Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Method G: Uncertainty Quantification for Distributed Data Problems Using Generalized Fiducial Inference

Randy C. S. Lai, Jan Hannig & Thomas C. M. Lee

To cite this article: Randy C. S. Lai, Jan Hannig & Thomas C. M. Lee (2021): Method G: Uncertainty Quantification for Distributed Data Problems Using Generalized Fiducial Inference, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2021.1923514

To link to this article: https://doi.org/10.1080/10618600.2021.1923514

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2021, VOL. 00, NO. 0, 1–12
https://doi.org/10.1080/10618600.2021.1923514

Taylor & Francis
Taylor & Francis Group

Method G: Uncertainty Quantification for Distributed Data Problems Using Generalized Fiducial Inference

Randy C. S. Lai[a,b], Jan Hannig[c], and Thomas C. M. Lee[b]

[a]Department of Mathematics & Statistics, University of Maine, Orono, ME; [b]Department of Statistics, University of California at Davis, Davis, CA; [c]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC

ABSTRACT
It is not unusual for a data analyst to encounter datasets distributed across several computers. This can happen for reasons such as privacy concerns, efficiency of likelihood evaluations, or just the sheer size of the whole dataset. This presents new challenges to statisticians as even computing simple summary statistics such as the median becomes computationally challenging. Furthermore, if other advanced statistical methods are desired, then novel computational strategies are needed. In this article, we propose a new approach for distributed analysis of massive data that is suitable for generalized fiducial inference and is based on a careful implementation of a "divide-and-conquer" strategy combined with importance sampling. The proposed approach requires only small amount of communication between nodes, and is shown to be asymptotically equivalent to using the whole dataset. Unlike most existing methods, the proposed approach produces uncertainty measures (such as confidence intervals) in addition to point estimates for parameters of interest. The proposed approach is also applied to the analysis of a large set of solar images. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2019
Revised March 2021

KEYWORDS
Divide and conquer; Importance sampling; Massive data; Parallel processing; Solar imaging

1. Introduction

The increased availability of cloud computing brings new challenges to practical data analysis. For example, advances in modern science and business allow the collection of massive datasets. An example is high-throughput sequencing in genetics that is capable of producing terabytes of data in a single experiment. Even if the dataset itself is not massive, there are other reasons that it needs to be analyzed in a distributive manner. For example, privacy concerns might require datasets to stay within the country or company of origin and share summary information only. Similarly, computational efficiency of MCMC algorithms sometimes deteriorates with the sample size so one might want to run multiple MCMC chains on different portions of the data.

This presents new challenges to statisticians as even computing simple summary statistics such as the median of such a dataset becomes computationally challenging. If other advanced statistical methods are required for analyzing such datasets, then novel computational strategies are needed. An appealing approach to analyzing a massive dataset is the so-called divide-and-conquer strategy. That is, if the dataset is first divided into manageable subsets, then each subset is analyzed separately, often on a parallel computer, and finally the results of the analyses are combined.

To efficiently combine the results from the various subgroups, one needs to account for the uncertainties in the estimates based on each of the subsets. Among the frequentist proposals, Kleiner et al. (2014) proposed a parallelized version of bootstrap, Chen and Xie (2014) proposed the use of confidence distributions, and Battey et al. (2015) performed distributed testing. In the Bayesian literature, many recent algorithms propose using the embarrassingly parallel approach with various modifications to assess with combination of the results afterward. Huang and Gelman (2005), Scott et al. (2016), Liu (2016), Neiswanger, Wang, and Xing (2014), and Leisen, Craiu, and Casarin (2016) decomposed the posterior distribution into smaller parts and approximate them with normal distributions. Another school of studies advocate combinations based on the inflation of the subset data likelihood; see, for example, Wang et al. (2015), Srivastava et al. (2015), and Entezari, Craiu, and Rosenthal (2018).

In this article, we propose a new approach that is suitable for generalized fiducial inference (GFI), which has proven to provide a distribution on the parameter space with good inferential properties without the need for a subjective prior specification (Hannig et al. 2016). Our parallel algorithm uses minimal amount of information swapping between workers to improve efficiency of the algorithm while maintaining the ability to run different MCMC algorithms on each worker. It does not require any normal distribution approximation as seen in Huang and Gelman (2005), Scott et al. (2016), Liu (2016), Neiswanger, Wang, and Xing (2014), and Leisen, Craiu, and Casarin (2016). We use a carefully implemented an importance sampling scheme to combine the results from various workers.

CONTACT Jan Hannig haiying.wang@uconn.edu Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3260.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

As we do not have to modify the likelihood function such as Wang et al. (2015), Srivastava et al. (2015), and Entezari, Craiu, and Rosenthal (2018), we could easily obtain fiducial samples from the subsets by using already established infrastructure for small datasets and do not require new implementation for generating samples from each subset. Our method produces uncertainty measures (such as confidence intervals) as well as point estimates for the parameters of interest. We prove consistency and asymptotic normality of the approximation scheme and provide numerical comparisons showing good performance of our algorithm. While the proposed method has been designed for GFI, it is also applicable for Bayes posteriors. We call our proposal *Method G*.

The rest of this article is organized as follows. First, some background material for GFI is provided in Section 2. Then the proposed methodology is developed in Section 3, which include theoretical backup and a practical algorithm. The finite sample performance of the proposed methodology is illustrated via numerical experiments in Section 4 and real data application in Section 5. Last, concluding remarks are offered in Section 6 while technical details are deferred to the appendix.

2. Background of Generalized Fiducial Inference

Fisher (1930) introduced fiducial inference in the hope to define a distribution on the parameter space when the Bayesian approach cannot be applied due to the lack of a suitable prior. Unfortunately, his fiducial proposal carried some defects and hence was not welcomed by the statistics community. Generalized fiducial inference (GFI) is an improved version of Fisher's idea that rectifies these defects. See Hannig et al. (2016) for an up-to-date review of GFI.

