathrm{var}_{\theta_0}[\Theta_0(\boldsymbol{z}_0; \ell_2)]. \tag{12}$$

The above bias and variance are defined with respect to a fixed θ_0 with random $\boldsymbol{z}_0$.

Theorem 3. The asymptotic bias and variance of $\hat{\theta}_0^{(c)}$ in estimating a fixed θ_0 are given by

$$\mathbb{E}_{\theta_0}[\hat{\theta}_0^{(c)}] - \theta_0 = B_0(\theta_0) + O_p(b^2),$$
$$\mathrm{var}_{\theta_0}[\hat{\theta}_0^{(c)}] = V_0(\theta_0) + O_p\left(\frac{1}{Kb^d}\right),$$

where the intrinsic bias B_0 and the intrinsic variance V_0 are defined in (12).

The proof is given in the supplemental material. In the conditional probabilities, $\Theta_0 = \mathbb{E}_\pi[\theta_0|\boldsymbol{z}_0]$, as a function of $\boldsymbol{z}_0$, is considered random under a given θ_0.

The bias and variance of $\hat{\theta}_0^{(c)}$ in terms of estimating a fixed θ_0 can therefore be decomposed into two parts. The first part (the intrinsic part) comes from the bias and variance of estimating $\Theta_0[\boldsymbol{z}_0]$ itself to θ_0 and the second part comes from estimating Θ_0 nonparametrically. Since $\boldsymbol{z}$ is observed with error, this is similar to error in variable problem where certain intrinsic bias cannot be avoided (Carroll, Ruppert, and Stefanski 1995; Wansbeek and Meijer 2000; Bound, Brown, and Mathiowetz 2001; Fuller 2009). Such intrinsic bias and variance are asymptotically linear of σ_z^2, which is the noise level of $\boldsymbol{z}_k$, as shown in Theorem 4. Especially, when σ_z^2 is exactly zero, all intrinsic terms vanish, and it reduces to the exact case when $\Theta_0 = \theta_0$.

Theorem 4. Suppose $g(\cdot)$ is second-order differentiable and the distribution of ϵ_k has finite higher moments. Then, for a fixed θ_0, when $\sigma_z^2 \to 0$,

$$B_0 \asymp \sigma_z^2, \quad V_0 \asymp \sigma_z^2.$$

The proof is given in the supplemental material.

Research in nonparametric regression with error in variable shows a slower convergence rate to recover the function $\theta_0 = g(\boldsymbol{\eta})$ at any given $\boldsymbol{\eta}$ (Stefanski and Carroll 1990; Fan and Truong 1993). Our problem is different. We focus on providing a point estimator of $\theta_0 = g(\boldsymbol{\eta}_0)$ without knowing $\boldsymbol{\eta}_0$, but its noisy version $\boldsymbol{z}_0$. Even if we know the function $g(\cdot)$ precisely, θ_0 is not known as we do not observe η_0. When considering an individual with fixed but unobserved $(\theta_0, \boldsymbol{\eta}_0)$, it is difficult to choose an optimal bandwidth by bias-variance optimization with the nonzero intrinsic terms in Theorem 3, because in this case the asymptotic mean squared error $(B_0 + O_p(b^2))^2 + V_0 + O_p((Kb^d)^{-1})$ may not have a local minimum. However, if we assume the target individual 0 is randomly chosen from the population, the target estimator Θ_0 is the estimator that minimizes the overall risk under squared loss, that is, a Bayes estimator, because it minimizes the squared loss pointwise for any $\boldsymbol{z}_0$. Furthermore, immediately from Theorem 1, $\hat{\theta}_0^{(c)}$ is a consistent estimator for Θ_0. The overall performance of $\hat{\theta}_0^{(c)}$ for all individuals of the population could be optimized by choosing a proper bandwidth b as stated in the following Theorem 5. It provides a way to optimize the bandwidth globally.

Theorem 5. Assume Assumptions 1–3 hold, then the estimator $\hat{\theta}_0^{(c)}$ has the following Bayes risk under squared loss:

$$\mathbb{E}[(\hat{\theta}_0^{(c)} - \theta_0)^2] = R_0 + O_p(b^4) + O_p\left(\frac{1}{Kb^d}\right),$$

where

$$R_0 = \mathrm{var}[\Theta_0 - \theta_0]$$

is the risk of the Bayes estimator $\Theta_0 = E_\pi[\theta|\boldsymbol{z}_0]$, and all above expectations is taken over all random variables assuming an empirical population distribution $\pi(\cdot)$ for θ_0. The optimal

choice of the bandwidth b is $b \asymp K^{1/(d+4)}$ with the corresponding overall risk $R_0 + O_p(K^{4/(d+4)})$.

The proof is given in the supplemental material.

The magnitude of the measurement error of $\boldsymbol{z}_k$, measured by σ_z^2, compared to that of the individual estimation error is crucial for the performance of the iGroup method. The bias and variance of iGroup estimator increase when σ_z^2 increases (see Theorem 4). And the asymptotic Bayes risk R_0 also depends on σ_z^2. When iGroup is based on unreliable $\boldsymbol{z}$, it could result in a worse estimator compared to the one without any grouping. This phenomenon will be demonstrated in Section 4.

Remark: Results in Theorems 3–5 can be generalized to the iGroup estimator $\tilde{\theta}_0^{(c)}$, which combines the objective functions, except that the target estimator changes from $\mathbb{E}_\pi[\theta|z_0]$ to $\arg\min_\theta \mathbb{E}_\pi[M(\theta)|z_0]$. As shown in (A.1) in the supplemental material, $\tilde{\theta}_0^{(c)}$ is asymptotically a kernel smoothing estimator with the same bias and variance rates.

3.3. Case 2: Without Exogenous Variables

In this case, we assume the exogenous variable $\boldsymbol{z}$ is not available. Our target estimator is $\Theta_0(\boldsymbol{x}; \ell_2) = \mathbb{E}_\pi[\theta_0|\boldsymbol{x}_0]$ under squared loss and is $\Theta_0(\boldsymbol{x}_0; L) = \arg\min_\theta \mathbb{E}_\pi[L(\theta, \theta_0)|\boldsymbol{x}_0]$ under a general loss function L. The iGroup estimation depends solely on $\hat{\theta}$. The weight function (5) used in Equations (1) and (2) now reduces to

$$w_2(\hat{\theta}_k, \hat{\theta}_0) = \frac{\int p(\hat{\theta}_k|\theta)p(\hat{\theta}_0|\theta)\pi(\theta)d\theta}{\int p(\hat{\theta}_k|\theta)\pi(\theta)d\theta \int p(\hat{\theta}_0|\theta)\pi(\theta)d\theta}, \tag{13}$$

where $\pi(\theta)$ corresponds to the unknown distribution of θ in the whole population. As discussed in Section 2.3, an estimation of this weight function can be achieved by kernel density estimation on the bootstrapped samples $(\hat{\theta}_k^{(1)}, \hat{\theta}_k^{(2)})$.

The weight function (13) is used to aggregated individual unbiased estimators to the posterior mean, and to aggregate objective functions $M : \Omega_\theta \times \Omega_\theta \to \mathbb{R}$ to the corresponding Bayes estimator under certain loss function, as shown in Theorems 6 and 7.

Theorem 6. Suppose $w_2(\hat{\theta}_k, \hat{\theta}_0)$ is defined as in Equation (13) and $\hat{\theta}_k$ is a sufficient and unbiased estimator of θ_k for all k, then as $K \to \infty$:

$$\hat{\theta}_0^{(c)} \to \Theta_0(\boldsymbol{x}_0; \ell_2) \quad \text{in probability.}$$

Furthermore, if $\mathbb{E}_{\hat{\theta}_0}[w_2^2(\hat{\theta}_k, \hat{\theta}_0)] < \infty$ for any fixed $\hat{\theta}_0$ and $\mathbb{E}_\pi[\hat{\theta}^2] < \infty$, then

$$\sqrt{K}(\hat{\theta}_0^{(c)} - \Theta_0) = O_p(1).$$

The proof is given in the supplemental material.

For the aggregated estimator (2), suppose the objective function $M : \Omega_\theta \times \Omega_\theta \to \mathbb{R}$ used satisfies

$$\int M(\theta, \hat{\theta})p(\hat{\theta}|\theta')d\hat{\theta} = L(\theta, \theta') + C(\theta'), \tag{14}$$

where L is nonnegative and $L(\theta, \theta) = 0$ for all θ, and C is constant with respect to θ. Then L is the loss function corresponding to M, under which the target estimator is

$$\Theta_0(\boldsymbol{x}_0; L) = \arg\min_\theta \int L(\theta, \theta_0)p(\hat{\theta}_0|\theta_0)\pi(\theta_0)d\theta_0.$$

For example, if the objective function M is the negative log-likelihood function $M(\theta, \hat{\theta}) = -\log p(\hat{\theta}|\theta)$, then the corresponding loss function $L(\theta, \theta')$ is the Kullback–Leibler divergence of the given parameters.

