ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

*i*Fusion: Individualized Fusion Learning

Jieli Shen, Regina Y. Liu & Min-ge Xie

To cite this article: Jieli Shen, Regina Y. Liu & Min-ge Xie (2019): *i*Fusion: Individualized Fusion Learning, Journal of the American Statistical Association, DOI: 10.1080/01621459.2019.1672557

To link to this article: https://doi.org/10.1080/01621459.2019.1672557

Accepted author version posted online: 24 Oct 2019.
Published online: 31 Oct 2019.

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 00, NO. 0, 1–17: Theory and Methods
https://doi.org/10.1080/01621459.2019.1672557

*i*Fusion: Individualized Fusion Learning

Jieli Shen[a], Regina Y. Liu[b], and Min-ge Xie[b]

[a]Deustche Bank, New York, NY; [b]Department of Statistics, Rutgers University, New Brunswick, NJ

ABSTRACT
Inferences from different data sources can often be fused together, a process referred to as "fusion learning," to yield more powerful findings than those from individual data sources alone. Effective fusion learning approaches are in growing demand as increasing number of data sources have become easily available in this big data era. This article proposes a new fusion learning approach, called "*i*Fusion," for drawing efficient individualized inference by fusing learnings from relevant data sources. Specifically, *i*Fusion (i) summarizes inferences from individual data sources as individual confidence distributions (CDs); (ii) forms a clique of individuals that bear relevance to the target individual and then combines the CDs from those relevant individuals; and, finally, (iii) draws inference for the target individual from the combined CD. In essence, *i*Fusion strategically "borrows strength" from relevant individuals to enhance the efficiency of the target individual inference while preserving its validity. This article focuses on the setting where each individual study has a number of observations but its inference can be further improved by incorporating additional information from similar studies that is referred to as its *clique*. Under the setting, *i*Fusion is shown to achieve oracle property under suitable conditions. It is also shown to be flexible and robust in handling heterogeneity arising from diverse data sources. The development is ideally suited for goal-directed applications. Computationally, *i*Fusion is parallel in nature and scales up easily for big data. An efficient scalable algorithm is provided for implementation. Simulation studies and a real application in financial forecasting are presented. In effect, this article covers methodology, theory, computation, and application for individualized inference by *i*Fusion.

ARTICLE HISTORY
Received May 2018
Accepted September 2019

KEYWORDS
Combining inferences; Combining information; Confidence distribution; Fusion learning; *i*Fusion; Individualized inference.

1. Introduction

Fusion learning refers to synthesizing statistical inferences from multiple data sources to yield a more powerful inference than those from individual data sources alone. It has become a highly researched area, partly driven by the increasing availability of data sources brought forth by the big data era. The challenges in fusion learning often stem from the volume, the complexity, and the heterogeneity of different data sources. Many approaches have been developed recently to address different aspects of fusion learning (e.g., Chen and Xie 2014; Kleiner et al. 2014; Liu, Liu, and Xie 2014; Yang et al. 2014; Liu, Liu, and Xie 2015; Tang, Zhou, and Song 2016; Liu, Liu, and Xie 2017; Zhu and Qu 2018). It should be stressed that fusion learning is different from data aggregation, as the former synthesizes inference results from different data sources while the latter aggregates all data. In many situations, akin to phenomena associated with Simpson's paradox, the latter can yield incorrect or misleading overall inference results. One such example is in Liu, Liu, and Xie (2017), which presents extremely low p-values from separate datasets of two different aircraft types indicating poor landing performance of both aircraft types, but a large p-value is obtained from the aggregated data, leading to a false conclusion of a good performance for both instead.

Given the inferences from multiple data sources, they can be combined through fusion learning to yield a more efficient overall inference. Can they also be combined to yield a more efficient inference for a specific individual data source or subject? Often, the inference based on the specific individual data source itself is valid, but it may be inefficient due to its limited sample size and ignoring information in other sources. A case in point is our collaborative project with the global consulting firm Dun & Bradstreet (D&B) which provides risk management services worldwide. It involves a practical dataset of time series from more than 10,000 companies. One objective of the project is to build a dynamic forecast model based on only the most recent 24 or 36 months data for each company. A natural approach is to construct a model, say an autoregressive model with exogenous variables as covariates, for each company, using its own time series data and relevant economic and market indices in the past two or three years. However, such individual company models tend to be unstable and inefficient due to the limited data size from each company. With the availability of the large database containing over 10,000 companies, there may exist a set of companies that share similar traits of the target company, whose information can be utilized to improve its analysis.

Motivated by the D&B project, we propose in this article a new fusion learning approach called *individualized fusion learning*, abbreviated as *i*Fusion, to strategically merge inference or information from relevant data sources to enhance inference efficiency for a target individual study or company. The proposed approach uses the tool of *confidence distributions*

CONTACT Min-ge Xie *mxie@stat.rutgers.edu* Department of Statistics, Rutgers University, New Brunswick, NJ 08903.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JASA*.

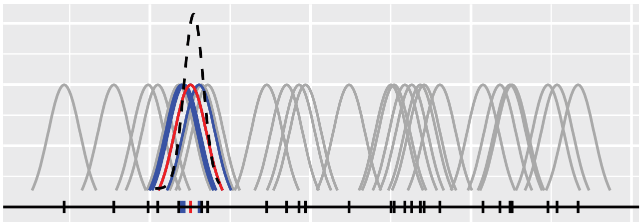

Figure 1. An illustration of the effect of individualized inference using *i*Fusion approach to combine inference results from the relevant individuals.

(CDs) (see a brief review of CDs in Section 2.1). Specifically, *i*Fusion is a three-step approach: Step (i) Analyze the data from each individual company and summarize each inference as a CD function. Step (ii) Identify a set of companies, referred to as a *clique*, whose inferences are relevant to that of the target company. Step (iii) Strategically combine the CD functions in the clique and use this combined CD to draw inference for the target company. The combining in Step (iii) uses a suitably chosen set of target-specific screen weights (detailed in Section 2.2). The choice of screen weights is crucial, and it is determined by bias-variance tradeoff to achieve the best allowable efficiency for the target inference. This article focuses mainly on the setting where an individual study has a number of observations but its inference can be further improved by incorporating additional information from similar studies in its clique. In Section 3 and for our purpose, a clique is defined as a set of studies that share the same or have similar model parameters as the target individual study; further discussions relating to grouping by covariates are provided the discussion section. Under the setting, *i*Fusion is shown to achieve the oracle property under regularity conditions. Overall, it is efficient, flexible, computationally scalable, and even robust for handling heterogeneity arising from diverse data sources.

Figure 1 presents a simple conceptual illustration of the effect of individualized inference using *i*Fusion approach to combine inference results from relevant individual studies. Each normal curve represents an individual inference result as a CD hovering around its true parameter value (marked as a bar on the x-axis). The red curve is a CD for the target individual. The peaked normal curve in the black dashed line represents the combined CD obtained by applying *i*Fusion to suitably combine the individual CD functions which are deemed relevant to the target individual, namely the individuals in the clique (colored blue). The other individuals, colored light gray, contribute negligibly, or even negatively to the inference of the target individual, and thus are excluded from the combining step of *i*Fusion.

There exist several approaches for making individualized inference. A common approach is to first cluster companies into different subgroups; for example, in Figure 1, one may apply an unsupervised learning method on point estimates of individual-specific parameters to learn the subgroups. The data in the same subgroup are then pooled to derive an overall inference for all the individuals in the same subgroup. Although this clustering approach leads to increased sample sizes in each subgroup, it has several shortcomings. For example, the formation of subgroups can be arbitrary as it depends on not only the number of clusters specified in the approach (which is known to be difficult to determine especially when the number of individuals is large), but also on the specifically chosen clustering method. Furthermore, this approach forces all the individuals in the same subgroup to have identical inference outcomes (e.g., parameter estimation or testing). Worse still, in a situation where there are no well-separated subgroups, the above subgroup approach, by imposing an artificial subgroup structure, can induce large biases in estimation and lead to invalid inference.

Bayesian hierarchical models can also be used for the D&B project. Here, a forecast model for a company would be assumed to be conditional on company-specific parameters that are further modeled through a prior or hierarchical prior distribution. Then, the resulting posterior distribution is used to make inference about individual company-specific parameters. See, for example, Gelman et al. (2013) and Gustafson, Hossain, and McCandless (2005) for discussions on Bayesian hierarchical models. However, a simple prior such as a Gaussian prior or a standard hierarchical prior may be insufficient to capture the underlying complexity of between-company heterogeneity. One may consider more complicated models and priors such as finite mixtures; the finite mixtures model faces the same difficulties in determining the number of mixture components, especially in the absence of well separated subgroups. Nonparametric Bayesian (NPB) approaches based on infinite mixtures though, for instance, Dirichlet process priors (see Grün and Leisch 2007; Hannah, Blei, and Powell 2011) may help overcome these difficulties of determining subgroups. The main challenge of these Bayesian approaches, however, is that they often rely on MCMC sampling schemes and need to analyze all companies altogether in each iteration. This is often computationally prohibitively intensive, especially for a large number of companies, unless certain scalable parallel computing platforms are involved.

The goal of *i*Fusion is similar to that of the Bayesian hierarchical methods in terms of improving inference efficiency of the target company by "borrowing information" from relevant others. *i*Fusion has the following methodological advantages: (i) inference validity in terms of frequentist properties is guaranteed by choosing properly the screen weights so that the information sharing is taken place only among relevant individuals; (ii) the proposed framework can be easily adapt to any forms of individual parameters, so *i*Fusion is essentially nonparametric and needs no assumptions of any priors on the underlying parameters; and (iii) it naturally fits in the "divide-and-conquer" scheme and can be scaled up to big data applications such as the D&B project, due to the fact that the first step of analyzing individual companies can be performed without accessing the entire dataset, which can be easily done by distributed or parallel implementation. All these make *i*Fusion particularly appealing, especially in big data applications.

We organize the rest of the article as follows. In Section 2.1, we briefly review CDs and show how CDs facilitate fusion learning in general. We describe in Section 2.2 a general *i*Fusion approach, and then show in Section 3 that *i*Fusion provides a proper and efficient inference for a target individual, and achieves the oracle property under some suitable regularity conditions. Section 4 extends *i*Fusion to heterogeneous data settings. Section 5 describes implementation details, including a scalable tuning algorithm. Sections 6 and 7 present, respectively, simulation studies and a real-data application to demonstrate the effectiveness of *i*Fusion. Section 8 contains some concluding remarks and discussions of other settings where the *i*Fusion methodology can be further developed.

2. Methodology

2.1. Confidence Distribution and Fusion Learning

Point or interval estimates are commonly used estimates for an unknown parameter in statistical analyses. This section presents a review of CD function which, being a sample-dependent distribution function defined on the parameter space, serves as a viable alternative.

Consider a simple normal example with $x_i \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$, $i = 1, \dots, n$ for a known σ, where the mean θ is the parameter of interest. Instead of using a point (say, $\bar{x}$) or an interval (say, $(1-\alpha)$ level confidence interval $(\bar{x} + \Phi^{-1}(\alpha/2)\sigma/n^{1/2}, \bar{x} + \Phi^{-1}(1-\alpha/2)\sigma/n^{1/2})$), we can also use a sample-dependent function $N(\bar{x}, \sigma^2/n)$ to estimate θ. Such a distribution estimate, referred to as a CD, provides meaningful answers for almost all questions related to statistical analysis, including point estimation, confidence interval, and p-value (see Xie and Singh 2013; Schweder and Hjort 2016 and references therein). Cox (2013) stated that a CD is to provide "simple and interpretable summaries of what can reasonably be learned from data (and an assumed model)." A CD may be conveniently defined as "a sample-dependent distribution that can represent confidence intervals or regions of all levels for parameters of interest" (Xie and Singh 2013). A formal definition of CD can be found in Xie and Singh (2013) and Schweder and Hjort (2016). If a CD is presented as a density function when appropriate, it is referred to as a confidence density or a CD density (see Efron 1993; Singh, Xie, and Strawderman 2007).

The rich information contained in a CD makes it an effective tool to synthesize information from multiple data sources. Singh, Xie, and Strawderman (2005) proposed a general framework for combining CDs for a scalar parameter from independent data sources and showed that the combined CD yields valid statistical inference so long as each individual CD is valid, regardless how they are obtained individually. Xie, Singh, and Strawderman (2011) showed that the general framework of CD combination can subsume almost all existing meta-analysis approaches as special cases. Singh, Xie, and Strawderman (2005) established a framework for combining univariate CDs by multiplying confidence density functions, which was extended by Liu, Liu, and Xie (2015) to fusion learning on multivariate common parameters and to heterogeneous study designs, adopted later by Tang, Zhou, and Song (2016) and others. A basic combining scheme is based on

$$h^{(c)}(\boldsymbol{\theta}; \mathcal{S}_1, \dots, \mathcal{S}_K) = \prod_{k=1}^{K} h_k(\boldsymbol{\theta}; \mathcal{S}_k), \tag{1}$$

where $h_k(\boldsymbol{\theta}; \mathcal{S}_k)$ is a confidence density function derived from the kth study or individual using only its dataset $\mathcal{S}_k$. Liu, Liu, and Xie (2015) showed that the point estimator obtained from the combined CD, $\hat{\boldsymbol{\theta}}^{(c)} = \text{argmax}_\theta\, h^{(c)}(\boldsymbol{\theta}; \mathcal{S}_1, \dots, \mathcal{S}_K)$, though using only individual summary statistics, enjoys the same efficiency achieved by the maximum likelihood estimator derived from the analysis of the full dataset.

Most existing work on combining information in the current literature assume that all the individual parameter values are the same or similar. This assumption seems too stringent in many real applications. Claggett, Xie, and Tian (2014) relaxed the assumption by allowing unstructured different study parameter values in a fixed-effects meta-analysis setup, but its development was only for quantiles of the set of study parameter values and not for individual study parameters $\boldsymbol{\theta}_k$'s.

2.2. *i*Fusion by Adaptive Combination of CDs

We now proceed to describe the *i*Fusion approach and articulate its broad applicability to general settings without enforcing any assumptions on the individual parameter values. Such flexibility makes *i*Fusion particularly useful for a broad range of problems in individualized inference.

Consider a collection of K individual subjects with a dataset $\mathcal{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_K\}$, where $\mathcal{S}_k$ contains samples of size n_k generated independently for the kth individual for $k = 1, \dots, K$, respectively. For ease of presentation, we assume in this article K is a (large) constant, although the *i*Fusion development can be extended to $K \to \infty$ with some modifications on the conditions; see further discussions on the case of $K \to \infty$ in Section 8. We further assume that $n_k/n \to r_k$ for some constant $r_k \in (0, 1)$ as $n \to \infty$, where $n = \sum_{k=1}^{K} n_k$ the sample size of the entire dataset. Suppose the features for the kth individual can be characterized by a p_k-dimensional parameter $\boldsymbol{\theta}_k \in \mathbb{R}^{p_k}$. Also, in this Sections 2 and 3 assume that the K individual models have a shared model design (so $p_1 = \cdots = p_K \equiv p$), but their unknown parameter values $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K\}$ can vary across individuals or equal/close to one another. In Section 4, we extend *i*Fusion to heterogeneous model designs with varying p_k's, under which the method developed in Liu, Liu, and Xie (2015) for varying p_k's can be viewed as a special case of *i*Fusion.

Without loss of generality, individual-1, and thus $\boldsymbol{\theta}_1$, are chosen as the target unless specified otherwise. For convenience, we will use the terms individual-1, model-1, and $\boldsymbol{\theta}_1$, interchangeably. The goal is to make a valid and efficient inference about $\boldsymbol{\theta}_1$.