Suppose we have $\mathbf{Y} = \{Y_1, \dots, Y_n\}$ iid continuous random variables from some distribution $F(y; \boldsymbol{\theta})$ with an unknown p-variate parameter $\boldsymbol{\theta}$ and parameter space Θ; that is, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \Re^p$. Denote the corresponding density function as $f(y; \boldsymbol{\theta})$. It is further supposed that the observation vector $\mathbf{Y}$ could be written as a mapping from a pivotal random vector $\mathbf{U} = \{U_1, \dots, U_n\}$ such that

$$\mathbf{Y} = \mathbf{G}(\mathbf{U}, \boldsymbol{\theta}). \tag{1}$$

Inverting this data-generating equation provides us with a generalized fiducial density $r(\boldsymbol{\theta})$: a distribution on the parameter space obtained without the need to define a prior distribution. Hannig et al. (2016) showed that the generalized fiducial density $r(\boldsymbol{\theta}; \mathbf{y})$ of $\boldsymbol{\theta}$ for a fixed observed data $\mathbf{Y} = \mathbf{y}$ is

$$r(\boldsymbol{\theta}; \mathbf{y}) = \frac{f(\mathbf{y}; \boldsymbol{\theta}) J(\mathbf{y}, \boldsymbol{\theta})}{\int f(\mathbf{y}; \boldsymbol{\theta}') J(\mathbf{y}, \boldsymbol{\theta}') d\boldsymbol{\theta}'} \overset{\text{def}}{=} \frac{1}{c(\mathbf{y})} f(\mathbf{y}; \boldsymbol{\theta}) J(\mathbf{y}, \boldsymbol{\theta}), \tag{2}$$

where

$$J(\mathbf{y}, \boldsymbol{\theta}) = D\left(\nabla_{\boldsymbol{\theta}} \mathbf{G}(\mathbf{u}, \boldsymbol{\theta}) \Big|_{u = \mathbf{G}^{-1}(\mathbf{y}, \boldsymbol{\theta})} \right). \tag{3}$$

The D function has two canonical forms derived in Hannig et al. (2016). The form of D depends on how we define neighborhoods of the observed data y. The first uses neighborhoods specified by the L_∞ norm (corresponding to observing discretized data) and the resulting D is $D_\infty(A) = \sum_{\mathbf{i}} |\det(A_{\mathbf{i}})|$. The sum spans over $\binom{n}{p}$ of p-tuples of indexes $\mathbf{i} = (1 \le i_1 < \cdots < i_p \le n)$. For any $n \times p$ matrix A, the sub-matrix $A_{\mathbf{i}}$ is the $p \times p$ matrix containing the rows $\mathbf{i} = (i_1, \dots, i_p)$ of A. The second form uses an L_2 norm and the corresponding D is $D_2(A) = (\det A^\top A)^{1/2}$ (the product of singular values). According to our experiences, these two canonical forms often yield similar results in practical applications and D_2 is less computational expensive than D_∞. Interested readers are referred to Hannig et al. (2016) for the exact assumptions under which (2) holds.

Suppose $\boldsymbol{\vartheta}$ follows the generalized fiducial density $r(\boldsymbol{\theta}; \mathbf{y})$. When using GFI to solve an inference problem, very often one seeks to evaluate the expectation of a function $h(\boldsymbol{\vartheta})$, which is defined as

$$E[h(\boldsymbol{\vartheta}) \mid \mathbf{y}] = \int_{\boldsymbol{\Theta}} h(\boldsymbol{\theta}') r(\boldsymbol{\theta}'; \mathbf{y}) \mathrm{d}\boldsymbol{\theta}'. \tag{4}$$

We are guilty of committing a small abuse of notation in Equation (4). The expectation is computed by using a generalized fiducial density and not a conditional density. However, just like a conditional expectation, it is a measurable function of the observed data.

As an illustration, consider the following example. Suppose $p > 1$ and it is of interest to compute the marginal generalized fiducial distribution of the first entry of $\boldsymbol{\theta}$, say θ_1. One can consider the expectation of the indicator function $h_t(\boldsymbol{\theta}) = 1\{\theta_1 \le t\}$. The (generalized fiducial) expectation would yield

$$E[h(\boldsymbol{\vartheta}) \mid \mathbf{y}] = \int_{\theta_1' \le t} r(\boldsymbol{\theta}'; \mathbf{y}) \mathrm{d}\boldsymbol{\theta}' = P(\vartheta_1 \le t | \mathbf{y}) \overset{\text{def}}{=} R_1(t). \tag{5}$$

This formulation will be useful to construct interval estimates of θ_1. For example, a lower 95% confidence interval could be obtained by inverting the function R_1 in Equation (5) at 0.95. Also, a two-sided 95% confidence interval could be similarly evaluated by inverting R_1 at 0.025 and 0.975. We remark that Equation (5) and more generally Equation (4) cannot be easily computed for most practical problems, and could be much more challenging for massive datasets.

It is very often of interest to provide a measure to summarize the evidence in the data $\mathbf{y}$ supporting the truthfulness of an assertion $A \subset \boldsymbol{\Theta}$ of the parameter space. GFI provides a straightforward way to express the amount of belief by the generalized fiducial distribution function:

$$R(A) = E[1\{\boldsymbol{\vartheta} \in A\} \mid \mathbf{y}] = \int_A r(\boldsymbol{\theta}'; \mathbf{y}) \mathrm{d}\boldsymbol{\theta}'. \tag{6}$$

This R function is a valid probability measure and, in many ways, could be viewed as a function similar to posterior distribution in the context of Bayesian inference.

3. Massive Data Problems

For massive data problems, where n is huge, the generalized fiducial density in Equation (2) could be difficult to evaluate or to obtain samples from. As mentioned before, one way to address this issue is to partition the whole dataset $\mathbf{Y}$ into K subsets $\{\mathbf{Y}_k\}_{k=1}^K$. For each k, the elements of $\mathbf{Y}_k$ are specified by an (nonempty) index set I_k via $\mathbf{Y}_k = \{Y_i, i \in I_k\}$, where $\{I_k\}_{k=1}^K$ form a partition of $\{1, \ldots, n\}$. From Equation (2), the generalized fiducial density of $\boldsymbol{\theta}$ based on observations $\mathbf{Y}_k = \mathbf{y}_k$ for the kth partition is given by

$$r_k(\boldsymbol{\theta}; \mathbf{y}_k) = \frac{f(\mathbf{y}_k; \boldsymbol{\theta}) J(\mathbf{y}_k, \boldsymbol{\theta})}{\int f(\mathbf{y}_k; \boldsymbol{\theta}') J(\mathbf{y}_k, \boldsymbol{\theta}') \mathrm{d}\boldsymbol{\theta}'} \overset{\text{def}}{=} \frac{1}{c(\mathbf{y}_k)} f(\mathbf{y}_k; \boldsymbol{\theta}) J(\mathbf{y}_k, \boldsymbol{\theta}). \tag{7}$$

Let n_k be the size of I_k. It is assumed that for all k, n_k is small enough so that samples of $\boldsymbol{\theta}$ can be generated from (7) using one single worker.

Combining Equations (2) and (7), the overall generalized fiducial density $r(\boldsymbol{\theta}; \mathbf{y})$ for the whole observed dataset $\mathbf{y}$ can be expressed as a product of generalized fiducial density $r_k(\boldsymbol{\theta}; \mathbf{y}_k)$ and the weights $\prod_{j \neq k} f(\mathbf{y}_j; \boldsymbol{\theta})$:

$$r(\boldsymbol{\theta}; \mathbf{y}) \propto \frac{J(\mathbf{y}; \boldsymbol{\theta})}{J(\mathbf{y}_k; \boldsymbol{\theta})} r_k(\boldsymbol{\theta}; \mathbf{y}_k) \prod_{j \neq k} f(\mathbf{y}_j; \boldsymbol{\theta}).$$

This formula decomposes the full density $r(\boldsymbol{\theta}; \mathbf{y})$ into parts of smaller densities $r_k(\boldsymbol{\theta}; \mathbf{y}_k)$'s. With this formula, we develop an algorithm to draw samples from the full density $r(\boldsymbol{\theta}; \mathbf{y})$ efficiently by drawing (reweighed) samples from those smaller densities $r_k(\boldsymbol{\theta}; \mathbf{y}_k)$'s. Ultimately, these samples will be used to approximate the generalized fiducial measure $R(A)$ defined in Equation (6).