Theorem 7. If for any given $\hat{\theta}$, $M(\theta, \hat{\theta})$ as a function of θ is convex and the second-order differentiable, then the combined estimator $\tilde{\theta}_0^{(c)}$ using the objective function M converges in probability to the target estimator under the loss function L as $K \to \infty$:

$$\hat{\theta}_0^{(c)} = \arg\min_\theta \sum_{k=0}^{K} w_2(\hat{\theta}_k, \hat{\theta}_0)M(\theta, \hat{\theta}_k) \xrightarrow{P} \Theta_0(\boldsymbol{x}_0; L).$$

Furthermore, if $\mathbb{E}_{\hat{\theta}_0}[w_2(\hat{\theta}_k, \hat{\theta}_0)M'_\theta(\theta_0, \hat{\theta})]^2 < \infty$ for any fixed $\hat{\theta}_0$,

$$\sqrt{K}(\tilde{\theta}_0^{(c)} - \Theta_0) = O_p(1).$$

The proof is given in the supplemental material.

The finite second moment conditions in Theorems 6 and 7 are satisfied in most cases. Both Theorems 6 and 7 assume an accurate estimation of the weight $w_2(\hat{\theta}_k, \hat{\theta}_0)$ (with an error rate smaller than $O_p(K^{-1/2})$. With the accurate weights $w_2(\hat{\theta}_k, \hat{\theta}_0)$, both iGroup estimators have faster convergence rates to the target estimator Θ_0 than the nonparametric one in Theorems 1.

When no accurate estimations for $w_2(\hat{\theta}_k, \hat{\theta}_0)$ are feasible, we proposed an approximate estimator for $w_2(\hat{\theta}_k, \hat{\theta}_0)$ in Section 2.3, using a set of bootstrap samples $(\hat{\theta}_k^{(1)}, \hat{\theta}_k^{(2)})$ for $k = 0, \ldots, K$. When $\boldsymbol{z}$ is not available, the integral $\int p(\hat{\theta}_k|\theta)p(\hat{\theta}_0|\theta)\pi(\theta)d\theta$ can be estimated by a kernel density estimator in a lower dimensional space:

$$\frac{1}{K+1}\sum_{j=0}^{K} \mathcal{K}_1\left(\frac{|\hat{\theta}_j^{(1)} - \hat{\theta}_k|}{b_1}\right)\mathcal{K}_2\left(\frac{|\hat{\theta}_j^{(2)} - \hat{\theta}_0|}{b_2}\right),$$

where $\mathcal{K}_1$ and $\mathcal{K}_2$ are two kernel functions with b_1, b_2 the corresponding bandwidths. The bootstrap estimation of the weight $w_2(\hat{\theta}_k, \hat{\theta}_0)$ has a nonparametric error rate $O_p(K^{-1/(d'+2)})$, where d' is the dimension of θ_0. This inaccuracy gives rise to the final error rate in Theorem 6 and 7 such that for $\hat{\theta}_0^{(c)}$ (or $\tilde{\theta}_0^{(c)}$) constructed based on $\hat{w}_2(\hat{\theta}_k, \hat{\theta}_0)$ with error rate $O_p(K^{-1/(d'+2)})$, $\hat{\theta}_0^{(c)} - \Theta_0(\boldsymbol{x}_0; \ell_2) = O_p(K^{-1/(d'+2)})$ and $\tilde{\theta}_0^{(c)} - \Theta_0(\boldsymbol{x}_0; L) = O_p(K^{-1/(d'+2)})$. Both are slower than $O_p(K^{-1/2})$.

The performance of the target estimator $\Theta_0(\boldsymbol{x}_0; \ell_2)$ in estimating θ_0 strongly depends on the accuracy of individual level $\hat{\theta}_k$. Define the bias and variance of the target estimator $\Theta_0(\boldsymbol{x}_0; \ell_2) = \mathbb{E}_\pi[\theta_0|\hat{\theta}_0]$ by

$$\begin{aligned} B_0(\theta_0) &= \mathbb{E}_{\theta_0}[\Theta_0(\boldsymbol{x}_0; \ell_2)] - \theta_0, \\ V_0(\theta_0) &= \mathrm{var}_{\theta_0}[\Theta_0(\boldsymbol{x}_0; \ell_2)]. \end{aligned} \tag{15}$$

Suppose $\hat{\theta}_0 = \theta_0 + \zeta_0$ with $\mathbb{E}[\zeta_0] = 0$ and $\mathbb{E}[\zeta_0^2] = \sigma_\theta^2$. Similar to Theorem 4, B_0 and V_0 are of order σ_θ^2 when $\sigma_\theta^2 \to 0$.

Theorem 8. Suppose ζ_0 has finite higher moments. Then, when $\sigma_\theta^2 \to 0$, the bias and variance of the target estimator $\Theta_0(\boldsymbol{x}_0; \ell_2)$ with respect to a fixed θ_0 are

$$B_0 \asymp \sigma_\theta^2, \quad V_0 \asymp \sigma_\theta^2,$$

where B_0 and V_0 are defined in Equation (15).

The proof is provided in the supplemental material.

When $\hat{\theta}_0$ is exact such that $\sigma_\theta = 0$, the target estimator equals to the true parameter θ_0 as the weight function $w_2(\hat{\theta}_k, \hat{\theta}_0)$ assigns zero weight for all other individuals except individual 0. Similar results hold for the target estimator $\Theta_0(\boldsymbol{x}_0; L)$.

3.4. Case 3: The Complete Case

When both $\hat{\theta}$ and $\boldsymbol{z}$ are available and reasonably accurate, we should use both information to improve the inference via grouping. Assuming $\hat{\theta}$ is sufficient for θ_0, the target estimator is $\Theta_0(\boldsymbol{x}_0, \boldsymbol{z}_0; \ell_2) = \mathbb{E}_\pi[\theta_0|\hat{\theta}_0, \boldsymbol{z}_0]$ under squared loss and $\Theta_0(\boldsymbol{x}_0, \boldsymbol{z}_0; L) = \arg\min_\theta \mathbb{E}_\pi[L(\theta, \theta_0)|\hat{\theta}_0, \boldsymbol{z}_0]$ under other loss function L. The following results are based on a combination of both information.

Theorem 9. Suppose $\hat{\theta}_k$ is a sufficient and unbiased estimator for θ_k, and $\hat{\theta}_0^{(c)}$ is a combined estimator as in Equation (1) with the weight functions (3)–(5), where $\mathcal{K}(\cdot)$ is a kernel function satisfying Assumption 3. Then under Assumptions (1) and (2)

$$\hat{\theta}_0^{(c)} \to \Theta_0(\boldsymbol{x}_0, \boldsymbol{z}_0; \ell_2) \quad \text{in probability.}$$

With the optimal bandwidth $\hat{b}$ chosen to be $\hat{b} \asymp K^{1/(d+4)}$, the optimal mean squared error is $\mathbb{E}[\hat{\theta}_0^{(c)} - \Theta_0]^2 \asymp K^{-4/(d+4)}$.

The proof is given in the supplemental material.

Let $M(\theta, \hat{\theta})$ be the corresponding objective function as defined in Equation (14). We have that the aggregated estimator (2) based on the objective function $M(\theta, \hat{\theta})$ converges to the target estimator $\Theta_0(\boldsymbol{x}_0, \boldsymbol{z}_0; L)$ as shown in the following Theorem 10.

Theorem 10. If for any given $\hat{\theta}$, $M(\theta, \hat{\theta})$ as a function of θ is convex and second-order differentiable, then under Assumptions (1) and (2), the combined estimator $\tilde{\theta}^{(c)}$ using the objective function M satisfying (14) converges to the target estimator:

$$\tilde{\theta}_0^{(c)} = \arg\min_\theta \sum_{k=1}^{K} w(\hat{\theta}_k, \boldsymbol{z}_k; \hat{\theta}_0, \boldsymbol{z}_0) M(\theta, \hat{\theta}_k) \xrightarrow{P} \Theta_0(\boldsymbol{x}_0, \boldsymbol{z}_0; L).$$

With the optimal bandwidth $\hat{b}$ chosen to be $\hat{b} \asymp K^{1/(d+4)}$, the optimal mean squared error is $\mathbb{E}[\tilde{\theta}_0^{(c)} - \Theta_0]^2 \asymp K^{-4/(d+4)}$.

The proof is given in the supplemental material.

Define the bias and variance of the target estimator $\Theta_0(\boldsymbol{x}_0, \boldsymbol{z}_0; \ell_2)$ as

$$\begin{aligned} B_0(\theta_0) &= \mathbb{E}_{\theta_0}[\Theta_0(\boldsymbol{x}_0, \boldsymbol{z}_0; \ell_2)] - \theta_0, \\ V_0(\theta_0) &= \mathrm{var}_{\theta_0}[\Theta_0(\boldsymbol{x}_0, \boldsymbol{z}_0; \ell_2)]. \end{aligned} \tag{16}$$

The asymptotic rate of B_0 and V_0 as σ_θ^2 or σ_z^2 approaches zero is shown in Theorem 11.

Theorem 11. Suppose $g(\cdot)$ is the second-order differentiable and ϵ_k and ζ_k have finite higher moments. If B_0 and V_0 are as defined in Equation (16), then

(i) for a fixed σ_z^2, when $\sigma_\theta^2 \to 0$,

$$B_0 \asymp \sigma_\theta^2, \quad V_0 \asymp \sigma_\theta^2.$$

(ii) for a fixed σ_θ^2, when $\sigma_z^2 \to 0$,

$$B_0 \asymp \sigma_z^2, \quad V_0 \asymp \sigma_z^2.$$

The proof is provided in the supplemental material. The bias and variance of the target estimator is of the order of the more accurate one between $\boldsymbol{z}_0$ and $\hat{\theta}_0$. Especially, when either is exact such that $\sigma_z^2 = 0$ or $\sigma_\theta^2 = 0$, the target estimator equals the true parameter θ_0.

