Received: 17 September 2019 | Revised: 4 December 2019 | Accepted: 28 December 2019

DOI: 10.1002/wics.1498

ADVANCED REVIEW

WIREs COMPUTATIONAL STATISTICS WILEY

# Individualized inference through fusion learning

**Chencheng Cai** | **Rong Chen** | **Min-ge Xie**

Department of Statistics, Rutgers University, New Brunswick, New Jersey

**Correspondence**
Min-ge Xie, Department of Statistics, Rutgers University, New Brunswick, NJ.
Email: mxie@stat.rutgers.edu

**Funding information**
National Science Foundation, Grant/Award Numbers: DMS-1737857, DMS-1812048, IIS-1741390

**Abstract**

Fusion learning methods, developed for the purpose of analyzing datasets from many different sources, have become a popular research topic in recent years. Individualized inference approaches through fusion learning extend fusion learning approaches to individualized inference problems over a heterogeneous population, where similar individuals are fused together to enhance the inference over the target individual. Both classical fusion learning and individualized inference approaches through fusion learning are established based on weighted aggregation of individual information, but the weight used in the latter is localized to the *target* individual. This article provides a review on two individualized inference methods through fusion learning, *i*Fusion and *i*Group, that are developed under different asymptotic settings. Both procedures guarantee optimal asymptotic theoretical performance and computational scalability.

This article is categorized under:
Statistical Learning and Exploratory Methods of the Data Sciences > Manifold Learning
Statistical Learning and Exploratory Methods of the Data Sciences > Modeling Methods
Statistical and Graphical Methods of Data Analysis > Nonparametric Methods
Data: Types and Structure > Massive Data

**KEYWORDS**

big data, fusion learning, heterogeneity, individualized inference

## 1 | INTRODUCTION

With the soaring development of data processing and data storage technologies, high volumed and complicated data become available in economics, health care, financial services, internet, and other fields. Fusion learning methods, developed for the purpose of analyzing datasets from many different sources, become a popular research topic in recent years. Specifically, fusion learning is a known and well-established statistical methodology that aggregates information learned from different studies, multiple sources, or distinct parts of a single study, to make a coherent *overall* inference (Chen & Xie, 2014; Liu, Liu, & Xie, 2014; Liu, Liu, & Xie, 2015; Yang, Liu, Wang, & Xie, 2016). However, instead of population-average, sometimes we may be interested in making inference for a specified individual, for example, in the precision medicine problems and individualized marketing strategies (Collins & Varmus, 2015; Liu & Meng, 2016; Qian & Murphy, 2011; Wang, Lagakos, Ware, Hunter, & Drazen, 2007; Yang, Miescke, & McCullagh, 2012; Zhao, Zeng, Rush, & Kosorok, 2012). Such a need in many applications gives rise to the research of *individualized inference*. Individualized inference methods through fusion learning broaden the scope of population-level fusion learning to inference problems of specific individuals. In particular, the goal of individualized inference through fusion learning is to improve the inference for the *specified* study/individual by combining inference results from different sources/individuals. Instead of using all available data, individualized inference methods through fusion learning prefer to select the individuals whose true parameter is the same or close to the target one's. The core of the classical fusion learning and the individualized version is the weighted aggregation of estimating functions from a set of individuals. While in classical fusion learning the weights are usually equal or the inverse of variance matrices of the parameter estimators from individuals, individualized inference through fusion learning assigns weights usually based on the similarity between individuals, which is the major difference between the classical fusion learning and the individualized one.

A classical approach that also exploits similar individuals in local neighborhoods is the k-nearest neighbor (k-NN) method (cf., e.g., Hall, Park, & Samworth, 2008). In the k-NN approach, information from $k$ similar individuals are gathered to make a local inference. The similarity is usually measured based on covariates that are assumed without measurement errors. Apart from k-NN, in individualized inference approaches through fusion learning, the covariates used in similarity assessment are often assumed with measurement errors. In addition, the individual-level point estimates of the parameter of interest, $\hat{\theta}$, can be utilized in identifying neighborhoods. Furthermore, fusion learning based individualized inference methods are usually kernel-based. A large number of other individuals are fused with weight functions that measures importance.

This article reviews an aggregation-based framework for fusion learning and discusses in detail two specific fusion learning based individualized inference approaches: *i*Fusion (Shen, Liu, & Xie, 2019) and *i*Group(Cai, Chen, & Xie, 2019). The two approaches are established under different asymptotic settings and are therefore applicable for different cases. *i*Fusion considers a population with a finite number of individuals, but each individual can have an infinite number of observations. *i*Group studies a population consists of an infinite number of individuals, each with a finite data size. Individualized inference with fusion learning provides a tool to make individualized inference in a heterogeneous population, which classical fusion learning cannot handle. Theoretically, both *i*Fusion and *i*Group have promising asymptotic performances; for example, minimizing mean squared error and other generic risk functions. Computationally, both are scalable for big data.