Obviously, data $\mathcal{S}_1$ can be analyzed directly under the assumed model-1, for which a number of statistical procedures may apply. For simplicity, we assume that $\boldsymbol{\theta}_1$ can be estimated consistently by a point estimator $\hat{\boldsymbol{\theta}}_1$, as $n \to \infty$, and $\hat{\boldsymbol{\theta}}_1$ follows asymptotically a normal distribution with an estimated variance $\hat{\Sigma}_1$. In other words, $\boldsymbol{\theta}_1$ is estimated by an asymptotic normal CD, $N(\hat{\boldsymbol{\theta}}_1, \hat{\Sigma}_1)$, with the corresponding confidence density given by

$$h_1(\boldsymbol{\theta}_1; \mathcal{S}_1) = \frac{1}{(2\pi)^{p/2}|\hat{\Sigma}_1|^{1/2}} \times \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1)^t \hat{\Sigma}_1^{-1} (\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1)\right\}. \tag{2}$$

If we use the likelihood approach, $\hat{\boldsymbol{\theta}}_1$ is then the maximum likelihood estimator of $\boldsymbol{\theta}_1$, namely, $\hat{\boldsymbol{\theta}}_1 = \arg\max_{\boldsymbol{\theta}} l_1(\boldsymbol{\theta}_1|\mathcal{S}_1)$, and an estimator of $\Sigma_1(\boldsymbol{\theta}_1)$ is $\hat{\Sigma}_1 = [-\partial^2 l_1(\boldsymbol{\theta}_1|\mathcal{S}_1)/\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_1^t]^{-1}|_{\boldsymbol{\theta}_1=\hat{\boldsymbol{\theta}}_1}$. Other estimation approaches may also be used, as long as the asymptotic normality is available under mild regularity conditions. We refer to this use of only $\mathcal{S}_1$ to make inference about $\boldsymbol{\theta}_1$ as the *individual approach*. As discussed in Section 1, without utilizing potentially useful information from other individuals in the dataset $\mathcal{S}$, such an individual approach may miss out the opportunity for improving efficiency. To improve the efficiency for $\hat{\boldsymbol{\theta}}_1$ of the individual approach, *i*Fusion adaptively borrows information from other relevant individuals. Specifically, it first conducts separately the K individual approaches to obtain the inference results as K confidence density functions $h_k(\boldsymbol{\theta}_k; \mathcal{S}_k)$, $k = 1, \ldots, K$, similar to (2). Next, it combines these confidence density functions using a set of screen weights, say, w_{1k} for $k = 1, \ldots, K$,

$$h_1^{(c)}(\boldsymbol{\theta}; \mathcal{S}_1, \ldots, \mathcal{S}_K) = \prod_{k=1}^{K} h_k(\boldsymbol{\theta}; \mathcal{S}_k)^{w_{1k}}, \tag{3}$$

where $h_k(\boldsymbol{\theta}; \mathcal{S}_k)$ is the confidence density function for $\boldsymbol{\theta}_k$ based on $\mathcal{S}_k$, and $w_{1k} \in [0, 1]$ is the screen weight for individual-k with respect to individual-1, with larger w_{1k} indicating the higher degree of relevance of individual-k to individual-1 (i.e., sharing more similar traits). Individuals very different from the target individual-1 will receive low screen weights and thus be virtually excluded. For convenience, from now on we omit $\mathcal{S}_k$ from $h_k(\cdot)$ and $\mathcal{S}_1, \ldots, \mathcal{S}_K$ from $h_1^{(c)}(\cdot)$ by setting $h_k(\boldsymbol{\theta}) = h_k(\boldsymbol{\theta}; \mathcal{S}_k)$ and $h_1^{(c)}(\boldsymbol{\theta}) = h_1^{(c)}(\boldsymbol{\theta}; \mathcal{S}_1, \ldots, \mathcal{S}_K)$. This combined $h_1^{(c)}(\boldsymbol{\theta})$ can then be used to derive a new point estimator of $\boldsymbol{\theta}_1$, namely,

$$\hat{\boldsymbol{\theta}}_1^{(c)} = \arg\max_{\boldsymbol{\theta}} \log h_1^{(c)}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{k=1}^{K} w_{1k} \log h_k(\boldsymbol{\theta}). \tag{4}$$

When the individual confidence density functions take the form of (2), some simple algebra shows

$$h_1^{(c)}(\boldsymbol{\theta}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_1^{(c)})^t \left(\sum_{k=1}^{K} w_{1k}\hat{\Sigma}_k^{-1}\right)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_1^{(c)})\right\}, \tag{5}$$

and

$$\hat{\boldsymbol{\theta}}_1^{(c)} = \left(\sum_{k=1}^{K} w_{1k}\hat{\Sigma}_k^{-1}\right)^{-1} \sum_{k=1}^{K} w_{1k}\hat{\Sigma}_k^{-1}\hat{\boldsymbol{\theta}}_k. \tag{6}$$

We establish the asymptotic properties of $\hat{\boldsymbol{\theta}}_1^{(c)}$ in Section 3.

The screen weights w_{1k}'s play a critical role in allowing *i*Fusion borrow efficiently information from other relevant individuals. The choice of w_{1k}'s hinges on the determination of the clique of relevant individuals that contribute to improving the inference for the target individual-1. Note that, incorporating the information from other individuals could potentially decrease variance of $\hat{\boldsymbol{\theta}}_1^{(c)}$ (due to the increased sample size), but could also introduce estimation bias. Intuitively, forming the clique for individual-1, say $\mathcal{C}_1$, should include those individuals that the resulting estimation bias can be offset by the resulting variance reduction. This consideration of bias-and-variance trade-off dictates how we form a clique. Specifically, the clique for individual-1 is constructed in two settings of and otherwise in Section 4, respectively.

To achieve the maximum efficiency gain by *i*Fusion, we require the screen weights w_{1k}'s to satisfy the following condition: for $k = 1, \ldots, K$, where $\mathcal{C}_1$ is the clique for individual-1,

$$w_{1k} = \begin{cases} 1 - a_k & \text{if } \boldsymbol{\theta}_k \in \mathcal{C}_1; \\ b_k & \text{otherwise.} \end{cases} \tag{7}$$

for some nonnegative $a_k, b_k = o_p(n^{-1/2})$. Under this requirement, we will be able to control the aforementioned bias-and-variance trade-off. Further theoretical details on the weight choices and clique $\mathcal{C}_1$ are given in Sections 3 and 4. Their empirical implementations are discussed Section 5.

3. Theoretical Properties of *i*Fusion

This section concerns a case where the K individual models have a shared model design with $p_1 = \cdots = p_K \equiv p$, but their parameter values, $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$, may vary. We assume that $n_k/n \to r_k \in (0, 1)$ for some constant r_k, as $n = \sum_{k=1}^{K} n_k \to \infty$. We define under this setup a *clique* for individual-1 as

$$\mathcal{C}_1 = \{\boldsymbol{\theta}_k : \boldsymbol{\theta}_k \in B_r(\boldsymbol{\theta}_1), k = 1, \ldots, K\}, \tag{8}$$

where $B_r(\boldsymbol{\theta}_1)$ is a ball centered at $\boldsymbol{\theta}_1$ with radius $r = o(n^{-1/2})$. The clique $\mathcal{C}_1$ always contains $\boldsymbol{\theta}_1$, and for any $\boldsymbol{\theta}_k \in \mathcal{C}_1$ and $k \neq 1$, it is indistinguishable from $\boldsymbol{\theta}_1$ by its $\sqrt{n}$-consistent estimates based on the current sample size. Two extreme cases are: (i) $|\mathcal{C}_1| = 1$, indicating that $\boldsymbol{\theta}_1$ is separated from all the other $\boldsymbol{\theta}_k$'s, or (ii) $|\mathcal{C}_1| = K$, indicating that all individual parameters are indistinguishable from one another. Between these two extremes is the general situation where $2 \le |\mathcal{C}_1| \le K - 1$ ($K \ge 3$), which implies a potentially suitable grouping effect around $\boldsymbol{\theta}_1$. An equivalent expression of (8) akin to the so-called "near tie" development is $\mathcal{C}_1 = \{\boldsymbol{\theta}_k : n^{1/2}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_1\|_2 = o(1), k = 1, \ldots, K\}$ (see, e.g., Xie, Singh, and Zhang 2009; Hall and Miller 2010; Claggett, Xie, and Tian 2014). It resembles a "local asymptotic" development (e.g., van deer Vaart 1998) by which "we study the local behavior around a fixed value of the target parameter through a sequence of $\sqrt{n}$-rated parameters" and "help measure the performance of an estimator in finer detail and ensure its performance in moderate sample size" (Claggett, Xie, and Tian 2014). Similar asymptotic considerations are also seen in the high-dimensional regression literature where it is assumed that the signal level grows at some rate of the sample size, among others.

In addition to the clique $\mathcal{C}_1$, we also define *boundary set* $\mathcal{B}_1$ and the *disperse set* $\mathcal{D}_1$ as

$$\begin{aligned} \mathcal{B}_1 = \{\boldsymbol{\theta}_k : n^{1/2}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_1\|_2 \to c, \text{ for some constant } c, \\ 0 < c < \infty, k = 1, \ldots, K\}, \end{aligned} \tag{9}$$

$$\mathcal{D}_1 = \{\boldsymbol{\theta}_k : n^{1/2}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_1\|_2 \to \infty, k = 1, \ldots, K\}, \tag{10}$$

respectively. Clearly, for individual-1, the set of K parameters can be partitioned into three disjoint sets, $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\} = \mathcal{C}_1 \cup \mathcal{B}_1 \cup \mathcal{D}_1$, and each $\boldsymbol{\theta}_k$ lies in one and only one of them. Let

$$d_1 = \min_k\{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k\|_2 : \boldsymbol{\theta}_k \in \mathcal{D}_1\} \tag{11}$$

be the minimal distance between $\boldsymbol{\theta}_1$ and any parameter inside the disperse set. By construction, we have $n^{1/2}d_1 \to \infty$. When $\mathcal{B}_1$ is empty, a $\boldsymbol{\theta}_k$ is either in $\mathcal{C}_1$ or $\mathcal{D}_1$, or equivalently

$$d_1 \equiv \min_k\{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k\|_2 : \boldsymbol{\theta}_k \notin \mathcal{C}_1\}. \tag{12}$$

We refer to $\mathcal{B}_1 = \emptyset$ or the equivalence (12) with $n^{1/2}d_1 \to \infty$ as the *separation condition*.

*i*Fusion is a local grouping approach adaptively designed for each target individual by balancing bias-variance trade-off described in Section 2.2. In the terms of the clique, boundary and disperse sets, including individuals in $\mathcal{C}_1$ for the inference of $\boldsymbol{\theta}_1$ incurs only negligible bias, but including individuals in $\mathcal{D}_1$ may incur nonnegligible bias. In reality, the membership of $\mathcal{C}_1$ is unknown, and we propose to develop a data-based screen method to identify studies inside $\mathcal{C}_1$.

We evaluate the performance of the *i*Fusion estimator in (4) using the *oracle estimator* as a benchmark, where the *oracle estimator* of $\boldsymbol{\theta}_1$, is defined by pretending the membership of $\mathcal{C}_1$ were completely known. The oracle estimator can be expressed in a mathematical form:

$$\hat{\boldsymbol{\theta}}_1^{(o)} = \arg\max_{\boldsymbol{\theta}} \log h_1^{(o)}(\boldsymbol{\theta}), \text{ where } h_1^{(o)}(\boldsymbol{\theta}) = \prod_{\boldsymbol{\theta}_k \in \mathcal{C}_1} h_k(\boldsymbol{\theta}). \quad (13)$$

Under the normal individual confidence densities, it is easy to see that

$$\hat{\boldsymbol{\theta}}_1^{(o)} = \left(\sum_{\boldsymbol{\theta}_k \in \mathcal{C}_1} \hat{\Sigma}_k^{-1} \right)^{-1} \sum_{\boldsymbol{\theta}_k \in \mathcal{C}_1} \hat{\Sigma}_k^{-1} \hat{\boldsymbol{\theta}}_k. \quad (14)$$

Lemma 1 states that $\hat{\boldsymbol{\theta}}_1^{(o)}$ is consistent, asymptotically normal, and efficient.

Lemma 1. Suppose that the membership of $\mathcal{C}_1$ is known. Then, as $n \to \infty$,

(i) $\hat{\boldsymbol{\theta}}_1^{(o)} = \boldsymbol{\theta}_1 + o_p(n^{-1/2})$;

(ii) $n^{1/2}(\hat{\boldsymbol{\theta}}_1^{(o)} - \boldsymbol{\theta}_1) \xrightarrow{d} N(\mathbf{0}, \Delta_1^{(o)})$, where $\Delta_1^{(o)} = \mathbb{E}[n(\sum_{\boldsymbol{\theta}_k \in \mathcal{C}_1} \hat{\Sigma}_k^{-1})^{-1}]$;

(iii) $\hat{\boldsymbol{\theta}}_1^{(o)}$ attains the optimal mean squared error (MSE) among all $\hat{\boldsymbol{\theta}}_1^{\mathcal{F}}$, given by

$$\hat{\boldsymbol{\theta}}_1^{\mathcal{F}} = \arg\max_{\boldsymbol{\theta}} \log \prod_{\boldsymbol{\theta}_k \in \mathcal{F}} h_k(\boldsymbol{\theta}_k), \text{ for any } \mathcal{F} \subseteq \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}. \quad (15)$$

A proof of Lemma 1 is given in Appendix A. Results in (i) and (ii) of Lemma 1 imply that the oracle estimator is consistent and asymptotic normal. Result (iii) of Lemma 1 further shows that the choice of $\mathcal{F} = \mathcal{C}_1$ yields the smallest asymptotic MSE, among all the estimators given in the form of (15). Note that the individual estimator $\hat{\boldsymbol{\theta}}_1$ itself is a special case of $\hat{\boldsymbol{\theta}}_1^{\mathcal{F}}$ with $\mathcal{F} = \{\boldsymbol{\theta}_1\}$. Furthermore, to achieve consistency and asymptotic normality, the individuals to be combined under the conventional meta-analysis or fusion learning methods are generally required to have the same parameter values. Here, $\mathcal{C}_1$ only requires the individuals to be combined having parameters sufficiently near the target parameter.

Theorem 1 states that our *i*Fusion estimator $\hat{\boldsymbol{\theta}}_1^{(c)}$ performs as well as the oracle approach asymptotically, even without knowing the memberships of $\mathcal{C}_1$. Specifically, Theorem 1 provides a sufficient condition on the screen weights, under which $\hat{\boldsymbol{\theta}}_1^{(c)}$ is a consistent estimate of $\boldsymbol{\theta}_1$, follows a normal distribution asymptotically, and moreover, achieves the same limiting covariance matrix and MSE as those of the oracle estimator $\hat{\boldsymbol{\theta}}_1^{(o)}$. A proof of Theorem 1 is given in Appendix A.

Theorem 1 (Oracle property). Suppose that w_{1k} satisfies (7), where $\mathcal{C}_1$ is defined in (8), and the separation condition also holds. Then, as $n \to \infty$, $\hat{\boldsymbol{\theta}}_1^{(c)}$ obtained from (4) possesses the following properties:

(i) $\hat{\boldsymbol{\theta}}_1^{(c)} = \boldsymbol{\theta}_1 + o_p(n^{-1/2})$;

(ii) $n^{1/2}(\hat{\boldsymbol{\theta}}_1^{(c)} - \boldsymbol{\theta}_1) \xrightarrow{d} N(\mathbf{0}, \Delta_1^{(o)})$, where $\Delta_1^{(o)} = \mathbb{E}[n(\sum_{\boldsymbol{\theta}_k \in \mathcal{C}_1} \hat{\Sigma}_k^{-1})^{-1}]$ and can be consistently estimated by $n(\sum_{k=1}^K w_{1k} \hat{\Sigma}_k^{-1})^{-1}(\sum_{k=1}^K w_{1k}^2 \hat{\Sigma}_k^{-1})(\sum_{k=1}^K w_{1k} \hat{\Sigma}_k^{-1})^{-1}$;

(iii) $\hat{\boldsymbol{\theta}}_1^{(c)}$ has the same MSE as the oracle estimator $\hat{\boldsymbol{\theta}}_1^{(o)}$, and thus attains the optimal MSE among all $\hat{\boldsymbol{\theta}}_1^{\mathcal{F}}$ defined in (15).

It is worth noting that condition (7) in Theorem 1 can be satisfied in different approaches. For instance, it is satisfied by the following data-driven and kernel-based screen weights

$$w_{1k} = \mathcal{K}\left(\frac{\|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k\|_2}{b_n} \right) \Big/ \mathcal{K}(0), \quad (16)$$

where b_n is a bandwidth parameter and $\mathcal{K}(\cdot)$ is a given kernel function. Different choices of kernel functions may require different choices of bandwidths and also result in some change in the finite-sample behaviors of w_{1k}. This point is discussed further in Section 8. To simplify our presentation, we use a uniform kernel $\mathcal{K}(\cdot) = \frac{1}{2}\mathbb{1}\{|\cdot| \le 1\}$ throughout the article.

Lemma 2 suggests that condition (7) can be satisfied when formula (16) is used with a suitably-chosen bandwidth b_n. Its proof is also deferred to Appendix A.

Lemma 2. The screen weights w_{1k}'s in (16) with $\mathcal{K}(\cdot) = \frac{1}{2}\mathbb{1}\{|\cdot| \le 1\}$ satisfies (7) if b_n satisfies

$$b_n/d_1 \to 0 \quad \text{and} \quad n^{1/2} b_n \to \infty. \quad (17)$$

We have assumed the separation condition in Theorem 1, under which *i*Fusion is shown to yield an estimator asymptotically equivalent to the oracle estimator, and thus the most efficient inference about $\boldsymbol{\theta}_1$. We now turn to the case that $\mathcal{B}_1 \neq \emptyset$ and the separation condition does not hold. Note that the parameters in $\mathcal{B}_1$ are not easy to separate from those in $\mathcal{C}_1$ by using data alone, and inclusion of an individual in $\mathcal{B}_1$ often reduces estimation standard deviation at the same rate as the bias it incurs. Theorem 2 quantifies precisely the performance of *i*Fusion under this setting.