3.5. Further Results on Risk Decomposition

Let $\hat{\theta}_0^{(c)}$ be an iGroup estimator as defined in Equation (1) based on information sets $\{\boldsymbol{z}\}$, $\{\hat{\theta}\}$, or $\{\hat{\theta}, \boldsymbol{z}\}$ as in Sections 3.2, 3.3, and 3.4, respectively. Let Θ_0 be the target estimator in any of the three cases: $\Theta_0(\boldsymbol{x}_0; \ell_2)$, $\Theta_0(\boldsymbol{z}_0; \ell_2)$, or $\Theta_0(\boldsymbol{x}_0, \boldsymbol{z}_0; \ell_2)$, depending on the information set used in $\hat{\theta}_0^{(c)}$. We have $\hat{\theta}_0^{(c)} \to \Theta_0$ in probability. When both $\hat{\theta}$ and $\boldsymbol{z}$ are available for all individuals, the overall risk of $\hat{\theta}_0^{(c)}$ under the prior $\pi(\theta)$ can be decomposed into three components as shown in Proposition 3 as an extension to Proposition 1.

Proposition 3. Suppose $\hat{\theta}_0^{(c)}$ is an iGroup estimator as defined in Equation (1) with the target estimator Θ_0. Then

$$R(\hat{\theta}_0^{(c)}) = R_{np}(\hat{\theta}_0^{(c)}) + R_{\text{target}}(\Theta_0),$$

where $R(\hat{\theta}_0^{(c)}) = \mathbb{E}[(\hat{\theta}_0^{(c)} - \theta_0)^2]$ is the overall risk of $\hat{\theta}_0^{(c)}$ under squared loss and prior $\pi(\theta_0)$, and

$$\begin{aligned} R_{np}(\hat{\theta}_0^{(c)}) &= \mathbb{E}[(\hat{\theta}_0^{(c)} - \Theta_0)^2], \\ R_{\text{target}}(\Theta_0) &= \mathbb{E}[(\Theta_0 - \theta_0)^2] \end{aligned}$$

are the risk components from the nonparametric estimation and the target estimator itself, respectively.

Furthermore, assuming both $\boldsymbol{x}$ and $\boldsymbol{z}$ are available, for $\Theta_0 = \Theta_0(\boldsymbol{x}_0; \ell_2)$ or $\Theta_0 = \Theta_0(\boldsymbol{z}_0; \ell_2)$, which only uses partial information, we have

$$R_{\text{target}}(\Theta_0) = R_{\text{inf}}(\Theta_0) + R_0,$$

where $R_{\text{inf}}(\Theta_0) = \mathbb{E}[(\Theta_0 - \Theta_0(\boldsymbol{x}_0, \boldsymbol{z}_0; \ell_2))^2]$ is the risk premium resulting from using partial information, and $R_0 = \mathbb{E}[(\Theta_0(\boldsymbol{x}_0, \boldsymbol{z}_0; \ell_2) - \theta_0)^2]$ is the overall risk of $\Theta_0(\boldsymbol{x}_0, \boldsymbol{z}_0; \ell_2)$.

The proof is provided in the supplemental material.

The decomposition in Proposition 3 reveals a guideline to optimize the iGroup estimator. The overall risk of iGroup estimator $\hat{\theta}_0^{(c)}$ can be decomposed into two parts: one from the nonparametric estimation of the target estimator and the other from the risk of the target estimator itself. The risk component R_{np} involves the bandwidth b in the nonparametric estimation.

The corresponding optimal bandwidth is chosen as in a high-dimensional kernel smoothing problem (see Theorems 1, 5, and 9), since the bandwidth does not appear in the other risk terms.

The risk component R_{target} evaluates the performance of the target estimator. Different choices in constructing iGroup weight correspond to different Θ_0's. Such difference is revealed by decomposing R_{target} into two parts: R_{inf} is the risk term arising from using partial information and R_0 is the risk of the target estimator $\Theta_0(\boldsymbol{x}_0, \boldsymbol{z}_0; \ell_2)$, which incorporates the full information set. Since R_{inf} obtains its minimum at $\Theta_0 = \Theta_0(\boldsymbol{x}_0, \boldsymbol{z}_0; \ell_2)$, it is always (asymptotically) optimal to use the full information set $\{\hat{\theta}, \boldsymbol{z}\}$ in grouping, if both are available as in the complete case. On the other hand, if $\hat{\theta}$ (or $\boldsymbol{z}$) is extremely noisy such that $\Theta_0 = \mathbb{E}_\pi[\theta_0|\boldsymbol{z}_0] \approx \mathbb{E}_\pi[\theta_0|\hat{\theta}_0, \boldsymbol{z}_0]$ (or $\Theta_0 = \mathbb{E}_\pi[\theta_0|\hat{\theta}_0] \approx \mathbb{E}_\pi[\theta_0|\hat{\theta}_0, \boldsymbol{z}_0]$, respectively), it is more practical to use $\boldsymbol{z}$ only (or $\hat{\theta}$ only, respectively) for grouping, since it will have similar performance but less computational cost, and finite sample variation.

The last risk component R_0 is the minimum overall risk one can achieve. In our approach, such a minimum risk can be asymptotically reached when both $\hat{\theta}$ and $\boldsymbol{z}$ are included in grouping and the number of individuals K approaches infinity. When $\hat{\theta}$ or $\boldsymbol{z}$ is exact, $\Theta_0(\boldsymbol{x}_0, \boldsymbol{z}_0; \ell_2) = \mathbb{E}_\pi[\theta_0|\hat{\theta}_0, \boldsymbol{z}_0] = \theta_0$ and R_0 is 0. In this case, all iGroup estimators in Equation (1) converges to θ_0. The three risk components of different iGroup models are compared in Table 1. Note that the rate of R_{np} for Case 2 assumes an accurate evaluation of the weight function $w_2(\hat{\theta}_k, \theta_0)$.

Similar to Proposition 3, the risk decomposition for the iGroup estimator $\tilde{\theta}_0^{(c)}$ in Equation (2) is provided in Proposition 4 as an extension to Proposition 2.

Proposition 4. Suppose the loss function L is as defined in Equation (14). The iGroup estimator $\tilde{\theta}_0^{(c)}$ is defined in Equation (2) with the target estimator Θ_0. If $L(\hat{\theta}, \theta)$ is the second-order partially differentiable with respect to $\hat{\theta}$ such that $L'(\hat{\theta}, \theta) = \partial L/\partial\hat{\theta}$ and $L''(\hat{\theta}, \theta) = \partial^2 L/\partial\hat{\theta}^2$, then

$$\tilde{R}(\tilde{\theta}_0^{(c)}) = \tilde{R}_{np}(\tilde{\theta}_0^{(c)}) + \tilde{R}_{\text{target}}(\Theta_0) + o(\mathbb{E}[(\tilde{\theta}_0^{(c)} - \Theta_0)^2]),$$

where $\tilde{R}(\tilde{\theta}_0^{(c)}) = \mathbb{E}[L(\tilde{\theta}_0^{(c)}, \theta_0)]$ is the overall risk of $\tilde{\theta}_0^{(c)}$ under loss L and prior $\pi(\theta)$, and

$$\tilde{R}_{np}(\tilde{\theta}_0^{(c)}) = \frac{1}{2}\mathbb{E}[L''(\Theta_0, \theta_0)(\tilde{\theta}_0^{(c)} - \Theta_0)^2],$$
$$\tilde{R}_{\text{target}}(\Theta_0) = \mathbb{E}[L(\Theta_0, \theta_0)],$$

are the risk components from the nonparametric estimation of the target estimator and the target estimator itself, respectively. Furthermore, assuming both $\boldsymbol{x}$ and $\boldsymbol{z}$ are available, for any $\Theta_0 = \Theta_0(\boldsymbol{z}_0; L)$ or $\Theta_0 = \Theta_0(\boldsymbol{x}_0; L)$, which only uses partial information, we have

$$\tilde{R}_{\text{target}}(\Theta_0) = \tilde{R}_{\text{inf}}(\Theta_0) + \tilde{R}_0,$$

where $\tilde{R}_0 = \mathbb{E}[L(\Theta_0(\boldsymbol{x}_0, \boldsymbol{z}_0; L), \theta_0)]$ is the overall risk of $\Theta_0(\boldsymbol{x}_0, \boldsymbol{z}_0; L)$ and $\tilde{R}_{\text{inf}}(\Theta_0) = \mathbb{E}[L(\Theta_0, \theta_0)] - \tilde{R}_0$ is the risk premium resulting from using partial information.

The proof is given in the supplemental material.

Table 1. Comparison of the three risk components in different iGroup cases.

	iGroup Set	R_{np}	R_{target}	
			R_{inf}	R_0
Case 1	$\{\boldsymbol{z}\}$	$\asymp K^{-4/(d+4)}$	> 0	
Case 2	$\{\hat{\theta}\}$	$\asymp K^{-1}$	> 0	same value
Case 3	$\{\hat{\theta}, \boldsymbol{z}\}$	$\asymp K^{-4/(d+4)}$	$= 0$	

3.6. Bandwidth Selection and Other Practical Guide

For real applications, the bandwidth b in the weight function (4) remains to be tuned. Ideally one would perform bandwidth selection to the target individual θ_0. However, cross-validation cannot be implemented to determine b with only one estimator $\hat{\theta}_0^{(c)}$ for a single individual. Instead, we consider a set Ω_0 around target individual 0 such that the bandwidth b is tuned to minimize the averaged risk over Ω_0.