## 2 | PREVIOUS WORK ON FUSION LEARNING AND INDIVIDUALIZED INFERENCE

The methodology of fusion learning can be traced back to the studies of meta-analysis. The terminology "meta-analysis" was named by Glass (1976) as the "analysis of analyses", where outcomes from different studies on the same object are combined to provide a more powerful inference result. There are two important factors in meta-analysis: what to combine and how to combine. Marden (1991) discusses $p$-value based meta-analysis approaches, where $p$-values as point summary information are combined with equal weights. Model-based meta-analysis approaches (Normand, 1999) including fixed-effects models and random-effects models are developed to deal with possibly unobserved heterogeneity. In fixed-effects models, the parameters of interest are assumed to be unknown and fixed, while in random-effects models, the parameters of interest are supposed to be generated randomly from a distribution controlled by some hyper-parameters. In both models, the summary information of different studies are combined with unequal weights, which usually relate to their precision. Xie, Singh, and Strawderman (2011) unify these meta-analysis models under the framework using confidence distribution (CD). We refer the aggregation-based meta-analysis approaches as "fusion learning."

The idea of data fusion or data aggregation has wide applications and has shown success in other areas of researches. In parallel to the statistical studies in meta-analysis, a fusion approach was investigated by Voorhees and Gupta (1995) to solve the collection fusion problem in researches of information retrieval, where an optimal inquiry strategy is proposed to combine query results from multiple independent databases. In statistical learning, ensemble learning fits a dataset with multiple models and combines the model outputs to generate a more accurate prediction (Opitz & Maclin, 1999; Polikar, 2006).

The individualized inference problem originates from the studies in precision medicine (or personalized medicine) (Hamburg & Collins, 2010; Insel, 2009). The goal of precision medicine is to provide an optimal treatment suggestion, which is tailored to the situation of a particular patient. Specifically, the optimality of a treatment is measured over the average effect of all controls with similar situations to the target patient. See Qian and Murphy (2011) for example. Van der Laan and Rose (2011) apply the targeted inference and targeted maximum likelihood estimate approaches proposed in van der Laan and Rubin (2006) to the causal inference and precision medicine problem to reduce the bias in estimating the treatment effect.

Analogue to doctor assigning a personalized treatment to the target patient, Liu and Meng (2016) pointed out the concept of individualized inference, where a statistician provides an individualized estimate for a target dataset utilizing information from others. As a comparison, the precision medicine focuses on finding the optimal personalized treatment assignment function in the context of causal inference, while individualized inference is a broader topic containing other inference problems.

## 3 | FUSION LEARNING BASED ON AGGREGATION

Suppose the sample dataset contains $K$ independent individuals. Each *individual* can be one study, an individual patient, or in some cases a subset of observations according to a partition of the whole sample, depending on the context of application. Let $\mathcal{S}_k$ be the dataset for individual $k$ and $\theta_k$ be the corresponding parameter of interest for individual $k$. A general framework of fusion learning includes the following three steps: (a) For each individual $k$, information about the parameter $\theta_k$ is summarized by an estimating function $m(\theta;\mathcal{S}_k)$. (b) The aggregated estimating function is calculated by a weighted average of individual functions

$$m^{(c)}(\theta)=\frac{\sum_{k=1}^{K}w_k m(\theta;\mathcal{S}_k)}{\sum_{k=1}^{K}w_k}, \tag{1}$$

where $m(\theta;\mathcal{S}_k)$ is the $k$th individual-level estimating function and $w_k \geqslant 0$ are generic weights for the aggregation. (c) A point estimation for $\theta$ is further inferred from the aggregated estimating function $m^{(c)}(\theta)$. For example, $m^{(c)}(\theta)$ can be obtained by minimizing the aggregated loss function. The weights can be either fixed or adaptive based on data. The estimating function $m(\theta;\mathcal{S}_k)$ can be log-likelihood function, pseudo-likelihood function, loss function, log CD function (Schweder & Hjort, 2016; Xie, Liu, Damaraju, & Olson, 2013), or the individual level point estimate $\hat{\theta}_k$. The purpose of aggregating different individuals is to enrich the dataset in estimating the target parameter. Depending on the target parameter, we classify fusion learning into two categories: the classical fusion learning, which aims to estimate a population-wise parameter of interest, and the fusion learning based individualized inference, which focuses more on estimating the parameter of a specified individual.

### 3.1 | Classical fusion learning

In classical fusion learning, it is usually assumed that all individuals share a common parameter of interest such that $\theta_1=\theta_2=\cdots=\theta_K=\theta$. The assumption is often imposed when the individuals are independent studies on the same object or the information aggregation is from random subsets of a single dataset. In the classical meta-analysis literature, this assumption corresponds to the fixed effects model assumption. This assumption is also slightly relaxed to assuming the parameters of each individual $\theta_k$, $k=1, \ldots, K$, are independent identical realizations from a common distribution with parameter $\theta$ (see, e.g., Normand, 1999).

