INFERENCE FOR EXTREME VALUE INDEX THROUGH INDIVIDUALIZED FUSION LEARNING

Hongfang Sun[1], Yu Chen[2], Tao Xu[3], and Zhengjun Zhang[4,5,6]

[1] *School of Mathematical Sciences, Ministry of Education Key Laboratory of NSLSCS, Nanjing Normal University*

[2] *School of Public Affairs, University of Science and Technology of China*

[3] *College of Computer and Information Sciences, Fujian Agriculture and Forestry University*

[4] *School of Data Science and Artificial Intelligence, Beijing University of Chinese Medicine*

[5] *School of Economics and Management, University of Chinese Academy of Sciences*

[6] *Department of Statistics, University of Wisconsin*

Supplementary Material

This supplementary material document contains four sections. Section S1 discusses the selection of the screening weights. The proofs of all theoretical results in the article are provided in Section S2. Sections S3 and S4 show the additional simulation study and real data analyses, respectively. Notations are inherited from the main manuscript.

S1 Details on the Screening Weight Selection

S1.1 Kernel-based screening weight

To achieve efficient performance in implementing the iFusion estimation, it is crucial to select appropriate screening weights w_{1j} for $j = 1, \cdots, N$. In this subsection, we present

Corresponding authors: Yu Chen, E-mail:cyu@ustc.edu.cn. Zhengjun Zhang, E-mail: zjz@stat.wisc.edu; ORCID: 0000-0003-2615-1539

an example and verify that it satisfies the corresponding requirement.

Recall that the sets including clique $\mathcal{C}$, boundary and disperse sets $\mathcal{B}, \mathcal{D}$, are designed according to the distances between relevant parameters and the target parameter γ_1, and are mainly introduced for theoretical exposition. Nevertheless, it often appears that the true values of all parameters are unknown, but the corresponding estimators are available. Hence, it is natural to construct a data-driven screening rule based on the estimated distances between the source-specific EVIs and the target EVI. Following the approach of Shen et al. (2020), we propose the following data-driven and kernel-based screening weights

$$w_{1j} = \mathcal{K}\left(\frac{|\widehat{\gamma}_1 - \widehat{\gamma}_j|}{h_n}\right) \Big/ \mathcal{K}(0), \tag{S1.1}$$

where h_n is a bandwidth parameter and $\mathcal{K}(\cdot)$ is a given kernel function. The authors further argued that different choices of kernel functions may require different choices of bandwidths and also result in some change in the finite-sample behaviors of w_{1j}. As suggested by Shen et al. (2020), we use a uniform kernel $\mathcal{K}(\cdot) = \frac{1}{2}\mathrm{I}\{|\cdot| \leq 1\}$ throughout the article. Under the uniform kernel $\mathcal{K}(\cdot) = \frac{1}{2}\mathrm{I}\{|\cdot| \leq 1\}$, Eq.(S1.1) reduces to the hard-threshold form

$$w_{1j} = \mathbf{I}\left(|\widehat{\gamma}_1 - \widehat{\gamma}_j| \leq h_n\right).$$

This screening rule induces a data-driven retained set

$$\mathbb{I}_{\widehat{\mathcal{R}}} := \{j : w_{1j} = 1\} = \{j : |\widehat{\gamma}_1 - \widehat{\gamma}_j| \leq h_n\}.$$

Therefore, the implemented inclusion/exclusion of sources is data-dependent (random) through $\mathbb{I}_{\widehat{\mathcal{R}}}$, while $\mathcal{C}, \mathcal{B}, \mathcal{D}$ remain oracle sets used for theoretical characterization. As for the selection of bandwidth h_n, Proposition S1 states the detailed restrictions for its order.

Proposition S1. *(Shen et al. (2020))*

The screening weights w_{1j} take the form of Eq.(S1.1) with $\mathcal{K}(\cdot) = \frac{1}{2}\mathrm{I}\{|\cdot| \leq 1\}$ and the

bandwidth h_n *satisfies* $h_n/d \to 0$ *and* $k^{1/2}h_n \to \infty$, *where* $d \equiv \min_j\{|\gamma_j - \gamma_1| : \gamma_j \notin \mathcal{C}\}$.

(i) If separation condition $\mathbf{C}(d)$ *holds, i.e.,* $\mathcal{B} = \varnothing$, *then* w_{1j} *satisfies condition* $\mathbf{C}_1(w)$.

(ii) If separation condition $\mathbf{C}(d)$ *does not hold, i.e.,* $\mathcal{B} \neq \varnothing$, *then* w_{1j} *satisfies condition* $\mathbf{C}_2(w)$.

The results (i) and (ii) in Proposition S1 correspond to Lemmas 2 and 3 in Shen et al. (2020), respectively. Notice that, as the separation condition holds or not, the weight w_{1j} has different results due to the expression of d.

S1.2 Bandwidth selection

Considering that the performance of the kernel-based screening weights is impacted by the choice of bandwidth, we adopt a convenient decomposition

$$h_n = hk^{-x} = \tau_k h,$$

where $h = O(1)$ is a constant tuning parameter and $\tau_k = k^{-x}$. In practice, we can determine the admissible range of x according to Proposition S1 and treat the unknown constant h as a tuning parameter that may impact the performance of iFusion under finite sample size. We propose the following cross-validation (CV) algorithm to empirically select the optimal value of h in the bandwidth.

Remark S1. The tuning procedure can be accelerated using standard strategies: (i) a quick prescreen that retains only a small fraction of sources with the smallest preliminary distances $|\widehat{\gamma}_j - \widehat{\gamma}_1|$; (ii) early stopping along $\mathcal{H}$ once the CV loss stops decreasing for several consecutive steps; and (iii) a distributed implementation where source-wise quantities are computed locally and only aggregated statistics are communicated.

Algorithm S1 Cross-validation selection of the bandwidth constant h

Require: Multi-source data $\{\mathcal{S}_j\}_{j=1}^N$, number of folds V, candidate set $\mathcal{H}$, kernel weight bandwidth $h_n = \tau_k h$, tail fraction $q \in (0,1)$.

Ensure: Selected bandwidth constant h^{cv}.

(Data splitting) For each $j = 1, \ldots, N$, randomly split $\mathcal{S}_j$ into V equally sized folds $\{\mathcal{S}_j^1, \ldots, \mathcal{S}_j^V\}$.

For each $v = 1, \ldots, V$, define the training subset $\mathcal{S}_j^{-v} := \mathcal{S}_j \setminus \mathcal{S}_j^v$.

(Threshold selection) For each $v = 1, \ldots, V$, determine the fold-specific threshold from the *target training fold*:

$$u_1^{-v} = \widehat{Q}_{1-q}\big(\mathcal{S}_1^{-v}\big)$$

where $\widehat{Q}_{1-q}(\cdot)$ is the empirical $(1-q)$-quantile; a natural choice is $q = k_1/n_1$.

for $h \in \mathcal{H}$ **do**

  for $v = 1, \ldots, V$ **do**

    (Fit on training folds) Apply iFusion to $\{\mathcal{S}_1^{-v}, \ldots, \mathcal{S}_N^{-v}\}$ using bandwidth $h_n = \tau_k h$ in the screening weights w_{1j}, and obtain $\widehat{\gamma}_1^{(c)}(h, v)$.

    (Validate on target fold) Compute the validation loss on $\mathcal{S}_1^v$:

$$\mathbb{L}(h, v) = \frac{1}{m_{1v}} \sum_{X_i^{(1)} \in \mathcal{S}_1^v} \Big(\log(X_i^{(1)}/u_1^{-v}) - \widehat{\gamma}_1^{(c)}(h, v) \Big)^2 \mathbf{I}(X_i^{(1)} > u_1^{-v})$$

    where $m_{1v} = \sum_{X_i^{(1)} \in \mathcal{S}_1^v} \mathbf{I}(X_i^{(1)} > u_1^{-v})$.

  end for

  Define fold weights $w_v = \dfrac{m_{1v}}{\sum_{r=1}^V m_{1r}}$, $v = 1, \ldots, V$. Compute the weighted mean loss

$$\overline{\mathbb{L}}_w(h) = \sum_{v=1}^V w_v \, \mathbb{L}(h, v).$$

end for

Choose $h^{\mathrm{cv}} = \arg\min_{h \in \mathcal{H}} \overline{\mathbb{L}}_w(h)$. **return** h^{cv}.

S1.3 Finite-sample study

In this subsection, a sensitivity analysis is first conducted to examine how the choice of the bandwidth constant affects the stability of the iFusion estimator, and then a simulation study is conducted to compare the performance of different bandwidth selection methods. The data-generating process, under Fréchet distribution, follows Case I in Section 6.1 of the main manuscript with group sizes $N_{\mathcal{C}} = 5, N_{\mathcal{B}} = 3$, and $N_{\mathcal{D}} = 2$, where balanced samples are considered.

Sensitivity analysis to bandwidth. Figure S1.1 presents finite-sample performance (squared bias, variance, and MSE) as functions of the bandwidth constant h in $h_n = hk^{-x}$. With k fixed, h_n increases monotonically in h and decreases monotonically in x. Therefore, bandwidth controls how aggressively information is borrowed across sources.

The results suggest sensitivity to the choice of h. When h is too small, the screening is overly strict and potentially useful sources are excluded, making iFusion close to individual-source behavior. When h is too large, screening becomes overly permissive and irrelevant sources may be included, which can inflate bias. This mechanism is reflected in the plots: for fixed x, increasing h generally leads to a U-shaped squared bias and a decreasing-then-flattening variance, so MSE decreases first and then rises mildly. The benchmark lines (Class._Hill and iFusion_O) help contextualize this pattern. In the $\mathcal{B} = \varnothing$ setting (top row), iFusion_C can match or even outperform iFusion_O when h is chosen in a suitable intermediate range. In contrast, when $\mathcal{B} \neq \varnothing$ (middle and bottom rows), iFusion_BC tends to have higher bias than iFusion_O, and under very large h, its bias may even exceed that of Class._Hill. On the variance side, however, iFusion_BC can reach a lower variance floor than iFusion_C.