When Ω_0 is chosen as the full set $\{1, 2, \ldots, K\}$, it is the global bandwidth selection scheme that usually used in kernel smoothing and machine learning. However, the bandwidth selected by such global optimization is not optimal for the particular target individual 0. A cross-validation set Ω_0 localized to individual 0 is more appreciated to tune this individualized local bandwidth. When tuning the bandwidth in w_1 over $\boldsymbol{z}_k$'s, such a set Ω_0 can be constructed based on $\boldsymbol{z}_0$ such as $\Omega_0(\boldsymbol{z}_0, \epsilon) = \{k \in \{1, \ldots, K\} : \|\boldsymbol{z}_0 - \boldsymbol{z}_k\| \leqslant \epsilon\}$.

Suppose $\hat{\theta}_k$'s are available and the individual estimators are aggregated to form an iGroup estimator as described in Equation (1). The goal is to choose a bandwidth b that minimizes the local risk function over Ω_0 (under squared loss) around θ_0

$$R_{\Omega_0}(b) = \mathbb{E}\left[\frac{1}{|\Omega_0|}\sum_{k\in\Omega_0}(\hat{\theta}_k^{(c)} - \theta_k)^2\right].$$

The cross-validation error we use is computed as

$$CV_{\Omega_0}(b) = \frac{1}{|\Omega_0|}\sum_{k\in\Omega_0}\left(\hat{\theta}_{(-k)}^{(c)} - \hat{\theta}_k\right)^2,$$

where $\hat{\theta}_{(-k)}^{(c)}$ is the leave-one-out estimator defined by

$$\hat{\theta}_{(-k)}^{(c)} = \frac{\sum_{l\neq k}\hat{\theta}_l w(l;k)}{\sum_{l\neq k} w(l;k)}. \tag{17}$$

It is worth to point out that although the cross-validation set Ω_0 is localized/individualized, the leave-one-out estimators (17) still use all individuals instead of limited to Ω_0.

It is seen in Proposition 5 that the leave-one-out cross-validation can estimate the local risk over Ω_0 up to a constant and hence be useful.

Proposition 5. Suppose $\hat{\theta}_k$ is an unbiased estimator for θ_k for all $k = 1, \ldots, K$ and the weight function $w(l;k)$ satisfies

$$\frac{w(k;k)}{\sum_{l\neq k} w(l;k)} = O\left(\frac{1}{K}\right). \tag{18}$$

Then

$$\mathbb{E}[CV_{\Omega_0}(b)] = R_{\Omega_0}(b) + C_{\Omega_0} + O\left(\frac{1}{K}\right),$$

where C_{Ω_0} is related to Ω_0 but is a constant with respect to b.

The proof is given in the supplemental material.

Remark I: A sufficient condition for the weight function to satisfy (18) is that the function is bounded. With bounded weights, we have

$$\frac{w(k;k)}{\sum_{l\neq k} w(l;k)} \to \frac{w(k;k)}{K\mathbb{E}w(\cdot;k)} = O\left(\frac{1}{K}\right).$$

Common kernels such as the boxed, Gaussian and Epanechnikov kernels satisfy this condition. Our choice of weight function (5) with a bounded kernel $\mathcal{K}$ satisfies the condition as well.

Remark II: Similar results hold for aggregating objective functions (2) as long as the objective function is convex and second-order differentiable, and a Taylor series expansion is available.

Beside the theoretical discussions on iGroup's asymptotic performance, there are many other factors that may affect the accuracy in real applications with finite number of individuals. First of all, the weight component $w_2(\cdot)$ is estimated from bootstrapped samples. It lowers the convergence rate since bootstrapped samples from finite population are usually correlated. Second, computing the full weight function requires a kernel density estimation in a high dimensional space. When K is finite, aggregating individuals with weights evaluated directly from a high dimensional space suffers from the lack of sample size. It often requires some feature selection procedures to reduce the dimension.

Therefore, when the weight estimation is not accurate and when the sample size is limited, the complete case may not be the best choice. In real application, we suggest using (local) cross-validation to tune the bandwidth and to choose the most appropriate weight formulation.

4. Simulations

4.1. iGroup With Noisy Exogenous Variables (Case 1 in Section 3.2)

In this example, the performance of using an exogenous variable z in iGroup is studied. Suppose, for each individual, the true parameter θ is a quadratic function of η:

$$\theta_k = g(\eta_k) = (\eta_k + 1)^2.$$

The relationship is set to a quadratic form because a continuous function of z can be approximated by a quadratic function within a small enough neighborhood of z_0. A population of size $K = 1000$ is generated with their η_k's following a Gaussian distribution $N(0.2, 1)$. For each individual k, let $\hat{\theta}_k$ be a sufficient unbiased estimator of θ_k using $\boldsymbol{x}_k$ such that $\hat{\theta}_k$ is directly generated with error $\epsilon \sim N(0, \tau^2 = 1)$ and there is no need to generate $\boldsymbol{x}_k$ explicitly. z_k is a noisy observation of η_k such that $z_k \sim N(\eta_k, \sigma^2)$.

More specifically, the dataset is generated by the following hierarchical structure:

$$\eta_k \sim N(0.2, 1), \quad \theta_k = (\eta_k + 1)^2,$$
$$\hat{\theta}_k \sim N(\theta_k, 1), \quad z_k \sim N(\eta_k, \sigma^2),$$

for $k = 1, \ldots, K$. The estimator in Equation (11) is used by setting $\mathcal{K}(\cdot)$ to the Gaussian kernel.

The parameter σ^2 controls the noise level in the observed z_k. Both individualized performance at $\theta_0 = 1$ and the overall performance over the population are studied at six choices of noise levels $\sigma = 0, 0.2, 0.4, 0.6, 0.8,$ and 1.0 with 1000 replications each.

The in-sample performance of the iGroup estimators are demonstrated in Figure 3. The first row shows the bias, variance and mean squared error for the individual at $\theta_0 = 1$, while the second row plots the overall performance by averaging individual performance over the population. Every curve represents a performance measure (bias, variance or MSE) as a function of the bandwidth b used in weight calculation in Equation (4) and six different curves distinguish different noise levels σ^2.

From Figure 3, it is seen that an increase in the noise level in $\boldsymbol{z}_k$ increases both the bias and variance of the iGroup estimator. When $\sigma > 0$, an intrinsic bias is observed for individual 0 when the bandwidth shrinks to zero, while at the population level, the average bias vanishes when the bandwidth shrinks to zero as the iGroup estimator converges to the target estimator $\Theta_0(\boldsymbol{z}_0; \ell_2) = \mathbb{E}_\pi[\theta_0|\boldsymbol{z}_0]$, whose expectation is $\mathbb{E}_\pi[\theta_0]$. Recall that the individual estimate $\hat{\theta}_k$ without grouping has a risk $\tau^2 = 1.0$ by the simulation design. It is marked on the right panels by the horizontal line. When the noise level σ exceeds 0.4, both the individual- and population-level risk are worse than using $\hat{\theta}_k$ directly without grouping. Smaller noise in z_k would significantly reduce the risk of the iGroup estimator.

In real applications, the performance plots such as Figure 3 are not available without knowing the true parameter. As suggested in Section 3.6, an optimal bandwidth can be selected by leave-one-out cross-validation. We simply use the global set $\Omega_0 = \{1, \ldots, K\}$ to tune the bandwidth. Figure 4 compares the mean square errors of three different estimators under different noise-level settings for σ^2. The individual-level estimator uses $\hat{\theta}_k$, which achieves a constant MSE at $\tau^2 = 1$. The population-level estimator uses the averaged estimator $(\sum_{k=1}^K \hat{\theta}_k)/K$, assuming population homogeneity. The iGroup estimator uses the estimator (11) and selects the optimal bandwidth by leave-one-out cross-validation over a grid of bandwidths. The population-level estimator is always the worst because the homogeneity population assumption is invalid in this simulation. The overall MSE of the iGroup estimator is a monotone increasing function of the noise level σ, because the intrinsic bias and variance increase with σ. The iGroup estimator outperforms the individual estimator when σ is below the threshold $\sigma = 0.35$. It also suggests that the iGroup method works better when more accurate exogenous variable z is used.