A class of fusion learning approaches is through aggregating the CD functions obtained from each individual. Specifically, a CD function is a sample-dependent distribution function defined on the parameter space. It can represent confidence intervals of all levels for a parameter of interest (Xie et al., 2013). CD contains much richer information than a point estimate or a specific confidence interval. Examples of CD include bootstrap distributions, likelihood functions, Bayesian posteriors, and so forth (cf., Schweder & Hjort, 2016; Xie et al., 2013). It has been shown that combining information through CDs can preserve all information across all individuals under some regularity conditions (cf., e.g., Xie et al., 2011; Liu et al., 2014; Schweder & Hjort, 2016).

By extending the classical methods of combining $p$-values, Singh, Xie, and Strawderman (2005) proposed a general framework for combining independent CDs using any given coordinate-wise monotonic function. Depending on the choice of the monotonic function, the general framework unifies almost all existing information combination approaches in meta-analysis and other fields (cf., Xie et al., 2011; Yang et al., 2016). A special class of the general framework corresponds to the aggregation in (1), which plays an important role in meta-analysis. Specifically, let the estimating function in (1) be

$$m(\theta;\mathcal{S}_k)=F_0^{-1}(H_k(\theta)), \tag{2}$$

where $F_0(\cdot)$ is a given cumulative distribution function and $H_k(\cdot)$ is a CD for $\theta$ induced from the dataset $\mathcal{S}_k$. Then, the combined estimating function $m^{(c)}(\theta)$ in (1) corresponds to the outcome obtained using (2.2) of Xie et al. (2011). Xie et al. (2011) show that sometimes the combination with (1) and (2) can be simplified to

$$m^{(c)}(\theta)=\sum_{k=1}^{K} w_k H_k(\theta), \tag{3}$$

when $F_0$ is chosen to be the cumulative distribution of the uniform distribution and $w_k$'s are normalized. Compared with (2), the formula (3) has a more straightforward interpretation: the CDs from different resources are combined together directly via a weighted average. Note that the combining weight $w_k$ used in the CD-based fusion learning framework is determined by the data $\mathcal{S}_k$. It can be either fixed or dependent on $\mathcal{S}_k$, for example, $w_k$ is proportional to the precision matrix of the parameter estimator of individual $k$. Specifically, in cases when all resources are symmetric in terms of methodology used, number of observations, and so forth, the weights are set to be equal. When there is no difference from individual to individual, the weighs are set to equal. When in fixed-effects models and random effects models, the optimal choice of weights is proportional to the precision matrix of each individual. Otherwise, whether fixed or adaptive weight is used depends on whether the precision is known. For instance, in fixed-effects models, if the precision matrix can be calculated exactly then a fixed weight is used. In random-effects models, the precision matrix often can only be estimated from the data and we typically would use data-adaptive weights.

The classical fusion learning framework is generic and powerful with a broad range of applications in challenging problem settings, including robust fusion learning (Xie et al., 2011), discrete data (Liu et al., 2014), heterogeneous individuals (Liu et al., 2015), and split-conquer-combine approaches (Chen & Xie, 2014). A detailed review on CD based fusion learning approaches with varies of examples is provided by Cheng, Liu, and Xie (2017).

There are also many fusion methods that do not directly utilize the CD framework. For instance, Gao and Carroll (2017) proposed a fusion scheme with pseudo-likelihood functions, where the integrated pseudo-likelihood function is an aggregation of individual pseudo-likelihoods.

$$\ell^{(c)}(\theta)=\sum_{k=1}^{K} w_k \ell_k(\theta_k;\mathcal{S}_k), \tag{4}$$

where $\ell_k(\theta_k;\mathcal{S}_k)$ is the pseudo-likelihood function of $k$th individual with observations $\mathcal{S}_k$. Data-structure-based practical strategies for choosing the weight $w_k$ in (4) are provided in Varin and Vidoni (2006) and Joe and Lee (2009). Further inference and variable selection are feasible with the combined pseudo-likelihood function $\ell^{(c)}$ as shown in Gao and Carroll (2017).

## 3.2 | Individualized inference through fusion learning

When potential heterogeneity exists in the population, the identical parameter assumption in fusion learning often does not hold. Instead of assuming a common parameter of interest for all individuals, individualized inference through fusion learning focuses on improving the inference efficiency of one specific study or individual by borrowing strength from similar studies or individuals. Specifically, suppose the individual of interest is marked as individual 0 with the parameter of interest $\theta_0$, the goal of fusion learning based individualized inference is to provide a better point estimator, through the aggregation in (1), than that obtained based only on the data of individual 0.

The main challenge is that bias arises when fusing a heterogeneous population. On one hand, aggregating too many other individuals brings extra bias due to heterogeneity. On the other hand, if few other individuals are fused, variance reduction is limited. Individual level study yielding an estimator, say $\hat{\theta}_0$, is the extreme case with no bias but also no variance reduction. The population-wise fusion learning, where all individuals are fused together with equal weights, is another extreme that maximizes variance reduction but may potentially have a large bias. Individualized version of fusion learning alleviates the problem by taking control over the aggregation weight $w_k$ through a similarity measure between $\mathcal{S}_k$ and $\mathcal{S}_0$. Particularly, the aggregation formula for individualized inference through fusion learning is

$$m_0^{(c)}(\theta) = \frac{\sum_{k=0}^{K} w_{0,k} m(\theta; \mathcal{S}_k)}{\sum_{k=0}^{K} w_{0,k}}, \tag{5}$$

where the extra subscript 0 indicates the target individual of such an aggregation. In fusion learning, the aggregation in (1) is constructed once and yields one point estimate for all individuals since they share the common parameter of interest. However, in individualized inference, the corresponding individualized aggregation (5) is constructed for each individual of interest.