3.1. Importance Sampling

Importance sampling is a general technique for approximating the expectation of a target distribution via the use of a proposal distribution (see, e.g., Geweke 1989). This subsection develops a naive version of importance sampling to approximate the generalized fiducial measure $R(A)$. The next subsection will discuss methods for improving this naive version.

For the moment consider using the subset density $r_k(\boldsymbol{\theta}; \mathbf{y}_k)$ as the proposal. An advantage of using $r_k(\boldsymbol{\theta}, \mathbf{y}_k)$ is that, it only requires a subset of data, and therefore $\mathbf{y}_k$, it would be computationally more feasible than sampling from the original generalized fiducial density $r(\cdot; \mathbf{y})$ based on the whole dataset $\mathbf{y}$.

Next, for each k, define a (un-normalized) proposal density function for $r(\boldsymbol{\theta}, \mathbf{y})$ as

$$\pi_k(\boldsymbol{\theta}) = c(\mathbf{y}_k) r(\theta; \mathbf{y}_k) = f(\mathbf{y}_k; \boldsymbol{\theta}) J(\mathbf{y}_k, \boldsymbol{\theta}). \tag{8}$$

A normalized version of $\pi_k(\boldsymbol{\theta})$ will then be used as the proposal distribution in the importance sampling algorithm. As similar to most Bayesian problems, MCMC techniques are often employed to obtain samples from this proposal.

Assume now we are able to draw T samples from $\pi_k(\boldsymbol{\theta})$ for each k. Denote the samples as $\{\boldsymbol{\theta}_{k,t}\}$ for $k = 1, \ldots, K$ and $t = 1 \ldots, T$. Also, for each k, define the importance weight function as

$$w_k(\boldsymbol{\theta}) = \frac{c(\mathbf{y}) r(\boldsymbol{\theta}; \mathbf{y})}{\pi_k(\boldsymbol{\theta})} = \frac{J(\mathbf{y}, \boldsymbol{\theta})}{J(\mathbf{y}_k, \boldsymbol{\theta})} \prod_{j \neq k} f(\mathbf{y}_j; \boldsymbol{\theta}). \tag{9}$$

Using those samples $\{\boldsymbol{\theta}_{k,t}, t = 1, \ldots, T\}$ generated from the kth subset, one can estimate $R(A)$ by $\hat{R}_k(A)$ via the usual importance sampling method

$$\hat{R}_k(A) = \frac{\sum_{t=0}^T 1\{\boldsymbol{\theta}_{k,t} \in A\} w_k(\boldsymbol{\theta}_{k,t})}{\sum_{t=0}^T w_k(\boldsymbol{\theta}_{k,t})}. \tag{10}$$

Combining all the $\hat{R}_k(A)$'s, one obtains the following improved estimate for $R(A)$:

$$\hat{R}(A) = \frac{1}{K} \sum_{k=1}^K \hat{R}_k(A). \tag{11}$$

Asymptotic normality of $\hat{R}_k(A)$ can be obtained by an application of the Markov chain central limit theorem (Jones 2004) as $T \to \infty$. This result is presented in Proposition 1. In what follows, we assume that K is fixed.

Proposition 1. If the chain $\{\boldsymbol{\theta}_{k,t}\}$ satisfies Assumption D1 in the appendix and $E\left[w_k(\boldsymbol{\vartheta})|\mathbf{y}\right]$ is finite, then the central limit theorem holds for $\hat{R}_k(A)$; that is,

$$\sqrt{T}[\hat{R}_k(A) - R(A)]|\mathbf{y} \xrightarrow{D} N(0, \sigma_k^2) \quad \text{as } T \to \infty,$$

where $\sigma_k^2 = a_k - 2R(A)c_k + R^2(A)b_k$, and

$$\begin{aligned} a_k &= \operatorname{var}_{\pi_k}[1\{\boldsymbol{\theta}_{k,0} \in A\} w_k(\boldsymbol{\theta}_{k,0})] \\ &\quad + 2\sum_{t=1}^{\infty} \operatorname{cov}_{\pi_k}[1\{\boldsymbol{\theta}_{k,0} \in A\} w_k(\boldsymbol{\theta}_{k,0}), 1\{\boldsymbol{\theta}_{k,t} \in A\} w_k(\boldsymbol{\theta}_{k,t})] \\ &\quad < \infty, \end{aligned}$$

$$b_k = \operatorname{var}_{\pi_k}[w_k(\boldsymbol{\theta}_{k,0})] + 2\sum_{t=1}^{\infty} \operatorname{cov}_{\pi_k}[w_k(\boldsymbol{\theta}_{k,0}), w_k(\boldsymbol{\theta}_{k,t})] < \infty,$$

$$\begin{aligned} c_k &= \sum_{t=0}^{\infty} \operatorname{cov}_{\pi_k}[1\{\boldsymbol{\theta}_{k,0} \in A\} w_k(\boldsymbol{\theta}_{k,0}), w_k(\boldsymbol{\theta}_{k,t})] \\ &= \sum_{t=0}^{\infty} \operatorname{cov}_{\pi_k}[1\{\boldsymbol{\theta}_{k,t} \in A\} w_k(\boldsymbol{\theta}_{k,t}), w_k(\boldsymbol{\theta}_{k,0})] < \infty. \end{aligned}$$

The proof follows the arguments from Geweke (1989) and Jones (2004), and is hence omitted to save space. This proposition guarantees that $\hat{R}_k(A)$ is a reasonable approximation of $R(A)$ as long as the proposal is chosen wisely. Furthermore, by averaging the $\hat{R}_k(A)$'s, the variability from the MCMC samples in $\hat{R}(A)$ is further reduced, resulting in an more accurate estimate for $R(A)$.

3.2. Improving Importance Weights

Amongst other factors, the overall speed of the above importance sampling algorithm relies on how fast one could compute the weights (9). The first term $J(\mathbf{y}, \boldsymbol{\theta})/J(\mathbf{y}_k, \boldsymbol{\theta})$ is the lengthy term to compute, as it involves the whole dataset $\mathbf{y}$. However, we would expect this ratio to be close to a constant as a function of θ, when compared to the likelihood function. Heuristically, speaking and using the notation immediately after Equation (3), when the L_∞ version is used the Jacobian is a U statistics that should converge to $E(\det A_{\mathbf{i}})$, and when the L_2 version is used each entry of the matrix $A^\top A$ is a sum of n numbers and the matrix $n^{-1}A^\top A$ should converge to its expectation by law of large numbers. In either case when n is large, both $J(\mathbf{y},\boldsymbol{\theta})$ and $J(\mathbf{y}_k,\boldsymbol{\theta})$ should be close to the same limiting function and their ratio should be close to a constant; this is particularly true when $\mathbf{y}_k$ is a representative sample of $\mathbf{y}$. We make this precise in Proposition 3.