4.2. Short Time Series (Case 2 in Section 3.3)

In this simulation study, the individualized grouping learning method is applied to a set of short time series without any

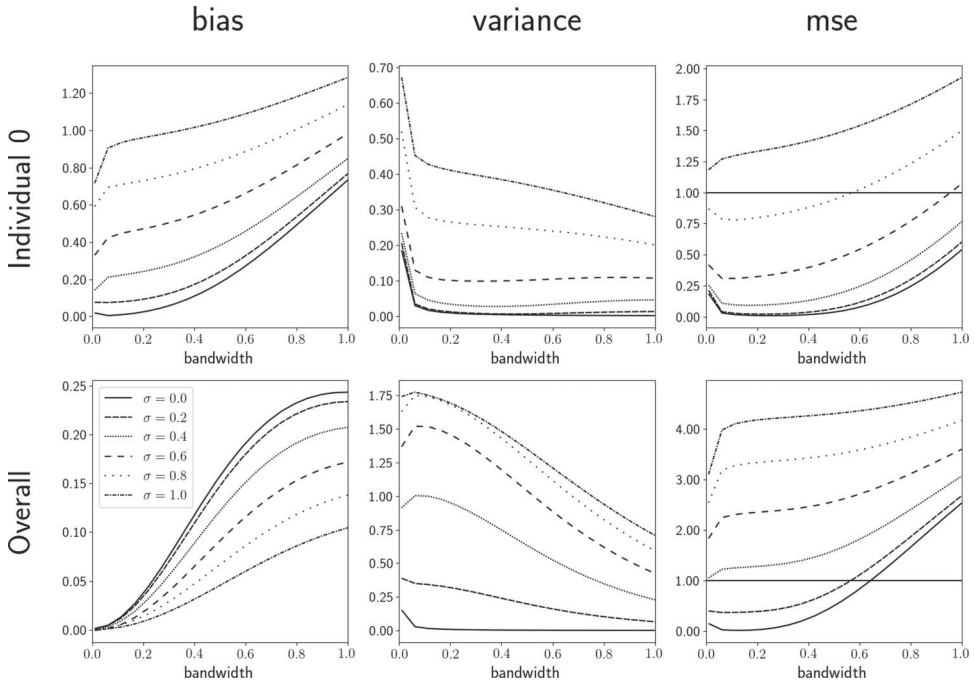

Figure 3. Bias, variance, and mean squared error as a function of bandwidth under different noise levels for individual 0 (top) and the population (bottom)

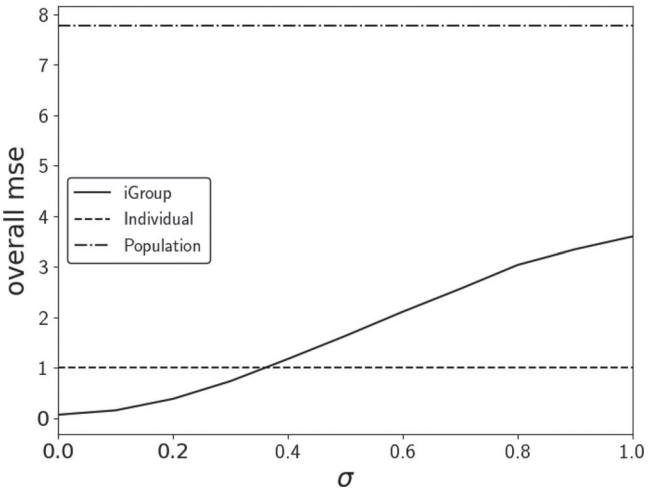

Figure 4. Overall MSE of three estimators: individual level, iGroup with cross-validation, and population level.

exogenous information. It is a simulation study for Case 2 in Section 3.3. Suppose we have $K = 200$ time series following an AR(1) model. Their AR coefficients $\theta_1, \ldots, \theta_{200}$ are drawn randomly from a beta-shaped distribution on $[-1, 1]$ such that

$$\frac{\theta_k + 1}{2} \sim \text{Beta}(4, 4), \quad k = 1, \ldots, 200. \tag{19}$$

The length of each time series is 10. They are generated from their stationary distributions

$$x_{k,0} \sim N\left(0, \frac{\sigma^2}{1 - \theta_k^2}\right),$$
$$x_{k,t} = \theta_k x_{k,t-1} + \epsilon_{k,t}, \quad k = 1, \ldots, 200, \ t = 1, \ldots, 10,$$

where $\epsilon_{k,t} \sim N(0, \sigma^2)$ and $\sigma = 3$.

Four estimators are used and their mean squared errors averaged over the 200 individual time series are compared. The individual level estimator is based on each time series of 10 observations and does not borrow any information from the others. It is an unbiased estimator for each individual. The iGroup1 estimator aggregates the log-likelihood functions according to Equation (2), where the weight function used in Equation (13), which is estimated by bootstrap samples. The bootstrap estimates are obtained based on multinomial samples of (x_{t-1}, x_t) pairs for each individual. The bandwidth used in estimating $w_2(\hat{\theta}_k, \hat{\theta}_0)$ in Equation (13) is chosen by cross-validation as in a kernel density estimation problem. The iGroup2 estimator aggregates individual level estimators by the weight function in Equation (13), the same weight function as in the iGroup1 estimator. These three methods do not use the true prior distribution. The fourth estimator, the oracle one, uses the posterior mean as the estimator with the true population prior (19) as the prior. The oracle estimator, which is the best point estimator for θ_0 given the prior information $\pi(\cdot)$, is the target estimator $\Theta_0(\boldsymbol{x}_0; \ell_2)$ for iGroup methods.

The simulation (including generating the data) is repeated 100 times. The boxplots of the mean squared errors of the four estimators are reported in the left panel of Figure 5. On average, the iGroup1 and iGroup2 estimators achieve smaller mean squared errors and smaller variances compared with the individual one. The oracle estimator is the best among those four with the smallest average error and variation. The iGroup estimators are quite close to the oracle one. The slight worse performance is due to the approximation error when constructing the weight functions. Between the two iGroup estimators, iGroup2 is slightly better than iGroup1 because the loss function used in iGroup2 is the squared loss, whose overall risk is minimized by aggregating $\hat{\theta}_k$ (See Theorem 6).

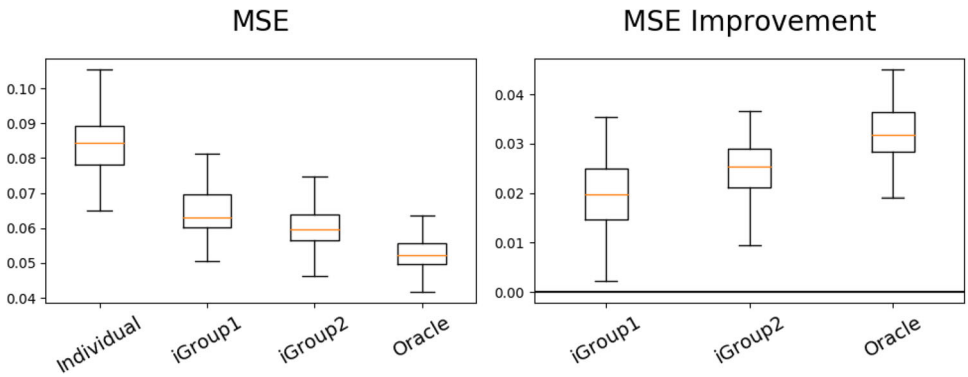

Figure 5. Comparison of the averaged MSE over 200 individuals on 100 replications for four estimators

Table 2. Mean squared error for the experiment in Section 4.3 in different configurations.

Configuration	n	$\tau^2 = \sigma_x^2/n$	σ	iGroup($\emptyset$)	iGroup($\hat{\theta}$)	iGroup(z)	iGroup($z, \hat{\theta}$)
1	5	0.20	0.10	0.200	0.163	**0.044**	0.154
2	5	0.20	0.15	0.200	0.163	**0.090**	0.163
3	5	0.20	0.20	0.200	0.163	**0.137**	0.170
4	5	0.20	0.30	0.200	**0.163**	0.200	0.179
5	10	0.10	0.10	0.100	0.089	**0.048**	0.059
6	10	0.10	0.15	0.100	0.089	0.089	**0.070**
7	10	0.10	0.20	0.100	0.089	0.099	**0.077**
8	10	0.10	0.30	0.100	0.089	0.100	**0.084**
9	20	0.05	0.10	0.050	0.046	0.044	**0.040**
10	20	0.05	0.15	0.050	0.046	0.050	**0.044**
11	20	0.05	0.20	0.050	0.046	0.050	**0.045**
12	20	0.05	0.30	0.050	**0.046**	0.050	0.047

The right panel in Figure 5 plots the improvement (difference) of the mean square errors of the iGroup estimators and the oracle estimator over the individual estimator for the 100 replications. It shows that in all experiment replications, the mean square errors of the iGroup estimators are uniformly better than the individual one. Estimation does benefit from individualized grouping in this case.

4.3. A Combined Case (Case 3 in Section 3.4)

In this simulation, we compare the performance of different iGroup estimators constructed on different information sets when both $\hat{\theta}$ and z are available as in Case 3 discussed in Section 3.4. Consider a population with $n = 1024$ individuals as follows:

$$\eta_k \sim N(0,1), \quad \theta_k = \sin(\pi\eta_k),$$
$$z_k \sim N(\eta_k, \sigma^2), \quad x_{k,1}, x_{k,2}, \ldots, x_{k,n} \sim N(\theta_k, \sigma_x^2),$$

for $k = 1, \ldots, 1024$. θ is the parameter of interest. Individual estimator used is

$$\hat{\theta}_k = \frac{1}{n}\sum_{i=1}^{n} x_{k,i} \text{ for } k = 1, \ldots, 1024.$$

Four approaches are investigated here as special cases of the iGroup method. iGroup($\emptyset$) is the individual estimation without grouping, that is, using $\hat{\theta}_k$ as the estimator. iGroup(z) uses the exogenous observation z only for grouping and an iGroup estimator is obtained by aggregating $\hat{\theta}$'s using $w_1(\boldsymbol{z}_k, \boldsymbol{z}_0)$ in Equation (4), where the bandwidth b is selected by leave-one-out cross-validation. iGroup($\hat{\theta}$) uses $\hat{\theta}_k$ only for grouping, using $w_2(\hat{\theta}, \hat{\theta}')$ in Equation (5) as the weight function. The weight is approximated by kernel density estimation on the bootstrapped samples with bandwidth selected by cross-validation. And lastly, iGroup($z, \hat{\theta}$) uses both z and $\hat{\theta}$ for calculating the weight function $w(\boldsymbol{z}_k, \hat{\theta}_k; \boldsymbol{z}_0, \hat{\theta}_0)$ in Equation (3) as discussed in Section 3.4, with the bandwidth selected by leave-one-out cross-validation.