Finite-Sample Comparison of Bandwidth Selection Methods. Moreover, we

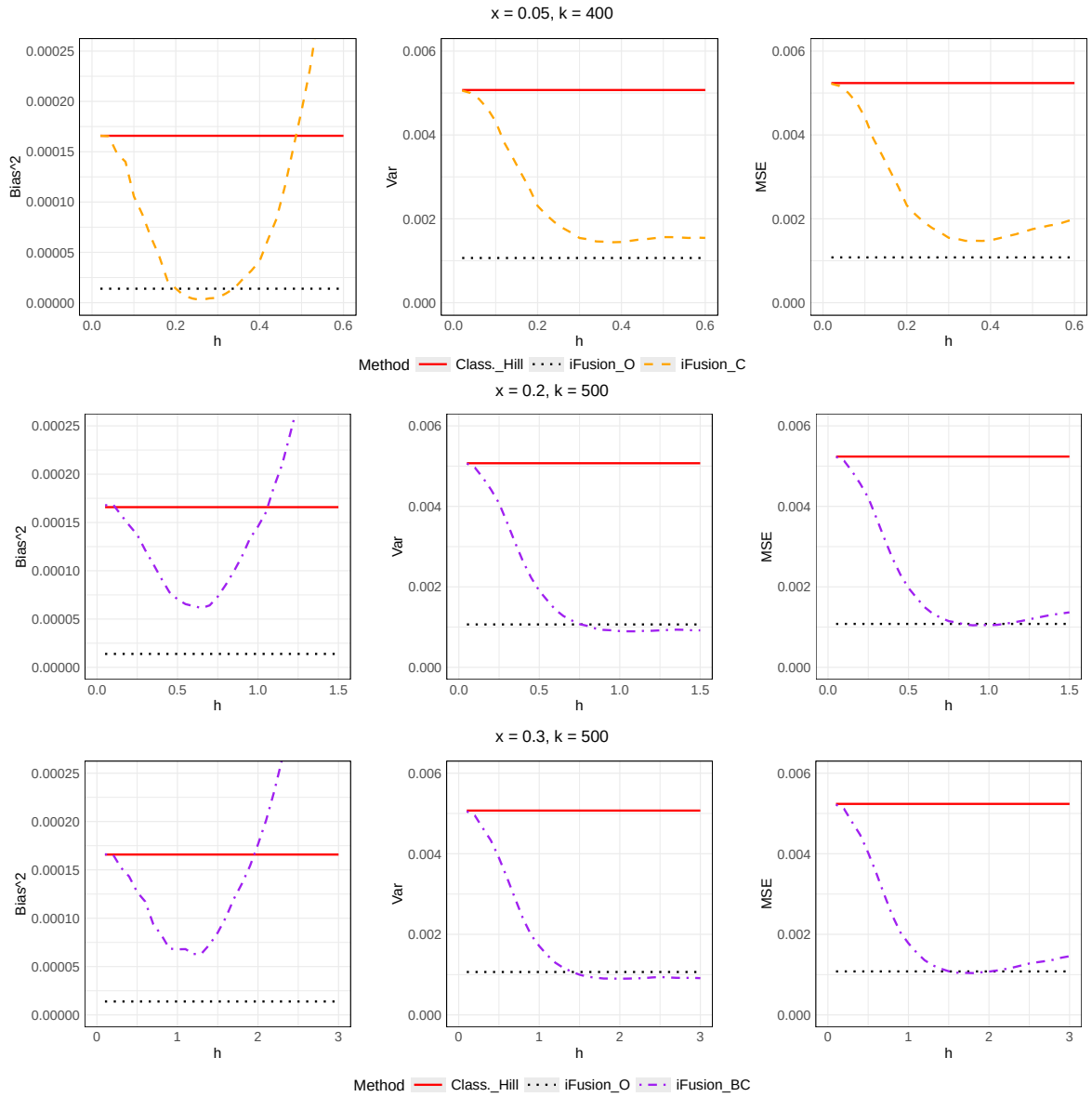

Figure S1.1: Finite-sample performance of EVI estimators for Fréchet distribution ($\gamma_1 = 0.5$) under different levels of h.

include a simulation comparison of bandwidth selection strategies for the iFusion method (iFusion_BC), namely $h_{\text{Full-Min}}$, $h_{\text{CV-Min}}$, and $h_{\text{Oracle-Min}}$. Here, $h_{\text{Full-Min}}$ is an in-sample benchmark obtained by fitting iFusion_BC on the full multi-source dataset and choosing the h that minimizes the empirical loss on the target source. In contrast, $h_{\text{CV-Min}}$ is the data-driven choice proposed in this paper, selected by V-fold cross-validation using training folds for fitting and the held-out target fold for validation, and then minimizing

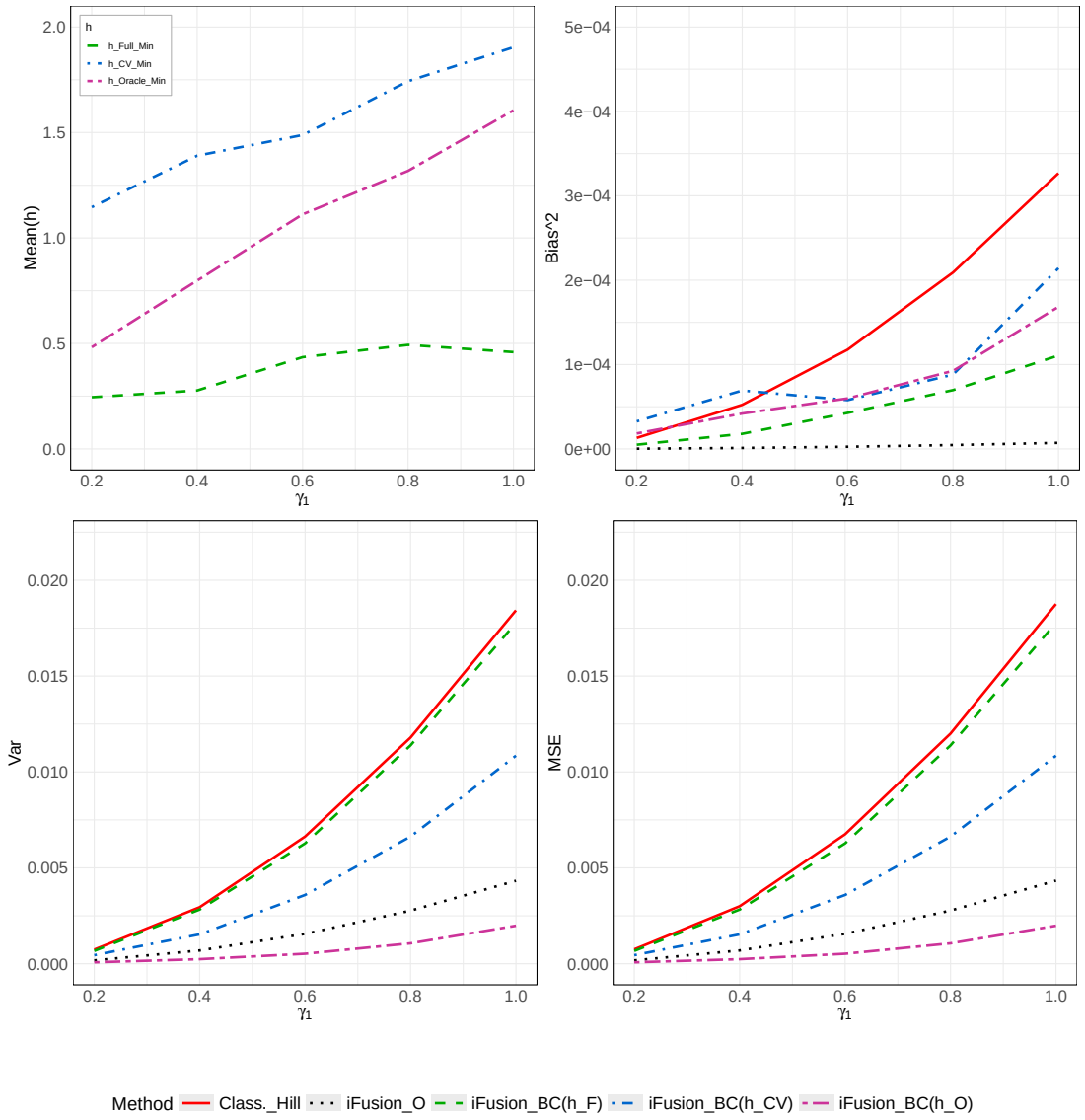

Figure S1.2: Finite-sample performance comparison of bandwidth selection methods for Fréchet distribution.

the aggregated validation loss across folds. Finally, $h_{\text{Oracle-Min}}$ is an oracle benchmark (available only in simulations) that minimizes $(\widehat{\gamma}_1(h) - \gamma_1)^2$ using the true EVI γ_1. It serves as an ideal reference for assessing how close the practical bandwidth selectors are to the optimal choice.

Figure S1.2 reports the mean of h under different selection methods, and the resulting squared bias, Var, and MSE, respectively. We observe that both $h_{\text{CV-Min}}$, and $h_{\text{Oracle-Min}}$

increase with γ_1, whereas $h_{\text{Full-Min}}$ remains relatively small and flat. This indicates that CV-Min and Oracle-Min methods tend to choose broader smoothing under heavier tails, while the Full-Min method is comparatively conservative. In terms of estimation accuracy, $h_{\text{Oracle-Min}}$ gives the best overall performance, $h_{\text{CV-Min}}$ is consistently second-best, and $h_{\text{Full-Min}}$ is the weakest among the three iFusion_BC variants, often yielding variance or MSE comparable to that of Class._Hill. This pattern suggests that overly small bandwidth leads to insufficient borrowing, while CV-based selection substantially improves practical performance and tracks the oracle trend.

Overall, these results confirm that bandwidth selection is important, and that the CV-Min method is an effective and implementable choice in practice when oracle tuning is unavailable.

S2 Proofs of the Main Results

Throughout this section, all limits are taken under condition $\mathbf{C}(n,k)$ in the main manuscript.

Proof of Theorem 1. Denote

$$\gamma_1^{(c)} = \Big(\sum_{j=1}^{N} w_{1j} k_j \widehat{\gamma}_j^{-2}\Big)^{-1} \sum_{j=1}^{N} w_{1j} k_j \widehat{\gamma}_j^{-2} \gamma_j = \sum_{j=1}^{N} \eta_j^{(c)} \gamma_j. \tag{S2.1}$$

Then we expand $|\widehat{\gamma}_1^{(c)} - \gamma_1|$ as

$$\begin{aligned} |\widehat{\gamma}_1^{(c)} - \gamma_1| &= |\big(\widehat{\gamma}_1^{(c)} - \gamma_1^{(c)}\big) + \big(\gamma_1^{(c)} - \gamma_1\big)| \le |\widehat{\gamma}_1^{(c)} - \gamma_1^{(c)}| + |\gamma_1^{(c)} - \gamma_1| \\ &=: \Gamma_1 + \Gamma_2. \end{aligned}$$

In the following, we will show that $\Gamma_1 = o_p(1)$, $\Gamma_2 = o_p(1)$, which then proves the original claim.

Note that under condition $\mathbf{C}_1(\gamma_j)$, each $\widehat{\gamma}_j$ is a consistent estimator of γ_j by the property of Hill estimator of EVI. For Γ_1, which can be rewritten as

$$\Gamma_1 = |\widehat{\gamma}_1^{(c)} - \gamma_1^{(c)}| = \sum_{j=1}^{N} \big|\eta_j^{(c)} \big(\widehat{\gamma}_j - \gamma_j\big)\big| \le \sum_{j=1}^{N} \big|\widehat{\gamma}_j - \gamma_j\big| = o_p(1),$$

where we used $0 \le \eta_j^{(c)} \le 1$ and the fact N is fixed. Now we turn to Γ_2. By the definitions of the sets $\mathcal{C}, \mathcal{B}, \mathcal{D}$ and the screening weights w_{1j}, we have

$$\begin{aligned}\Gamma_2 =& |\gamma_1^{(c)} - \gamma_1| = \sum_{j=1}^{N} \left|\eta_j^{(c)}(\gamma_j - \gamma_1)\right| \\ \le& \sum_{j\in\mathbb{I}_{\mathcal{C}}} \left|\gamma_j - \gamma_1\right| + \sum_{j\in\mathbb{I}_{\mathcal{B}}} \left|\gamma_j - \gamma_1\right| + \sum_{j\in\mathbb{I}_{\mathcal{D}}} \left|\eta_j^{(c)}(\gamma_j - \gamma_1)\right| \\ \le& \, o(k^{-1/2}) + O(k^{-1/2}) + \sum_{j\in\mathbb{I}_{\mathcal{D}}} o_p(k^{-1/2})\left|\gamma_j - \gamma_1\right| \\ =& o_p(1).\end{aligned}$$

Indeed, for $j \in \mathbb{I}_{\mathcal{D}}$, condition $\mathbf{C}_1(w)$ gives $w_{1j} = b_j = o\left(k^{-1/2}\right)$, we have

$$\eta_j^{(c)} = \frac{w_{1j} k_j \hat{\gamma}_j^{-2}}{\sum_{\ell=1}^{N} w_{1\ell} k_\ell \hat{\gamma}_\ell^{-2}} = \frac{b_j O(k) O_p(1)}{O_p(k)} = o_p\left(k^{-1/2}\right), \quad j \in \mathbb{I}_{\mathcal{D}}.$$

Therefore, we complete the proof.

□

Remark S2.1 The same consistency argument also applies under condition $\mathbf{C}_2(w)$, as stated in the corresponding remark in the main manuscript. Under $\mathbf{C}_2(w)$, the clique and boundary sources are retained together, while the disperse sources are still downweighted; hence the only change is in the boundary term, which remains of order $O(k^{-1/2})$ and is therefore negligible for consistency.