There are two individualized inference approaches that are fusion learning based in the literature. They focuse on two different settings.

The *i*Fusion approach proposed by Shen et al. (2019) considers the asymptotic settings when the effective sample size for each individual $n_k$ increases to infinite but the proportion converges to some value between 0 and 1 such that $n_k / \sum_k n_k = O_p(1)$. Note that the effective sample size is formally defined with the variance of the individual level estimator $\hat{\theta}_k$ such that $1/n_k \propto \mathrm{Var}(\hat{\theta}_k)$. Especially, for estimators of $\sqrt{n}$ error rate, $n_k$ equals the number of observations for individual $k$. It is shown that under the settings, an individual's inference can be further improved by incorporating additional information from similar individuals, which is referred as its clique group. To be more specific, their approach aggregates individual log CD functions according to (5) by choosing a weight function $w_{0,k}$ which converges to an indicator function of the clique group.

The *i*Group approach (Cai et al., 2019) investigated a different asymptotic scheme when each individual has a finite number of observations, but the number of individuals $K$ approaches infinite. Different from *i*Fusion where every consistent individual estimator $\hat{\theta}_k$ is asymptotically unbiased with diminishing variance, $\hat{\theta}_k$ under the *i*Group setting comes with a nondiminishing error due to the finite sample size. In *i*Group, either individual level estimators $\hat{\theta}_k$ or individual level M-estimating functions can be aggregated through (5). Loss functions, log-likelihood functions, and other objective functions can all be viewed as the individual level M-estimating function. The weight function used is constructed based on both individual estimators $\hat{\theta}_k$ and some exogenous variable $\boldsymbol{z}_k$ that helps in measuring similarity. Cai et al. (2019) show that *i*Group approach can minimize the overall risk given the target individual.

Both *i*Fusion and *i*Group methods are established through the individualized aggregation (5), though they consider different asymptotic schemes and construct $m(\theta; \mathcal{S}_k)$ and $w_{0,k}$ in different ways. Computationally, both methods can be paralleled in nature and scales up easily for big data applications. More detailed reviews on the two methods are discussed in the following subsections.

### 3.2.1 | *i*Fusion approach

Shen et al. (2019) proposed the *i*Fusion (abbr. for individualized fusion) approach to make inference for an individual study or subject, motivated by an application to build a dynamic forecast model based on the most recent 24–36 months data for each of more than 10,000 companies in a time series dataset. Fitting a unique model for each company with its own data results in unstable and inefficient models due to the small data size. For each target individual, *i*Fusion aims to find a group of similar individuals that share similar traits of the target one. The group is known as the clique group for the target individual. The statistical inference based on the clique group improves the analysis of the target individual in terms of variance.

More specifically, recall $n_k$ is the effective sample size for individual $k$ and $n = \sum_{k=0}^{K} n_k$ is the total number of observations. Assuming $n_k/n \to r_k \in (0, 1)$ for some $r_k = O(1)$ as $n \to \infty$, *i*Fusion defines a *clique* for the target individual 0 as

$$\mathcal{C}_0 = \left\{ \theta_k : n^{1/2} \|\theta_k - \theta_0\|_2 = o(1), \quad k = 0, 1, \ldots, K \right\}, \tag{6}$$

where the true parameters $\{\theta_0, \ldots, \theta_K\}$ are assumed to vary as the total sample size $n$ changes. The *boundary* set $\mathscr{B}_0$ and the *disperse* set $\mathcal{D}_0$ can be defined in a similar way:

$$\mathcal{B}_0 = \left\{ \theta_k : n^{1/2} \|\theta_k - \theta_0\|_2 \to c, \quad \text{for some constant } c \in (0, \infty), \quad k = 0, 1, \ldots, K \right\}, \tag{7}$$

$$\mathcal{D}_0 = \left\{ \theta_k : n^{1/2} \|\theta_k - \theta_0\|_2 \to \infty, \quad k = 0, 1, \ldots, K \right\}. \tag{8}$$

The three sets $\mathcal{C}_0$, $\mathcal{B}_0$, and $\mathcal{D}_0$ form a partition of the set of the true parameters of all $K$ individuals $\{\theta_0, \ldots, \theta_K\}$, where each $\theta_k$ is a series as $n \to \infty$. Inclusion of the individuals in $\mathcal{C}_0$ improves the inference of $\theta_0$ at the cost of negligible bias, while inclusion of the individuals in $\mathcal{D}_0$ may incur non-negligible bias. In reality, the membership of all those three sets is unknown and it is usually difficult to separate parameters in $\mathcal{C}_0$ from those in $\mathcal{B}_0$. Shen et al. (2019) point out that if the weight satisfies the property that