Several different (n, σ, σ_x) configurations are studied. The mean square errors are reported in Table 2. The smallest MSE across the different methods is shown in bold face for each configuration. From Table 2, it is seen that in Configurations 6 to 11, using both $\boldsymbol{z}$ and $\hat{\theta}$ outperforms the other three methods. However, it is worth to point out that it is not always the best. When z is relatively accurate and $\hat{\theta}$ is not so as in Configurations 1, 2, 3 and 5, using $\boldsymbol{z}$ alone is better than involving $\hat{\theta}$ in the grouping. The reason is that the weight function used in the estimation is an approximation based on bootstrap sampling, which is not accurate when the sample size n is too small (as discussed in Section 3.6). It is also intuitive since using inaccurate $\hat{\theta}_k$ for grouping may reduce the grouping quality. When $\boldsymbol{z}$ is quite noisy as in Scenarios 4 and 12, using $\hat{\theta}$ only is better than using the complete information set. Note that when the bandwidth in $w_1(\boldsymbol{z}_k, \boldsymbol{z}_0)$ shrinks to zero, iGroup(z) reduces to the individual estimator and the complete estimator iGroup($z, \hat{\theta}$) reduces to iGroup($\hat{\theta}$). However, due to the randomness from finite sample size and possible overfitting, iGroup($\hat{\theta}$) or iGroup(z) sometimes performs better.

In conclusion, we suggest the following brief guideline in choosing iGroup models. When $\hat{\theta}$ is relatively inaccurate and the bootstrap method has unignorable error, it is better not to use $\hat{\theta}$ in grouping. When $\boldsymbol{z}$ is relatively inaccurate, it is better to

either use $\hat{\theta}$ only or use the full model. But when using the full model, the bandwidth needs to be tuned carefully around zero. When both $\hat{\theta}$ and $\boldsymbol{z}$ are considerably accurate, it is beneficial to consider both in grouping.

5. VaR Analysis Based on Fama–French factors

In this example we use iGroup to improve the estimation of VaR in stock returns. Denote the return of stock k in day t as $r_{t,k}$. The one-day VaR of $r_{t,k}$, denoted as $\text{VaR}(t,k)$, is defined as the smallest quantity v such that the probability of the event $r_{t+1,k} \leqslant -v$ is no greater than a predetermined confidence level α (for example, 1%). Statistically, $-v$ is the α quantile of $r_{t+1,k}$. VaR is widely used in quantitative finance and risk management to estimate the possible losses in worse cases (e.g., 1% lower quantile) due to adverse market moves. Typically $\text{VaR}(t,k)$ is estimated (predicted) based on all observations from $t-S+1$ to t (the look-back period). When a parametric (time series) model is used, $\text{VaR}(t,k)$ is often estimated as the α-quantile of the one-step ahead predictive distribution. It requires a strong model assumption. Nonparametrically, $\text{VaR}(t,k)$ is often estimated as the α-quantile of the marginal distribution, assuming the distribution does not change within the look-back period. The nonparametric approach is usually difficult to use because it requires a large sample size to estimate small quantiles accurately, but the market conditions change over time hence one cannot use a long look-back period.

We compare the three iGroup approaches with three baseline approaches.

5.1. iGroup Estimation

The three Fama–French factors (MKT, SMB, and HML) are widely used to describe the behavior of stock returns (Fama and French 1993), extending the celebrated capital asset pricing model of Markowitz (1952). Here, MKT is the excess market return, the average return of all stocks in the market; SMB is the size factor, measured by the difference of returns of the portfolios consisting of the smallest 30% and the largest 30% of stocks in the market, respectively, in terms of their market value; and HML is the book-to-market ratio factor or value risk factor, measured by the difference of returns of the portfolios consisting of the highest 30% and the lowest 30% of stocks, respectively, in terms of the ratio of their book value to market value. Book value is the total accounting value of all assets of the company, while the market value is the total outstanding share of stock times its current market price.

The Fama–French three-factor model (Fama and French 1993) assumes

$$\begin{aligned} r_{t,k} &= \alpha_{t,k} + r_f + b_{0,t,k}(\text{MKT}_t - r_f) + b_{1,t,k}SMB_t \\ &\quad + b_{2,t,k}HML_t + \epsilon_{t,k}, \\ \epsilon_{t,k} &\sim \mathcal{N}(0, \sigma_k^2), \end{aligned}$$

where r_f is the risk-free interest rate. The Fama–French model links the return of individual stock to the three factors. The corresponding coefficients reflect certain characteristics of the behavior of the stock. For example, if $b_{0,t,k}$ is large, then the return of stock k tends to amplify the market return (both positive and negative returns). If $b_{1,t,k}$ is positive, then the stock k tend to have positive return when the current condition favors small stocks (when SMB is positive, or small stocks outperforms large stocks). Although typically large stocks have positive $b_{1,t,k}$, it is not always true. Hence the $b_{1,t,k}$ provides a more direct link between the size factor and the return of the stock than the size of the stock itself. Similarly, $b_{2,t,k}$ provides a link between the book-to-market ratio and the return of the stock. If $b_{2,t,k}$ is positive, then the stock tends to do worst when the market favors growth stock (when HML is negative, or growth stocks with small book value but large market value outperforms the value stocks). The daily Fama-French factor data can be found at Professor French's publicly available data library.

Here we assume the Fama-French coefficients $b_0, b_1,$ and b_2 vary over time slowly and can be estimated using data in the look-back window $(t-S+1, t)$. Corresponding to iGroup framework demonstrated in Figure 1, we view the true Fama–French coefficients as $\boldsymbol{\eta}$ and view the estimated ones (with error) as the exogenous variable $\boldsymbol{z}$ since we assume that the one-day VaR is related to the Fama-French coefficients but the exact relationship is yet known. A misspecification of the relationship between Fama–French coefficients and the one-day VaR may lead to unsatisfactory results as we will show later with a typical quantile regression model. We use the empirical α-quantile of the returns in the look-back window $(t-S+1, t)$ as $\hat{\theta}$ such that

$$\hat{\theta}_{t,k} = Q_\alpha\left(\bigcup_{s=0}^{S-1}\{r_{t-s,k}\}\right), \tag{20}$$

where $Q_\alpha(\cdot)$ is the empirical α quantile given a set of observations.

We consider three iGroup estimators: iGroup(z), iGroup($\hat{\theta}$), and iGroup($z, \hat{\theta}$), as specified and discussed in Section 3. All of them take the general formulation of

$$\begin{aligned} \widehat{\text{VaR}}(t,k) &= Q_\alpha^{(w)}\left(\bigcup_{l=1}^{K}\bigcup_{s=0}^{S-1}\{(r_{t-s,l}, w(l;k)\}\right) \\ &= \arg\min_\theta \sum_{l=1}^{K} M_k(\theta;t)w(l;k), \end{aligned} \tag{21}$$

where $Q_\alpha^{(w)}(\cdot)$ is the empirical α quantile estimator from a weighted sample and

$$M_k(\theta;t) = \sum_{s=0}^{S-1} |r_{t-s,k} - \theta| \left(\alpha \mathbf{1}_{\{r_{t-s,k}>\theta\}} + (1-\alpha)\mathbf{1}_{\{r_{t-s,k}\leqslant\theta\}}\right)$$

is the quantile estimation equation. Depending on the information set used, different weight functions $w(l;k)$ are used, either based on solely on z, solely on $\hat{\theta}$ or on both as discussed in Section 3.1. We need two weight functions w_1 and w_2.

Using the Fama–French coefficients as z, the weight function $w_1(\cdot)$ here is chosen to be a Gaussian kernel

$$w_1(\boldsymbol{z}_{t,l}; \boldsymbol{z}_{t,k}) \propto \exp\left(-\frac{\|\boldsymbol{z}_{t,l} - \boldsymbol{z}_{t,k}\|_2^2}{2b^2}\right),$$

where the features $\boldsymbol{z}_{t,k} = (b_{0,t,k}, b_{1,t,k}, b_{2,t,k})$ are the estimated Fama–French coefficients of stock k using the returns in the S days before day t. The bandwidth b is the parameter to be tuned.

For the weight function $w_2(\cdot)$, the stochastic distance between $\hat{\theta}_{t,k}$'s are computed using a slightly modified version of the bootstrap method proposed in Section 2.3, since the bootstrap samples of the extreme values are not stable. Instead of using the standard bootstrap samples, we obtain two values of the quantiles $\hat{\theta}_{t,k}$ and $\hat{\theta}'_{t,k}$ from the same stock, but with two different time frames. Specifically, for stock k at time t, in addition to one quantile value as in Equation (20) using data $\{r_{t-s,k}, s = 0, \ldots, S-1\}$, another quantile value $\hat{\theta}'$ is obtained using data $\{r_{t-s,k}, s = 0.5S, \ldots, 1.5S\}$. The time frame in calculating $\hat{\theta}'$ has $0.5S$ days overlap with $\hat{\theta}$ for stability purpose. With two samples $\hat{\theta}_{t,k}$ and $\hat{\theta}'_{t,k}$, the weight $w_2(\hat{\theta}_l, \hat{\theta}_k)$ can be obtained with Equation (5) in association with Equation (6).