Proof of Theorem 2. (i) Defined

$$\gamma_1^{(o)} = \Big(\sum_{j\in\mathbb{I}_{\mathcal{C}}} k_j \widehat{\gamma}_j^{-2}\Big)^{-1} \sum_{j\in\mathbb{I}_{\mathcal{C}}} k_j \widehat{\gamma}_j^{-2} \gamma_j = \sum_{j\in\mathbb{I}_{\mathcal{C}}} \eta_j^{(o)} \gamma_j.$$

Then

$$k^{1/2}(\widehat{\gamma}_1^{(o)} - \gamma_1) = k^{1/2}(\widehat{\gamma}_1^{(o)} - \gamma_1^{(o)}) + k^{1/2}(\gamma_1^{(o)} - \gamma_1).$$

In the first term, $\eta_j^{(o)} \xrightarrow{p} c_j\gamma_j^{-2} / \sum_{j\in\mathbb{I}_{\mathcal{C}}} c_j\gamma_j^{-2} =: \eta_j$. By the asymptotic normality of the Hill estimator and Slutsky's Theorem, we have

$$k^{1/2}\eta_j^{(o)}\big(\widehat{\gamma}_j - \gamma_j\big) \xrightarrow{d} \eta_j c_j^{-1/2} \cdot N\Big(\frac{\lambda_j}{1-\rho_j}, \gamma_j^2\Big),$$

which implies

$$k^{1/2}(\widehat{\gamma}_1^{(o)} - \gamma_1^{(o)}) \xrightarrow{d} N\big(\mu_1^{(o)}, \Delta_1^{(o)}\big),$$

where

$$\mu_1^{(o)} = \sum_{j\in\mathbb{I}_{\mathcal{C}}} \eta_j c_j^{-1/2}\Big(\frac{\lambda_j}{1-\rho_j}\Big) = \big(\sum_{j\in\mathbb{I}_{\mathcal{C}}} c_j\gamma_j^{-2}\big)^{-1} \sum_{j\in\mathbb{I}_{\mathcal{C}}} c_j^{1/2}\gamma_j^{-2}\Big(\frac{\lambda_j}{1-\rho_j}\Big),$$

$$\Delta_1^{(o)} = \sum_{j\in\mathbb{I}_{\mathcal{C}}} \eta_j^2 c_j^{-1}\gamma_j^2 = \big(\sum_{j\in\mathbb{I}_{\mathcal{C}}} c_j\gamma_j^{-2}\big)^{-2} \sum_{j\in\mathbb{I}_{\mathcal{C}}} c_j^2\gamma_j^{-4}\cdot c_j^{-1}\gamma_j^2 = \big(\sum_{j\in\mathbb{I}_{\mathcal{C}}} c_j\gamma_j^{-2}\big)^{-1}.$$

For the second term, $k^{1/2}(\gamma_1^{(o)} - \gamma_1) = o_p(1)$, which leads to

$$k^{1/2}(\widehat{\gamma}_1^{(o)} - \gamma_1) \xrightarrow{d} N\big(\mu_1^{(o)}, \Delta_1^{(o)}\big).$$

(ii) For any nonempty index set $\mathcal{F} \subseteq \mathbb{I}_{\mathcal{C}}$, under the Gaussian working CD, the estimator defined in Eq. (4.8) can be written as

$$\widehat{\gamma}_1^{\mathcal{F}} = \left(\sum_{j\in\mathcal{F}} k_j\widehat{\gamma}_j^{-2}\right)^{-1} \sum_{j\in\mathcal{F}} k_j\widehat{\gamma}_j^{-2}\widehat{\gamma}_j$$

Define $\eta_j^{\mathcal{F}} = \frac{c_j\gamma_j^{-2}}{\sum_{\ell\in\mathcal{F}} c_\ell\gamma_\ell^{-2}}$, $j \in \mathcal{F}$. By the consistency of $\widehat{\gamma}_j$ and the condition $k_j/k \to c_j$, the empirical weights satisfy

$$\frac{k_j\widehat{\gamma}_j^{-2}}{\sum_{\ell\in\mathcal{F}} k_\ell\widehat{\gamma}_\ell^{-2}} \xrightarrow{p} \eta_j^{\mathcal{F}}, \quad j \in \mathcal{F}.$$

By the same argument as in the proof of part (i), we have

$$k^{1/2}\left(\widehat{\gamma}_1^{\mathcal{F}} - \gamma_1\right) \xrightarrow{d} N\left(\mu_{\mathcal{F}}, \Delta_{\mathcal{F}}\right),$$

where

$$\mu_{\mathcal{F}} = \sum_{j\in\mathcal{F}} \eta_j^{\mathcal{F}} c_j^{-1/2}\left(\frac{\lambda_j}{1-\rho_j}\right), \text{ and } \Delta_{\mathcal{F}} = \sum_{j\in\mathcal{F}} \left(\eta_j^{\mathcal{F}}\right)^2 c_j^{-1}\gamma_j^2.$$

Substituting the expression of $\eta_j^{\mathcal{F}}$, we obtain $\Delta_{\mathcal{F}} = \left(\sum_{j\in\mathcal{F}} c_j\gamma_j^{-2}\right)^{-1}$. For the oracle estimator, $\mathcal{F} = \mathbb{I}_{\mathcal{C}}$, and therefore $\Delta_1^{(o)} = \left(\sum_{j\in\mathbb{I}_{\mathcal{C}}} c_j\gamma_j^{-2}\right)^{-1}$. Since $c_j\gamma_j^{-2} > 0$ for all j, for any nonempty $\mathcal{F} \subseteq \mathbb{I}_{\mathcal{C}}$, we have

$$\Delta_1^{(o)} = \left(\sum_{j\in\mathbb{I}_{\mathcal{C}}} c_j\gamma_j^{-2}\right)^{-1} \le \left(\sum_{j\in\mathcal{F}} c_j\gamma_j^{-2}\right)^{-1} = \Delta_{\mathcal{F}}.$$

Thus, $\widehat{\gamma}_1^{(o)} = \widehat{\gamma}_1^{\mathbb{I}_c}$ attains the minimum first-order asymptotic variance among all estimators $\widehat{\gamma}_1^{\mathcal{F}}$ with $\varnothing \neq \mathcal{F} \subseteq \mathbb{I}_{\mathcal{C}}$.

Moreover, in the bias-negligible case where $\lambda_j = 0,\ j \in \mathbb{I}_{\mathcal{C}}$, we have $\mu_{\mathcal{F}} = 0$ for every nonempty $\mathcal{F} \subseteq \mathbb{I}_{\mathcal{C}}$. Therefore, the first-order asymptotic MSE associated with the limiting distribution is $\mathrm{AMSE}\left(\widehat{\gamma}_1^{\mathcal{F}}\right) = \Delta_{\mathcal{F}}/k$. Consequently, the oracle estimator also attains the minimum first-order asymptotic MSE in the bias-negligible case. This completes the proof.

□

Proof of Theorem 3. The proof is similar to that of Theorem 2.

(i) Here we still use the auxiliary term $\gamma_1^{(c)}$ defined as Eq.(S2.1). Then, following the similar techniques as the proof for part (i) of Theorem 2, we have

$$k^{1/2}(\widehat{\gamma}_1^{(c)} - \gamma_1) = k^{1/2}(\widehat{\gamma}_1^{(c)} - \gamma_1^{(c)}) + k^{1/2}(\gamma_1^{(c)} - \gamma_1).$$

We first analyze the first term by decomposing it into two parts:

$$\begin{aligned} k^{1/2}(\widehat{\gamma}_1^{(c)} - \gamma_1^{(c)}) &= k^{1/2} \sum_{j=1}^{N} \eta_j^{(c)} \left(\widehat{\gamma}_j - \gamma_j\right) \\ &= \sum_{j \in \mathbb{I}_{\mathcal{C}}} \eta_j^{(c)} k^{1/2} \left(\widehat{\gamma}_j - \gamma_j\right) + \sum_{j \notin \mathbb{I}_{\mathcal{C}}} \eta_j^{(c)} k^{1/2} \left(\widehat{\gamma}_j - \gamma_j\right) \\ &=: \Lambda_{11} + \Lambda_{12}. \end{aligned}$$

For Λ_{11}, under condition $\mathbf{C}_1(w)$, we have $w_{1j} = 1 - a_j = 1 - o(k^{-1/2})$ for $j \in \mathbb{I}_{\mathcal{C}}$. It follows that $\left|\eta_j^{(c)} - \eta_j^{(o)}\right| = o_p(k^{-1/2})$. Together with $k^{1/2}(\widehat{\gamma}_j - \gamma_j) = O_p(1)$, we obtain

$$\begin{aligned} \Lambda_{11} &= \sum_{j \in \mathbb{I}_{\mathcal{C}}} \eta_j^{(c)} k^{1/2} \left(\widehat{\gamma}_j - \gamma_j\right) \\ &= \sum_{j \in \mathbb{I}_{\mathcal{C}}} \eta_j^{(o)} k^{1/2} \left(\widehat{\gamma}_j - \gamma_j\right) + \sum_{j \in \mathbb{I}_{\mathcal{C}}} \left(\eta_j^{(c)} - \eta_j^{(o)}\right) k^{1/2} \left(\widehat{\gamma}_j - \gamma_j\right) \\ &= \sum_{j \in \mathbb{I}_{\mathcal{C}}} \eta_j^{(o)} k^{1/2} \left(\widehat{\gamma}_j - \gamma_j\right) + o_p(1) \\ &= k^{1/2} \left(\widehat{\gamma}_1^{(o)} - \gamma_1^{(o)}\right) + o_p(1). \end{aligned}$$

Therefore, by the proof of Theorem 2(i) and Slutsky's theorem,

$$\Lambda_{11} \xrightarrow{d} N\left(\mu_1^{(o)}, \Delta_1^{(o)}\right).$$

But for $j \notin \mathbb{I}_{\mathcal{C}}$, $\eta_j^{(c)} = o(k^{-1/2})$, that yields $\Lambda_{12} = o_p(1)$. Next, we can do a similar decomposition for the second term,

$$\begin{aligned}\Lambda_2 :=& k^{1/2}(\gamma_1^{(c)} - \gamma_1) = \sum_{j\in\mathbb{I}_{\mathcal{C}}} \eta_j^{(c)} k^{1/2}\big(\gamma_j - \gamma_1\big) + \sum_{j\notin\mathbb{I}_{\mathcal{C}}} \eta_j^{(c)} k^{1/2}\big(\gamma_j - \gamma_1\big)\\ =& \sum_{j\in\mathbb{I}_{\mathcal{C}}} O_p(1) k^{1/2} o(k^{-1/2}) + \sum_{j\notin\mathbb{I}_{\mathcal{C}}} o_p(k^{-1/2}) k^{1/2}\big(\gamma_j - \gamma_1\big)\\ =& o_p(1).\end{aligned}$$

Here, the term $\eta_j^{(c)}$ in the first summation is $O_p(1)$, since it is a normalized fusion weight satisfying $0 \le \eta_j^{(c)} \le 1, \quad \sum_{j=1}^N \eta_j^{(c)} = 1$.

Combining the results of $\Lambda_{11}, \Lambda_{12}$, and Λ_2, we obtain

$$k^{1/2}(\widehat{\gamma}_1^{(c)} - \gamma_1) \xrightarrow{d} N(\mu_1^{(o)}, \Delta_1^{(o)}).$$

(ii) From the proof of part (i), we have

$$k^{1/2}\left(\widehat{\gamma}_1^{(c)} - \gamma_1\right) = \Lambda_{11} + \Lambda_{12} + \Lambda_2,$$

where

$$\Lambda_{11} = k^{1/2}\left(\widehat{\gamma}_1^{(o)} - \gamma_1^{(o)}\right) + o_p(1), \quad \Lambda_{12} = o_p(1), \quad \Lambda_2 = o_p(1).$$

It follows that

$$k^{1/2}\left(\widehat{\gamma}_1^{(c)} - \gamma_1\right) = k^{1/2}\left(\widehat{\gamma}_1^{(o)} - \gamma_1^{(o)}\right) + o_p(1).$$

Recall that

$$\gamma_1^{(o)} = \sum_{j\in\mathbb{I}_{\mathcal{C}}} \eta_j^{(o)} \gamma_j, \quad \sum_{j\in\mathbb{I}_{\mathcal{C}}} \eta_j^{(o)} = 1,$$

where $\eta_j^{(o)} \ge 0$. By the definition of the clique set, we have

$$k^{1/2}\left|\gamma_1^{(o)} - \gamma_1\right| = \left|\sum_{j\in\mathbb{I}_{\mathcal{C}}} \eta_j^{(o)} k^{1/2}\left(\gamma_j - \gamma_1\right)\right| \le \max_{j\in\mathbb{I}_{\mathcal{C}}} k^{1/2}\left|\gamma_j - \gamma_1\right| = o(1).$$

Hence,

$$k^{1/2}\left(\widehat{\gamma}_1^{(o)}-\gamma_1\right)=k^{1/2}\left(\widehat{\gamma}_1^{(o)}-\gamma_1^{(o)}\right)+o_p(1).$$

Combining the preceding results gives

$$k^{1/2}\left(\widehat{\gamma}_1^{(c)}-\gamma_1\right)=k^{1/2}\left(\widehat{\gamma}_1^{(o)}-\gamma_1\right)+o_p(1),$$

or equivalently,

$$k^{1/2}\left(\widehat{\gamma}_1^{(c)}-\widehat{\gamma}_1^{(o)}\right)=o_p(1). \tag{S2.2}$$

Thus, $\widehat{\gamma}_1^{(c)}$ and $\widehat{\gamma}_1^{(o)}$ have the same first-order limiting distribution and, in particular, the same first-order asymptotic variance. By Theorem 2(ii), $\widehat{\gamma}_1^{(o)}$ attains the minimum first-order asymptotic variance among estimators $\widehat{\gamma}_1^{\mathcal{F}}$ with $\varnothing \neq \mathcal{F} \subseteq \mathbb{I}_{\mathcal{C}}$. Therefore, $\widehat{\gamma}_1^{(c)}$ inherits the first-order asymptotic variance optimality of the oracle estimator.