$$w_{0,k} = \begin{cases} 1 + o_p\left(n^{-1/2}\right) & \text{if } \theta_k \notin \mathcal{D}_0; \\ o_p\left(n^{-1/2}\right) & \text{otherwise.} \end{cases} \tag{9}$$

for all $k$, then the $i$Fusion estimator obtained by $\hat{\theta}_0^{(c)} = \arg\max_\theta m_0^{(c)}(\theta)$ is consistent with a $\sqrt{n}$-rate bias brought by the individuals in the boundary set $\mathcal{B}_0$. Especially, when $\mathcal{B}_0 = \emptyset$, $i$Fusion approach fuses all individuals in the clique set $\mathcal{C}_0$ achieving the performance of the oracle estimator, where the oracle estimator knows the exact membership of all parameters and attains the optimal mean squared error among all possible fusions. One common choice of the weight function satisfying condition (9) is the kernel based weights such that

$$w_{0,k} \propto \mathcal{K}\left( \frac{\|\hat{\theta}_0 - \hat{\theta}_k\|_2}{b_n} \right), \tag{10}$$

where $\mathcal{K}(\cdot)$ is a given kernel function and $b_n$ is a sequence of bandwidths that satisfies

$$b_n / d_1 \to 0 \quad \text{and} \quad n^{1/2} b_n \to \infty,$$

where $d_1 = \min_k \left\{ \|\theta_0 - \theta_k\|_2 : \theta_k \in \mathcal{D}_0 \right\}$ is the minimal distance between $\theta_0$ and any parameter in the disperse set.

In essence, $i$Fusion enhances the efficiency of the target individual inference by combining CD functions from individuals in its clique set and boundary set. Members of the clique set and the boundary set can be asymptotically selected by any individualized weight function that satisfies the condition (9). Theoretically, $i$Fusion is shown to achieve the oracle property under the absence of boundary set.

### 3.2.2 | *i*Group approach

Cai et al. (2019) proposed the $i$Group approach under the asymptotic setting when each individual has a finite number of observations but the number of individuals goes to infinity. Under this setting, the error in $\hat{\theta}_k$ cannot be ignored due to the finite sample size. $i$Fusion with the weight function in (10) is not applicable for this setting, because as $b_n \to 0$ the weight in (10) tends to select the $\hat{\theta}_k$'s that is close to $\hat{\theta}_0$, which cannot offset the error $\hat{\theta}_0 - \theta_0$, especially when $\hat{\theta}_0$ is inaccurate with a large variance.

$i$Group considers the hierarchical structure shown in Figure 1, where $\pi(\cdot)$ is the population distribution of the true parameters $\{\theta_0, \ldots, \theta_K\}$ as $K \to \infty$ and is usually unknown. $p(\cdot;\theta_k)$ is a probability model generating the observations $\boldsymbol{x}_k$ under parameter $\theta_k$. The $\boldsymbol{z}$ model depicts how the exogenous variable $\boldsymbol{z}_k$ is generated through a latent parameter $\boldsymbol{\eta}_k$ with a continuous function $g(\cdot)$. $\boldsymbol{z}$ model works as an exogenous model, which does not show up in the estimation of $\hat{\theta}$ but closeness between $\boldsymbol{z}$'s implies closeness between $\theta$'s. Under this setting, only $\boldsymbol{x}_k$ and $\boldsymbol{z}_k$ are observed and individual level estimator $\hat{\theta}_k$ is estimated from $\boldsymbol{x}_k$.

$i$Group approach is an individualized aggregation based approach as well, but the weight construction is different from the one in $i$Fusion. The weight $w_{0,k}$ in (5) for $i$Group can be constructed based on $\hat{\theta}$, $\boldsymbol{z}$ or both, depending on their availability. A typical weight function for $i$Group has two parts:

**FIGURE 1** Hierarchical structure and parameter diagram

| $\boldsymbol{x}$ model | $\boldsymbol{z}$ model | diagram |
|---|---|---|
| $\theta_k \sim \pi(\cdot),$ <br> $\boldsymbol{x}_k \mid \theta_k \sim p(\cdot;\theta_k),$ | $\theta_k = g(\boldsymbol{\eta}_k),$ <br> $\boldsymbol{z}_k \mid \boldsymbol{\eta}_k \sim q(\cdot;\boldsymbol{\eta}_k).$ | $\begin{array}{ccc} \theta_k & \longleftarrow & \boldsymbol{\eta}_k \\ \downarrow & & \downarrow \\ \boldsymbol{x}_k & & \boldsymbol{z}_k \end{array}$ |

$$w_{0,k} = w_1(\boldsymbol{z}_0, \boldsymbol{z}_k) w_2\left(\hat{\theta}_0, \hat{\theta}_k \mid \boldsymbol{z}_0, \boldsymbol{z}_k\right).$$

The first part $w_1$ measures the similarity between $\boldsymbol{z}_0$ and $\boldsymbol{z}_k$ through a kernel function $\mathcal{K}_1(\cdot)$:

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

where $b_1$ is the bandwidth. The second part $w_2$ calculates the similarity between $\hat{\theta}_0$ and $\hat{\theta}_k$ conditioned on $\boldsymbol{z}_0$ and $\boldsymbol{z}_k$ with the noises in $\hat{\theta}_0$ and $\hat{\theta}_k$ considered:

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

Theoretically, Cai et al. (2019) show that the *i*Group estimator converges to the Bayes estimator under unknown population "prior" $\pi$, which minimizes the expected loss for a given target individual with observations $\boldsymbol{x}_0$ and $\boldsymbol{z}_0$. In particular, if $m(\theta;\mathcal{S}_k) = \hat{\theta}_k$ is aggregated as in (5), $m_0^{(c)}$ converges to the posterior mean $\mathbb{E}_\pi[\theta \mid \boldsymbol{x}_0, \boldsymbol{z}_0]$ with asymptotic normality. When a general estimating function is aggregated, the *i*Group estimator $\hat{\theta}_0^{(c)} = \mathrm{argmax}_\theta m_0^{(c)}(\theta)$ converges in probability to the Bayes estimator, $\arg\max_\delta \mathbb{E}[L(\delta,\theta) \mid \boldsymbol{x}_0, \boldsymbol{z}_0]$, under certain loss function $L$. For example, when the log-likelihood function is aggregated, the corresponding loss function $L$ is the Kullback–Leibler divergence.

In summary, *i*Group makes individualized inference on the target individual 0 by aggregating information from similar individuals with a soft-threshold weight design, which takes both $\boldsymbol{x}$, $\boldsymbol{z}$ and their errors into consideration. *i*Group estimator is shown to be asymptotically optimal in terms of minimizing the expected loss function given the observed data. In practice, the bandwidth $b_1$ can be estimated by cross-validation within a local set of individuals and the weight function $w_2(\cdot)$ can be approximated by bootstrapping as proposed in Cai et al. (2019).

## 4 | EXAMPLE IN FINANCE

In this example, we illustrate the methodology of *i*Group with an application in estimating the value at risk in stock returns. Specifically, denote the return of stock $k$ in day $t$ as $r_{t,k}$. The one-day value at risk (VaR) of $r_{t,k}$, denoted as $\widehat{VaR}_{t,k}$, is defined as the smallest quantity $v$ such that the probability of the event $r_{t,k} < -v$ is no greater than a predetermined confidence level $\alpha$ (for example, 1%). Statistically, $-v$ is the $\alpha$ quantile of $r_{t,k}$. In quantitative finance and risk management, VaR is widely used to estimate the possible losses in worse cases (e.g., 1% lower quantile) due to adverse market moves. In practice, it is usually difficult to estimate the value of risk. On one hand, it requires a large size of data to estimate small quantiles accurately. On the other hand, the market conditions change over time, which limits the available sample size for one company. The large number of stocks in the market and the limited number of observations for each stock coincide with the setting of *i*Group.

In this application, we consider the daily return of 490 stocks in S&P 500 for 2016. Three approaches to estimate VaR are compared in estimating the 1-day VaR with $\alpha = 0.01$.

*Individual*: A naive method to estimate 0.01 VaR for stock $k$ is to use the empirical quantile of past 100 days observations $r_{t-1,k}, \ldots, r_{t-100,k}$ such that

$$\widehat{VaR}(t,k) = \min\{r_{t-1,k}, \ldots, r_{t-100,k}\}.$$

*Classical fusion learning*: The second approach assumes all stocks are the same and aggregate them equally. The value-at-risk could then be estimated by pooling historical returns of all stocks. In this case, the estimator is

$$\widehat{VaR}(t,k) = Q_{0.01}\left(\bigcup_{l=1}^{K}\bigcup_{s=1}^{100}\{r_{t-s,l}\}\right),$$

where $Q_\alpha(A)$ is the empirical 0.01 quantile estimator given a set of observations $A$.

*iGroup*: The third approach is an application of the *i*Group approach. Before applying any aggregation, we use their Fama–French factors as the variable $\boldsymbol{z}_k$. Specifically, assume on each day, each stock return follows the Fama–French three factor model (Fama & French, 1993):

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

where $MKT$, $SMB$, and $HML$ are the three Fama–French factors, and $b_{0,k,t}$, $b_{1,k,t}$, and $b_{2,k,t}$ are the corresponding coefficients for the stock labeled $k$ at time $t$. The three coefficients characterize stocks by their sensitivity to the corresponding factors. In this model, we assume the Fama–French coefficients $b_0$, $b_1$, $b_2$ vary over time slowly. Therefore, the Fama–French coefficients could be used as the variable $\boldsymbol{z}$ in the *i*Group framework. To be more specific, the *i*Group estimator is

$$\widehat{VaR}(t,k) = Q_{0.01}^{(w)}\left(\bigcup_{l=1}^{K}\bigcup_{s=1}^{S}\{(r_{t-s,l}, w(\boldsymbol{z}_{t,l};\boldsymbol{z}_{t,k}))\}\right),$$

where $Q_{0.01}^{(w)}(\cdot)$ is the empirical 0.01 quantile estimator from a weighted sample and $\boldsymbol{z}_{t,k} = (b_{0,t,k}, b_{1,t,k}, b_{2,t,k})$ are the Fama–French coefficients of stock $k$ fitted using the returns in the past 100 days. The weight function here is chosen to be a Gaussian kernel

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

The bandwidth $b$ is the parameter to be tuned.