Motivated by this, we propose approximating the original weight function (9) by ignoring the first term, which gives the following improved weight function

$$\tilde{w}_k(\boldsymbol{\theta}) = \frac{J(\mathbf{y}_k,\boldsymbol{\theta})}{J(\mathbf{y},\boldsymbol{\theta})} w_k(\boldsymbol{\theta}) = \prod_{j\neq k} f(\mathbf{y}_j;\boldsymbol{\theta}). \tag{12}$$

With this $R(A)$ can be estimated, in a similar fashion as in Equation (10), with

$$\tilde{R}_k(A) = \frac{\sum_{t=0}^{T} 1\{\boldsymbol{\theta}_{k,t} \in A\}\tilde{w}_k(\boldsymbol{\theta}_{k,t})}{\sum_{t=0}^{T} \tilde{w}_k(\boldsymbol{\theta}_{k,t})}. \tag{13}$$

We have the following proposition immediately.

Proposition 2. If the chain $\{\boldsymbol{\theta}_{k,t}\}$ satisfies Assumption D1 in the appendix and if $E\left[\frac{\tilde{w}_k^2(\boldsymbol{\vartheta})}{w_k(\boldsymbol{\vartheta})}|\mathbf{y}\right]$ is finite, then

$$\sqrt{T}[\tilde{R}_k(A) - R_k^*(A)]|\mathbf{y} \xrightarrow{D} N(0,\sigma_k^2) \quad \text{as } T \to \infty,$$

where

$$R_k^*(A) = E\left[1\{\boldsymbol{\theta}_{k,t} \in A\}\frac{J(\mathbf{y}_k,\boldsymbol{\vartheta})}{J(\mathbf{y},\boldsymbol{\vartheta})}|\mathbf{y}\right] \Big/ E\left[\frac{J(\mathbf{y}_k,\boldsymbol{\vartheta})}{J(\mathbf{y},\boldsymbol{\vartheta})}|\mathbf{y}\right],$$

$$\sigma_k^2 = (a_k - 2R_k^*(A)c_k + R_k^*(A)^2 b_k) \Big/ \left\{E\left[\frac{J(\mathbf{y}_k,\boldsymbol{\vartheta})}{J(\mathbf{y},\boldsymbol{\vartheta})}|\mathbf{y}\right]\right\}^2,$$

and a_k, b_k and c_k are defined in Proposition 1.

The major idea behind the proof of Proposition 2 is very similar to that of Proposition 1, and therefore is omitted for brevity. This proposition indicates that $\tilde{R}_k(A)$ is converging to $R_k^*(A)$ as $T \to \infty$ and hence $\tilde{R}_k(A)$ is a biased estimator of $R(A)$. This bias is introduced when $w_k(\boldsymbol{\theta})$ are replaced by $\tilde{w}_k(\boldsymbol{\theta})$ in order to obtain higher computational speed.

The next proposition shows that the bias in $\tilde{R}_k(A)$ is asymptotically negligible, providing a theoretical support of the use of $\tilde{w}_k(\boldsymbol{\theta})$. The convergence in probability below is with respect to the distribution of the data $\mathbf{y}$. The proof is given in the appendix.

Proposition 3. Let $\hat{\boldsymbol{\theta}}_n$ be the maximum likelihood estimate of $\boldsymbol{\theta}$. Suppose Assumptions E1 and E2 in the appendix hold. Then as $n \to \infty$,

$$\sqrt{n}E\left[\left|\frac{J(\mathbf{y}_k,\boldsymbol{\vartheta})}{J(\mathbf{y},\boldsymbol{\vartheta})} - \frac{J(\mathbf{y}_k,\hat{\boldsymbol{\theta}}_n)}{J(\mathbf{y},\hat{\boldsymbol{\theta}}_n)}\right| |\mathbf{y}\right] \xrightarrow{P} 0 \tag{14}$$

and

$$R_k^*(A) = R(A) + o_p(n^{-1/2}). \tag{15}$$

Now we are ready to present our main theoretical result.

Algorithm 1 Direct Implementation

```
1. Partition the data $\mathbf{y}$ into $K$ subsets $\mathbf{y}_1, \ldots, \mathbf{y}_K$. Each subset $\mathbf{y}_k$
   becomes the input for one of the $K$ parallel jobs.
2. For $k = 1, \ldots, K$, the $k$th worker generates a sample of $\boldsymbol{\vartheta}$ of
   size $T$ from (8) and returns the result to the main node.
3. The collected samples are broadcasted to all workers and each
   worker computes its relevant portion of $\tilde{w}_k(\boldsymbol{\vartheta})$ in (12) and
   returns the result to the main node.
4. Combine the results from the workers to obtain $\tilde{w}_k(\boldsymbol{\vartheta})$ and
   calculate $\tilde{R}_k(A)$ using (13).
5. Average all the $\tilde{R}_k(A)$'s and obtain the final estimate $\tilde{R}(A)$ as
   in (16).
```

Proposition 4. Under the conditions of Propositions 2 and 3, we have, for all $\varepsilon > 0$,

$$P\left\{\sqrt{n}|\tilde{R}_k(A) - R(A)| > \varepsilon \,\Big|\, \mathbf{y}\right\} \xrightarrow{P} 0$$

as $T \to \infty$, $n \to \infty$ and $n/T \to 0$.

Proposition 4 indicates that the fiducial probability of an assertion set $A \subset \boldsymbol{\Theta}$ can be approximated by $\tilde{R}_k(A)$ with high accuracy. Note that this asymptotic result holds when both n and T go to infinity, with T diverging at a faster rate than n. These conditions are required since we have to ensure that the approximation error due to the importance sampling procedure is comparable to the bias introduced in the weight function $\tilde{w}_k(\boldsymbol{\theta})$ in Equation (12). The proof of this proposition is given in the appendix.

Since from Proposition 4 each $\tilde{R}_k(A)$ is a consistent estimator, it is natural to define our final estimator of $R(A)$ as $\tilde{R}(A)$

$$\tilde{R}(A) = \frac{1}{K}\sum_{k=1}^{K}\tilde{R}_k(A). \tag{16}$$

Note that the averaging operation further reduces the variability in the estimation in $\tilde{R}(A)$. Once $\tilde{R}(A)$ is obtained, it can be used to conduct statistical inference about the parameters of interest, in a similar manner as with a posterior distribution in the Bayesian context.