The three iGroup estimators: iGroup(z), iGroup($\hat{\theta}$), and iGroup($z, \hat{\theta}$) are then constructed using Equation (21) with $w(k,l)$ being $w_1(z_k, z_l)$, $w_2(\hat{\theta}_k, \hat{\theta}_l)$, and $w_1(z_k, z_l)w_2(\hat{\theta}_k, \hat{\theta}_l)$, respectively.

5.2. Baseline Methods

We compare the iGroup methods with the following three based methods. The quantile regression method is built upon the same Fama–French model and it is used compare with iGroup($\hat{\theta}$) and iGroup($z, \hat{\theta}$). The other two methods, individual VaR estimation and Market level VaR approach, are compared with iGroup(z).

Quantile Regression: For comparison, we use a quantile regression version of the Fama–French model to obtain the VaR prediction. Assume on each day t and stock k, the α-quantile of the excess return $r_{t,k} - r_f$ follows

$$Q_\alpha(k,t) = \alpha_{t,k} + b_{0,t,k}(\mathrm{MKT}_{t-1} - r_f) + b_{1,t,k}\mathrm{SMB}_{t-1} + b_{2,t,k}\mathrm{HML}_{t-1}.$$

In the above we use the Fama–French factors at $t-1$ in order to perform one-day prediction of the quantile. By assuming the quantile regression model changes slowly over the past S days, the quantile regression can be estimated (Koenker and Bassett 1978; Koenker 2005). The estimated one-day VaR (for day $t+1$) at day t is given by

$$\widehat{\mathrm{VaR}}(t,k) = -x_t^T \hat{\beta}_{t,k},$$

where $x_t = (1, \mathrm{MKT}_t - r_f, \mathrm{SMB}_t, \mathrm{HML}_t)$ contains the constant 1 and the three Fama–French factors at day t.

Individual VaR estimation using empirical quantiles: A naive method to estimate VaR is to use the empirical quantile of $r_{t,k}, \ldots, r_{t-S+1,k}$. When α is set to be 1% and $S = 100$, we have $\widehat{\mathrm{VaR}}(t,k) = \min\{r_{t,k}, r_{t-1,k}, ..., r_{t-99,k}\}$. Such a quantile estimation is not very accurate. On one hand, when S is small and there is not enough observations, the empirical quantile is not defined. On the other hand, S cannot be very large as the market changes over time and so does the distribution of returns. The individual estimator is an extreme version of *i*group when the bandwidth b shrinks to 0.

Market-level VaR: Another approach assumes homogeneity among all stocks. The VaR could then be estimated by pooling historical returns of all stocks. In this case, the estimator is

$$\widehat{\mathrm{VaR}}(t,k) = Q_\alpha\left(\bigcup_{l=1}^{K}\bigcup_{s=0}^{S-1}\{r_{t-s,l}\}\right),$$

Table 3. Prediction errors for the individual estimation, market level estimation, quantile regression estimation and iGroup estimations. The right three methods use Fama-French factors, while the left three do not.

Method	Individual	Market	iGroup($\hat{\theta}$)	Quantile Reg	iGroup(z)	iGroup($z, \hat{\theta}$)
RMSE ($\times 10^{-3}$)	9.61	13.4	6.63	29.8	5.75	5.54

where $Q_\alpha(A)$ is the empirical α quantile estimator given a set of observations A. Pooling observations from other stocks bring a significant bias if the homogeneity assumption is not valid.It can be viewed as an extreme case of iGroup estimation when the bandwidth b approaches ∞.

5.3. Performance Comparison

In this study, we use $\alpha = 0.01$, $S = 100$, and $K = 491$ stocks in the S&P 500 index with new additions and drop-offs during the year removed. The prediction error is measured over 250 trading days in the year 2016 for 491 stocks using

$$\mathrm{RMSE} = \left[\frac{1}{491}\sum_{k=1}^{491}\left(\frac{1}{250}\sum_{t=1}^{250}\mathbf{1}_{\{r_{t+1,k} \leqslant \widehat{\mathrm{VaR}}(t,k)\}} - 0.01\right)^2\right]^{1/2},$$

where $\widehat{\mathrm{VaR}}(t,k)$ is based on returns $\{r_{t,k}, \ldots, r_{t-99,k}, k = 1, \ldots, 491\}$.

The RMSEs of six aforementioned models are shown in Table 3. The bandwidth used in the weight function $w_1(\cdot)$ is tuned to achieve minimum RMSE. Specifically, iGroup(z) uses $b = 0.05$ and iGroup($z, \hat{\theta}$) uses $b = 0.08$. In the kernel density estimation involved in approximating the distance between $\hat{\theta}_{t,k}$'s to obtain the weight function $w_2(\cdot)$, both iGroup($\hat{\theta}$) and iGroup($z, \hat{\theta}$) choose the bandwidth according to Scott's rule of thumb (Scott 2015).

Among the three methods that do not use Fama-French factors (Individual estimation, market estimation and iGroup($\hat{\theta}$)), iGroup($\hat{\theta}$) achieves the minimum RMSE. Among the methods that use Fama–French factors, both iGroup(z) and iGroup($z, \hat{\theta}$) have a substantial decrease in RMSE compared to the quantile regression method. It shows that using the Fama–French factors directly and linearly to estimate the quantile is not sufficient. The iGroup estimator uses these factors indirectly and nonparametrically. Both z and $\hat{\theta}$ helps in identifying the cliques of the stocks as iGroup($z, \hat{\theta}$) outperforms both iGroup(z) and iGroup($\hat{\theta}$).

6. Conclusion and Discussion

In conclusion, the proposed iGroup method provides an effective tool for efficient inference in a heterogeneous population. The approach is essentially nonparametric. It has several special features: (i) The grouping idea can facilitate and answer some inference questions that are otherwise difficult or impossible to address such as estimating variance/quantile when each individual has only one observation. (ii) It reduces the standard error of the estimator by pooling together individuals with similar characteristics. (iii) The grouping can take a nonstandard exogenous variable $\boldsymbol{z}$ into consideration, as long as

a similarity/distance measure is defined. (iv) Noisy exogenous variable $\boldsymbol{z}$ can contribute to grouping as well. (v) A useful weight function measuring similarity between $\hat{\theta}$'s is designed with statistical interpretation. (vi) The method can be extended to a wide range of estimating methods, which optimizes an objective function, such as regularized least-square estimation, generalized moment estimation, etc. (vii) The bandwidth can be tuned by leave-one-out cross-validation either globally or locally.

In addition, we showed the asymptotic performance and theoretical properties of the method, which assess the accuracy and efficiency of iGroup and provide practical guidance in implementation. More specifically, when the loss function is given and weight function is properly constructed by our approach, the iGroup estimator converges to the Bayes estimator that minimizes the overall risk without knowing the prior. Computationally, as the group construction and inference procedure are identical for all individuals, the iGroup method can be easily parallelized for large dataset.

In Theorems 2, 7, and 10, we assumed a quite strong sufficient condition on the objective functions $M_k(\theta)$ or $M(\theta, \hat{\theta})$ such that the minimum point of the aggregated objective function will converge to the true value. Instead of assuming the second-order differentiability and convexity, other sufficient conditions can also guarantee the convergence of the minimum point (Van der Vaart 2000). But most of them depends on the explicit formula of kernel $\mathcal{K}$ and the objective function $M_k(\theta)$.

As an individualized inference method, the iGroup approach is closely related to kernel methods, since they follow the same principle of borrowing the information from other individuals with similar features. In this article, we focus on a more general class of individualized inference problems, though our approach remains under the principle of finding "similar" individuals. However, the setting, the goal and theoretic support needed in our development are quite different and they are not a straight forward extension of the standard kernel smoothing method. First, our goal is to obtain individualized inference of a set of parameters based on two sets of distinct information: the observations $\boldsymbol{x}_k$ directly linked to the parameters and the other features $\boldsymbol{z}_k$ of the individual that are indirectly related to the parameters, a problem that is significantly different from nonparametric regression and has a much wider range of applications. When using $\boldsymbol{z}_k$ only, our estimator is similar to kernel smoothing, though its theoretical properties are different due to the noise in $\boldsymbol{z}_k$. More importantly, using $\hat{\theta}_k = \theta(\boldsymbol{x}_k)$ requires a completely different new weight function, even though the estimator is in a weighted average form. To our best knowledge, the idea has not been explored in the literature. Second, our approach is more flexible. We can either pool the estimators directly, as in Equation (1), or pool the estimation equations or objective functions, as in Equation (2). Third, the variables inside the kernel function may have errors ("measurement errors," sometimes large and nonignorable), and these errors could greatly impact the performance of the kernel method. We need new theory to support our development. As illustrated in Figure 2, when $\hat{\theta}_0$ is away from θ_0, a direct use of the standard kernel method results in pooling a wrong set of "neighbors" and thus biased estimation. Unlike recent articles that assume $\hat{\theta}_0 \to \theta_0$ (e.g., Shen, Liu, and Xie 2020), we face a more difficult problem allowing the bias $|\hat{\theta}_0 - \theta_0| \not\to 0$. We use the distribution of $\hat{\theta}_0|\theta_0$ (instead of the point estimator $\hat{\theta}_0$), and the techniques of an empirical Bayes method to help pool the correct information to improve inference. We also investigate the cases when exogenous variables are available and study how the information in exogenous variables can help us in the inference.