Theorem 2. Assume that the screen weights w_{1k}'s satisfies

$$w_{1k} = \begin{cases} 1 - a_k & \text{if } \boldsymbol{\theta}_k \notin \mathcal{D}_1; \\ b_k & \text{otherwise} \end{cases} \quad (18)$$

for some nonnegative $a_k, b_k = o_p(n^{-1/2})$, for $k = 1, \ldots, K$. Then, $\hat{\boldsymbol{\theta}}_1^{(c)}$ obtained from (4) possesses the following properties: as $n \to \infty$,

(i) $\hat{\boldsymbol{\theta}}_1^{(c)} = \boldsymbol{\theta}_1 + O_p(n^{-1/2})$;

(ii) $n^{1/2}(\hat{\boldsymbol{\theta}}_1^{(c)} - \boldsymbol{\theta}_1 - \boldsymbol{B}_1^{(c)}) \xrightarrow{d} N(\mathbf{0}, \Delta_1)$, where $\boldsymbol{B}_1^{(c)} = (\sum_{\boldsymbol{\theta}_k \notin \mathcal{D}_1} \hat{\Sigma}_k^{-1})^{-1}(\sum_{\boldsymbol{\theta}_k \in \mathcal{B}_1} \hat{\Sigma}_k^{-1}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_1))$, and $\Delta_1 = \mathbb{E}[n(\sum_{\boldsymbol{\theta}_k \notin \mathcal{D}_1} \hat{\Sigma}_k^{-1})^{-1}]$; and

(iii) $\mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{(c)}) \leq \mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{\mathcal{F}})$, provided that $\mathcal{D}^{\mathcal{F}} \neq \emptyset$, or

$$\sum_{\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2} \in \mathcal{B}_1} (\boldsymbol{\theta}_{k_1} - \boldsymbol{\theta}_1)^t \hat{\Sigma}_{k_1}^{-1} \left(\sum_{\boldsymbol{\theta}_k \notin \mathcal{D}_1} \hat{\Sigma}_k^{-1} \right)^{-2} \times \hat{\Sigma}_{k_2}^{-1}(\boldsymbol{\theta}_{k_2} - \boldsymbol{\theta}_1) + \mathrm{tr}\left\{ \left(\sum_{\boldsymbol{\theta}_k \notin \mathcal{D}_1} \hat{\Sigma}_k^{-1} \right)^{-1} \right\}$$
$$\leq \sum_{\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2} \in \mathcal{B}^{\mathcal{F}}} (\boldsymbol{\theta}_{k_1} - \boldsymbol{\theta}_1)^t \hat{\Sigma}_{k_1}^{-1} \left(\sum_{\boldsymbol{\theta}_k \in \mathcal{F}} \hat{\Sigma}_k^{-1} \right)^{-2} \times \hat{\Sigma}_{k_2}^{-1}(\boldsymbol{\theta}_{k_2} - \boldsymbol{\theta}_1) + \mathrm{tr}\left\{ \left(\sum_{\boldsymbol{\theta}_k \in \mathcal{F}} \hat{\Sigma}_k^{-1} \right)^{-1} \right\}, \tag{19}$$

for any given $\mathcal{F} = \mathcal{C}^{\mathcal{F}} \cup \mathcal{B}^{\mathcal{F}} \cup \mathcal{D}^{\mathcal{F}}$ with $\mathcal{C}^{\mathcal{F}} \subseteq \mathcal{C}_1, \mathcal{B}^{\mathcal{F}} \subseteq \mathcal{B}_1$, and $\mathcal{D}^{\mathcal{F}} \subseteq \mathcal{D}_1$.

As in (i) of Theorem 1, (i) of Theorem 2 suggests that the *i*Fusion estimator is consistent. Results (ii) and (iii) of Theorem 2 are similar to results (ii) and (iii) of in Theorem 1, although they include a bias correction term $\boldsymbol{B}_1^{(c)}$ that involves with unknown parameter values in the boundary set $\mathcal{B}_1$. Note that, if the parameter values in $\boldsymbol{B}_1^{(c)}$ are substituted with their corresponding $\sqrt{n}$-consistent estimators, the limiting distribution $n^{1/2}(\hat{\boldsymbol{\theta}}_1^{(c)} - \boldsymbol{\theta}_1 - \boldsymbol{B}_1^{(c)})$ becomes nonnormal. Regardless, *the results in Theorem 2 still suggest potential gains with smaller MSE by the iFusion estimator*.

To establish the claims in Theorem 2, Lemma 3 shows that (16) can be used to obtain a w_{1k} that satisfies (18), even if $\mathcal{B}_1 \neq \emptyset$ and the separation condition does not hold. The proof of Lemma 3 is similar to that of Lemma 2 and thus omitted.

Lemma 3. When $\mathcal{B}_1 \neq \emptyset$, w_{1k} given by (16) satisfies (18), provided that (17) also holds.

4. Extension to Heterogeneous Model Designs

In this section, we extend *i*Fusion to more complex study designs, where "the estimable model parameters may be different from one individual model to another" (see, e.g., Simmonds and Higgins 2007; Liu, Liu, and Xie 2015). In particular, we assume that the estimable parameter of the kth study $\boldsymbol{\theta}_k \in \mathbb{R}^{p_k}$, for $k = 1, \ldots, K$, and the vector length $p_1, \ldots, p_K$ may be different. To help quantify the difference and also make a connection between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_k$, we introduce for each k a latent vector $\mathbf{v}_k \in \mathbb{R}^q$ and assume that there is a mapping $B_k(\cdot) : \mathbb{R}^q \to \mathbb{R}^{p_k}$ from $\mathbf{v}_k$ to $\boldsymbol{\theta}_k$, where all the latent vectors $\mathbf{v}_k$'s have the same dimension q. We further partition $\mathbf{v}_k = (\boldsymbol{\psi}_k^t, \boldsymbol{\xi}_k^t)^t$ with $\boldsymbol{\psi}_k \in \mathbb{R}^{q-p}$ and $\boldsymbol{\xi}_k \in \mathbb{R}^p$. We assume for some individuals $\boldsymbol{\xi}_k$ may be the same or sufficiently close to $\boldsymbol{\xi}_1$. The task now is to extend the *i*Fusion method to this setting to improve the individual inference for the target parameter $\boldsymbol{\theta}_1$. Note that the setting considered in Sections 2.2 and 3 can be viewed as a special case here with: $p_k \equiv q \equiv p$, $B_k(\cdot)$ being the identity mapping, and $\boldsymbol{\xi}_k$ being identical to $\boldsymbol{\theta}_k$.

We use a linear regression model similar to those in Simmonds and Higgins (2007) and Liu, Liu, and Xie (2015) to illustrate heterogeneous individual model designs considered here.

Example 1. Consider K independent clinical trials (studies) conducted on different subpopulations given by the following linear model:

$$Y_{ik} = \alpha_k + \beta_k x_{ik} + \gamma_k z_{ik} + \varepsilon_{ik},\ i = 1, \ldots, n_k,\ k = 1, \ldots, K, \tag{20}$$

where Y_{ik} is the response for the ith observation from the kth subpopulation, x_{ik} is the treatment status (1/0 for treatment/control), z_{ik} is the drug dosage, with errors $\varepsilon_{ik} \overset{\text{iid}}{\sim} N(0, \sigma_k^2)$. Here, α_k is a study-specific intercept, and β_k and γ_k are study-specific regression coefficients corresponding to the treatment and drug dosage, respectively. Consider the following two scenarios:

Scenario I. Suppose the intercept α_k is subpopulation-specific and $\alpha_k \neq \alpha_1$, $k \neq 1$, but some of the treatment effects (β_k, γ_k), $k \neq 1$, are the same or close to (β_1, γ_1). We hope to borrow information from these studies to improve the inference of $\boldsymbol{\theta}_1 = (\alpha_1, \beta_1, \gamma_1)^t$. In this case, $\mathbf{v}_k \equiv \boldsymbol{\theta}_k = (\alpha_k, \beta_k, \gamma_k)^t$ and $\psi_k = \alpha_k$, $\boldsymbol{\xi}_k = (\beta_k, \gamma_k)^t$, for $k = 1, \ldots, K$. It is clear that we should devise our clique for subpopulation-1 using $\boldsymbol{\xi}_k$, rather than $\boldsymbol{\theta}_k$.

Scenario II. Continuing from Scenario I, suppose additionally in some of the clinical studies, say k, the drug dosage is not part of the research goal and thus is held constant $z_{ik} \equiv z_k$, with a fixed known constant z_k. This reduces individual Model-k to $Y_{ik} = (\alpha_1 + \gamma_k z_k) + \beta_k x_{ik} + \varepsilon_{ik}$. The estimable parameters in Model-k are $\boldsymbol{\theta}_k = (\alpha_k + \gamma_k z_k, \beta_k)^t$ rather than $(\alpha_k, \beta_k, \gamma_k)^t$. Using the mapping and notations we have introduced, we can rewrite $\boldsymbol{\theta}_k = B_k \mathbf{v}_k$ where $B_k = \left(\begin{smallmatrix} 1 & 0 & z_k \\ 0 & 1 & 0 \end{smallmatrix}\right)$ is a 2×3 matrix, $\mathbf{v}_k = (\alpha_k, \beta_k, \gamma_k)^t$ and $\psi_k = \alpha_k$, $\boldsymbol{\xi}_k = (\beta_k, \gamma_k)^t$. The question now is whether we can still borrow information from those $k \in \mathcal{K}$ studies whose $\boldsymbol{\xi}_k = (\beta_k, \gamma_k)^t$ are the same or close to $\boldsymbol{\xi}_1 = (\beta_1, \gamma_1)^t$ to improve the inference for $\boldsymbol{\theta}_1$.

To extend *i*Fusion under such heterogeneous model designs to make inference for the target parameter $\boldsymbol{\theta}_1$, we need a new combining formula and definitions of clique, boundary and disperse sets. Specifically, we use $\boldsymbol{\xi}_k$ rather than $\boldsymbol{\theta}_k$ to define the clique, boundary and disperse sets with respect to individual-1, and treat the parameter $\boldsymbol{\psi}_1$ as a nuisance parameter. Specifically,

$$\begin{aligned}\tilde{\mathcal{C}}_1 &= \{\boldsymbol{\xi}_k : n^{1/2}\|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_k\|_2 = o(1),\ k = 1, \ldots, K\} \\ &= \{\boldsymbol{\xi}_k : \boldsymbol{\xi}_k \in B_r(\boldsymbol{\xi}_1),\ k = 1, \ldots, K\},\end{aligned}$$

is the clique set, where $B_r(\boldsymbol{\xi}_1)$ is a ball centered at $\boldsymbol{\xi}_1$ with radius $r = o(n^{-1/2})$. We define

$$\begin{aligned}\tilde{\mathcal{B}}_1 &= \{\boldsymbol{\xi}_k : n^{1/2}\|\boldsymbol{\xi}_k - \boldsymbol{\xi}_1\|_2 \to c, \text{ for some constant } c, \\ &\qquad 0 < c < \infty, k = 1, \ldots, K\}, \\ \tilde{\mathcal{D}}_1 &= \{\boldsymbol{\xi}_k : n^{1/2}\|\boldsymbol{\xi}_k - \boldsymbol{\xi}_1\|_2 \to \infty,\ k = 1, \ldots, K\}.\end{aligned} \tag{21}$$

Furthermore, we denote by $\boldsymbol{\eta}_k = (\boldsymbol{\psi}_1^t, \ldots, \boldsymbol{\psi}_K^t, \boldsymbol{\xi}_k^t)^t$ and let $A_k \in \mathbb{R}^{p_k \times \{K(q-p)+p\}}$ be the matrix that maps $\boldsymbol{\eta}_k$ to $\boldsymbol{\theta}_k$. We have $\boldsymbol{\theta}_k = A_k \boldsymbol{\eta}_k$. We also extend our combining formula in Section 3 to the current setup as follows

$$h_1^{(c)}(\boldsymbol{\eta}) = \prod_{k=1}^{K} h_k(A_k \boldsymbol{\eta})^{w_{1k}}, \tag{22}$$

with the screen weights w_{1k}, where $\boldsymbol{\eta} = (\boldsymbol{\psi}_1^t, \ldots, \boldsymbol{\psi}_K^t, \boldsymbol{\xi}^t)^t$. A point estimator of $\boldsymbol{\theta}_1$ is then

$$\hat{\boldsymbol{\theta}}_1^{(c)} = A_1 \hat{\boldsymbol{\eta}}_1^{(c)}, \quad \text{where } \hat{\boldsymbol{\eta}}_1^{(c)} = \arg\max_{\boldsymbol{\eta}} \log h_1^{(c)}(\boldsymbol{\eta}). \tag{23}$$

Inference for $\boldsymbol{\theta}_1$ can then be made using $\hat{\boldsymbol{\theta}}_1^{(c)} = A_1 \hat{\boldsymbol{\eta}}_1^{(c)}$ following a procedure similar to that in Section 3 for the homogeneous model design.

If $\tilde{\mathcal{C}}_1$ were known, we could write $h_1^{(o)}(\boldsymbol{\eta}) = \prod_{\boldsymbol{\xi}_k \in \tilde{\mathcal{C}}_1} h_k(A_k \boldsymbol{\eta})$ and define the oracle estimator of $\boldsymbol{\theta}_1$ as

$$\hat{\boldsymbol{\theta}}_1^{(o)} = A_1 \hat{\boldsymbol{\eta}}_1^{(o)}, \quad \text{where } \hat{\boldsymbol{\eta}}_1^{(o)} = \arg\max_{\boldsymbol{\eta}} \log h_1^{(o)}(\boldsymbol{\eta}).$$

Similar to Lemma 1, the oracle estimator $\hat{\boldsymbol{\theta}}_1^{(o)}$ can be shown to be consistent, asymptotically normally distributed and attain the smallest asymptotic MSE among all estimators of $\boldsymbol{\theta}_1$ given by $\hat{\boldsymbol{\theta}}^{\mathcal{F}} = A_1 \hat{\boldsymbol{\eta}}^{\mathcal{F}}$, for $\mathcal{F} \subseteq \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K\}$, where $\hat{\boldsymbol{\eta}}^{\mathcal{F}} = \arg\max_{\boldsymbol{\eta}} \prod_{\boldsymbol{\xi}_k \in \mathcal{F}} h_k(A_k \boldsymbol{\eta})$.

Theorems 3 and 4 show that *i*Fusion in the extended framework to heterogeneous model designs retains similar desirable properties established in Section 3. Specifically, they give the asymptotic properties of $\hat{\boldsymbol{\theta}}_1^{(c)}$, respectively, when $\tilde{\mathcal{B}}_1 = \emptyset$ and when $\tilde{\mathcal{B}}_1 \neq \emptyset$. Theorem 3 shows that the *i*Fusion estimator achieves the oracle property, namely $\hat{\boldsymbol{\theta}}_1^{(c)}$ is a consistent estimate of $\boldsymbol{\theta}_1$, asymptotically normally distributed for suitably chosen w_{1k}'s. Moreover, it has the same limiting covariance matrix and MSE as those of $\hat{\boldsymbol{\theta}}_1^{(o)}$, showing once again that no loss of efficiency is incurred by *i*Fusion.

Theorem 3 (Oracle property). Suppose that w_{1k} satisfies, for $k = 1, \ldots, K$,

$$w_{1k} = \begin{cases} 1 - a_k & \text{if } \boldsymbol{\xi}_k \in \tilde{\mathcal{C}}_1; \\ b_k & \text{otherwise} \end{cases}$$

for some nonnegative $a_k, b_k = o_p(n^{-1/2})$, and $n^{1/2} \min\{\|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_k\|_2 : \boldsymbol{\xi}_k \notin \tilde{\mathcal{C}}_1\} \to 0$. Then, $\hat{\boldsymbol{\theta}}_1^{(c)}$ obtained from (23) possesses the following properties: as $n \to \infty$,

(i) $\hat{\boldsymbol{\theta}}_1^{(c)} = \boldsymbol{\theta}_1 + o_p(n^{-1/2})$;

(ii) $n^{1/2}(\hat{\boldsymbol{\theta}}_1^{(c)} - \boldsymbol{\theta}_1) \xrightarrow{d} N(\mathbf{0}, \Delta_1^{(o)})$, where $\Delta_1^{(o)} = \mathbb{E}[nA_1 (\sum_{\boldsymbol{\xi}_k \in \tilde{\mathcal{C}}_1} A_k^t \hat{\Sigma}_k^{-1} A_k)^{-1} A_1^t]$ and can be consistently estimated by $nA_1(\sum_{k=1}^K w_{1k} A_k^t \hat{\Sigma}_k^{-1} A_k)^{-1} (\sum_{k=1}^K w_{1k}^2 A_k^t \hat{\Sigma}_k^{-1} A_k)(\sum_{k=1}^K w_{1k} A_k^t \hat{\Sigma}_k^{-1} A_k)^{-1} A_1^t$;

(iii) $\hat{\boldsymbol{\theta}}_1^{(c)}$ has the same MSE as the oracle estimator $\hat{\boldsymbol{\theta}}_1^{(o)}$.