3.3. *Practical Algorithms*

This subsection presents two practical algorithms that implement the above results. The first one is a straightforward and direct implementation, and is listed in Algorithm 1.

The effectiveness of Algorithm 1 depends on the importance weights and the effective sample sizes of the importance samplers. For an implementation of K workers and each worker stores n_k observations, the relative efficiency (Kong 1992; Liu 1996) for each worker is approximately in the order of $\exp(-\tau K)\}$, where the constant τ depends on the likelihood being considered. For large values of K, the corresponding number of fiducial samples has to increase exponentially in order to achieve the same estimation accuracy. To address this issue, we propose a modified algorithm, which is given as Algorithm 2.

Algorithm 2 Improved Implementation: Method G

```
1. Partition the data $\mathbf{y}$ into $K$ subsets $\mathbf{y}_1, \dots, \mathbf{y}_K$. Each subset $\mathbf{y}_k$ becomes the input for one of the $K$ parallel jobs. Here $K$ is chosen as a power of 2.
2. For $k = 1, \dots, K$, the $k$-th worker generates a sample of $\boldsymbol{\vartheta}$ of size $T$ from (8) and returns the result to the main node.
3. Repeat the following until one subset is left:
    (a) For any two subsets, say $k_i$ and $k_j$,
        i. Compute parallelly the weights as $\widetilde{w}_{k_i}(\boldsymbol{\vartheta}) = f(\mathbf{y}_{k_j}; \boldsymbol{\vartheta})$.
        ii. Return the weights to the main node.
        iii. At the main node, resample the sample of $\boldsymbol{\vartheta}$ from subset $k_i$ with weights $\widetilde{w}_{k_j}$ and resample the sample of $\boldsymbol{\vartheta}$ from subset $k_j$ with weights $\widetilde{w}_{k_j}$.
        iv. Merge the two samples in the previous step into form a single sample of $\boldsymbol{\vartheta}$.
        v. Group the subsets $\mathbf{y}_{k_i}$ and $\mathbf{y}_{k_j}$ together.
    (b) Repeat (a) with another pair of subsets until there are only half the original number of subsets remain.
4. With the combined sample of $\boldsymbol{\vartheta}$, compute $\tilde{R}(A)$ by using $\tilde{R}_k(A) = T^{-1} \sum_{t=0}^{T} 1\{\boldsymbol{\vartheta}_{k,t} \in A\}$ and (16).
```

With Algorithm 2, the effective number of workers is reduced to $\log K$ as the number of operations has decreased logarithmically. Therefore the relative efficiency of the importance samplers increases to some polynomial order of K. Algorithm 2 also reduces the number of evaluations of the likelihood functions since the evaluations are now only required by the merging fiducial samples. In contrast, Algorithm 1, the evaluations are required by all the workers for each of the fiducial samples.

In all the numerical work to be reported below, only Algorithm 2 was used.

4. Simulations

To investigate the feasibility and the empirical performance of the proposed approach, we consider simulated data from two different models: a mixture of two normal distributions and a linear regression model with p covariates and Cauchy distributed errors. For each of the models, we will construct fiducial confidence intervals for the parameters. Their nominal and empirical coverage will be presented. We will vary the simulation settings using different sample sizes n, different number of workers K, and also different number of parameters.

In each simulation setup, we first generate n observations from the underlying model and then divide them randomly into K groups. Each group of observations will be sent to a parallel worker for further processing. Each of the K workers will then perform a MCMC procedure to sample from T particles using Equation (8), while $K = 1$ corresponds to the GFI on the full dataset. In our simulation, the Metropolis–Hastings algorithm is implemented for this purpose and T is chosen to be 10,000 for all cases. Each setting is then repeated 100 times to obtain the empirical coverages for the one-sided fiducial confidence intervals. The widths of fiducial confidence intervals are also reported in the Cauchy Regression setting.

4.1. Mixture of Normals

The density of Y is $f_Y(y) = \gamma\phi(y; \mu_1, \sigma) + (1-\gamma)\phi(y; \mu_2, \sigma)$, where $\phi(y; \mu, \sigma)$ is the normal density with mean μ and variance σ^2. For simplicity, we assume that $\sigma = 1$ is known. The value of (μ_1, μ_2, γ) is $(-1, 1, 0.6)$. Note $\mu_1 < \mu_2$ so identifiability is ensured. Three values of $n = 10^5, 2 \times 10^5$, and 4×10^5, and six values of $K = 1, 2, 4, 6, 16$, and 32 are used.

For the cases $n = 10^5$ and $n = 4 \times 10^5$, the empirical coverages for all $100(1-\alpha)$% lower sided fiducial confidence intervals for the parameters μ_1 and γ are shown, respectively, in Figures 1 and 2. The dotted lines are the theoretical confidence

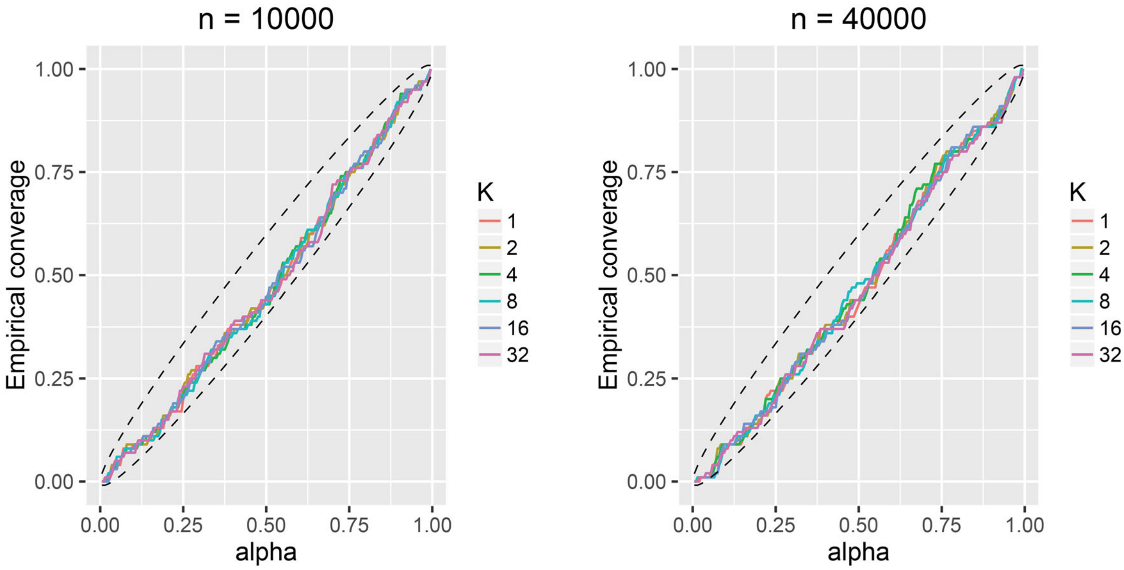

Figure 1. The empirical coverages for the lower sided fiducial confidence intervals for μ_1 for different number of observations and number of workers. The results for $n = 2 \times 10^5$ and μ_2 are similar and hence omitted.