In the bias-negligible case, it follows directly from Theorem 2(ii) and the asymptotic equivalence established in Eq. (S2.2) that $\widehat{\gamma}_1^{(c)}$ inherits the first-order asymptotic MSE optimality of the oracle estimator.

□

Proof of Theorem 4. The result can be proved following the same arguments in the proof of Theorem 3,

$$\begin{aligned}
&k^{1/2}(\widehat{\gamma}_1^{(c)}-\gamma_1)=k^{1/2}(\widehat{\gamma}_1^{(c)}-\gamma_1^{(c)})+k^{1/2}(\gamma_1^{(c)}-\gamma_1)\\
=&\Big(\sum_{j\in\mathbb{I}_{\mathcal{C}}}+\sum_{j\in\mathbb{I}_{\mathcal{B}}}+\sum_{j\in\mathbb{I}_{\mathcal{D}}}\Big)\left(\eta_j^{(c)}k^{1/2}(\widehat{\gamma}_j-\gamma_j)\right)+\Big(\sum_{j\in\mathbb{I}_{\mathcal{C}}}+\sum_{j\in\mathbb{I}_{\mathcal{B}}}+\sum_{j\in\mathbb{I}_{\mathcal{D}}}\Big)\left(\eta_j^{(c)}k^{1/2}(\gamma_j-\gamma_1)\right)\\
=:&\Lambda_{1C}+\Lambda_{1B}+\Lambda_{1D}+\Lambda_{2C}+\Lambda_{2B}+\Lambda_{2D}.
\end{aligned}$$

Let $\mathbb{I}_{\mathcal{BC}}=\mathbb{I}_{\mathcal{C}}\cup\mathbb{I}_{\mathcal{B}}$. Under condition $\mathbf{C}_2(w)$ and the consistency of $\widehat{\gamma}_J$, for $j\in\mathbb{I}_{\mathcal{BC}}$,

$$\widehat{\eta}_j^{(c)}\xrightarrow{p}\frac{c_j\gamma_j^{-2}}{\sum_{j\in\mathbb{I}_{\mathcal{BC}}}c_j\gamma_j^{-2}}=:\eta_j^*,$$

whereas $\widehat{\eta}_j^{(c)}=o_p(k^{-1/2})$ for $j\in\mathbb{I}_{\mathcal{D}}$. Based on the aforementioned proof, it is easy to conclude that $\Lambda_{1D}=o_p(1),\Lambda_{2C}=o_p(1),\Lambda_{2D}=o_p(1)$, $\Lambda_{2B}\xrightarrow{p}\sum_{j\in\mathbb{I}_{\mathcal{B}}}\eta_j^*B_j=:B^{(c)}$ with

$B_j := \lim_{k\to\infty} k^{1/2}(\gamma_j - \gamma_1)$. By the asymptotic normality of the source-specific Hill estimators and Slutsky's theorem,

$$\Lambda_{1C} + \Lambda_{1B} \xrightarrow{d} N\big(\mu_1, \Delta_1\big)$$

where

$$\mu_1 = \sum_{j\in\mathbb{I}_{\mathcal{C}}\cup\mathbb{I}_{\mathcal{B}}} \eta_j^* c_j^{-1/2}\Big(\frac{\lambda_j}{1-\rho_j}\Big), \quad \Delta_1 = \sum_{j\in\mathbb{I}_{\mathcal{C}}\cup\mathbb{I}_{\mathcal{B}}} (\eta_j^*)^2 c_j^{-1}\gamma_j^2.$$

Together with these items, it follows that

$$k^{1/2}\left(\widehat{\gamma}_1^{(c)} - \gamma_1\right) = \Lambda_{1C} + \Lambda_{1B} + B^{(c)} + o_p(1)$$

and hence

$$k^{1/2}\left(\widehat{\gamma}_1^{(c)} - \gamma_1\right) - B^{(c)} \xrightarrow{d} N\left(\mu_1, \Delta_1\right).$$

□

Proof of Theorem 5. Recall that $\widehat{\boldsymbol{\gamma}}_c = (\widehat{\gamma}_j)_{j\in\mathbb{I}_C}$, $\boldsymbol{\gamma}_c = (\gamma_j)_{j\in\mathbb{I}_C}$. By the consistency of the EVI estimators $\widehat{\gamma}_j$, the uniform consistency of the tail-copula estimators $\widehat{R}(\cdot,\cdot)$, and the continuity of the covariance function, we have $\widehat{\Sigma}_c \xrightarrow{p} \Sigma_c$. Since Σ_c is positive definite, the continuity of matrix inversion gives $\widehat{V}_c = \widehat{\Sigma}_c^{-1} \xrightarrow{p} \Sigma_c^{-1} = V_c$. Consequently,

$$\mathbf{1}^\top \widehat{V}_c \mathbf{1} \xrightarrow{p} \mathbf{1}^\top V_c \mathbf{1} > 0.$$

Define

$$\gamma_1^{(d)} = \frac{\sum_{j\in\mathbb{I}_{\mathcal{C}}}\big(\sum_{i\in\mathbb{I}_{\mathcal{C}}}\widehat{V}_{ij}\big)\gamma_j}{\sum\sum_{i,j\in\mathbb{I}_{\mathcal{C}}}\widehat{V}_{ij}} = \frac{\mathbf{1}^\top\widehat{\boldsymbol{V}}_c\boldsymbol{\gamma}_c}{\mathbf{1}^\top\widehat{\boldsymbol{V}}_c\mathbf{1}}.$$

Then

$$\sqrt{k}(\widetilde{\gamma}_1^{(o)} - \gamma_1) = \sqrt{k}(\widetilde{\gamma}_1^{(o)} - \gamma_1^{(d)}) + \sqrt{k}(\gamma_1^{(d)} - \gamma_1) \tag{S2.3}$$

For the second term,

$$\sqrt{k}\left(\gamma_1^{(d)} - \gamma_1\right) = \frac{\mathbf{1}^\top\widehat{V}_c\sqrt{k}\left(\boldsymbol{\gamma}_c - \gamma_1\mathbf{1}\right)}{\mathbf{1}^\top\widehat{V}_c\mathbf{1}}.$$

By the definition of the clique set and the fixed-N setting, $\max_{j\in\mathbb{I}_{\mathcal{C}}}\sqrt{k}\,|\gamma_j-\gamma_1|=o(1)$, and hence $\sqrt{k}\,(\boldsymbol{\gamma}_c-\gamma_1\mathbf{1})=o(1)$. Together with $\widehat{\boldsymbol{V}}_c\xrightarrow{p}\boldsymbol{V}_c$ and $\mathbf{1}^\top\widehat{\boldsymbol{V}}_c\mathbf{1}\xrightarrow{p}\mathbf{1}^\top\boldsymbol{V}_c\mathbf{1}>0$, we obtain

$$\sqrt{k}\left(\gamma_1^{(d)}-\gamma_1\right)=o_p(1).$$

For the first term, by the joint asymptotic normality in Proposition 2, restricted to the clique index set, and Slutsky's theorem, we have

$$\sqrt{k}\big(\widetilde{\gamma}_1^{(o)}-\gamma_1^{(d)}\big)\xrightarrow{d}N\big(\nu_1^{(o)},\Sigma_1^{(o)}\big),$$

where

$$\nu_1^{(o)}=\frac{\sum_{j\in\mathbb{I}_{\mathcal{C}}}\big(\sum_{i\in\mathbb{I}_{\mathcal{C}}}V_{ij}\big)\mu_j}{\mathbf{1}^\top\boldsymbol{V}_c\mathbf{1}}=\frac{\mathbf{1}^\top\boldsymbol{V}_c\boldsymbol{\mu}_c}{\mathbf{1}^\top\boldsymbol{V}_c\mathbf{1}},\quad \Sigma_1^{(o)}=\frac{1}{\mathbf{1}^\top\boldsymbol{V}_c\mathbf{1}}.$$

The proof is completed by Eq.(S2.3). □

Proof of Corollary 1. The two conclusions follow by combining Theorem 5 with arguments analogous to those used in the proofs of Theorems 3 and 4; the details are therefore omitted. □

S3 Supplement to the Simulation

S3.1 Sensitivity analysis to the tail sample size k

In this subsection, we examine the sensitivity of the proposed method to variations in the tail sample size k_1 for the target source. We consider the $N=10$ data sources with $N_{\mathcal{C}}=5, N_{\mathcal{B}}=3, N_{\mathcal{D}}=2$ and balanced samples, i.e., $n_1=\cdots=n_{10}=500, k_1=\cdots=k_{10}\in[30,90]$. Then the total tail sample size $k=10k_1$. We focus on the Fréchet distribution with target EVI $\gamma_1=0.5$. The other settings are kept the same as Case I in Section 6.

Figure S3.3 reports the finite-sample performance in terms of squared bias, variance, and MSE as k_1 varies. The results show a bias-variance trade-off. For most methods, as k_1 increases, the bias increases while the variance decreases. Among the compared estimators, the classical Hill estimators have the largest variance and MSE over the whole range. The information-borrowing methods are substantially more stable. In particular, iFusion_O achieves the lowest MSE across a wide range of k_1. iFusion_C and iFusion_BC also improve over the baseline but become more sensitive when k_1 is very large due to increasing bias. Notably, the non-monotone (U-shape) squared bias pattern appears in iFusion_O. We attribute this behavior to its adaptive weights with inverse-variance structure (i.e., $\eta_j^{(o)} \propto \widehat{\sigma}_j^{-2}$). When k_1 is small, the classical Hill estimator $\widehat{\gamma}_j$ may have relatively small bias but large variance. This high-variance can degrade the estimation accuracy of iFusion_O by affecting the random weights $\eta_j^{(o)}$.

Overall, the results support the conclusion that the proposed iFusion framework is robust to moderate changes in k_1 in finite samples; and choosing an intermediate tail sample size is important in practice to balance bias and variance and to fully exploit the advantage of information borrowing.

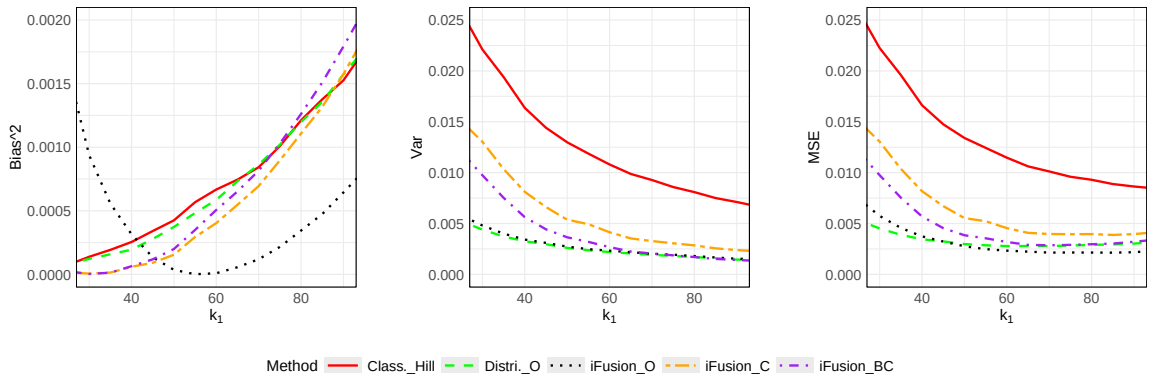

Figure S3.3: Finite-sample performance of EVI estimators for Fréchet distribution ($\gamma_1 = 0.5$) under different levels of k_1.