The iGroup approach has its connection to the empirical Bayes approach (Robbins 1956), where the prior is unknown, but a Bayes estimator is constructed. Although an unknown population distribution for θ is assumed to be $\pi(\theta)$ viewed as the prior, it does not appear explicitly in either $\hat{\theta}_0^{(c)}$ or $\tilde{\theta}_0^{(c)}$ in our approach. And we showed in Section 3 that under mild conditions, the iGroup estimators converge to certain Bayes estimators under the unknown prior. In the empirical Bayes, the prior is usually estimated by either discretization or deconvolution. But the iGroup approach is different. The unknown $\pi(\theta)$ is not directly estimated and it is not needed. The prior information is taken into consideration by taking a (weighted) average of sample estimators or sample objective functions. And the weight function $w_2(\cdot)$, which is related to $\pi(\theta)$ in close form, is approximated using the bootstrap method in Section 2.3.

One of the proposed weight functions $w_1(\cdot)$ in Equation (4) is kernel-based. It is well known that a direct use of a kernel method suffers "curse of dimensionality" in the presence of many predictors and especially in high dimensional situations can be problematic (see, e.g., Wasserman 2010). In such situations, we may need to consider dimension reduction techniques in such situations, including feature selection, construction of linear combinations of features as in single-index and multiple-index models, and the use of principle component analysis of the features, and many others. As a method using the kernel function, the approach is also sensitive to the choice of the metric defining the neighborhood. Unfortunately, the choices of the metric and dimensional reduction techniques also depend on specific application at hand. Data-driven selection criteria such as cross-validation measures and various regularized methods can be used for determining the optimal choices of the difference components in iGroup. The needed research is out of the scope of this article.

We note that the VaR example is fundamentally different from standard kernel smoothing and k-NN, though some pooling operations appear to be similar. First, the proposed framework inspired us to consider the construction and use of exogenous variable $\boldsymbol{z}_t$. In the VaR example, we assumed β_k being the Fama–French coefficients and $\boldsymbol{z}_k$ be an estimate of these coefficients. We want to mention that although estimation of VaR has been studied extensively in finance and risk management, we have not seen any similar pooling estimator. Second, all the exogenous variables used for pooling in the example are estimated with errors, while in standard kernel smoothing and k-NN, the grouping variables used are typically assumed to be observed without errors. Third, the VaR example is an estimating problem by writing down the objective function (or the estimating function) $M_k(\theta; t)$. The iGroup estimator is equivalent to optimize the aggregated objective function as in Equation (2). In standard kernel smoothing, only the observed

responses or the estimates, instead of objective functions, are smoothed.

Acknowledgments

The authors wish to thank the editor, associate editor, and two referees for their insightful comments and suggestions.

Funding

Chen's research is supported in part by National Science Foundation grants DMS-1737857, IIS-1741390, CCF-1934924, and DMS-2027855. Xie's research is supported in part by National Science Foundation grants DMS-1737857, DMS-1812048, DMS-2015373 and DMS-2027855.

Supplemental Material

The supplemental material contains all the proofs for our theoretical results in in Section 3.

ORCID

Rong Chen http://orcid.org/0000-0003-0793-4546

References

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (1998), "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications," *SIGMOD Rec.*, 27, 94–105. [622]

Antoniak, C. E. (1974), "Mixtures of Dirichlet Processes With Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 2, 1152–1174. [622]

Binder, D. A. (1978), "Bayesian Cluster Analysis," *Biometrika*, 65, 31–38. [622]

Bound, J., Brown, C., and Mathiowetz, N. (2001), "Measurement Error in Survey Data," in *Handbook of Econometrics*, Vol. 5. eds. Z. Griliches and M. D. Intriligator, North Holland, Amsterdam: Elsevier; pp. 3705–3843. [628]

Carroll, R., Ruppert, D., and Stefanski, L. (1995), "Nonlinear Measurement Error Models," in *Measurement Error in Nonlinear Models: Monographs on Statistics and Applied Probability*, Vol. 63. New York: Chapman & Hall/CRC. [628]

Chiu, S.-T. (1991), " Bandwidth Selection for Kernel Density Estimation," *The Annals of Statistics*, 19, 1883–1905. [626]

Collins, F. S., and Varmus, H. (2015), "A New Initiative on Precision Medicine," *New England Journal of Medicine*, 372, 793–795. [622]

Diaconis, P., and Freedman, D. (1986), "On the Consistency of Bayes Estimates," *The Annals of Statistics*, 14, 1–26. [627]

Duda, R. O., and Hart, P. E. (1973), *Pattern Classification and Scene Analysis*, New York: Wiley. [622]

Fama, E. F., and French, K. R. (1993), "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics*, 33, 3–56. [635]

Fan, J., and Truong, Y. K. (1993), "Nonparametric Regression With Errors in Variables," *The Annals of Statistics*, 21, 1900–1925. [628]

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230. [622]

Figueiredo, M. A. T., and Jain, A. K. (2000), "Unsupervised Learning of Finite Mixture Models," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24, 381–396. [622]

Fuller, W. A. (2009), *Measurement Error Models*, Vol. 305. New York: Wiley. [628]

Gan, G., Ma, C., and Wu, J. (2007), *Data Clustering: Theory, Algorithms, and Applications*, Vol. 20, Philadelphia: Society for Industrial and Applied Mathematics. [622]

Jain, A. K. (2010), "Data Clustering: 50 Years Beyond k-Means," *Pattern Recognition Letters*, 31, 651–666. [622]

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999), "Data Clustering: A Review," *ACM Computing Surveys (CSUR)*, 31, 264–323. [622]

Koenker, R. (2005), *Quantile Regression*, Cambridge: Cambridge University Press. [636]

Koenker, R., and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, 33–55. [636]

Liao, T. W. (2005), "Clustering of Time Series Data – A Survey," *Pattern Recognition*, 38, 1857–1874. [622]

Lindsay, B. G. (1995), "Mixture Models: Theory, Geometry and Applications," in *NSF-CBMS Regional Conference Series in Probability and Statistics*, 5, Hayward: Institute of Mathematical Statistics, pp. i–163. [622]

Liu, K., and Meng, X. L. (2016), "There is Individualized Treatment. Why Not Individualized Inference?" *Annual Review of Statistics and Its Application*, 3, 79–111. [622]

Lo, A. Y. (1984), "On a Class of Bayesian Nonparametric Estimates: I. Density Estimates," *The Annals of Statistics*, 12, 351–357. [622]

Markowitz, H. (1952), "Portfolio Selection," *Journal of Finance*, 7, 77–91. [635]

Ng, R. T., and Han, J. (1994), "Efficient and Effective Clustering Methods for Spatial Data Mining," in *Proceedings of the 20th International Conference on Very Large Data Bases* (VLDB '94), eds. J. B. Bocca, M. Jarke, C. Zaniolo. Santiago de Chile, Chile: Morgan Kaufmann Publishers Inc; pp. 144–155. [622]

Qian, M., and Murphy, S. A. (2011), "Performance Guarantees for Individualized Treatment Rules," *The Annals of Statistics*, 39, 1180. [622]

Robbins, H. (1956), "An Empirical Bayes Approach to Statistics," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, ed. J. Neyman. Berkeley: University of California Press; pp 157–163. [637]

Scott, D. W. (2015), *Multivariate Density Estimation: Theory, Practice, and Visualization,* Hoboken, New Jersey: Wiley. [636]

Shen, J., Liu, R. Y., and Xie, M. (2020), "iFusion: Individualized Fusion Learning," *Journal of the American Statistical Association*, 115, 1251–1267. [622,623,637]

Stefanski, L. A., and Carroll, R. J. (1990), "Deconvolving Kernel Density Estimators," *Statistics*, 21(2):169–184. [628]

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2005), "Sharing Clusters Among Related Groups: Hierarchical Dirichlet Processes," in *Advances in Neural Information Processing Systems*, Vol. 17. Cambridge, UK: MIT Press, pp. 1385–1392. [622]

Van der Vaart, A. W. (2000), *Asymptotic Statistics*, Vol. 3, Cambridge: Cambridge University Press. [637]

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., and Drazen, J. M. (2007), "Statistics in Medicine-Reporting of Subgroup Analyses in Clinical Trials," *New England Journal of Medicine*, 357, 2189–2194. [622]

Wansbeek, T. J., and Meijer, E. (2000), *Measurement Error and Latent Variables in Econometrics*, Vol. 37, North-Holland, Amsterdam: Elsevier. [628]

Wasserman, L. (2010), *All of Nonparametric Statistics*, New York: Springer Publishing Company, Incorporated. [628,637]

Xu, R., and Wunsch, D. (2005), "Survey of Clustering Algorithms," *IEEE Transactions on Neural Networks*, 16, 645–678. [622]

Yang, J., Miescke, K., and McCullagh, P. (2012), "Classification Based on a Permanental Process With Cyclic Approximation," *Biometrika*, 99, 775–786. [622]

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012), "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 107, 1106–1118. [622]