Suppose that $\hat{\boldsymbol{\theta}}_k$ is partitioned into $\hat{\boldsymbol{\theta}}_k = (\boldsymbol{\psi}_k^t, \boldsymbol{\xi}_k^t)^t$, similar to that described in Scenario I in Example 1. Let $\hat{\boldsymbol{\psi}}_1^{(c)}$ be the part of $\hat{\boldsymbol{\theta}}_1^{(c)}$ that estimates $\boldsymbol{\psi}_1$. An interesting byproduct of Theorem 3 is that $\hat{\boldsymbol{\psi}}_1^{(c)}$ actually improves upon $\hat{\boldsymbol{\psi}}_1$, the corresponding subpart of the individual estimate $\hat{\boldsymbol{\theta}}_1 (= \arg\max_{\boldsymbol{\theta}} \log h_1(\boldsymbol{\theta}))$. Assume, without loss of generality, that $\boldsymbol{\psi}_1$ is a scalar. This interesting finding can be summarized in the following corollary.

Corollary 1. Under the assumptions of Theorem 3, $\text{var}(\hat{\psi}_1^{(c)}) \leq \text{var}(\hat{\psi}_1)$, asymptotically.

It shows that there is efficiency gain in the joint approach over the individual approach, as the estimation of other individuals can contribute to improve the estimation of $\boldsymbol{\psi}_1$. This may seem counterintuitive at first, as other individuals contain no direct information on $\boldsymbol{\psi}_1$. However, $\boldsymbol{\psi}_1$ and $\boldsymbol{\xi}_1$ are often correlated and through this hidden correlation the improvement of the estimation of $\boldsymbol{\xi}_1$ can be passed on to the estimation of $\boldsymbol{\psi}_1$, and vice versa. As pointed out in Liu, Liu, and Xie (2015), "this phenomenon of borrowing strength is not yet well appreciated in conventional meta-analysis and the individual-specific parameters are generally reported as the final estimators." Although this advantage is observed in Liu, Liu, and Xie (2015) where $\boldsymbol{\xi}_1 = \cdots = \boldsymbol{\xi}_K$ (in our notation), *i*Fusion shows the same advantage even when $\boldsymbol{\xi}_k$'s are not identical.

Theorem 4 can be viewed as an extension of Theorem 2, showing that *i*Fusion has potential gain in efficiency in heterogeneous model designs even when $\tilde{\mathcal{B}}_1 \neq \emptyset$ and the separation condition does not hold.

Theorem 4. Suppose that w_{1k} satisfies

$$w_{1k} = \begin{cases} 1 - a_k & \text{if } \boldsymbol{\xi}_k \notin \tilde{\mathcal{D}}_1; \\ b_k & \text{otherwise} \end{cases}$$

for some nonnegative $a_k, b_k = o_p(n^{-1/2})$, for $k = 1, \ldots, K$. Then, $\hat{\boldsymbol{\theta}}_1^{(c)}$ obtained from (23) possesses the following properties: as $n \to \infty$,

(i) $\hat{\boldsymbol{\theta}}_1^{(c)} = \boldsymbol{\theta}_1 + O_p(n^{-1/2})$;

(ii) $n^{1/2}(\hat{\boldsymbol{\theta}}_1^{(c)} - \boldsymbol{\theta}_1 - \boldsymbol{B}_1^{(c)}) \xrightarrow{d} N(\mathbf{0}, \Delta_1)$, where $\boldsymbol{B}_1^{(c)} = A_1 (\sum_{\boldsymbol{\xi}_k \notin \tilde{\mathcal{D}}_1} A_k^t \hat{\Sigma}_k^{-1} A_k)^{-1} (\sum_{\boldsymbol{\xi}_k \in \tilde{\mathcal{B}}_1} A_k^t \hat{\Sigma}_k^{-1} A_k (\boldsymbol{\eta}_k - \boldsymbol{\eta}_1))$, and $\Delta_1 = \mathbb{E}[nA_1 (\sum_{\boldsymbol{\xi}_k \notin \tilde{\mathcal{D}}_1} A_k^t \hat{\Sigma}_k^{-1} A_k)^{-1} A_1^t]$; and

(iii) $\text{MSE}(\hat{\boldsymbol{\theta}}_1^{(c)}) \leq \text{MSE}(\hat{\boldsymbol{\theta}}_1^{\mathcal{F}})$, provided that $\tilde{\mathcal{D}}^{\mathcal{F}} \neq \emptyset$ or

$$\sum_{\boldsymbol{\xi}_{k_1}, \boldsymbol{\xi}_{k_2} \in \tilde{\mathcal{B}}_1} (\boldsymbol{\eta}_{k_1} - \boldsymbol{\eta}_1)^t A_{k_1}^t \hat{\Sigma}_{k_1}^{-1} A_{k_1} \left(\sum_{\boldsymbol{\xi}_k \notin \tilde{\mathcal{D}}_1} A_k^t \hat{\Sigma}_k^{-1} A_k \right)^{-2} \times A_{k_2}^t \hat{\Sigma}_{k_2}^{-1} A_{k_2} (\boldsymbol{\eta}_{k_2} - \boldsymbol{\eta}_1) + \text{tr}\left\{ \left(\sum_{\boldsymbol{\xi}_k \notin \tilde{\mathcal{D}}_1} A_k^t \hat{\Sigma}_k^{-1} A_k \right)^{-1} \right\}$$

$$\leq \sum_{\xi_{k_1},\xi_{k_2}\in\tilde{\mathcal{B}}^{\mathcal{F}}} (\boldsymbol{\eta}_{k_1}-\boldsymbol{\eta}_1)^t A_{k_1}^t \hat{\Sigma}_{k_1}^{-1} A_{k_1} \left(\sum_{\xi_k\in\mathcal{F}} A_k^t \hat{\Sigma}_k^{-1} A_k\right)^{-2} \times A_{k_2}^t \hat{\Sigma}_{k_2}^{-1} A_{k_2}(\boldsymbol{\eta}_{k_2}-\boldsymbol{\eta}_1) + \mathrm{tr}\Big\{\big(\sum_{\xi_k\in\mathcal{F}} A_k^t \hat{\Sigma}_k^{-1} A_k\big)^{-1}\Big\},$$

for any $\mathcal{F} = \tilde{\mathcal{C}}^{\mathcal{F}} \cup \tilde{\mathcal{B}}^{\mathcal{F}} \cup \tilde{\mathcal{D}}^{\mathcal{F}}$ with $\tilde{\mathcal{C}}^{\mathcal{F}} \subseteq \tilde{\mathcal{C}}_1$, $\tilde{\mathcal{B}}^{\mathcal{F}} \subseteq \tilde{\mathcal{B}}_1$, and $\tilde{\mathcal{D}}^{\mathcal{F}} \subseteq \tilde{\mathcal{D}}_1$.

The proofs of Theorems 3 and 4 are similar to their counterparts in Section 3 with only slight modification to account for A_k and are thus omitted.

To end the section, we comment on a use of the kernel-based screen weight similar to (16) that was considered in Section 3. We now suppose that either $\boldsymbol{\xi}_k$ or a subvector of $\boldsymbol{\xi}_k$, say $\boldsymbol{\zeta}_k$, is estimable in the kth study, as described in the two scenarios of Example 1. In this case, we can directly substitute $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_k$ in (16) with $\hat{\boldsymbol{\xi}}_1$ and $\hat{\boldsymbol{\xi}}_k$, respectively, if $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_k$ are both estimable parameters; Otherwise, we substitute $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_k$ with $\hat{\boldsymbol{\zeta}}_1$ and $\hat{\boldsymbol{\zeta}}_k$, respectively, with an additional assumption that close in $\boldsymbol{\zeta}_1$ and $\boldsymbol{\zeta}_k$ implies close in $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_k$. With this substitution, we can show that the results of Lemmas 2 and 3 still hold under the same conditions, following similar proofs for the two previous lemmas with slight modifications.

5. A Scalable Algorithm and Empirical Selection of Screen Weights

In this section, we cover several computational issues concerning the implementation of *i*Fusion and the empirical selection of screen weights used in our data analysis.

For a chosen kernel function, the performance of the kernel-based weights is impacted by the choice of bandwidth. To assess such a finite-sample impact and to ensure good large sample performance, it is convenient to decompose $b_n = \tau_n b$, where $b = O(1)$ is a constant. In practice, we may set τ_n according to the conditions stated in Lemma 2 so that w_{1k} behave well asymptotically. In our development, we treat the unknown constant b as a tuning parameter that may impact the performance of *i*Fusion under finite sample size: a very large b would lead to "diluted" inference due to inclusion of irrelevant individuals while a very small b would essentially lead to the same result as the individual approach, gaining no efficiency. To this end, we use a cross-validation tuning algorithm to assess b from the data itself. This is elaborated below in greater generality.

In a more practical setting, $\boldsymbol{\theta}_k$ is often a vector with its components $(\theta_{k1},\ldots,\theta_{kp})$ measured in different scales. To avoid using a multivariate kernel and multiple bandwidths, we can use the distance norm in (16), although the norm may potentially be unduly influenced by one or a few components. Ideally, the screen weights should be scale-invariant and most, if not all, components should contribute to the screening in defining cliques. To reflect this consideration and improve finite sample performance, we modify (16) as follows:

$$w_{1k} = \mathcal{K}\left(\frac{\|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k\|_{(\hat{\Sigma}_1+\hat{\Sigma}_k)^{-1}}}{\tau_{\bar{n}_{1k}} b\cdot(\bar{n}_{1k}p)^{1/2}}\right) \Big/ \mathcal{K}(0), \qquad (24)$$

where $\|\mathbf{x}-\mathbf{y}\|_S = \sqrt{(\mathbf{x}-\mathbf{y})^t S(\mathbf{x}-\mathbf{y})}$ is the Mahalanobis distance w.r.t. matrix S and $\bar{n}_{1k}$ is the geometric average sample size of n_1 and n_k. Here, $\hat{\Sigma}_1$ and $\hat{\Sigma}_k$ are the variances of individual CDs for Individual-1 and k, respectively. Note that: (i) this modified version has the same asymptotically behavior to that of (16) so all the claims in Lemmas 2 and 3 apply without further modification; (ii) if $\boldsymbol{\theta}_k \in \mathcal{C}_1$, then $(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k)^t(\hat{\Sigma}_1+\hat{\Sigma}_k)^{-1}(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k)$ follows approximately a chi-squared distribution of $2p$ degrees of freedom, making the quantity inside $\mathcal{K}(\cdot)$ in (24) more stable than that in (16).

We propose the following cross-validation algorithm to empirically select the constant b in (24):

1. For each $k = 1,\ldots,K$, randomly split the data $\mathcal{S}_k$ into V equally sized folds $\{\mathcal{S}_k^1,\ldots,\mathcal{S}_k^V\}$. Denote by $\mathcal{S}_k^{-v} = \mathcal{S}_k/\mathcal{S}_k^v$ the subset data that exclude $\mathcal{S}_k^v$, for $v = 1,\ldots,V$.
2. For a given b, let $\hat{\boldsymbol{\theta}}_1^{(c)}(b,v)$ be the combined estimator from applying *i*Fusion to $\{\mathcal{S}_1^{-v},\mathcal{S}_2^{-v},\ldots,\mathcal{S}_K^{-v}\}$ with $b_n = \tau_{\bar{n}_{1k}} b$ included in the calculation of w_{1k}.
3. Compute the loss of $\hat{\boldsymbol{\theta}}_1^{(c)}(b,v)$ using subset data $\mathcal{S}_1^v$, denoted by $\mathbb{L}(b,v)$. For example, in Simulation I in Section 6, we use the average quadratic loss $\mathbb{L}(b,v) = \frac{1}{|\mathcal{S}_1^v|}\sum_{Y_{i1}\in\mathcal{S}_1^v}\{Y_{i1} - \hat{\boldsymbol{\theta}}_1^{(c)}(b,v)\}^2$.
4. Repeat Steps 2 and 3 for $v = 1,\ldots,V$. Compute the average loss over the V folds, $\bar{\mathbb{L}}(b) = \frac{1}{V}\sum_{v=1}^V \mathbb{L}(b,v)$, and the standard deviation of $\{\mathbb{L}(b,1),\ldots,\mathbb{L}(b,V)\}$, denoted by $\mathrm{std}(\mathbb{L}(b))$.
5. Repeat Steps 2 to 4 along a path of b (denoted by $\mathcal{P}$). Let $b^* = \arg\min_{b\in\mathcal{P}} \bar{\mathbb{L}}(b)$. Choose b as $b^{\mathrm{cv}} = \mathrm{median}\{b : \bar{\mathbb{L}}(b) \leq \bar{\mathbb{L}}(b^*) + \frac{c}{\sqrt{V}}\cdot \mathrm{std}(\mathbb{L}(b^*)), b \in \mathcal{P}\}$, for some $c \geq 0$.

In Step 5, rather than the global minimizer b^*, we choose the median of the b's which corresponds to a loss no greater than the minimum by one standard error of it, up to a constant multiplier c that can accommodate the inherent randomness in $\bar{\mathbb{L}}(b^*)$. Empirically, we have used $c = 1$ in our numerical study and found this cross-validation method performs reasonably well under various settings. This particular choice of c is similar to the one standard error rule that has been widely used in cross-validation (see, e.g., Hastie, Tibshirani, and Friedman 2001). In principle, b can be different for different studies. This difference can also be empirically accommodated by using a scaled $a_k b$. For example, $a_k = \hat{\sigma}_k/\hat{\sigma}_1$ and $\hat{\sigma}_k$ is an estimated variance of $\hat{\theta}_k$ in the univariate case, for $k = 1,\ldots,n$.

Obviously, the tuning algorithm introduces extra computational cost, but the algorithm can be accelerated by a number of strategies, especially for large data.

Strategy I. In Step 1, when K is huge, a quick prescreen can be carried out using the ranks of $\{\|\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_1\|_2\}_{k=1}^K$, since the computation of l_2 norms can be vectorized in most programmings, and thus quickly done. In contrast, the Mahalanobis distance involved in (24) requires inverting a matrix and has to be computed for each individual data source. Denote the ranks of $\{\|\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_1\|_2\}_{k=1}^K$ by $\{u_k\}_{k=1}^K$. We set $w_{1k} = 0$ if $u_k > u^*$ for a prespecified $u^* \in \{1,\ldots,K\}$. As a result, only a portion of the individuals can proceed to the next steps. The choice of u^* can depend on both the number of individual subjects in the project and the available computing resources.

Strategy II. In Step 5, it is often not necessary to search the full path $\mathcal{P}$. Loss functions such as the empirical quadratic loss are typically bowl-shaped (with noise) as a function of b, due to the bias-variance tradeoff. Hence, we may begin with some small b, then gradually increase it until the loss stops decreasing. Specifically, let $b^{m*} = \arg\min_{1,\dots,m} \bar{\mathbb{L}}(b_m)$ that corresponds to the running minimum average loss by the mth value in $\mathcal{P}$. Stop the search if $\bar{\mathbb{L}}(b)$ exceeds $\bar{\mathbb{L}}(b^{m*}) + \frac{\varepsilon}{\sqrt{V}} \cdot \text{std}(\mathbb{L}(b^{m*}))$ for consecutively rounds, and then choose $b^{\text{cv}} = \text{median}\{b_{m'} : \bar{\mathbb{L}}(b_{m'}) \le \bar{\mathbb{L}}(b^{m*}) + \frac{\varepsilon}{\sqrt{V}} \cdot \text{std}(\mathbb{L}(b^{m*})), m' \le m\}$.

Strategy III. The design of this algorithm, together with the framework of *i*Fusion, easily allows implementing *i*Fusion in a distributed fashion and is thus particularly suited for the case that individual datasets are stored in different computer clusters. In this case, a central coordinator will (i) collect the individual confidence density functions that are independently computed using $\mathcal{S}_k^{-v}$ on each cluster, (ii) compute a combined estimator $\hat{\theta}_1^{(c)}(b, v)$ according to some choice of b and return it to each cluster, and (iii) tally the losses with combined $\hat{\theta}_1^{(c)}(b, v)$ that are again independent evaluated on $\mathcal{S}_k^{v}$ on each computer. Repeating (i), (ii), and (iii) through different (b, v), the algorithm can scale up to big dataset that are too large for storage or processing in a single computer.

Finally, in real applications, different components of the parameter vector $\boldsymbol{\theta}_i$ may have different interpretations, scales, or units. We address these differences by enhancing the distance measure $\|\cdot\|_2$ in (16) or $\|\cdot\|_S$ in (24) to reflect the component-wise differences. Alternatively, we can consider using a kernel function on each component, $\prod_{l=1}^{p} \mathcal{K}\left(\frac{\hat{\theta}_{1l} - \hat{\theta}_{kl}}{b_{nl}}\right)$, with different element-wise bandwidths b_{nl}, although tuning multiple bandwidths will require extra computing efforts.