S3.2 Sensitivity analysis under weaker source separation

In the main simulation setting of Case I, the clique, boundary, and disperse sources are generated with perturbation orders $O(k^{-2})$, $O(k^{-1/2})$, and $O(k^{-1/10})$, respectively. This design provides a clear separation among the three source regimes. To further examine the sensitivity of the proposed method to the separation among clique, boundary, and disperse sources, we conduct two additional experiments under the Fréchet distribution. We fix the target EVI at $\gamma_1 = 0.5$ and consider $N = 10$ sources with group sizes $N_C = 5$, $N_B = 3$, and $N_D = 2$. The remaining simulation settings are kept the same as in the corresponding Fréchet setting considered above.

First, we examine the sensitivity to the separation between the clique and boundary sources by modifying only the clique perturbation. Specifically, we fix $\gamma_1 = 0.5$ and replace the original clique setting with $\gamma_j = \gamma_1 + U_j/k^{\theta_C}$, $j \in \mathbb{I}_C \backslash \{1\}$, where $\theta_C \in \{1.1,\ 1.3,\ 1.5,\ 1.7,\ 1.9\}$. The boundary and disperse sources are generated as before. Smaller values of θ_C correspond to weaker finite-sample separation between the clique and boundary sources, while larger values make the clique sources closer to the target. Second, we conduct a complementary analysis by modifying only the disperse perturbation. Specifically, the clique and boundary sources are kept unchanged, while the original disperse setting is replaced by $\gamma_j = \gamma_1 + U_j/k^{\theta_D}$, $j \in \mathbb{I}_D$, where $\theta_D \in \{0.1,\ 0.2,\ 0.3,\ 0.4\}$. Larger values of θ_D correspond to disperse sources that are closer to the boundary regime in finite samples, and hence lead to a more challenging screening problem.

Figures S3.4 and S3.5 report the squared biases, variances, and MSEs of the estimators under the two sensitivity settings. Figure S3.4 shows that the classical Hill estimator has the largest variance and MSE throughout the considered range of θ_C. The oracle benchmark remains stable, while the data-driven iFusion estimators continue to provide substantial variance reduction. As θ_C increases, the clique sources become closer

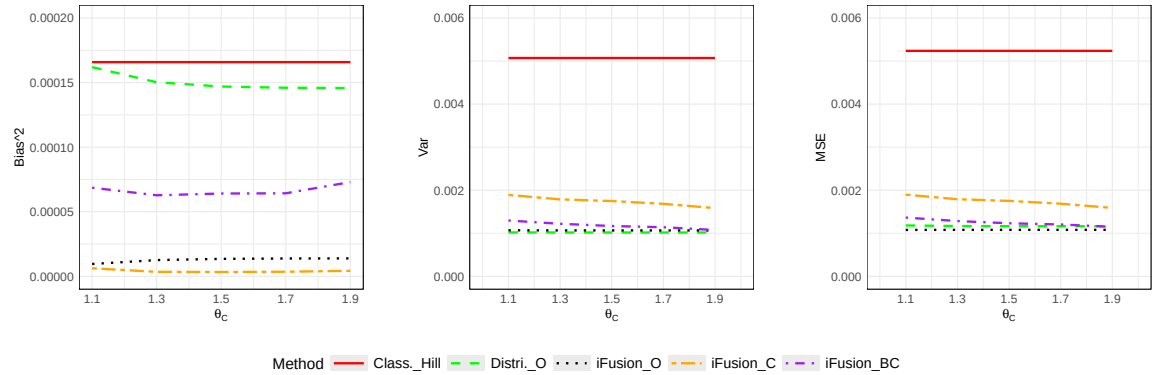

Figure S3.4: Finite-sample performance of EVI estimators for Fréchet distribution ($\gamma_1 = 0.5$) under varying clique perturbation exponent θ_C.

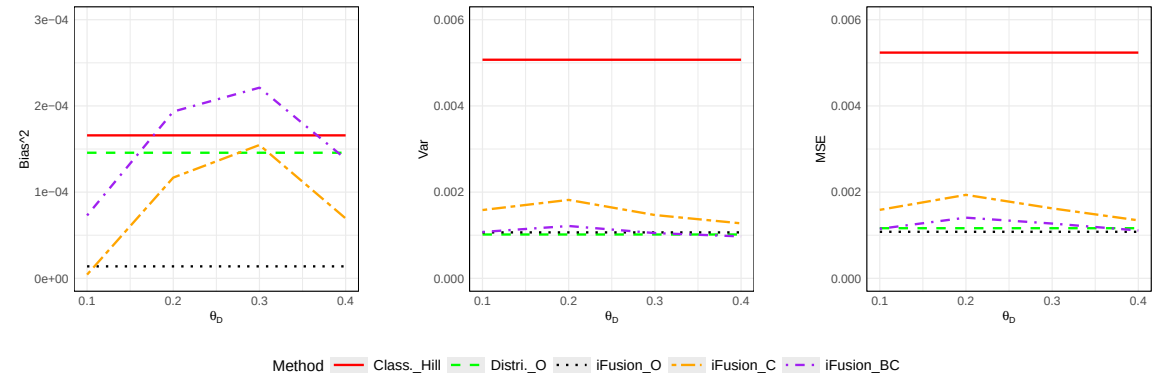

Figure S3.5: Finite-sample performance of EVI estimators for Fréchet distribution ($\gamma_1 = 0.5$) under varying disperse perturbation exponent θ_D.

to the target, and the MSEs of iFusion_C and iFusion_BC gradually approach that of iFusion_O. Even when θ_C is small, corresponding to weaker clique–boundary separation, the proposed iFusion estimators still outperform the classical Hill estimator in terms of MSE. Figure S3.5 presents the complementary sensitivity analysis for the disperse sources. As θ_D increases, the dispersed sources move closer to the boundary regime, which makes the screening problem more challenging. Although the finite-sample squared biases of the data-driven iFusion estimators may fluctuate mildly, their variances and MSEs remain substantially smaller than those of the classical Hill estimator across all considered values of θ_D. Overall, these two sensitivity analyses provide finite-sample evidence that the proposed iFusion estimators remain effective under the moderate reductions in

clique–boundary and disperse–boundary separation considered here, with their variances and MSEs remaining below those of the classical Hill estimator.

S3.3 The larger-N case

In this subsection, we extend the simulation study by increasing the number of data sources from $N = 20$ to $30, 40, 50$ with different combinations of $(N_{\mathcal{C}}, N_{\mathcal{B}}, N_{\mathcal{D}})$, which correspond to the number of elements in sets $\mathcal{C}, \mathcal{B}$ and $\mathcal{D}$, respectively. The data-generating process is similar to Case I, and we focus on Fréchet distribution with target EVI $\gamma_1 = 0.5$.

In general, as N increases, the methods that incorporate information borrowing across sources tend to exhibit reduced variance. However, if N increases while $N_{\mathcal{C}}$ remains fixed, the additional sources are primarily from the boundary/disperse sets, and they are screened out in principle; consequently, the variance reduction may diminish, and in finite samples the variance can even increase due to potential mis-screening. The comparison between the $N = 30$ and $N = 40$ settings provides empirical support for this point. Therefore, to fully realize the advantage of iFusion, it is not sufficient to increase the number of data sources N alone; the proportion of the clique must also be sufficiently large. Comparing the four scenarios in Tables S3.1 and S3.2, we observe that under $N = 50$, both the distributed framework and the iFusion framework achieve near-best performance. This is largely because N is sufficiently large and the clique proportion $N_{\mathcal{C}}/N$ is relatively high in this case, which enables more effective information borrowing and improves estimator stability.

Table S3.1: Estimation results of $\gamma_1 = 0.5$ for Fréchet distribution with balanced samples

	N	$(N_{\mathcal{C}}, N_{\mathcal{B}}, N_{\mathcal{D}})$	Class._Hill	Distri._All	Distri._O	iFusion_O	iFusion_C	iFusion_BC
BIAS	20	(10, 5, 5)	0.0127	0.1158	0.0129	-0.0052	0.0045	0.0085
	30	(20, 5, 5)	0.0127	0.0782	0.0136	-0.0054	0.0004	0.0023
	40	(20, 10, 10)	0.0122	0.1075	0.0136	-0.0052	0.0064	0.0073
	50	(30, 10, 10)	0.0116	0.0869	0.0134	-0.0059	0.0030	0.0040
Var ($\times 10^{-2}$)	20	(10, 5, 5)	0.5065	0.0470	0.0503	0.0541	0.0848	0.0579
	30	(20, 5, 5)	0.5028	0.0262	0.0263	0.0286	0.0425	0.0335
	40	(20, 10, 10)	0.4977	0.0225	0.0258	0.0282	0.0545	0.0361
	50	(30, 10, 10)	0.5084	0.0163	0.0174	0.0190	0.0350	0.0256
MSE ($\times 10^{-2}$)	20	(10, 5, 5)	0.5226	1.3881	0.0669	0.0569	0.0868	0.0652
	30	(20, 5, 5)	0.5189	0.6375	0.0447	0.0315	0.0425	0.0340
	40	(20, 10, 10)	0.5125	1.1772	0.0444	0.0309	0.0585	0.0414
	50	(30, 10, 10)	0.5218	0.7715	0.0354	0.0224	0.0359	0.0272

Table S3.2: Estimation results of $\gamma_1 = 0.5$ for Fréchet distribution with unbalanced samples

	N	$(N_{\mathcal{C}}, N_{\mathcal{B}}, N_{\mathcal{D}})$	Class._Hill	Distri._All	Distri._O	iFusion_O	iFusion_C	iFusion_BC
BIAS	20	(10, 5, 5)	0.0087	0.1152	0.0120	-0.0065	0.0100	0.0128
	30	(20, 5, 5)	0.0104	0.0781	0.0131	-0.0062	0.0063	0.0076
	40	(20, 10, 10)	0.0106	0.1075	0.0131	-0.0062	0.0123	0.0120
	50	(30, 10, 10)	0.0103	0.0871	0.0131	-0.0063	0.0080	0.0086
Var ($\times 10^{-2}$)	20	(10, 5, 5)	1.3654	0.0542	0.0579	0.0555	0.2810	0.2145
	30	(20, 5, 5)	1.4024	0.0301	0.0302	0.0281	0.2280	0.1964
	40	(20, 10, 10)	1.3756	0.0256	0.0303	0.0282	0.2355	0.1782
	50	(30, 10, 10)	1.3431	0.0195	0.0203	0.0187	0.1656	0.1405
MSE ($\times 10^{-2}$)	20	(10, 5, 5)	1.3730	1.3810	0.0724	0.0598	0.2910	0.2310
	30	(20, 5, 5)	1.4132	0.6404	0.0473	0.0319	0.2319	0.2022
	40	(20, 10, 10)	1.3869	1.1815	0.0474	0.0320	0.2505	0.1927
	50	(30, 10, 10)	1.3537	0.7783	0.0376	0.0227	0.1720	0.1478

S3.4 The independent case

We consider the absolute Student-t distribution, which belongs to the max-domain of attraction of Fréchet distribution and satisfies the second-order condition.

- Absolute Student-t: $|t(\alpha)|$ with the extreme value index $\gamma = 1/\alpha$ and second-order parameter $\rho = -2/\alpha$.