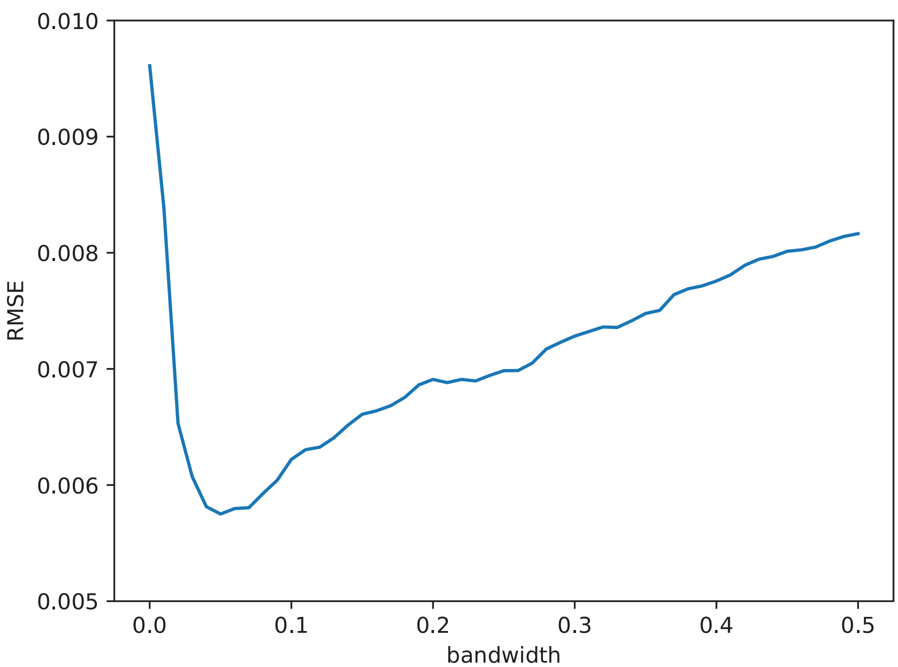


**FIGURE 2** Prediction error (RMSE) as a function of bandwidth

**TABLE 1** Prediction error for three candidate models

| Method | Individual estimation | Fusion learning estimation | *i*Group estimation |
|---|---|---|---|
| RMSE | $9.61 \times 10^{-3}$ | $1.34 \times 10^{-2}$ | $5.75 \times 10^{-3}$ |

Note that, the weighted empirical quantile function used in *i*Group estimation is equivalent to aggregating the following objective function

$$M_k(\theta;t) = \sum_{s=1}^{S} |r_{t-s,k} - \theta| \left(0.01\mathbf{1}_{\{r_{t-s,k} > \theta\}} + 0.99\mathbf{1}_{\{r_{t-s,k} < \theta\}}\right).$$

In this study, we use $K = 490$ stocks. The prediction error is measured over 250 trading days in the year 2016 for 490 stocks using

$$RMSE = \left[\frac{1}{490}\sum_{k=1}^{490}\left(\frac{1}{250}\sum_{t=1}^{250}\mathbf{1}_{\left\{r_{t,k} < \widehat{VaR}(t,k)\right\}} - 0.01\right)^2\right]^{1/2},$$

where $\widehat{VaR}(t,k)$ is based on returns $\{r_{t-1,\,k}, ..., r_{t-100,\,k},\ k = 1, ..., 490\}$.

Figure 2 shows the RMSE curve as a function of the bandwidth $b$. The bandwidth controls the bias-variance tradeoff. It is seen from the figure that the V-shaped RMSE curve decreases at the beginning and achieves a minimal value at approximately $b = 0.05$ with minimum RMSE being $5.75 \times 10^{-3}$. The RMSEs of each model are shown in Table 1. The *i*Group estimator improves the accuracy significantly.

## 5 | CONCLUSIONS

Individualized inferences through fusion learning provide a general framework to make individualized inferences with techniques in fusion learning. Based on the individualized aggregation (5), both *i*Fusion and *i*Group are shown to improve individualized inference by extracting useful information from a potential heterogeneous population. *i*Fusion and *i*Group are applicable for different asymptotic settings, and both are proven to have promising theoretical results. By focusing on the local structure around the target individual, fusion learning based individualized inference algorithms can be easily paralleled and is computationally scalable. Overall, with the availability of big data and with the increasing demands for personalized inference, individualized inference approaches through fusion learning can be a powerful tool to make efficient individualized inferences with sound theoretical support.

### CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

### AUTHOR CONTRIBUTION

**Chencheng Cai**: Methodology; writing-original draft; writing-review and editing. **Rong Chen**: Methodology; supervision, writing-review and editing. **Min-ge Xie**: Methodology; supervision, writing-review and editing.