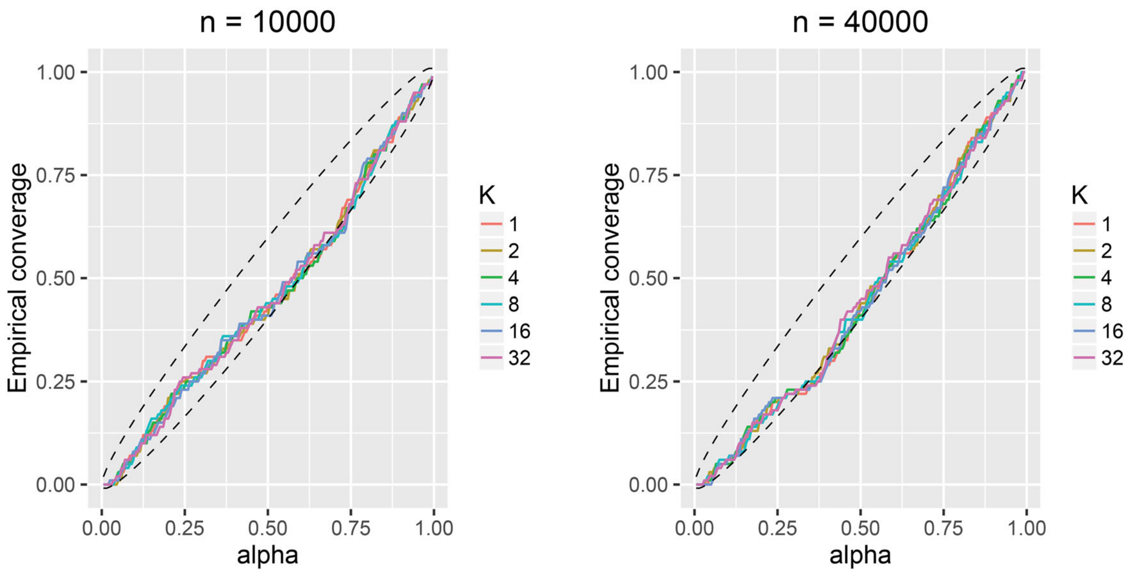

Figure 2. Similar to Figure 1 but for the parameter γ.

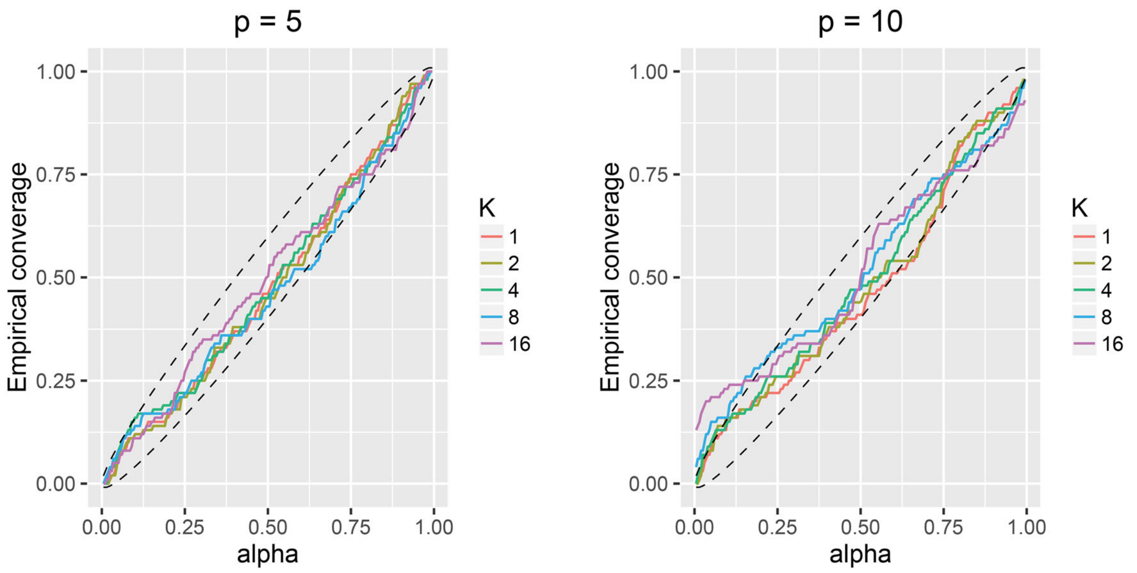

Figure 3. The empirical coverages for the lower sided fiducial confidence intervals for β_1 for different number of covariates and number of workers. The results for $p = 7$ and for β_2 and β_3 are similar and hence omitted.

interval for the empirical coverages: $\alpha \pm 1.96\sqrt{\alpha(1-\alpha)/100}$. From these figures, one can see that the proposed method performs very well with empirical coverages agreeing with the nominal coverages at all levels.

4.2. *Cauchy Regression*

The model is $Y = \beta_0 + \boldsymbol{\beta}\mathbf{X} + \sigma W$, where $\beta_0 \in \Re$, $\boldsymbol{\beta} \in \Re^p$ and $\sigma > 0$. The error distribution of W is standard Cauchy and the design matrix $\mathbf{X}$ is multivariate normal with zero mean, unit variance and pairwise correlation $\rho = 0.1$. The following parameter values are used: $\sigma = 1$, $\beta_0 = 0$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \ldots) = (1, 1, 1, 0, 0, \ldots)$; that is, all slope coefficients are zero except the first three.

The empirical coverages for the $100(1-\alpha)\%$ lower sided fiducial confidence intervals for the parameters β_1 and β_4 are shown, respectively, in Figures 3 and 4. The number of observations is fixed at $n = 10^5$ while $K = 1, 2, 4, 8$ and 16, and $p = 5, 7$ and 10. Similar to the previous subsection, the dotted lines are the theoretical confidence interval for the empirical coverages: $\alpha \pm 1.96\sqrt{\alpha(1-\alpha)/100}$. As with the previous subsection, the proposed method produced very good results in terms of empirical coverage. Two-sided 95% fiducial confidence intervals are also computed from each simulated dataset. The median widths and their standard deviations are reported in Table 1. The results suggest that the proposed algorithm produces not only confidence intervals with correct empirical coverage, but also it produces confidence intervals with widths similar to that produced by the full data algorithm. The table also confirms our expectation that the width of the confidence intervals is more variable when K increases because of the extra randomness due to importance sampling.