Case S(I): Generate random data: $X_{ij} \sim |t(\alpha_j)|$, for $i = 1, \ldots, n_j,\ j = 1, \cdots, 20$, where $\alpha_j = 1/\gamma_j$. Assume that $U_j \overset{iid}{\sim} \text{Unif}(1/2, 1)$ and γ_j take values as follows: (i)$\gamma_1 = \gamma^\star$ and $\gamma_j = \gamma^\star + U_j/k^2$ for $j = 2, \ldots, 10$, this forms a clique whose parameter values are asymptotically tied to the target value; (ii) $\gamma_j = \gamma^\star + U_j/\sqrt{k}$ for $j = 11, \ldots, 15$; (iii) $\gamma_j = \gamma^\star + U_j/k^{1/10}$ for $j = 16, \ldots, 20$, where $\gamma^\star \in \{1/4,\ 1/3,\ 1/2,\ 1\}$.

We adopt the above data-generating process and repeat the simulation 1000 times for $n = 5000$. The specific settings follow those in Section 6.1 for Fréchet distribution. The corresponding results are reported in Tables S3.3-S3.4.

Several insights have been obtained from the analysis of the results for the Fréchet distribution and the absolute Student-t distribution. We acknowledge the validity of the distri._O method, especially in scenarios where the sample sizes are balanced across sources. We attempt to analyze the underlying reasons. Recall that in the present setup, the distributed Hill estimator in this context takes the form of a simple average, i.e., $\widehat{\gamma}_{\text{DH}}^{(o)} = \sum_{j \in \mathbb{I}_{\mathcal{C}}} \widehat{\gamma}_j / N_{\mathcal{C}}$. When the parameters γ_j ($j \in \mathbb{I}_{\mathcal{C}}$) are equal or sufficiently close to γ_1, using equal weights of $1/N_{\mathcal{C}}$ may be preferable to screening weights, as the latter may introduce additional uncertainty in finite samples. Both intuitively and theoretically, the distributed estimator is indeed a desirable choice for the case where the unknown parameters and effective sample sizes are approximately equal across sources.

Table S3.3: Estimation results of γ_1 for Student-t distributions with balanced samples

	$\gamma^\star$	Class._Hill	Distri._All	Distri._O	iFusion_O	iFusion_C	iFusion_BC
BIAS	1/4	0.1008	0.1787	0.1001	0.0895	0.0937	0.0963
	1/3	0.0746	0.1585	0.0755	0.0627	0.0700	0.0720
	1/2	0.0363	0.1320	0.0409	0.0224	0.0333	0.0353
	1	0.0056	0.1058	0.0071	-0.0290	0.0036	-0.0011
Var ($\times 10^{-2}$)	1/4	0.2037	0.0222	0.0195	0.0208	0.0376	0.0262
	1/3	0.2773	0.0290	0.0293	0.0315	0.4455	0.0277
	1/2	0.5600	0.0419	0.0537	0.0593	0.1168	0.0783
	1	1.9603	0.1323	0.1950	0.2100	0.8064	0.6447
MSE ($\times 10^{-2}$)	1/4	1.2207	3.2155	1.0224	0.8227	0.9149	0.9531
	1/3	0.8340	2.5398	0.5988	0.4247	0.5358	0.5455
	1/2	0.6920	1.7852	0.2214	0.1096	0.2278	0.2032
	1	1.9634	1.2521	0.2000	0.2940	0.8077	0.6449

Notes: $\gamma\star$, the true value of the target parameter. Similar notes apply to the following tables.

Table S3.4: Estimation results of γ_1 for Student-t distributions with unbalanced samples

	$\gamma^\star$	Class._Hill	Distri._All	Distri._O	iFusion_O	iFusion_C	iFusion_BC
BIAS	1/4	0.1040	0.1786	0.1002	0.0891	0.0996	0.1009
	1/3	0.0805	0.1582	0.0757	0.0625	0.0751	0.0757
	1/2	0.0380	0.1315	0.0415	0.0231	0.0412	0.0417
	1	0.0135	0.1056	0.0079	-0.0285	0.0194	0.0160
Var ($\times 10^{-2}$)	1/4	0.5457	0.0265	0.0233	0.0218	0.1186	0.0838
	1/3	0.7074	0.0324	0.0338	0.0319	0.1349	0.0900
	1/2	1.3545	0.0479	0.0637	0.0601	0.2891	0.2209
	1	4.5884	0.1589	0.2242	0.2103	2.5169	2.2120
MSE ($\times 10^{-2}$)	1/4	1.6274	3.2166	1.0271	0.8162	1.1100	1.1017
	1/3	1.3549	2.5337	0.6070	0.4223	0.6995	0.6624
	1/2	1.4990	1.7771	0.2356	0.1133	0.4587	0.3945
	1	4.6066	1.2747	0.2304	0.2916	2.5546	2.2375

S3.5 The tail-dependent cases

Parallel to the Student-t and Fréchet distributions in the simulation for the independent case, we further consider the Multivariate-t and Multivariate Pareto distributions with the following forms and data generation process.

- Multivariate t distribution $\mathcal{T}(v, \boldsymbol{\mu}, \Sigma)$ with pdf

$$f(\boldsymbol{x}) = \frac{\Gamma\left(\frac{v+d}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} \frac{1}{(v\pi)^{\frac{d}{2}}} \frac{1}{\sqrt{\det(\Sigma)}} \left(1 + \frac{1}{\nu}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right)^{-\frac{d+v}{2}},$$

$$\boldsymbol{x} = (x_1, \cdots, x_d) \in \mathbb{R}^d,$$

  where $v > 0$ is the degrees of freedom, $\boldsymbol{\mu} \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$ is positive definite. Here the marginal tail indices $\gamma_1 = \gamma_2 = \cdots = \gamma_d = 1/v$.

- Multivariate Pareto distribution $\mathcal{MP}(\alpha, \boldsymbol{\mu}, \boldsymbol{\sigma})$ with survival function

$$\overline{F}(x_1, \ldots, x_d) = \left[1 + \sum_{i=1}^{d} \left(\frac{x_i - \mu_i}{\sigma_i}\right)\right]^{-\alpha}, x_i > \mu_i, i = 1, 2, \ldots, d,$$

  where $\alpha > 0$ is the shape parameter, and $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\sigma} \in \mathbb{R}^d_+$ are the location and scale vector, respectively. The marginal tail indices $\gamma_1 = \gamma_2 = \cdots = \gamma_d = 1/\alpha$.

Case S(II): Generate random data $(X_1, \ldots, X_d) \sim \mathcal{T}(v, \boldsymbol{\mu}, \Sigma)$ with $v = 1/\gamma^\star, \boldsymbol{\mu} = \mathbf{0}_d$, and $\Sigma = diag(1, \cdots, 1)_{d\times d}$ or

$$\Sigma = \begin{bmatrix} 1 & \sigma & \cdots & \sigma \\ \sigma & 1 & \cdots & \sigma \\ \vdots & \vdots & \ddots & \vdots \\ \sigma & \sigma & \cdots & 1 \end{bmatrix} \tag{S3.1}$$

where $\gamma^\star \in \{1/4,\ 1/3,\ 1/2,\ 1\}$ and $d = 5$. Hence, the group size $N = N_{\mathcal{C}} = 5$.

Case S(III): Generate random data $(X_1, \ldots, X_d) \sim \mathcal{MP}(\alpha, \boldsymbol{\mu}, \boldsymbol{\sigma})$ with $\alpha = 1/\gamma^\star, \boldsymbol{\mu} = \mathbf{0}_d$ and $\boldsymbol{\sigma} = \mathbf{1}_d$, where $\gamma^\star \in \{0.4,\ 0.6,\ 0.8,\ 1.0\}$ and $d = 5$ which means $N = N_{\mathcal{C}} = 5$.

For consistency with the finite-sample setting in Section 6, we repeat the simulation 1000 times and consider the sample for each individual as follows.

Balanced: $n_1 = \cdots = n_5 = 500$, $k_1 = \cdots = k_5 = 50$;

Unbalanced: $(n_1, \cdots, n_5) = (200, 300, 500, 700, 800)$ and $(k_1, \cdots, k_5) = 0.1 \cdot (n_1, \cdots, n_5)$.

For Cases S(II) and S(III), the extreme value indices for each marginal distribution are completely equal due to the properties of the distributions. Hence the situation is relatively simple, i.e., $\mathcal{C} = \{\gamma_j : \ j = 1, \ldots, N\}$, and $\mathbb{I}_{\mathcal{C}} = \{1, 2, \ldots, N\}$, where $N = d = 5$. For Multivariate t distribution $\mathcal{T}(v, \boldsymbol{\mu}, \Sigma)$, the scale matrix $\Sigma \in \mathbb{R}^{d \times d}$ defines the dispersion and dependencies between variables, which plays a role similar to the covariance matrix in the multivariate normal distribution. We assume Σ takes the form in Eq.(S3.1), and then the correlation between any two random variables is enhanced by increasing the value of σ. To assess the effect of accounting for dependence in estimation, we repeated the simulation with the same parameters but with $\sigma = 0$ or $\sigma = 0.8$ for the matrix Σ. Results under unbalanced cases are reported in Tables S3.5-S3.6. For the Multivariate Pareto distribution $\mathcal{MP}(\alpha, \boldsymbol{\mu}, \boldsymbol{\sigma})$, we set the location and scale parameters as standard forms to simplify our computation. The results related to balanced and unbalanced cases are shown in Tables S3.7-S3.8.

Table S3.5: Estimation results of γ_1 for Multivariate t distributions with unbalanced samples ($\sigma = 0$)

	$\gamma^\star$	Class._Hill	Bench._O	Distri._O	iFusion_O		iFusion	
					ind.	dep.	ind.	dep.
BIAS	1/4	0.1034	0.0988	0.1013	0.0915	0.0871	0.0954	0.0921
	1/3	0.0816	0.0748	0.0770	0.0667	0.0610	0.0705	0.0658
	1/2	0.0459	0.0397	0.0422	0.0286	0.0187	0.0341	0.0261
	1	-0.0006	0.0034	0.0058	-0.0132	-0.0387	-0.0046	-0.0199
Var ($\times 10^{-2}$)	1/4	0.5048	0.0544	0.0693	0.0556	0.0595	0.1297	0.1305
	1/3	0.7304	0.0928	0.1224	0.1003	0.1032	0.1914	0.1927
	1/2	1.4424	0.2208	0.2763	0.2237	0.2233	0.3973	0.3945
	1	4.5951	1.0306	1.2717	1.0534	1.0317	1.9479	1.8994
MSE ($\times 10^{-2}$)	1/4	1.5743	1.0308	1.0956	0.8931	0.8182	1.0402	0.9785
	1/3	1.3959	0.6519	0.7151	0.5448	0.4752	0.6880	0.6254
	1/2	1.6528	0.3781	0.4540	0.3056	0.2582	0.5135	0.4627
	1	4.5952	1.0317	1.2750	1.0708	1.1812	1.9501	1.9390

Table S3.6: Estimation results of γ_1 for Multivariate t distributions with unbalanced samples ($\sigma = 0.8$)

	$\gamma^\star$	Class._Hill	Bench._O	Distri._O	iFusion_O		iFusion	
					ind.	dep.	ind.	dep.
BIAS	1/4	0.1007	0.0990	0.1011	0.0941	0.0854	0.0956	0.0884
	1/3	0.0804	0.0751	0.0774	0.0690	0.0580	0.0717	0.0620
	1/2	0.0459	0.0363	0.0403	0.0283	0.0126	0.0309	-0.0047
	1	0.0037	0.0043	0.0076	-0.0079	-0.0383	-0.0005	-0.0214
Var ($\times 10^{-2}$)	1/4	0.5282	0.0867	0.1126	0.0911	0.0919	0.1324	0.1306
	1/3	0.7516	0.1358	0.1702	0.1348	0.1361	0.1922	0.1938
	1/2	1.4475	0.2904	0.3659	0.2927	0.2882	0.4076	0.3928
	1	4.8429	1.2584	1.5263	1.2468	1.1854	2.1179	2.0619
MSE ($\times 10^{-2}$)	1/4	1.5423	1.0670	1.1356	0.9766	0.8215	1.0464	0.9122
	1/3	1.3984	0.6992	0.7699	0.6103	0.4724	0.7059	0.5785
	1/2	1.6581	0.4224	0.5284	0.3728	0.3041	0.5031	0.3950
	1	4.8442	1.2603	1.5321	1.2530	1.3324	2.117 8	2.1077