6. Simulation Studies

This section shows the simulation studies under three different settings: (I) with some subgroup structures, (II) where subgroup analysis is not applicable, and (III) with a heterogeneous study design considered in Section 4. We compare results from *i*Fusion, the oracle approach (assuming the clique is known), and other competing methods such as the commonly used combination after clustering (in Simulation I) and NPB method (in Simulation II).

Simulation I. We generate random data: $Y_{ik} \sim N(\theta_k, 1)$, for $i = 1, \dots, n_k$, $k = 1, \dots, 9$, where θ_k assumes values as follows: (i) $\theta_k = 0$ for $k = 1, 2, 3$; this forms a clique with equal parameter values. (ii) $\theta_k = d + U_k/n_k$ for $k = 4, 5, 6$, where $U_k \overset{\text{iid}}{\sim} U[-1, 1]$; this forms a clique according to (8) but with varying parameter values. (iii) $\theta_k = (k - 5)d$ for $k = 7, 8, 9$. Here, d is proportional to the minimum distance between the parameters that are, respectively, inside and outside a clique, as defined in (12). In this simulation, we set $d = 3n_k^{-1/6}$.

In the individual approach, a CD for θ_k is $N(\hat{\theta}_k, \hat{\sigma}_k^2)$, with a point estimate $\hat{\theta}_k = \bar{Y}_{\cdot k} = \sum_{i=1}^{n_k} Y_{ik}/n_k$ and a $(1 - \alpha)$ asymptotic confidence interval $\hat{\theta}_k \pm z_{\alpha/2}\hat{\sigma}_k$. Here, $\hat{\sigma}_k^2 = \sum_{i=1}^{n_k}(Y_{ik} - \bar{Y}_{\cdot k})^2/(n_k - 1)$ and $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. We run *i*Fusion using these CDs from individuals for each θ_k, with screen weights tuned according to the numerical setup in column 1 of Table 1. The *i*Fusion method with (4) then yield a point estimate $\hat{\theta}_k^{(c)}$ and an $(1 - \alpha)$ asymptotic confidence interval $\hat{\theta}_k^{(c)} \pm z_{\alpha/2}\hat{\sigma}_k^{(c)}$, where $(\hat{\sigma}_k^{(c)})^2 = (\sum_{k=1}^{K} w_{1k}^2 \hat{\sigma}_k^{-2})/(\sum_{k=1}^{K} w_{1k}\hat{\sigma}_k^{-2})^2$. The oracle approach is also performed on each θ_k, where the screen weights match the membership of the clique. For example, $\boldsymbol{w}_{1,1:9} = (1, 1, 1, 0, 0, 0, 0, 0, 0)$ for targeting individual-1, and $\boldsymbol{w}_{8,1:9} = (0, 0, 0, 0, 0, 0, 0, 1, 0)$ for individual-8.

Table 1. Numeric setting of the *i*Fusion tuning algorithm in simulation studies.

	Simulation I	Simulation II	Simulation III
Bandwidth search path	$\{0.1, 0.2, \dots, 5\}$	$\{0.1, 0.2, \dots, 5\}$	$\{0.1, 0.2, \dots, 5\}$
CV-folds	5	5	5
Early stopping rounds	5	5	5
ϵ	0.5	0.5	0.5
Kernel	Uniform	Uniform	Uniform
Loss	l_2	l_2	l_2
Prescreen survival rate	–	1%	–

We repeat the simulation 500 times for $n_k = 40$ and $n_k = 400$ to represent moderate and large sample sizes, respectively. We compare the performance of the traditional method of using only individual data, the proposed *i*Fusion method and the oracle method described in Section 3 as the benchmark for comparison. We also include a modified *i*Fusion method that incorporates a bootstrap calibration to improve the finite-sample performance, especially when sample size is only moderate. More details of this calibration is provided in Section A.6 of Appendix A.

Table 2 reports MSE of the point estimate, empirical coverage probability and average length of the nominal 95% confidence interval obtained by these four methods. When the clique has size greater than one (individuals-1–6), *i*Fusion always returns point estimates with significantly reduced MSE, confidence intervals which are narrower but still retain approximately the desired coverage probabilities. When the sample size is large $n_k = 400$, the results from *i*Fusion and the oracle approach are the same, thus support the claims in Theorem 1. The coverage probabilities from *i*Fusion, under moderate $n_k = 40$, are slightly lower than the individual and oracle approaches. This is expected, due to additional uncertainty in the screen weights, but the results using the calibrated *i*Fusion method show that it can effectively overcome the potential under-coverage issue for small/moderate sample size cases. Finally, for an individual with a clique size one by itself (individuals-7–9), no information can be borrowed from their neighbors. In this case, all three approaches yield similar or same results, so *i*Fusion does not alter the inference when there is no clique to borrow information from.

Table 2 also provides a comparison with a popular subgroup analysis approach. To implement this method, we first use k-means clustering on $\hat{\theta}_k^{(c)}$ to divide the individuals into J groups/clusters, and use the pooled data within each cluster to make inference for all individuals within the cluster. The number of clusters need be determined in advance and $J = 4, 5, 6$ are used in our experiments, where $J = 5$ is the true number of subgroups. For individuals-, 1–6, the subgroup analysis approach works okay and only slightly worse for $k = 6$.

Table 2. Simulation I results—MSE of point estimates, empirical coverage, and average length of 95% confidence intervals.

		$n_k = 40$							$n_k = 400$						
		Indiv	*i*Fusion	*i*Fusion^c	Oracle	4-Sub	5-Sub	6-Sub	Indiv	*i*Fusion	*i*Fusion^c	Oracle	4-Sub	5-Sub	6-Sub
MSE	θ_1	0.025	0.012	0.012	0.008	0.009	0.013	0.019	0.003	0.001	0.001	0.001	0.001	0.001	0.002
	θ_2	0.026	0.011	0.011	0.008	0.009	0.014	0.020	0.003	0.001	0.001	0.001	0.001	0.001	0.002
	θ_3	0.023	0.010	0.010	0.008	0.009	0.012	0.018	0.003	0.001	0.001	0.001	0.001	0.001	0.002
	θ_4	0.023	0.011	0.011	0.008	0.016	0.013	0.015	0.002	0.001	0.001	0.001	0.001	0.001	0.002
	θ_5	0.025	0.012	0.012	0.008	0.016	0.012	0.018	0.003	0.001	0.001	0.001	0.001	0.001	0.002
	θ_6	0.023	0.011	0.011	0.008	0.016	0.012	0.016	0.002	0.001	0.001	0.001	0.001	0.001	0.002
	θ_7	0.025	0.028	0.028	0.025	0.387	0.157	0.072	0.002	0.002	0.002	0.002	0.153	0.060	0.031
	θ_8	0.026	0.027	0.027	0.026	0.678	0.291	0.111	0.002	0.002	0.002	0.002	0.318	0.110	0.047
	θ_9	0.026	0.026	0.026	0.026	0.332	0.158	0.058	0.002	0.002	0.002	0.002	0.148	0.047	0.019
Coverage	θ_1	0.942	0.928	0.948	0.944	0.942	0.910	0.912	0.946	0.950	0.954	0.950	0.950	0.938	0.922
	θ_2	0.940	0.928	0.950	0.944	0.942	0.910	0.898	0.946	0.950	0.954	0.950	0.950	0.938	0.918
	θ_3	0.942	0.930	0.950	0.944	0.942	0.926	0.924	0.938	0.948	0.952	0.950	0.950	0.948	0.926
	θ_4	0.958	0.936	0.952	0.956	0.910	0.936	0.936	0.946	0.952	0.958	0.952	0.950	0.942	0.916
	θ_5	0.932	0.944	0.954	0.960	0.914	0.940	0.910	0.940	0.950	0.954	0.950	0.948	0.932	0.924
	θ_6	0.954	0.942	0.956	0.956	0.910	0.936	0.916	0.954	0.952	0.958	0.952	0.950	0.940	0.914
	θ_7	0.944	0.940	0.950	0.944	0.480	0.768	0.878	0.946	0.946	0.952	0.946	0.462	0.766	0.854
	θ_8	0.916	0.906	0.932	0.916	0.042	0.558	0.800	0.948	0.948	0.952	0.948	0.000	0.626	0.812
	θ_9	0.944	0.944	0.952	0.944	0.478	0.748	0.894	0.950	0.950	0.960	0.950	0.486	0.810	0.902
Length	θ_1	0.613	0.354	0.383	0.351	0.351	0.384	0.441	0.196	0.113	0.116	0.113	0.113	0.123	0.141
	θ_2	0.616	0.355	0.383	0.351	0.351	0.388	0.440	0.196	0.113	0.116	0.113	0.113	0.122	0.138
	θ_3	0.618	0.356	0.384	0.351	0.351	0.387	0.438	0.196	0.113	0.116	0.113	0.113	0.122	0.141
	θ_4	0.618	0.353	0.382	0.351	0.348	0.370	0.419	0.196	0.113	0.116	0.113	0.113	0.120	0.137
	θ_5	0.614	0.351	0.380	0.351	0.348	0.372	0.426	0.196	0.113	0.116	0.113	0.113	0.120	0.139
	θ_6	0.616	0.351	0.380	0.351	0.348	0.369	0.420	0.196	0.113	0.116	0.113	0.113	0.118	0.134
	θ_7	0.619	0.618	0.649	0.619	0.517	0.582	0.605	0.196	0.196	0.200	0.196	0.166	0.185	0.190
	θ_8	0.610	0.599	0.632	0.610	0.439	0.539	0.587	0.196	0.196	0.200	0.196	0.138	0.176	0.187
	θ_9	0.617	0.617	0.648	0.617	0.525	0.578	0.607	0.196	0.196	0.201	0.196	0.167	0.187	0.193

NOTE: *i*Fusion^c indicates that bootstrap calibration is applied to the raw *i*Fusion confidence intervals. The subgroup approach uses *k*-means clustering to divide the individuals into J ($J = 4$, 5, or 6) subgroups and then combines individual confidence densities within each subgroup. The "implied" number of subgroups in this example is 5.

For individuals-7–9, all with no groups, the subgroup performs significantly worse. This is because the clustering algorithm sometimes incorrectly groups, say individual-8, with other individuals even when the correct J is used, thus leads to overly aggressive inference.

Simulation II. We generate 6000 datasets according to the regression model

$$Y_{ik} = \alpha_k + \beta_k x_{ik} + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0, 1),$$
$$\text{for } i = 1, \ldots, n_k \text{ and } k = 1, \ldots, 6000, \qquad (25)$$

where the true parameter values of $\{\boldsymbol{\theta}_k = (\alpha_k, \beta_k)^t, k = 1, \ldots, 6000\}$ are spreaded along a circle of radius $R = 500$. Specifically, the 6000 $\boldsymbol{\theta}$ values are obtained by (i) generating 1200 points evenly distributed along the circle $\{(\alpha, \beta) : \alpha^2 + \beta^2 = R^2\}$, (ii) replicating each point four times to obtain 6000 points in total, and (iii) adding to each point a small random perturbation. More precisely, the true $(\alpha_k, \beta_k) = (500\cos(\left\lfloor \frac{k-1}{5} \right\rfloor \frac{2\pi}{1200}) + \frac{U_{k1}}{n}, 500\sin(\left\lfloor \frac{k-1}{5} \right\rfloor \frac{2\pi}{1200}) + \frac{U_{k2}}{n})$, where $U_{kj} \overset{\text{iid}}{\sim} U[-1, 1]$, for $j = 1, 2$ and $k = 1, \ldots, 6000$. The setup suggests a clique of size five for each individual in a circular structure, where a subgroup analysis is generally not applicable. Finally, we simulate x_{ik} independently from $N(0, 1.5^2)$ and then Y_{ik} from (25).

For the kth individual regression, $N(\hat{\boldsymbol{\theta}}_k, \hat{\sigma}_k^2 (X_k^t X_k)^{-1})$ is an asymptotic CD for $\boldsymbol{\theta}_k$, where $\hat{\boldsymbol{\theta}}_k$ is the least square estimate of $\boldsymbol{\theta}_k$, X_k the design matrix and $\hat{\sigma}_k^2$ a consistent estimate of σ_k^2. The *i*Fusion and oracle approaches then follow similarly as in Simulation I, except that, as discussed in Section 5, a prescreen procedure is applied to *i*Fusion to exclude irrelevant individual datasets.

We calculate marginal coverage and length of the confidence intervals for α_k and β_k separately. The results on the coverage probability of the confidence region for $\boldsymbol{\theta}_k$ are similar and thus omitted. Table 3 reports the summary results for individuals-1500, 3000, and 4500, as representatives of the entire 6000 individuals. It compares the performance of individual-data method, the *i*Fusion and bootstrap calibrated *i*Fusion methods, and the oracle method, with 500 repeated simulations, again for $n_k = 40$ and 400. In all cases, *i*Fusion returns out point estimates with significantly smaller MSE and lengths of confidence intervals than the individual approach. It is close to the oracle approach under moderate sample size, and yields almost exactly the same results under large sample size. Overall, these numerical studies demonstrate well our theoretical claims under the setting of multivariate parameters and big data.

We also carry out a NPB approach to make individualized inference about $\boldsymbol{\theta}_k$. We use the `DPlmm` function in the R package `DPpackage` by Jara, Hanson, and Quintana (2011). The function estimates a linear mixed-effects model with a Dirichlet process mixture prior for the distribution of the random effects, and is suitable here, as both regression intercept and slope are treated as random effects. In each random simulation, the MCMC samples for the target parameter can be extracted to compute posterior mean and credible interval. Their frequentist properties can be then examined against the true value of $\boldsymbol{\theta}_k$ based on 500 simulations. However, this NPB approach is time-consuming even for a single random simulation, because it

Table 3. Simulation II results—MSE of point estimates, empirical coverage, and average length of 95% confidence intervals.

		$n_k = 40$					$n_k = 400$				
		Indiv	*i*Fusion	*i*Fusion^c	Oracle	NPB	Indiv	*i*Fusion	*i*Fusion^c	Oracle	NPB
MSE	α_{1500}	0.025	0.007	0.007	0.005	0.005	0.002	0.0005	0.0005	0.0005	–
	β_{1500}	0.017	0.004	0.004	0.002	0.002	0.001	0.0002	0.0002	0.0002	–
	α_{3000}	0.031	0.009	0.009	0.007	0.006	0.003	0.0005	0.0005	0.0005	–
	β_{3000}	0.011	0.004	0.004	0.002	0.002	0.001	0.0002	0.0002	0.0002	–
	α_{4500}	0.026	0.007	0.007	0.005	0.005	0.002	0.0005	0.0005	0.0005	–
	β_{4500}	0.021	0.006	0.006	0.003	0.003	0.001	0.0002	0.0002	0.0002	–
Coverage	α_{1500}	0.952	0.934	0.954	0.938	0.938	0.970	0.950	0.954	0.950	–
	β_{1500}	0.932	0.910	0.944	0.926	0.944	0.974	0.966	0.970	0.966	–
	α_{3000}	0.920	0.908	0.940	0.916	0.932	0.942	0.942	0.950	0.942	–
	β_{3000}	0.948	0.924	0.942	0.934	0.942	0.946	0.954	0.960	0.954	–
	α_{4500}	0.920	0.926	0.952	0.938	0.954	0.958	0.938	0.958	0.938	–
	β_{4500}	0.928	0.916	0.944	0.934	0.944	0.910	0.950	0.956	0.950	–
Length	α_{1500}	0.628	0.282	0.310	0.272	0.276	0.196	0.088	0.090	0.088	–
	β_{1500}	0.472	0.183	0.201	0.175	0.178	0.131	0.058	0.059	0.058	–
	α_{3000}	0.630	0.290	0.322	0.276	0.282	0.196	0.088	0.090	0.088	–
	β_{3000}	0.415	0.187	0.207	0.177	0.182	0.125	0.058	0.059	0.058	–
	α_{4500}	0.608	0.282	0.311	0.269	0.276	0.197	0.088	0.090	0.088	–
	β_{4500}	0.523	0.221	0.244	0.210	0.216	0.127	0.059	0.060	0.059	–

NOTE: *i*Fusion^c indicates that bootstrap calibration is applied to the raw *i*Fusion confidence intervals. The nonparametric Bayesian (NPB) approach is applied on a subset of individual datasets that have survived the *i*Fusion prescreen procedure. The case of $n_k = 400$ is not run for the NPB approach due to computational limit.

simultaneously estimates all the individual parameters rather than just a specific target individual parameter. The computing is unattainable in our computing environment (2000 MCMC iterations for a single random run; 2018 MacBook Pro with a 2.3 GHz Intel Core i5 processor). As a compromise, we restrict the analysis to a subset data with only 30 neighboring individuals for $n_k = 40$. (The analysis for $n_k = 400$ is terminated as the computing would seem to last forever.) In each random simulation, the last 1000 of the total 2000 MCMC samples are used to compute posterior means and credit intervals. (Despite this much reduced sample size, it still takes around 15 sec for a single run; in comparison, *i*Fusion less than a second for the same run.) As for the performance, the NBP approach works as well as the oracle approach and even slightly outperforms in terms of the coverage probability, noting that the oracle approach used in our simulation uses asymptotic formulas. To produce outputs comparable to the *i*Fusion and oracle approaches, the NPB approach will impose a huge burden in computing time and data storage.