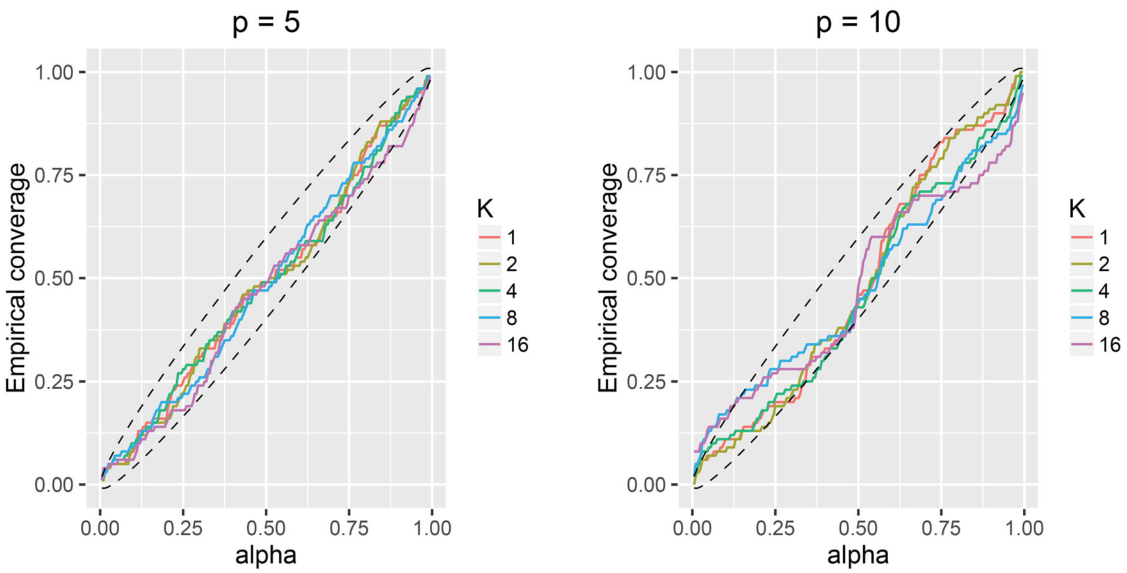

Figure 4. Similar to Figure 3 but for the parameter β_4. The results for $\beta_j, j > 4$ are similar and hence omitted.

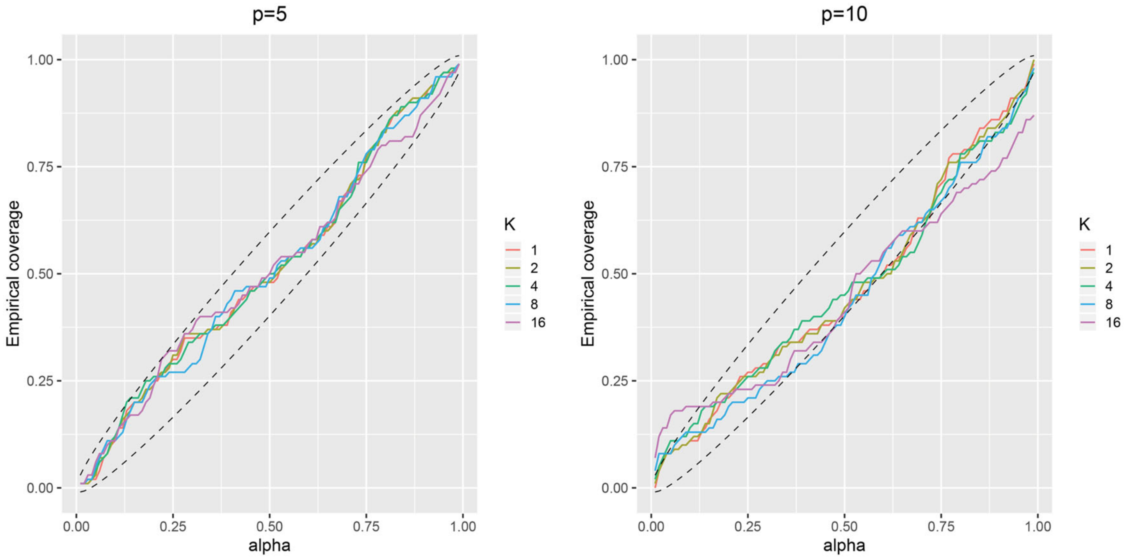

Figure 5. Similar to Figure 3 but for $n = 10^3$.

Table 1. Median width of 95% fiducial confidence intervals for the Cacuhy regression model ($n = 10^5$ and $p = 5$).

	β_1		β_4	
K	Median	sd	Median	sd
1	0.0477	8.94e-04	0.0472	8.68e-04
2	0.0475	1.09e-03	0.0474	1.15e-03
4	0.0475	1.68e-03	0.0474	1.65e-03
8	0.0473	3.24e-03	0.0471	3.42e-03

Table 2. Median width of 95% fiducial confidence intervals for the Cacuhy regression model ($n = 10^3$ and $p = 5, 10$).

		β_1		β_4	
p	K	Median	sd	Median	sd
5	1	0.1539	0.0089	0.1510	0.0090
	2	0.1531	0.0095	0.1513	0.0098
	4	0.1534	0.0106	0.1518	0.0102
	8	0.1527	0.0151	0.1513	0.0166
	16	0.1531	0.0192	0.1517	0.0238
10	1	0.1560	0.0103	0.1556	0.0092
	2	0.1552	0.0129	0.1532	0.0136
	4	0.1539	0.0219	0.1523	0.0205
	8	0.1541	0.0420	0.1542	0.0405
	16	0.1503	0.0627	0.1644	0.0659

Some additional simulations were conducted to examine the performance of the proposal when n is relatively smaller ($= 10^3$) to show the effects of p ($= 5, 10$) and K ($= 1, 2, 4, 6, 8,$ and 16). The empirical coverages of this study are reported in Figure 5 and are very similar to Figures 3 and 4 except when $p = 10$ and $K = 16$. When $K = 16$, each worker only receives $\approx 1000/16 = 60$ observations. The poor result would be explained by the small sample sizes and the relatively large dimension $p = 10$. The median widths and their standard deviations of two-sided 95% fiducial confidence intervals are shown in Table 2. Except for the case of ($p = 10, K = 16$), it shows that the our