Table S3.7: Estimation results of γ_1 for Multivariate Pareto distributions with balanced samples

	$\gamma^\star$	Class._Hill	Bench._O	Distri._O	iFusion_O		iFusion	
					ind.	dep.	ind.	dep.
BIAS	0.4	0.1669	0.1655	0.1679	0.1554	0.1407	0.1566	0.1434
	0.6	0.1171	0.1126	0.1158	0.1007	0.0755	0.1014	0.0799
	0.8	0.0862	0.0775	0.0820	0.0657	0.0299	0.0671	0.0336
	1.0	0.0521	0.0483	0.0530	0.0363	-0.0095	0.0380	-0.0010
Var ($\times 10^{-2}$)	0.4	0.5683	0.1781	0.1789	0.1839	0.2021	0.1969	0.2131
	0.6	0.9195	0.3737	0.3760	0.3821	0.4201	0.4075	0.4264
	0.8	1.5216	0.7909	0.8034	0.8110	0.8744	0.8412	0.9016
	1.0	2.1375	1.1851	1.2036	1.2164	1.3045	1.2481	1.3417
MSE ($\times 10^{-2}$)	0.4	3.3524	2.9175	2.9981	2.5988	2.1830	2.6483	2.2683
	0.6	2.2907	1.6406	1.7172	1.3956	0.9904	1.4356	1.0644
	0.8	2.2658	1.3921	1.4755	1.2420	0.9638	1.2910	1.0148
	1.0	2.4092	1.4188	1.4840	1.3482	1.3135	1.3928	1.3418

Table S3.8: Estimation results of γ_1 for Multivariate Pareto distributions with unbalanced samples

	$\gamma^\star$	Class._Hill	Bench._O	Distri._O	iFusion_O		iFusion	
					ind.	dep.	ind.	dep.
BIAS	0.4	0.1707	0.1670	0.1698	0.1556	0.1478	0.1605	0.1549
	0.6	0.1187	0.1126	0.1153	0.0985	0.0860	0.1083	0.1001
	0.8	0.0824	0.0801	0.0831	0.0634	0.0454	0.0712	0.0606
	1.0	0.0433	0.0462	0.0491	0.0273	0.0029	0.0348	0.0207
Var ($\times 10^{-2}$)	0.4	1.3225	0.1526	0.1917	0.1620	0.1735	0.3728	0.3735
	0.6	2.3165	0.3379	0.4315	0.3445	0.3450	0.7550	0.7428
	0.8	3.6102	0.5721	0.7398	0.5854	0.5824	1.3874	1.3493
	1.0	4.7528	0.9177	1.1265	0.9482	0.9780	1.8673	1.8276
MSE ($\times 10^{-2}$)	0.4	4.2356	2.9418	3.0758	2.5838	2.3578	2.9485	2.7728
	0.6	3.7265	1.6055	1.7617	1.3139	1.0850	1.9283	1.7446
	0.8	4.2892	1.2138	1.4297	0.9879	0.7888	1.8937	1.7171
	1.0	4.9400	1.1311	1.3680	1.0226	0.9788	1.9886	1.8704

Some points similar to those discussed in Section 6.2 of the main manuscript are omitted here for brevity. However, we would like to highlight two special observations and insights as follows.

For Case S(II), we increase the correlation between $X^{(i)}$ and $X^{(j)}$ by the parameter σ in matrix Σ, and compare the results reported in Tables S3.5 and S3.6. The results of the bias do not show a clear trend, but the variance of all the estimates behaves uniformly with an increasing tendency, which leads to a corresponding increase in the MSE in Table S3.6. This implies that the precision of the estimator is inevitably influenced by the dependence among sources, even though our proposed method has already accounted for the dependence structure.

It is readily seen that there is no obvious advantage in considering the dependence across the individuals, and even the results of bias tend to be slightly worse when $\gamma = 1$. This is most likely because a data-driven estimator of the R-function may contribute to estimation uncertainty.

S4 Supplement to the Real Data Analysis

S4.1 Additional results for the ozone data

This subsection provides additional empirical results for the ozone data analysis in Section 7 of the main manuscript. We report (i) a variance sensitivity analysis that varies the proportion of missing observations at each site to evaluate estimation stability under data sparsity, and (ii) detailed out-of-sample comparisons for predicting extreme events using high-quantile forecasting and quantile loss.

Variance sensitivity analysis under data missingness. To mimic sparse measurements at a target site, we randomly delete a proportion ρ_1 of its observations while keeping the remaining sites unchanged, and then recompute the EVI estimators and their

variance estimates under the four competing methods. Following the same experimental protocol employed for site S1, we extended this sensitivity analysis to the remaining five monitoring sites (S2–S6) to assess the generalizability of our findings. Figure S4.6 displays the average estimated variances for sites S2–S6 across a range of deletion ratios. The overall pattern is consistent across all sites: the Class._Hill estimator exhibits the most pronounced variance inflation as ρ_1 increases, whereas iFusion remains substantially more stable and consistently yields lower variance than Distri._C, providing further empirical support for the variance-reduction property established in our theoretical analysis.

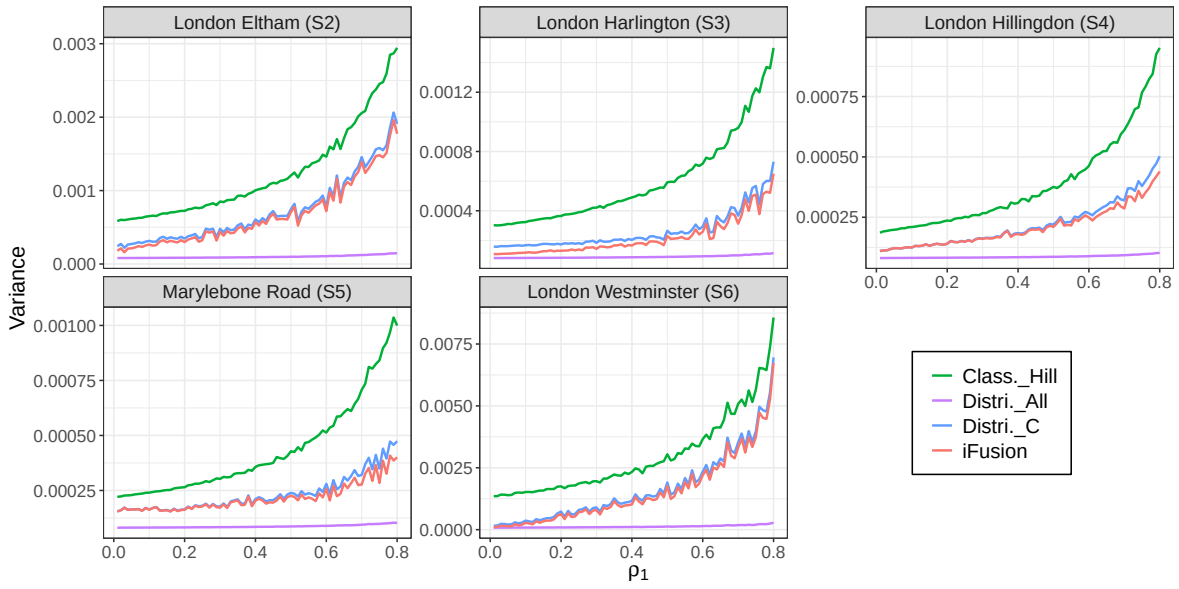

Figure S4.6: Variance comparison for sites S2–S6 under different deletion ratios.

Extreme-event prediction via high-quantile forecasting. In addition to the OPE analysis in the main manuscript, we further evaluate the predictive performance for extreme events via the quantile loss (QL) function. Specifically, we divide the observations of the given site into a training set and a testing set. Using the training set, we obtain the EVI ($\widehat{\gamma}$) estimated by different methods, and predict the quantiles of the ozone concentration at a high quantile level τ by the Weissman estimator (Weissman, 1978) $\widehat{Q}_Y(\tau) = Y_{n_1-k^\star,n_1}\left[k^\star/(n_1(1-\tau))\right]^{\widehat{\gamma}}$, where $Y_{n_1-k^\star,n_1}$ is the $(k^\star+1)$-th largest order

statistic in the training set $\{Y_i\}_{i=1}^{n_1}$. We then calculate the average quantile loss as follows:

$$\overline{\mathrm{QL}}(\rho_3) = \frac{1}{n_2}\sum_{i=1}^{n_2}(Y_i^\star - \widehat{Q}_Y(\tau))(\tau - I(Y_i^\star - \widehat{Q}_Y(\tau) < 0)),$$

where $\{Y_i^\star\}_{i=1}^{n_2}$ is the testing set, and $\rho_3 = n_2/(n_1+n_2)$ is the out-of-sample proportion. For each site, Table S4.9 presents the averaged quantile losses across varying out-of-sample proportions (ranging from 10% to 30%). The iFusion method consistently achieves the lowest QL values across all six sites for both the 99% and 99.5% levels, indicating superior accuracy in capturing the tail behavior of ozone concentrations compared to the classical Hill and distributed estimators. Figures S4.7 and S4.8 further display the site-wise quantile losses as functions of the out-of-sample fraction ρ_3 at the 99% and 99.5% quantile levels, respectively. These figures show that the relative advantage of iFusion is robust across different train/test splits, rather than being driven by a particular testing proportion.

Table S4.9: Average extreme quantile forecasting performance.

Site	Quantile Level: 99%				Quantile Level: 99.5%			
	Class._Hill	Distri._All	Distri._C	iFusion	Class._Hill	Distri._All	Distri._C	iFusion
S1	62.7930	60.9077	61.2268	**59.8313**	104.2009	109.3821	103.5216	**102.4923**
S2	81.8715	73.6392	75.2747	**72.8465**	77.8848	75.6504	76.1762	**75.0634**
S3	70.5072	69.1396	69.1046	**68.9832**	92.9325	93.2573	94.3595	**91.9252**
S4	64.7321	62.9680	63.2316	**62.9276**	86.3861	83.8652	84.0504	**83.4918**
S5	62.3963	61.7284	61.8476	**61.6201**	77.8291	74.4472	74.7174	**73.4021**
S6	88.5897	88.4147	87.7774	**87.1030**	74.7801	71.5178	72.5050	**71.3469**

Notes: The table reports averaged out-of-sample quantile losses at $\tau \in \{99\%, 99.5\%\}$ *for sites S1–S6, averaged over testing fractions* $\rho_3 \in \{0.10, 0.15, 0.20, 0.25, 0.30\}$*; smaller values indicate better prediction.*

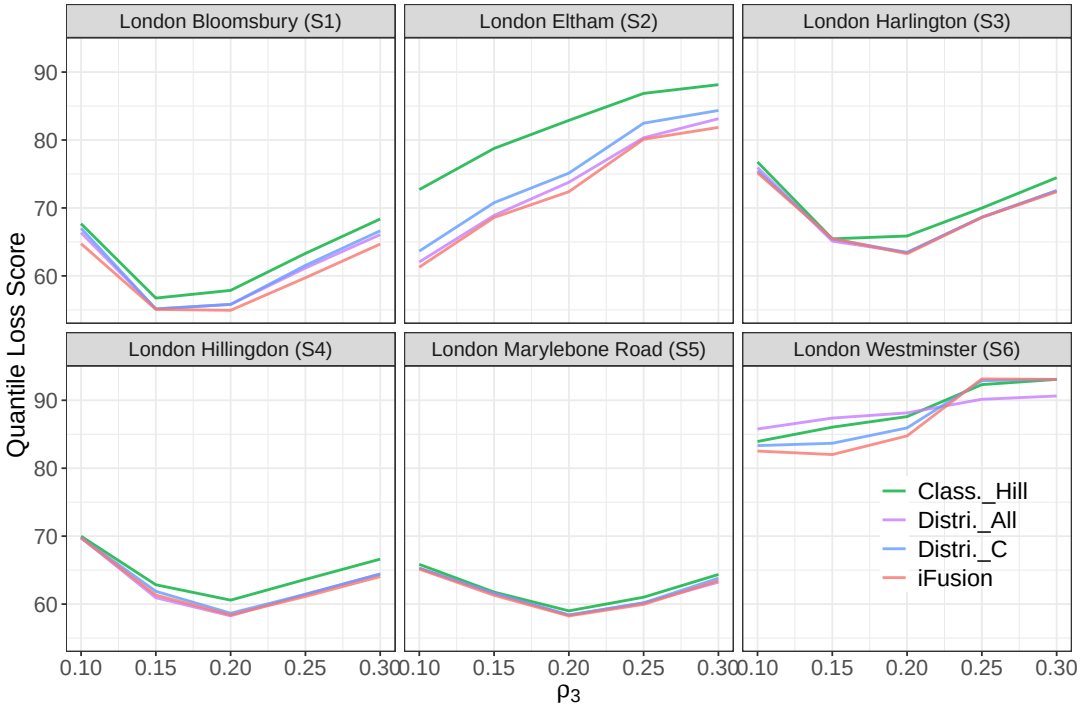

Figure S4.7: Extreme-quantile prediction at $\tau = 99\%$. The figure shows out-of-sample quantile losses across testing fractions ρ_3 for sites S1–S6.