Simulation III. To study the performance of *i*Fusion under a heterogeneous design described in Section 4, we generate $K = 4$ regression datasets from (20) with the following setup: In each regression, x_{ik} is 1 or 0 with equal probability, z_{ik} assumes three levels: 1, 2, 5, and each level is assigned with roughly $n_k/3$ observations. The regression parameters are $\alpha_1 = -1 + U_{11}/n_k, \alpha_2 = U_{21}/n_k, \alpha_3 = 1+U_{31}/n_k, \alpha_4 = 2+U_{41}/n_k, \beta_1 = 1 + U_{12}/n_k, \beta_1 = 1 + U_{22}/n_k, \beta_3 = 1 + U_{32}/n_k, \beta_4 = -1 + U_{42}/n_k, \gamma_1 = -1 + U_{13}/n_k, \gamma_2 = -1 + U_{23}/n_k, \gamma_3 = -1 + U_{33}/n_k, \gamma_1 = -1 + U_{43}/n_k$, where $U_{kj} \overset{\text{iid}}{\sim} U[-1, 1]$ for $k = 1, \ldots, 4$ and $j = 1, 2, 3$. The configuration follows the Scenario I of Example 1, where (β_k, γ_k) are approximately the same, up to a constant of order $O(1/n_k)$, for $k = 1, 2, 3$. The cliques are defined based on (β_k, γ_k) but not α_k. Individual-1 and individual-4 are our targets of interest, one for demonstrating the efficiency and validity of *i*Fusion when $|\tilde{\mathcal{C}}_1| = 3$ and the other for $|\tilde{\mathcal{C}}_4| = 1$. Also, we set $\sigma_k \equiv 1$ and let $n_k = 40$ or 400. For the oracle approach, we set $\boldsymbol{w}_{1,1:4} = (1, 1, 1, 0)$ and $\boldsymbol{w}_{4,1:4} = (0, 0, 0, 1)$.

Table 4. Simulation III results—MSE of point estimates, empirical coverage, and average length of 95% confidence intervals.

		$n_k = 40$				$n_k = 400$			
		Indiv	*i*Fusion	*i*Fusion^c	Oracle	Indiv	*i*Fusion	*i*Fusion^c	Oracle
MSE	α_1	0.105	0.055	0.055	0.052	0.013	0.006	0.006	0.006
	β_1	0.117	0.045	0.045	0.038	0.010	0.004	0.004	0.004
	γ_1	0.008	0.003	0.003	0.003	0.001	0.0003	0.0003	0.0003
	α_4	0.109	0.109	0.109	0.109	0.010	0.010	0.010	0.010
	β_4	0.109	0.109	0.109	0.109	0.010	0.010	0.010	0.010
	γ_4	0.007	0.007	0.007	0.007	0.001	0.001	0.001	0.001
Coverage	α_1	0.934	0.946	0.954	0.950	0.938	0.952	0.954	0.952
	β_1	0.940	0.934	0.946	0.938	0.944	0.942	0.952	0.942
	γ_1	0.934	0.944	0.958	0.946	0.960	0.944	0.952	0.944
	α_4	0.948	0.948	0.968	0.948	0.948	0.948	0.954	0.948
	β_4	0.942	0.942	0.960	0.942	0.958	0.958	0.964	0.958
	γ_4	0.958	0.958	0.968	0.958	0.950	0.950	0.950	0.950
Length	α_1	1.248	0.908	0.988	0.901	0.432	0.297	0.306	0.297
	β_1	1.319	0.748	0.814	0.735	0.392	0.226	0.233	0.226
	γ_1	0.346	0.213	0.232	0.210	0.113	0.066	0.068	0.066
	α_4	1.347	1.347	1.452	1.347	0.415	0.415	0.427	0.415
	β_4	1.244	1.244	1.341	1.244	0.391	0.391	0.402	0.391
	γ_4	0.348	0.348	0.375	0.348	0.116	0.116	0.119	0.116

NOTE: *i*Fusion^c indicates that bootstrap calibration is applied to the raw *i*Fusion confidence intervals.

Table 4 reports the summary statistics of MSEs, coverage probabilities and lengths of confidence intervals, all based on 500 repeated random simulations. For individual-1 where $|\mathcal{C}_1| = 3$, it shows that *i*Fusion outperforms the individual approach in two aspects. First, *i*Fusion is more efficient in making inference for β_1 and γ_1, achieving smaller MSEs and length of confidence intervals. In fact, *i*Fusion is approximately oracle. These observations agree with Theorem 3. The second, and a more intriguing, result is the inference on α_1: the MSE of the point estimator from *i*Fusion is much smaller than that from the individual approach, even though α_1 is not shared by other

α_k's. This clearly highlights the power of fusing learning. This also supports numerically the claim in Corollary 1. It appears that the improvement in estimating β_1 and γ_1 by *i*Fusion is channeled to bring about improvement in estimating α_1. Given that individual-4 forms a clique by itself, all three approaches obtain the same result as expected.

7. Real Data Example

Fama–French model is a widely used model to describe portfolio returns in asset pricing and portfolio management (Fama and French 1993). A Fama–French three-factor model for the kth portfolio over time $t = 1, \ldots, T$ is

$$r_{tk} = \alpha_k + \beta_{\text{mkt},k} r_{t,\text{mkt}} + \beta_{\text{smb},k} r_{t,\text{smb}} + \beta_{\text{hml},k} r_{t,\text{hml}} + \varepsilon_{tk}, \quad \text{for } k = 1, \ldots, K. \tag{26}$$

Here, r_{tk} is the excessive return on the kth portfolio over the risk-free rate at time t; $r_{t,\text{mkt}}$ is the excessive return on the market portfolio; $r_{t,\text{smb}}$ ("small minus big") is the return on a portfolio long small-capitalization stocks and short large-capitalization stocks; $r_{t,\text{hml}}$ ("high minus low") is the return on a portfolio long high book-to-price stocks and short low book-to-price stocks (i.e., value stocks vs. growth stocks). These are calculated with combinations of portfolios composed by ranked stocks and available historical market data. Additionally, the idiosyncratic errors ε_{tk} are serially uncorrelated and homoscedastic; α_k is known as Jensen's alpha and may account for any market inefficiency and friction. Due to its strong performance across multiple markets, the Fama–French model and its variants have enjoyed popularity in finance applications (see Fama and French 1993, 2012, 2014; Cakici, Fabozzi, and Tan 2013).

In this section, we analyze daily price returns in the year of 2016 for individual stock in Russell 3000 Index using the Fama–French three-factor model, and compare the *i*Fusion method with the individual approach, with each stock being an individual subject. The stocks in the index cover 3000 largest publicly held companies in United States as measured by total market capitalization, and represents approximately 98% of the American public equity market. In our analysis, the prices of each individual stock are obtained from Yahoo Finance from 2016/01/01 to 2016/12/31, and Fama–French factors as well as the risk-free rate for the same period are downloaded from Kenneth French's website. Furthermore, we narrow down our set of stocks for study by excluding those with absolute daily returns greater than 30% in any single day of the year. This helps us exclude potential data errors or idiosyncratic issues such as stock split/reverse split and focus on methodologies. We end up with 2558 such eligible stocks. Different from simulations, the underlying parameter values are unknown in real data analysis. It is impractical to use the same performance metrics such as MSE and coverage as we did in Section 6. Instead, we compare the forecasting ability on out-of-sample data via rolling prediction. The idea is that the more efficiently a model we can estimate, the more accurate forecasts we can expect from using the model.

In our analysis, we first obtain the least squares estimation of (26) based only on each individual stock data and given a fixed tick (time) window size of 60. We choose 60 heuristically corresponding the number of trading days in roughly three months. Then, for a target stock (say, Stock-k), *i*Fusion is applied to obtain a combined estimate of the model parameters. Using the combined estimate, we obtain the h-day forward factors and the h-day forward excessive return of the target stock based on the model in (26). We roll the window h day forward starting from day-tick $60 + h$ and repeat the same steps, until reaching the end of entire time period. For the target stock and each rolling window, the forecasted stock excessive returns and their realized/observed values are recorded, leading to the rolling mean squared prediction error (RMSPE):

$$\text{RMSPE}_k = \frac{1}{S} \sum_{s=1}^{S} (\hat{r}_{sk} - r_{sk})^2.$$

Here, S is the number of available rolling windows, $\hat{r}_{s,k}$ is the forecasted return, and $r_{s,k}$ is the observed return. The RMSPE based solely on the individual stock is also calculated for comparison. We repeated the computation for each of the target stock $k = 1, \ldots, 2558$.

The h-day forward factors (i.e., the time $t + h$ realized factor returns) are typically unknown by time t, though their values themselves can be estimated. For example, Hu (2003) projected the forward factor returns using their historical marks together with a number of macroeconomic variables. In out setting, the availability of macroeconomic data is limited at daily frequency. A more practical approach is to regress stock excessive returns directly on the time-lagged factors:

$$r_{t+h,k} = \alpha_k + \beta_{\text{mkt},k} r_{t,\text{mkt}} + \beta_{\text{smb},k} r_{t,\text{smb}} + \beta_{\text{hml},k} r_{t,\text{hml}} + \varepsilon_{tk}. \tag{27}$$

In our numerical study, we consider both models (26) and (27). The forecasting using (27) and the computation of RMSPE_k are the same as those using (26), excepted that the prediction is now based on (27) and h-day factors up to the current time.

Figure 2 reports the histogram of relative RMSPEs, that is, the ratio of RMSPE from *i*Fusion over that from the individual stock data approach for every stock, based on one-day forward ($h = 1$) forecasting. The two histograms correspond to settings (26) and (27), respectively, from the left to right. In both settings, *i*Fusion improves prediction accuracy for 99% of all the stocks, with the average reduction in PRMSE by 3%. Note that there is always a random error associated with a future observation and it adds a sizable base to the RMSPE calculation. Thus, the reduction in RMSPE is usually not as much as the reduction in MSE of parameter estimates seen in the simulation examples. Nevertheless, Figure 2 provides a clear evidence that *i*Fusion can help improve inference for forecasting by borrowing information from other relevant stocks.

8. Concluding Remarks and Further Discussions

This article introduces *i*Fusion as a statistical learning approach for making efficient individualized inference by borrowing "sharable" information from relevant individuals (or individual data sources), under both homogeneous and heterogeneous model designs. When there exist moderate number of observations for each individual study and there are similar individuals in the entire data source, *i*Fusion is shown to improve significantly the efficiency of each individual inference, by controlling

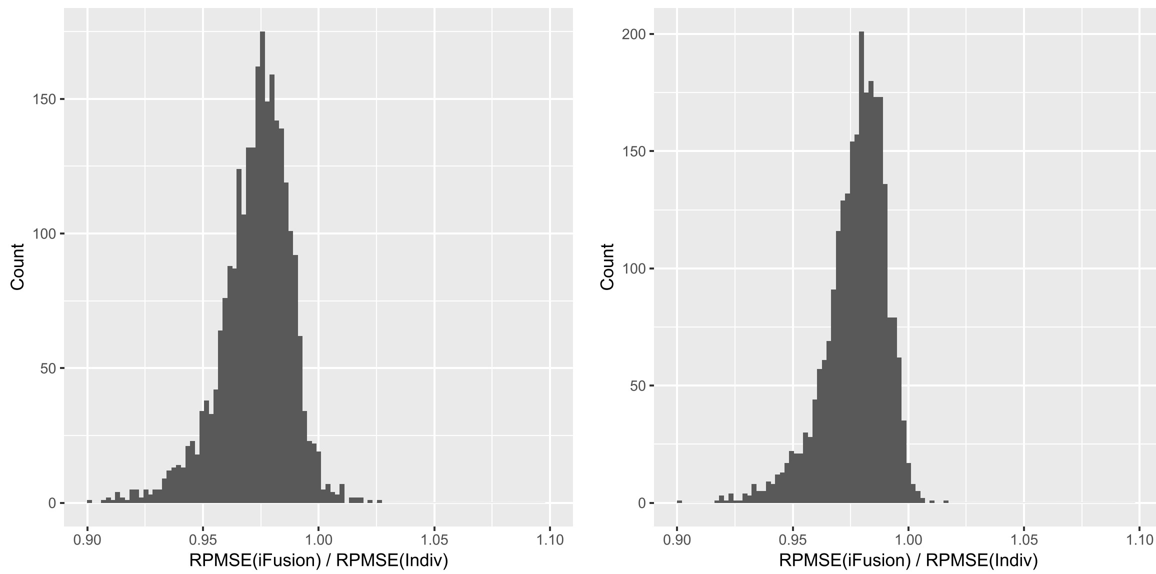

Figure 2. Ratio of one-day-ahead RMSPE from *i*Fusion to the individual approach for 2558 individual stocks under setting (26) (left) and (27) (right), respectively. The RMSPEs by *i*Fusion versus the conventional analysis of using only individual stock data are less than 1 for more than 99% of the 2558 stocks, with median reductions around 3%; It indicates that *i*Fusion method can often improve the forecasting accuracy by a meaningful amount by incorporating information from other relevant stocks.

the bias while reducing variance. Otherwise, *i*Fusion achieves the same efficiency as the individual inference based on the individual data source. Furthermore, under suitable conditions, the *i*Fusion method can achieve the oracle property to have the best asymptotic efficiency afforded by the entire data source.

In using CD as the inference tool in our development, *i*Fusion naturally inherits many desirable properties from CD. In particular, the validity of the combined CD, in terms of providing appropriate frequentist inference, relies solely on the individual CDs, regardless how they are obtained (Singh, Xie, and Strawderman 2005; Xie and Singh 2013), for example, through likelihood methods, or other frequentist, fiducial, or Bayesian approaches. Such a feature affords *i*Fusion with great versatility and applicability to combining individual CDs even if they are derived from different paradigms. More important, in many settings, point or interval estimates are undefined or unavailable, but CDs as distribution estimates remain available and can be used to carry out the inference. For example, using CDs Liu, Liu, and Xie (2014) carried out a fusion learning of multiple clinical studies that include some zero-total events where point estimates of odds-ratio are not well-defined but CDs are. This further attests to the broad applicability of *i*Fusion.

Another desirable feature of *i*Fusion is its scalability to big data applications. Developed under the frequentist framework, *i*Fusion allows the construction of confidence density functions independently for each individual, without being burdened by other individuals and any nuisance or less relevant information. This agrees with the so-called "division of labor" feature described in Efron (1986) and Wasserman (2007). Efron (1986) and Wasserman (2007) observed that in the Bayesian approach "statistical problems need to be solved as one coherent whole, including assigning priors and conducting analyses with nuisance parameters," while a frequentist approach can focus directly on the target parameter without estimating nuisance parameters. Compared to the full Bayesian approach which requires running a large-scale simulation using an MCMC algorithm, the "divide-and-conquer" nature of *i*Fusion makes it scale better to big data settings.

Given the availability of (asymptotic) confidence density functions for the individuals under consideration, *i*Fusion applies to a general inference framework that covers a wide range of problems. Although the numerical examples in Sections 6 and 7 only demonstrate the effectiveness of *i*Fusion for simple linear models, we stress that *i*Fusion is readily applicable to more complex models such as time series models, survival models, and high-dimensional models. Consider for instance a set of high-dimensional linear regressions corresponding to multiple individual subjects or data sources. Here, asymptotic confidence densities for the individual regression coefficients can be obtained by the de-biased lasso procedure (see Javanmard and Montanari 2014; van de Geer et al. 2014; Zhang and Zhang 2014), and then the combined estimate and inference for a target individual can be obtained through *i*Fusion. This procedure naturally extends the divide-and-conquer strategies for high-dimensional regression with multiple datasets (see Chen and Xie 2014; Kleiner et al. 2014; Battet et al. 2015; Tang, Zhou, and Song 2016) from the perspective of an overall inference for all data to individualized inference.

Among many goal-directed applications, *i*Fusion is ideally suited for *precision medicine*. Precision medicine tailors medical treatments to each individual patient rather than a treatment for the "average" or subgroup of patients. In the latter, patients are divided into subgroups by one or few baseline characteristics and subsequent analysis is conducted within the subgroup (see, e.g., Wang et al. 2007). This partitioning of patients has natural interpretations and seems perfectly logical, but it lacks statistical guarantees for the combined inference of model parameters within the subgroup. In comparison, *i*Fusion makes inference directly on the parameter space with statistical justifications. It may be worthwhile to consider combining the two procedures with ways to retain the merits of both and gain even more efficiency. One possibility is to partition the individuals into different subgroups according to their features, and then apply *i*Fusion within subgroups. It is important to note that *i*Fusion is different from the machine-learning-based methods used in precision medicine to assign an individualized treatment to each patient (see, e.g., Murphy 2005; Qian and Murphy 2011; Zhao et al. 2011; Goldberg and Kosorok 2012; Tian et al. 2014). Rather *i*Fusion can provide an effective alternative and compliment to the machine learning approach to provide and improve inference for treatment-decision for individual patient.