### FURTHER READING

Cheng et al. (2017) review the CD based fusion learning approaches with several challenging problem settings. For readers interested in precision medicine, we would refer to Qian and Murphy (2011), Zhao et al. (2012), Yang et al. (2012), Collins and Varmus (2015), and Wang et al. (2007). Targeted inference and targeted maximum-likelihood-estimate approach proposed in van der Laan and Rubin (2006) is used to estimate certain parameter of interest. Applications of targeted learning in causal inference and in precision medicine can be found in Bembom and van der Laan (2007) and Van der Laan and Rose (2011).

### ORCID

*Min-ge Xie* https://orcid.org/0000-0002-9946-4715

## REFERENCES


Bembom, O., & van der Laan, M. J. (2007). A practical illustration of the importance of realistic individualized treatment rules in causal inference. *Electronic Journal of Statistics*, *1*, 574–596.

Cai, C., Chen, R., & Xie, M. (2019). Individualized group learning. Retrieved from https://arxiv.org/abs/1906.05533

Chen, X., & Xie, M. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, *24*(4), 1655–1684.

Cheng, J. Q., Liu, R. Y., & Xie, M. (2017). Fusion learning. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, & J. L. Teugels (Eds.), *Wiley StatsRef: Statistics reference online* (pp. 1–8). Hoboken, NJ: American Cancer Society. https://doi.org/10.1002/9781118445112.stat07922

Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, *372*(9), 793–795.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3–56.

Gao, X., & Carroll, R. J. (2017). Data integration with high dimensionality. *Biometrika*, *104*(2), 251–272.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–8.

Hall, P., Park, B. U., & Samworth, R. J. (2008). Choice of neighbor order in nearest-neighbor classification. *The Annals of Statistics*, *36*(5), 2135–2152.

Hamburg, M. A., & Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine*, *363*(4), 301–304.

Insel, T. R. (2009). Translating scientific opportunity into public health impact: A strategic plan for research on mental illness. *Archives of General Psychiatry*, *66*(2), 128–133.

Joe, H., & Lee, Y. (2009). On weighting of bivariate margins in pairwise likelihood. *Journal of Multivariate Analysis*, *100*(4), 670–685.

Liu, D., Liu, R., & Xie, M. (2014). Exact meta-analysis approach for discrete data and its application to $2 \times 2$ tables with rare events. *Journal of the American Statistical Association*, *109*, 1450–1465.

Liu, D., Liu, R. Y., & Xie, M. (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: Efficiency and robustness. *Journal of the American Statistical Association*, *110*(509), 326–340.

Liu, K., & Meng, X.-L. (2016). There is individualized treatment. Why not individualized inference? *Annual Review of Statistics and its Application*, *3*(1), 79–111.

Marden, J. I. (1991). Sensitive and sturdy *p*-values. *The Annals of Statistics*, *19*(2), 918–934.

Normand, S.-L. T. (1999). Meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine*, *18*(3), 321–359.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, *11*, 169–198.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, *6*(3), 21–45.

Qian, M., & Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *The Annals of Statistics*, *39*(2), 1180–1210.

Schweder, T., & Hjort, N. L. (2016). Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions. In *Cambridge series in statistical and probabilistic mathematics*. Cambridge, England: Cambridge University Press.

Shen, J., Liu, R., & Xie, M. (2019). iFusion: Individualized fusion learning. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2019.1672557

Singh, K., Xie, M., & Strawderman, W. E. (2005). Combining information from independent sources through confidence distributions. *The Annals of Statistics*, *33*(1), 159–183.

Van der Laan, M. J., & Rose, S. (2011). *Targeted learning: Causal inference for observational and experimental data*. New York, NY: Springer Science & Business Media.

van der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. *U.C. Berkeley Division of Biostatistics Working Paper Series* (Working Paper 213).

Varin, C., & Vidoni, P. (2006). Pairwise likelihood inference for ordinal categorical time series. *Computational Statistics & Data Analysis*, *51*, 2365–2373.

Voorhees, E. M., & Gupta, N. (1995). Learning collection fusion strategies. In *Proceedings of the 18th International Conference on Research and Development in Information Retrieval*. Citeseer.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., & Drazen, J. M. (2007). Statistics in medicine-reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, *357*(21), 2189–2194.

Xie, M., Liu, R. Y., Damaraju, C. V., & Olson, W. H. (2013). Incorporating external information in analyses of clinical trials with binary outcomes. *The Annals of Applied Statistics*, *7*(1), 342–368.

Xie, M., Singh, K., & Strawderman, W. E. (2011). Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association*, *106*(493), 320–333.

Yang, G., Liu, D., Wang, J., & Xie, M. (2016). Meta-analysis framework for exact inferences with application to the analysis of rare events. *Biometrics*, *72*(4), 1378–1386.

Yang, J., Miescke, K., & McCullagh, P. (2012). Classification based on a permanental process with cyclic approximation. *Biometrika*, *99*(4), 775–786.

Zhao, Y., Zeng, D., Rush, A. J., & Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, *107*(499), 1106–1118.


**How to cite this article:** Cai C, Chen R, Xie M. Individualized inference through fusion learning. *WIREs Comput Stat*. 2020;12:e1498. https://doi.org/10.1002/wics.1498