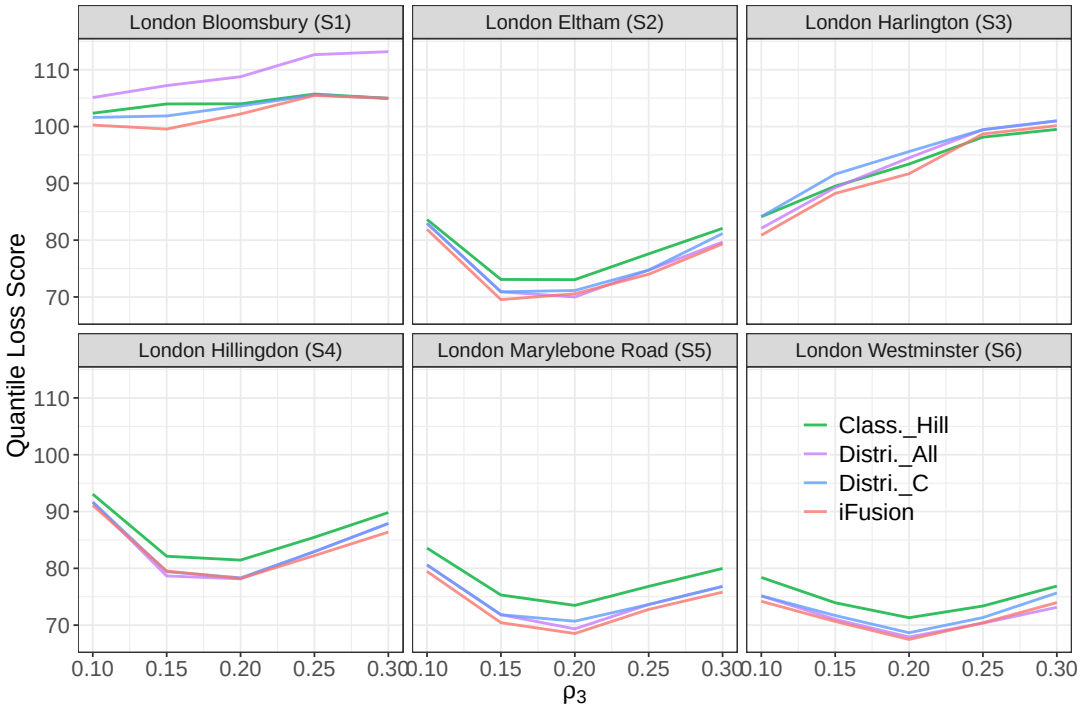

Figure S4.8: Extreme quantile prediction at $\tau = 99.5\%$. The figure shows out-of-sample quantile losses across testing fractions ρ_3 for sites S1–S6.

S4.2 Application to the car-insurance data

In this subsection, we analyze the car insurance dataset previously studied in Chen et al. (2022) and Daouia et al. (2024). The dataset contains the car insurance claims in the five states of the United States during January and February 2011. Specifically, the five states and their corresponding sample sizes are Iowa ($n_1 = 2601$), Kansas ($n_2 = 798$), Missouri ($n_3 = 3150$), Nebraska ($n_4 = 1703$), and Oklahoma ($n_5 = 882$), yielding a total sample size of $n = \sum_{j=1}^{5} n_j = 9134$. The heavy-tailedness and the equality of extreme value indices have been checked in Chen et al. (2022) and Daouia et al. (2024). We do not re-examine the heavy-tailed property of the data in this paper. As for the equality of extreme value indices, our approach does not rely on explicit testing; instead, it is implicitly reflected through the data-driven weights in the iFusion estimator.

Although this dataset has been analyzed by the above literature, we apply our method to explore its applicability under the proposed framework. To satisfy the requirement for sample balance, Chen et al. (2022) sampled 700 observations from each state and conducted their analysis based on this subsample. In contrast, such preprocessing is unnecessary in our framework, as the iFusion method can directly accommodate unbalanced sample sizes. Additionally, unlike the work of Chen et al. (2022) and Daouia et al. (2024), the iFusion method can be employed directly without testing the equality of extreme value indices. Instead, the screening weights automatically reflect the degree of similarity between the target index and the others. Thus, a major advantage of our method is its simplicity and broad applicability.

Naturally, the five states in the dataset can be treated as the five distinct data sources. For the sake of convenience, we denote the extreme value indices (and their estimators) for observations of the five states (Iowa, Kansas, Missouri, Nebraska, and Oklahoma) by γ_j (and $\widehat{\gamma}_j$), where $j = 1, \ldots, 5$. Furthermore, we take the index for Iowa, γ_1, as the

target parameter. As an initial step, we obtain classical Hill estimators based solely on the observations from each individual state. To compute the estimators for each data source, the effective sample size k_j is selected from an equally spaced grid over the interval $[0.05n_j, 0.25n_j]$ for $j = 1, 2, \ldots, 5$. The top panel in Figure S4.9 displays the result of the estimators of the EVI for the five states across varying rates of effective sample size. As the rate increases, we observe that the estimates become more stable, with reduced variability across states. Moreover, the estimates for all five states tend to converge within a relatively narrow range when the effective sample size exceeds 15% of the total. This observation suggests that the extreme value indices of the five states are likely to be similar, providing empirical support for the assumption of index equality discussed earlier.

If the original individual-level data can be shared across sources, we may combine all data sources into a single dataset and compute the Hill estimator under the same tail-sample fraction. The resulting estimator then serves as a benchmark against which we evaluate the performance of the other methods. A more realistic assumption, as adopted in prior studies such as Chen et al. (2022) and Daouia et al. (2024), is that individual-level data cannot be shared due to privacy concerns or regulatory restrictions. Nevertheless, the sources may be willing to share summary statistics or pre-computed estimators to facilitate collaborative risk analysis. In this context, it implies that we have access to the initial estimators $\widehat{\gamma}_j$ and the (effective) sample sizes from all five states, which makes it feasible to implement the iFusion method.

The bottom panel of Figure S4.9 presents the estimation results of the EVI for Iowa under different methods. Here, the Hill estimator computed from the full dataset $\mathbb{S}$ is treated as the benchmark. Compared with the classical Hill estimator based solely on Iowa's data, both the distributed and iFusion estimators yield estimates that are closer

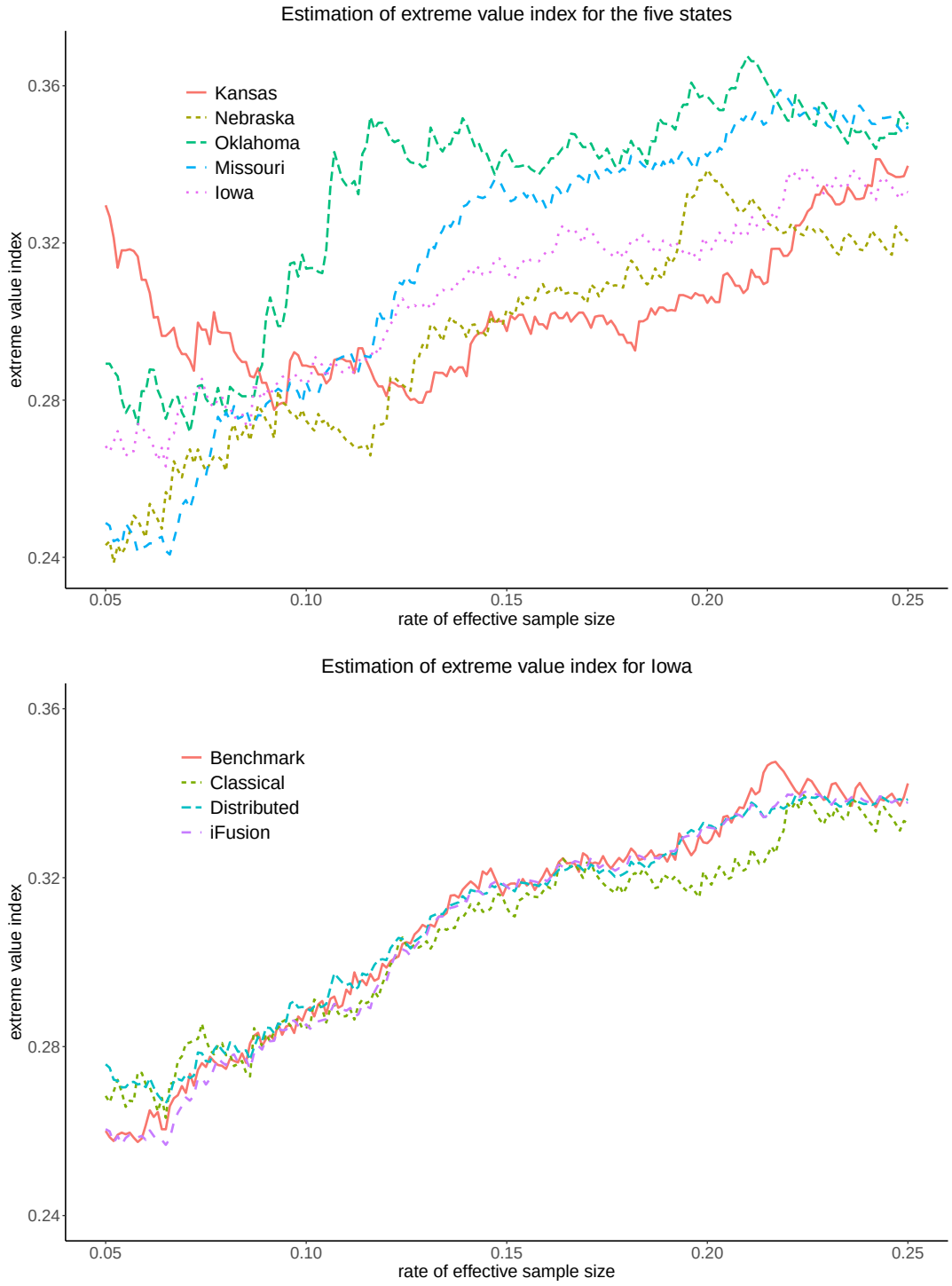

Figure S4.9: Top panel: the estimation of extreme value index for the five states; Bottom panel: the estimation of EVI for Iowa under different methods.

to the pooled-data benchmark, showing greater agreement with the pooled-data analysis. Notably, the iFusion estimator consistently aligns more closely with the benchmark curve than the distributed estimator, particularly in the effective sample size range between 0.10 and 0.20. Furthermore, the iFusion curve exhibits less fluctuation, suggesting improved stability in finite-sample settings. These results suggest that the data-driven weighting mechanism of iFusion produces estimates that align more closely with the pooled-data benchmark and are less sensitive to the choice of the effective sample size than those obtained using the simple averaging strategy.

Bibliography

Chen, L., D. Li, and C. Zhou (2022). Distributed inference for the extreme value index. *Biometrika 109*(1), 257–264.

Daouia, A., S. A. Padoan, and G. Stupfler (2024). Optimal weighted pooling for inference about the tail index and extreme quantiles. *Bernoulli 30*(2), 1287–1312.

Shen, J., R. Y. Liu, and M.-g. Xie (2020). iFusion: Individualized fusion learning. *Journal of the American Statistical Association 115*(531), 1251–1267.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association 73*(364), 812–815.