Although the theoretical development in the article is illustrated using asymptotic normal CDs, the *i*Fusion approach can be applied directly to the case where the CD obtained in some studies are nonnormal. In this case, most theoretical results (e.g., consistency and reduction in MSE) still hold under some mild conditions; see Singh, Xie, and Strawderman (2005) and Xie, Singh, and Strawderman (2011) for discussions on weighted combining of nonnormal CDs. Also, the use of adaptive screening weights in *i*Fusion is similar to those used in Hu and Zidek (2002) and Wang and Zidek (2005) in the context of weighted likelihood and also those in robust meta-analysis development in Xie, Singh, and Strawderman (2011) and Claggett, Xie, and Tian (2014). However, here *i*Fusion focused on screening out studies that are different from the target individual. In addition, since a (normalized) likelihood is often a CD function (Xie and Singh 2013), the weighted likelihood with weights tailored to the target individual can be viewed as a special case of *i*Fusion developed in this article and the *i*Fusion development covers broader cases than likelihood procedures including CDs obtained from quasi-likelihood inference methods, *p*-value functions and even a Bayesian or fiducial inference procedure.

So far, *i*Fusion in the article is developed under the asymptotic setting that $n_k/n \to r_k$ for some constant $r_k \in (0,1)$ and K is large but finite. The development can possibly be extended to the case with $n_k \to \infty$ and $K \to \infty$ in principle, although some notations and conditions in Sections 2–4 may need to be strengthen to accommodate $K \to \infty$. The development cannot be extended to the case that each individual study has only one or a limited number observations with $n_k = O(1)$. To form a clique and borrow information from other individuals in this case, a stringent assumption such as "dense assumption" that there are infinite many individuals in a small neighborhood of the target individual is needed. As a result, a different development related to empirical Bayes methods (e.g., Zhang 2003) can be utilized. A separate research is currently underway.

Finally, we comment on the choice of kernel functions when implementing the kernel screen weights. Generally, *i*Fusion is still applicable if we use functions other than uniform, such as (i) Epanechnikov kernel $\frac{3}{4}(1-u^2)\mathbb{1}\{|u| \le 1\}$; (ii) quartic kernel $\frac{15}{16}(1-u^2)^2\mathbb{1}\{|u| \le 1\}$; (iii) Gaussian kernel $\frac{1}{\sqrt{2\pi}}e^{-u^2/2}$. However, to achieve the same convergence rate of w_{1k} as required in Lemma 2, stronger regularity conditions on b_n may be needed for some of these kernel function. For instance, if $\mathcal{K}(\cdot)$ is the Epanechnikov or the quartic kernel, then w_{1k} given by (16) satisfies (7) if $b_n/d_1 \to 0$ and $n^{1/4}b_n \to \infty$. As for the Gaussian kernel, the regularity condition becomes $(b_n/d_1)^2 \log n \to 0$ and $n^{1/4}b_n \to \infty$. Note that both conditions are stronger than that for the uniform kernel (see (17)). Our empirical observations indicate that, with finite sample size, the uniform kernel is more effective than the others, which also agrees with the rate discussion above.

Appendix A

A.1. Proof of Lemma 1

(i) We begin by showing that $\mathbb{P}\left(n^\alpha \|\hat{\boldsymbol\theta}_1^{(o)} - \boldsymbol\theta_1\|_2 \ge \varepsilon\right) \to 0$ for any $\alpha \in (0,1/2)$ and $\varepsilon > 0$. Define $\boldsymbol\theta_1^{(o)} = (\sum_{\boldsymbol\theta_k \in \mathcal{C}_1} \hat\Sigma_k^{-1})^{-1} \sum_{\boldsymbol\theta_k \in \mathcal{C}_1} \hat\Sigma_k^{-1}\boldsymbol\theta_k$. Then

$$\boldsymbol\theta_1^{(o)} - \boldsymbol\theta_1 = \left(\sum_{\boldsymbol\theta_k \in \mathcal{C}_1} \hat\Sigma_k^{-1}\right)^{-1} \sum_{\boldsymbol\theta_k \in \mathcal{C}_1} \hat\Sigma_k^{-1}(\boldsymbol\theta_k - \boldsymbol\theta_1) = \boldsymbol{o}_p(n^{-1/2}). \tag{A.1}$$

Note that each $\hat{\boldsymbol\theta}_k$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol\theta_k$. It then follows that

$$\begin{aligned}\mathbb{P}\left(n^\alpha \|\hat{\boldsymbol\theta}_1^{(o)} - \boldsymbol\theta_1\|_2 \ge \varepsilon\right) &\le \mathbb{P}\left(n^\alpha \|\hat{\boldsymbol\theta}_1^{(o)} - \boldsymbol\theta_1^{(o)}\|_2 \ge \varepsilon/2\right) \\ &\quad + \mathbb{P}\left(n^{1/2} \|\boldsymbol\theta_1^{(o)} - \boldsymbol\theta_1\|_2 \ge \varepsilon/2\right) \\ &\le \mathbb{P}\left(O_p(n^{\alpha-1/2}) \ge \varepsilon/2\right) \\ &\quad + \mathbb{P}\left(o_p(1) \ge \varepsilon/2\right) \to 0.\end{aligned}$$

(ii) Let $n^{1/2}(\hat{\boldsymbol\theta}_1^{(o)} - \boldsymbol\theta_1) = n^{1/2}(\hat{\boldsymbol\theta}_1^{(o)} - \boldsymbol\theta_1^{(o)}) + n^{1/2}(\boldsymbol\theta_1^{(o)} - \boldsymbol\theta_1)$. The first term, $n^{1/2}(\hat{\boldsymbol\theta}_1^{(o)} - \boldsymbol\theta_1^{(o)}) \xrightarrow{d} N(\mathbf{0}, n(\sum_{\boldsymbol\theta_k \in \mathcal{C}_1} \hat\Sigma_k^{-1})^{-1})$, where $\lim_{n\to\infty} n(\sum_{\boldsymbol\theta_k \in \mathcal{C}_1} \hat\Sigma_k^{-1})^{-1} = \Delta_1^{(o)}$, and the second, $n^{1/2}(\boldsymbol\theta_1^{(o)} - \boldsymbol\theta_1) = \boldsymbol{o}_P(1)$ following (A.1), altogether leading to $n^{1/2}(\hat{\boldsymbol\theta}_1^{(o)} - \boldsymbol\theta_1) \xrightarrow{d} N(\mathbf{0}, \Delta_1^{(o)})$.

(iii) Following simple calculations,

$$\begin{aligned}\mathrm{MSE}(\hat{\boldsymbol\theta}_1^{\mathcal{F}}) = &\sum_{\boldsymbol\theta_{k_1}, \boldsymbol\theta_{k_2} \in \mathcal{F}} (\boldsymbol\theta_{k_1} - \boldsymbol\theta_1)^t \hat\Sigma_{k_1}^{-1} \left(\sum_{\boldsymbol\theta_k \in \mathcal{F}} \hat\Sigma_k^{-1}\right)^{-2} \\ &\times \hat\Sigma_{k_2}^{-1}(\boldsymbol\theta_{k_2} - \boldsymbol\theta_1) + \mathrm{tr}\left\{\left(\sum_{\boldsymbol\theta_k \in \mathcal{F}} \hat\Sigma_k^{-1}\right)^{-1}\right\}.\end{aligned} \tag{A.2}$$

If $\mathcal{F} \subseteq \mathcal{C}_1$, then the first term on the right hand side of (A.2) is of order $o(n^{-1})$ and is dominated by the trace term, thus $\mathrm{MSE}(\hat{\boldsymbol\theta}_1^{\mathcal{F}}) = O(n^{-1})$. On the other hand, if any $\boldsymbol\theta_k \notin \mathcal{C}_1$ (i.e., $\boldsymbol\theta_k \in \mathcal{D}_1$, since $\mathcal{B}_1 = \emptyset$) is included in $\mathcal{F}$, then $\mathrm{MSE}(\hat{\boldsymbol\theta}_1^{\mathcal{F}})$ is dominated by the first term, and $n\mathrm{MSE}(\hat{\boldsymbol\theta}_1^{\mathcal{F}}) \to \infty$. Thus, the MSE-optimal $\mathcal{F}$ should be a subset of $\mathcal{C}_1$. Now, because $\mathrm{tr}\{(A+B)^{-1}\} \le \mathrm{tr}\{A^{-1}\}$ for any two positive definite matrices A and B, $\mathrm{tr}\left\{(\sum_{\boldsymbol\theta_k \in \mathcal{C}_1} \hat\Sigma_k^{-1})^{-1}\right\} \le \mathrm{tr}\left\{(\sum_{\boldsymbol\theta_k \in \mathcal{F}} \hat\Sigma_k^{-1})^{-1}\right\}$ for $\forall \mathcal{F} \subseteq \mathcal{C}_1$. Therefore, the choice of $\mathcal{F} = \mathcal{C}_1$ affords the smallest asymptotic MSE for $\hat{\boldsymbol\theta}_1^{(o)}$ among all estimators in the form of $\hat{\boldsymbol\theta}_1^{\mathcal{F}}$.

A.2. Proof of Theorem 1

(i) Define

$$\boldsymbol\theta_1^{(c)} = \left(\sum_{k=1}^K w_{1k}\hat\Sigma_k^{-1}\right)^{-1} \sum_{k=1}^K w_{1k}\hat\Sigma_k^{-1}\boldsymbol\theta_k. \tag{A.3}$$

Then, $\boldsymbol{\theta}_1^{(c)} - \boldsymbol{\theta}_1 = (\sum_{k=1}^K w_{1k}\hat{\Sigma}_k^{-1})^{-1}\sum_{k=1}^K w_{1k}\hat{\Sigma}_k^{-1}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_1) = (\sum_{k=1}^K w_{1k}\hat{\Sigma}_k^{-1})^{-1}\Big(\sum_{\boldsymbol{\theta}_k\notin\mathcal{C}_1} w_{1k}\hat{\Sigma}_k^{-1}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_1) + \sum_{\boldsymbol{\theta}_k\in\mathcal{C}_1} w_{1k}\hat{\Sigma}_k^{-1}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_1)\Big) = \Big(\sum_{k=1}^K w_{1k}(n\hat{\Sigma}_k^{-1})\Big)^{-1}\Big(\sum_{\boldsymbol{\theta}_k\notin\mathcal{C}_1} o_p(n^{-1/2})(n\hat{\Sigma}_k^{-1})(\boldsymbol{\theta}_k - \boldsymbol{\theta}_1) + \sum_{\boldsymbol{\theta}_k\in\mathcal{C}_1}(1+o_p(n^{-1/2}))(n\hat{\Sigma}_k^{-1})\boldsymbol{o}_p(n^{-1/2})\Big) = \boldsymbol{o}_p(n^{-1/2})$. Since $\hat{\boldsymbol{\theta}}_1^{(c)}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}_1^{(c)}$, we have, for any $\alpha \in (0, 1/2)$ and $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}\Big(n^\alpha\|\hat{\boldsymbol{\theta}}_1^{(c)} - \boldsymbol{\theta}_1\|_2 \ge \varepsilon\Big) &\le \mathbb{P}\Big(n^\alpha\|\hat{\boldsymbol{\theta}}_1^{(c)} - \boldsymbol{\theta}_1^{(c)}\|_2 \ge \varepsilon/2\Big) \\ &\quad+\mathbb{P}\Big(n^{1/2}\|\boldsymbol{\theta}_1^{(c)} - \boldsymbol{\theta}_1\|_2 \ge \varepsilon/2\Big) \\ &\le \mathbb{P}\Big(O_p(n^{\alpha-1/2}) \ge \varepsilon/2\Big) \\ &\quad+\mathbb{P}\Big(o_p(1) \ge \varepsilon/2\Big) \to 0.\end{aligned}$$

(ii) Using similar proof as that for part (ii) of Lemma 1, we obtain

$$n^{1/2}(\hat{\boldsymbol{\theta}}_1^{(c)} - \boldsymbol{\theta}_1^{(c)}) \xrightarrow{d} N\left(\mathbf{0}, n\left(\sum_{k=1}^K w_{1k}\hat{\boldsymbol{\Sigma}}_k^{-1}\right)^{-1}\left(\sum_{k=1}^K w_{1k}^2\hat{\boldsymbol{\Sigma}}_k^{-1}\right) \times\left(\sum_{k=1}^K w_{1k}\hat{\boldsymbol{\Sigma}}_k^{-1}\right)^{-1}\right),$$

where the covariance matrix converges to $\Delta_1^{(o)}$ in probability, and $n^{1/2}(\boldsymbol{\theta}_1^{(c)} - \boldsymbol{\theta}_1) = \boldsymbol{o}_p(1)$.

(iii) Note that, asymptotically, $\mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{(c)}) = \mathrm{tr}\Big\{\mathrm{var}(\hat{\boldsymbol{\theta}}_1^{(c)})\Big\}$ and $\mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{(o)}) = \mathrm{tr}\Big\{\mathrm{var}(\hat{\boldsymbol{\theta}}_1^{(o)})\Big\}$. Note also that part ii) of Theorem 1 shows that $\hat{\boldsymbol{\theta}}_1^{(c)}$ and $\hat{\boldsymbol{\theta}}_1^{(o)}$ have the same limiting covariance matrix. Therefore, $\mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{(c)}) = \mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{(o)})$, asymptotically.

A.3. Proof of Lemma 2

If $\boldsymbol{\theta}_k \notin \mathcal{C}_1$, $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k\|_2 \ge d_1$. Then, for any $\varepsilon > 0$ and b_n satisfying (17), as $n \to \infty$,

$$\begin{aligned}&\mathbb{P}\Big(n^{1/2}\mathbb{1}\{\|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k\|_2/b_n \le 1\} \le \varepsilon\Big) \\ &= \mathbb{P}\Big(\|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k\|_2/b_n > 1\Big) \\ &= \mathbb{P}\Big((\|(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) + (\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k) - (\boldsymbol{\theta}_k - \boldsymbol{\theta}_1)\|)/b_n \ge 1\Big) \\ &\ge \mathbb{P}\Big((\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_1\|_2 - \|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\|_2 - \|\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_h\|_2)/b_n \ge 1\Big) \\ &\ge \mathbb{P}\Big(\big(1 - \frac{O(n^{-1/2})}{d_1}\big)\frac{d_1}{b_n} \ge 1\Big) \to 1.\end{aligned}$$

If $\boldsymbol{\theta}_k \in \mathcal{C}_1$, $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k\|_2 = o(n^{-1/2})$. Then, for any $\forall\varepsilon > 0$ and b_n satisfying (17), as $n \to \infty$,

$$\begin{aligned}&\mathbb{P}\Big(n^{1/2}|\mathbb{1}\{\|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k\|_2/b_n \le 1\} - 1| \le \varepsilon\Big) \\ &= \mathbb{P}\Big(\|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k\|_2/b_n \le 1\Big) \\ &= \mathbb{P}\Big(\|(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) + (\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k) - (\boldsymbol{\theta}_k - \boldsymbol{\theta}_1)\|_2/b_n \le 1\Big) \\ &\ge \mathbb{P}\Big((\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\|_2 + \|\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k\|_2 + \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_1\|_2)/b_n \le 1\Big) \\ &= \mathbb{P}\Big(\frac{O(n^{-1/2})}{b_n} \le 1\Big) \to 1.\end{aligned}$$

A.4. Proof of Theorem 2

We only prove part (iii) here, since parts (i) and (ii) can be proved following the same arguments in the proof of Theorem 1. Define $\boldsymbol{\theta}^{(c)}$ as (A.3). For large n, $\hat{\boldsymbol{\theta}}_1^{(c)} \approx (\sum_{\boldsymbol{\theta}_k\notin\mathcal{D}_1}\hat{\Sigma}_k^{-1})^{-1}\sum_{\boldsymbol{\theta}_k\notin\mathcal{D}_1}\hat{\Sigma}_k^{-1}\hat{\boldsymbol{\theta}}_k$ and $\boldsymbol{\theta}_1^{(c)} \approx (\sum_{\boldsymbol{\theta}_k\notin\mathcal{D}_1}\hat{\Sigma}_k^{-1})^{-1}\sum_{\boldsymbol{\theta}_k\notin\mathcal{D}_1}\hat{\Sigma}_k^{-1}\boldsymbol{\theta}_k$. Then, asymptotically,

$$\begin{aligned}\mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{(c)}) &= (\boldsymbol{\theta}_1^{(c)} - \boldsymbol{\theta}_1)^t(\boldsymbol{\theta}_1^{(c)} - \boldsymbol{\theta}_1) + \mathrm{tr}\Big\{\mathrm{var}(\hat{\boldsymbol{\theta}}_1^{(c)})\Big\} \\ &= \sum_{\boldsymbol{\theta}_{k_1},\boldsymbol{\theta}_{k_2}\notin\mathcal{D}_1}(\boldsymbol{\theta}_{k_1} - \boldsymbol{\theta}_1)^t\hat{\Sigma}_{k_1}^{-1}\left(\sum_{\boldsymbol{\theta}_k\notin\mathcal{D}_1}\hat{\Sigma}_k^{-1}\right)^{-2} \\ &\quad\times\hat{\Sigma}_{k_2}^{-1}(\boldsymbol{\theta}_{k_2} - \boldsymbol{\theta}_1) + \mathrm{tr}\left\{\left(\sum_{\boldsymbol{\theta}_k\notin\mathcal{D}_1}\hat{\Sigma}_k^{-1}\right)^{-1}\right\} \\ &= \sum_{\boldsymbol{\theta}_{k_1},\boldsymbol{\theta}_{k_2}\in\mathcal{B}_1}(\boldsymbol{\theta}_{k_1} - \boldsymbol{\theta}_1)^t\hat{\Sigma}_{k_1}^{-1}\left(\sum_{\boldsymbol{\theta}_k\notin\mathcal{D}_1}\hat{\Sigma}_k^{-1}\right)^{-2} \\ &\quad\times\hat{\Sigma}_{k_2}^{-1}(\boldsymbol{\theta}_{k_2} - \boldsymbol{\theta}_1) + \mathrm{tr}\left\{\left(\sum_{\boldsymbol{\theta}_k\notin\mathcal{D}_1}\hat{\Sigma}_k^{-1}\right)^{-1}\right\}.\end{aligned}$$

The last equation holds because if $\boldsymbol{\theta}_{k_1}$ or $\boldsymbol{\theta}_{k_2} \in \mathcal{C}_1$, then the squared bias vanishes as $n \to \infty$.

Now, for any $\mathcal{F} \subseteq \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$, if $\mathcal{D}^{\mathcal{F}} \ne \emptyset$, then (A.2) implies $n\mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{\mathcal{F}}) \to \infty$. Since $\mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{(c)}) = O(n^{-1})$, $\mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{(c)}) < \mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{\mathcal{F}})$, asymptotically. If $\mathcal{D}^{\mathcal{F}} = \emptyset$, then (A.2) implies

$$\begin{aligned}\mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{\mathcal{F}}) &= \sum_{\boldsymbol{\theta}_{k_1},\boldsymbol{\theta}_{k_2}\in\mathcal{C}^{\mathcal{F}}\cup\mathcal{B}^{\mathcal{F}}}(\boldsymbol{\theta}_{k_1} - \boldsymbol{\theta}_1)^t\hat{\Sigma}_{k_1}^{-1}\left(\sum_{\boldsymbol{\theta}_k\in\mathcal{F}}\hat{\Sigma}_k^{-1}\right)^{-2} \\ &\quad\times\hat{\Sigma}_{k_2}^{-1}(\boldsymbol{\theta}_{k_2} - \boldsymbol{\theta}_1) + \mathrm{tr}\left\{\left(\sum_{\boldsymbol{\theta}_k\in\mathcal{F}}\hat{\Sigma}_k^{-1}\right)^{-1}\right\} \\ &= \sum_{\boldsymbol{\theta}_{k_1},\boldsymbol{\theta}_{k_2}\in\mathcal{B}^{\mathcal{F}}}(\boldsymbol{\theta}_{k_1} - \boldsymbol{\theta}_1)^t\hat{\Sigma}_{k_1}^{-1}\left(\sum_{\boldsymbol{\theta}_k\in\mathcal{F}}\hat{\Sigma}_k^{-1}\right)^{-2} \\ &\quad\times\hat{\Sigma}_{k_2}^{-1}(\boldsymbol{\theta}_{k_2} - \boldsymbol{\theta}_1) + \mathrm{tr}\left\{\left(\sum_{\boldsymbol{\theta}_k\in\mathcal{F}}\hat{\Sigma}_k^{-1}\right)^{-1}\right\}.\end{aligned}$$

This shows the asymptotic equivalence between $\mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{(c)}) \le \mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{\mathcal{F}})$ and (19) if $\mathcal{D}^{\mathcal{F}} = \emptyset$.

A.5. Proof of Corollary 1

Since $\mathrm{var}(\hat{\boldsymbol{\eta}}_1^{(c)}) = (\sum_{\boldsymbol{\xi}_k\in\tilde{\mathcal{C}}_1} A_k^t\hat{\Sigma}_k^{-1}A_k)^{-1}$ asymptotically, it suffices to show that $\{(\sum_{\boldsymbol{\xi}_k\in\tilde{\mathcal{C}}_1} A_k^t\hat{\Sigma}_k^{-1}A_k)^{-1}\}_{1,1} \le \{\hat{\Sigma}_1\}_{1,1}$. For simplicity, we assume further, without loss of generality, that $K = 2$ and ψ_2 is a scalar as well. If $\boldsymbol{\xi}_2 \notin \tilde{\mathcal{C}}_1$ then the equality holds. If $\boldsymbol{\xi}_2 \in \tilde{\mathcal{C}}_1$, we need to show $\{(A_1^t\hat{\Sigma}_1^{-1}A_1 + A_2^t\hat{\Sigma}_2^{-1}A_2)^{-1}\}_{1,1} \le \{\hat{\Sigma}_1\}_{1,1}$. Partition $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$, respectively, as

$$\hat{\Sigma}_1 = \begin{pmatrix} a_1 & \boldsymbol{b}_1^t \\ \boldsymbol{b}_1 & C_1\end{pmatrix}, \quad \hat{\Sigma}_2 = \begin{pmatrix} a_2 & \boldsymbol{b}_2^t \\ \boldsymbol{b}_2 & C_2\end{pmatrix},$$

where C_1 and C_2 are $q \times q$ matrices. By definition,

$$A_1 = \begin{pmatrix} 1 & 0 & \mathbf{0}^t_{q\times 1} \\ \mathbf{0}^t_{q\times 1} & \mathbf{0}^t_{q\times 1} & I_q \end{pmatrix},$$
$$A_2 = \begin{pmatrix} 0 & 1 & \mathbf{0}^t_{q\times 1} \\ \mathbf{0}^t_{(q\times 1} & \mathbf{0}^t_{q\times 1} & I_q \end{pmatrix},$$

where I_q is an identity matrix of size $q \times q$. Some linear algebra with blockwise matrix inversion formula gives

$$\{(A_1^t \hat{\Sigma}_1^{-1} A_1 + A_2^t \hat{\Sigma}_2^{-1} A_2)^{-1}\}_{1,1} = a_1 - \boldsymbol{b}_1^t C_1^{-1} \boldsymbol{b}_1 + \boldsymbol{b}_1^t C_1^{-1} (C_1^{-1} + C_2^{-1})^{-1} C_1^{-1} \boldsymbol{b}_1.$$

Following Lemma A.3 in Liu, Liu, and Xie (2015): for the two $q\times q$ positive definite matrices W_1 and W_2 and $q\times 1$ vector $\boldsymbol{v}$, $\boldsymbol{v}^t(W_1+W_2)^{-1}\boldsymbol{v} \le \boldsymbol{v}W_1^{-1}\boldsymbol{v}$, we then obtain $\{(A_1^t \hat{\Sigma}_1^{-1} A_1 + A_2^t \Lambda_2 \hat{\Sigma}_2^{-1})^{-1}\}_{1,1} \le a_1 = \{\hat{\Sigma}_1\}_{1,1}$.

A.6. Correcting iFusion Confidence Intervals Using Bootstrap

We illustrate the model with scalar θ_k's. For vector parameters, we perform the correction on each dimension. Let $w_{11}, \dots, w_{1K}$ be the screen weights tuned by the algorithm in Section 5. Let $\hat{\theta}_1^{(c)}$ be the corresponding *i*Fusion estimator and $\widehat{\text{std}}(\hat{\theta}_1^{(c)})$ be its estimated standard deviation. For an asymptotic normal individual CD with n is large, an approximate $1-\alpha$ confidence interval of θ_1 is given by $\hat{\theta}_1^{(c)} \pm z_{\alpha/2}\widehat{\text{std}}(\hat{\theta}_1^{(c)})$. This confidence interval may result in coverage probability less than $1-\alpha$ when the sample size is moderate, due to both the approximation and the uncertainty associated with the screen weights. We intend to use a bootstrap calibration to find a constant c_α, $c_\alpha \ge 1$, so that $\hat{\theta}_1^{(c)} \pm c_\alpha z_{\alpha/2}\widehat{\text{std}}(\hat{\theta}_1^{(c)})$ will have a better coverage rate.

We bootstrap each individual dataset $\mathcal{S}_k$ to get a bootstrapped dataset $\mathcal{S}_k^b$, from which we get an individual bootstrap CD. Using the screen weights $w_{11}, \dots, w_{1K}$ obtained from the original data and the individual CDs, we obtain the confidence interval following the same way as described in the article. Repeat the above procedure for B times we obtain B confidence intervals $\hat{\theta}_{1,b}^{(c)} \pm z_{\alpha/2}\widehat{\text{std}}(\hat{\theta}_{1,b}^{(c)})$, $b = 1, \dots, B$. The empirical c_α is chosen to be

$$c_\alpha = \min_c \left\{ c \ge 1 \Big| \frac{1}{B}\sum_{b=1}^{B} 1\{\hat{\theta}_1^{(c)} \in [\hat{\theta}_{1,b}^{(c)} \pm c z_{\alpha/2}\widehat{\text{std}}(\hat{\theta}_{1,b}^{(c)})]\} \ge 1-\alpha \right\}.$$

Funding

The authors gratefully acknowledge the support from the National Science Foundation through grant #DMS151348, #DMS 1737857, #IIS-1741390, and #DMS-1812048. They also thank Professor Lingsong Xue for sharing data. The first author acknowledges the generous graduate support from Rutgers University. The views and opinions expressed in this article are those of the authors and do not necessarily reflect the views of Deustche Bank.

References

Battet, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. (2015), "Distributed Estimation and Inference With Statistical Guarantees," arXiv no. 1509.05457. [13]

Cakici, N., Fabozzi, F. J., and Tan, S. (2013), "Size, Value, and Momentum in Emerging Stock Returns," *Emerging Markets Review*, 16, 46–65. [12]

Chen, X., and Xie, M. (2014), "A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data," *Statistica Sinica*, 24, 1655–1684. [1,13]

Claggett, B., Xie, M., and Tian, L. (2014), "Meta Analysis With Fixed, Unknown, Study-Specific Parameters," *Journal of the American Statistical Association*, 109, 1667–1671. [3,4,14]

Cox, D. R. (2013), "Discussion of 'Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review," *International Statistical Review*, 81, 40–41. [3]

Efron, B. (1986), "Why Isn't Everyone a Bayesian," *The American Statistician*, 40, 262–266. [13]

——— (1993), "Bayes and Likelihood Calculations From Confidence Intervals," *Biometrika*, 80, 3–26. [3]

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. [12]

——— (2012), "Size, Value, and Momentum in International Stock Returns," *Journal of Financial Economics*, 105, 457–472. [12]

——— (2014), "A Five-Factor Asset Pricing Model," *Journal of Financial Economics*, 116, 1–22. [12]

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis* (3rd ed.), Boca Raton, FL: Chapman & Hall/CRC. [2]

Goldberg, Y., and Kosorok, M. R. (2012), "Q-Learning With Censored Data," *The Annals of Statistics*, 40, 529. [14]

Grün, B., and Leisch, F. (2007), "Finite Mixtures of Generalized Linear Regression Models," Tech. Rep., Department of Statistics, University of Munich. [2]

Gustafson, P., Hossain, S., and McCandless, L. (2005), "Innovative Bayesian Methods for Biostatistics and Epidemiology," in *Handbook of Statistics*, Bayesian Thinking, Modeling and Computation (Vol. 25), eds. Dey, D. K. and C. R. Rao, New York: Elsevier, pp. 763–792 [2]

Hall, P., and Miller, H. (2010), "Bootstrap Confidence Intervals and Hypothesis Tests for Extrema of Parameters," *Biometrika*, 97, 881–892. [4]

Hannah, L. A., Blei, D. M., and Powell, W. B. (2011), "Dirichlet Process Mixtures of Generalized Linear Models," *Journal of Machine Learning Research*, 12, 1923–1953. [2]

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, New York: Springer-Verlag. [8]

Hu, F., and Zidek, J. V. (2002), "The Weighted Likelihood," *The Canadian Journal of Statistics*, 30, 347–371. [14]

Hu, Q. (2003), "Forecasting Ability of the Fama and French Three-Factor Model—Implications for Capital Budgeting." [12]

Jara, A., Hanson, T. E., and Quintana, F. A. (2011), "DPpackage: Bayesian Semi- and Nonparametric Modeling in R," *Journal of Statistical Software*, 40, 1–30. [10]

Javanmard, A., and Montanari, A. (2014), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *Journal of Machine Learning Research*, 15, 2869–2909. [13]

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014), "A Scalable Bootstrap for Massive Data," *Journal of the Royal Statistical Society*, Series B, 76, 795–816. [1,13]

Liu, D., Liu, R., and Xie, M. (2014), "Exact Meta-Analysis Approach for Discrete Data and Its Application to 2 × 2 Tables With Rare Events," *Journal of the American Statistical Association*, 109, 1450–1465. [1,13]

——— (2015), "Multivariate Meta-Analysis of Heterogeneous Studies Using Only Summary Statistics: Efficiency and Robustness," *Journal of the American Statistical Association*, 110, 326–340. [1,3,6,7,16]

——— (2017), "Nonparametric Fusion Learning: Synthesize Inferences From Diverse Sources Using Depth Confidence Distribution," Preprint. [1]

Murphy, S. A. (2005), "A Generalization Error for Q-Learning," *Journal of Machine Learning Research*, 6, 1073–1097. [14]

Qian, M., and Murphy, S. A. (2011), "Performance Guarantees for Individualized Treatment Rules," *The Annals of Statistics*, 39, 1180. [14]

Schweder, T., and Hjort, N. (2016), *Confidence, Likelihood and Probability*, Cambridge, UK: Cambridge University Press. [3]

Simmonds, M. C., and Higgins, J. P. T. (2007), "Covariate Heterogeneity in Meta-Analysis: Criteria for Deciding Between Meta-Regression and Individual Patient Data," *Statistics in Medicine*, 26, 2982–2999. [6]

Singh, K., Xie, M., and Strawderman, W. E. (2005), "Combining Information From Independent Sources Through Confidence Distributions," *The Annals of Statistics*, 33, 159–183. [3,13,14]

——— (2007), "Confidence Distribution (CD)—Distribution Estimator of a Parameter," *IMS Lecture Notes-Monograph Series*, 54, 132–150. [3]

Tang, L., Zhou, L., and Song, P. X.-K. (2016), "Method of Divide-and-Combine in Regularised Generalised Linear Models for Big Data", arXiv no. 1611.06208. [1,3,13]

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014), "A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates," *Journal of the American Statistical Association*, 109, 1517–1532. [14]

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166–1202. [13]

van deer Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics (Vol. 3), Cambridge: Cambridge University Press. [4]

Wang, R., Stephen, M., Lagakos, W., Ware, J. H., Hunter, D. J., and Drazen, J. M. (2007), "Reporting of Subgroup Analyses in Clinical Trials," *Statistics in Medicine*, 357, 2189–2194. [13]

Wang, X., and Zidek, J. V. (2005), "Selecting Likelihood Weights by Cross-Validation," *The Annals of Statistics*, 33, 462–500. [14]

Wasserman, L. (2007), "Why Isn't Everyone a Bayesian?," in *The Science of Bradley Efron*, eds. C. N. Morris and R. J. Tibshirani, New York: Springer, pp. 260–261. [13]

Xie, M., and Singh, K. (2013), "Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review," *International Statistical Review*, 81, 3–39. [3,13,14]

Xie, M., Singh, K., and Strawderman, W. E. (2011), "Confidence Distributions and a Unifying Framework for Meta-Analysis," *Journal of the American Statistical Association*, 106, 320–333. [3,14]

Xie, M., Singh, K., and Zhang, C. (2009), "Confidence Intervals for Population Ranks in the Presence of Ties and Near Ties," *Journal of the American Statistical Association*, 104, 775–788. [4]

Yang, G., Liu, D., Liu, R., Xie, M., and Hoaglin, D. (2014), "A Confidence Distribution Approach for an Efficient Network Meta-Analysis," *Statistical Methodology*, 20, 105–125. [1]

Zhang, C. (2003), "Compound Decision Theory and Empirical Bayes Methods," *The Annals of Statistics*, 31, 379–390. [14]

Zhang, C., and Zhang, S. S. (2014), "Confidence Intervals for Low-Dimensional Parameters in High-Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [13]

Zhao, Y., Zeng, D., Socinski, M. A., and Kosorok, M. R. (2011), "Reinforcement Learning Strategies for Clinical Trials in Nonsmall Cell Lung Cancer," *Biometrics*, 67, 1422–1433. [14]

Zhu, X., and Qu, A. (2018), "Cluster Analysis of Longitudinal Profiles With Subgroups," *Electronic Journal of Statistics*, 12, 171–193. [1]