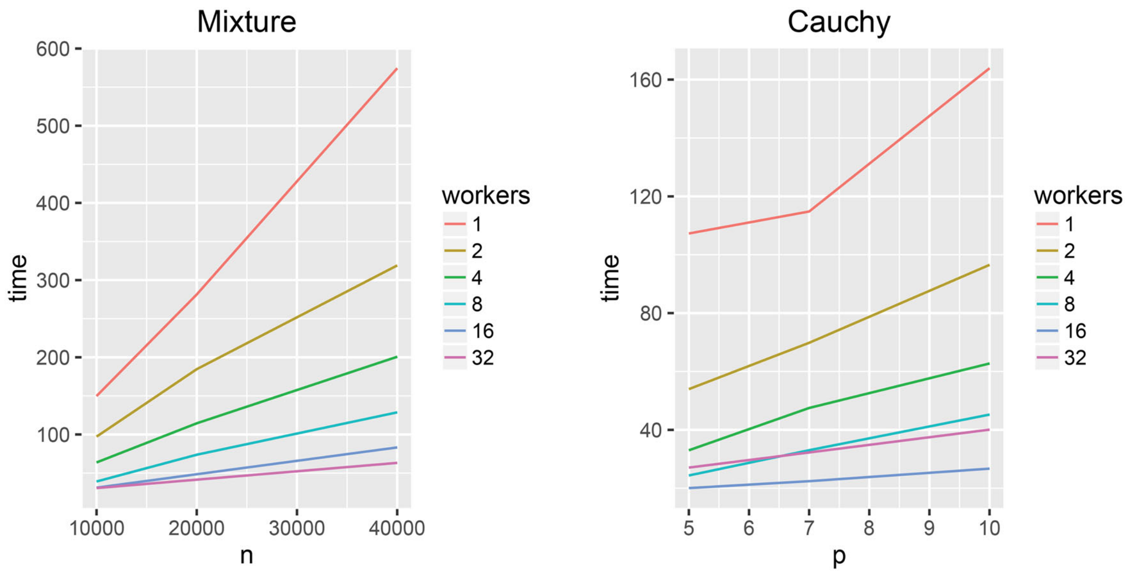

Figure 6. The elapsed time required for the normal mixture and Cauchy models for different number of workers *K*.

new algorithm is able to produce confidence intervals of similar quality as the full data GFI, in both coverages and widths.

4.3. Efficiency of Different Worker Sizes

One primary goal of this article is to develop a scalable solution to reduce the computational time required in performing generalized fiducial inference. The computational times in the above normal mixture and Cauchy models are reported in Figure 6. It can be seen that the proposed Method G is more time-efficient than the full data version of GFI when the sample size (for the normal mixture model) or the number of covariates (for the Cauchy model) increases.

Intuitively, one would believe that more workers would lead to a larger reduction of computational time. This is partially true, as if the number of workers exceeds the maximum beneficial optimal value, the total computational time and statistical optimality may rebound; see Cheng and Shang (2015) for an interesting discussion. A major reason is that there is a tradeoff between the actual computation cost and the cost for data transfer and communication among the workers in this divide-and-conquer strategy. For the Cauchy example, the total computational cost for the case of 32 workers was "unsurprisingly" more than that of the case of 16 workers. It is probably because more time was spent in data manipulation and allocation than in the real computations.

In the above simulation studies, different models and different error distributions were carefully chosen with the hope to represent some commonly encountered practical scenarios. However, just as any other simulation studies, the above numerical experiments by no means are sufficient to cover all cases that one may encounter in practice, and therefore caution must be exercised in drawing conclusions from the empirical results. Despite this, the following conclusion is appropriate. GFI can be parallelized to handle massive data problems with Algorithm 2 and the resulting statistical inferences are asymptotically correct. The performance of Algorithm 2 depends on the total sample size and worker sample sizes. Simulation results show that the fiducial confidence intervals produced by the algorithm have very reliable and attractive frequentist properties.

5. Data Analysis: Solar Flares

In this section, the methodology developed above is applied to help understand the occurrences of solar flares. The data were collected by the instrument atmospheric imaging assembly (AIA) mounted on the solar dynamics observatory (SDO). As stated in the official NASA website, SDO was designed to help study the influence of the Sun on the Earth and near-Earth space. SDO was launched in 2010.

The instrument AIA captures images of the Sun in eight different wavelengths every 12 sec, providing more than 70,000 high-resolution images per day; see Figure 7 for two examples. The image size is 4096 × 4096 pixels, which amounts to a total of 1.5 terabytes compressed data per day. An uncompressed and preprocessed version of the data can be obtained from Schuh et al. (2013). Here, each image was partitioned into 64 × 64 squared and equi-sized sub-images, each consists of 64 × 64 pixels. For each sub-image, 10 summary statistics were computed, such as the average and the standard deviation of the pixel values.

A solar flare is a sudden eruption of high-energy radiation from the Sun's surface, which could cause disturbances on communication and power systems on the Earth. In those images captured by AIA, such solar flares are characterized by extremely bright spots; see the right panel of Figure 7 for an example.

Wandler and Hannig (2012) provided a solution for estimating extremes using GFI. Their approach is based on modeling values over large threshold using a generalized Pareto distribution (Pickands III 1975). The data-generating equation is

$$Y_i = a + \frac{\sigma}{\gamma}\left(U_i^{-\gamma} - 1\right), \; i = 1, \dots, n,$$

where U_i are iid $U(0, 1)$, a is a known threshold, σ is a scale parameter, and γ is a shape parameters. The corresponding

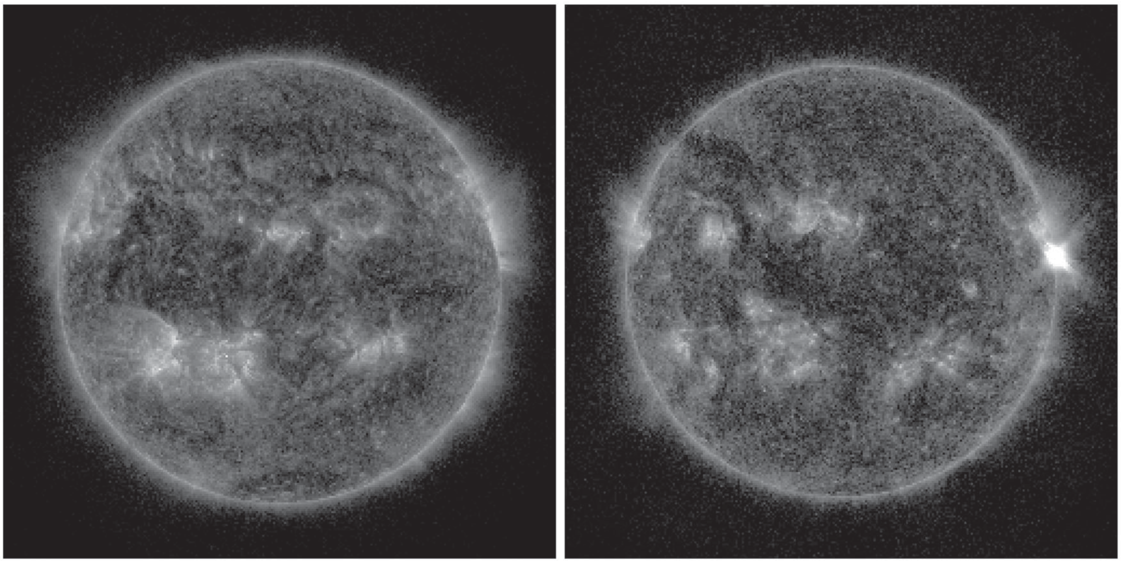

Figure 7. Two images of the Sun captured by AIA. There is a solar flare occurring in the right image, near the right end of the equator of the Sun. The white spot intensity has a value of 253 on a 8-bit scale from 0 (black) to 255 (white).

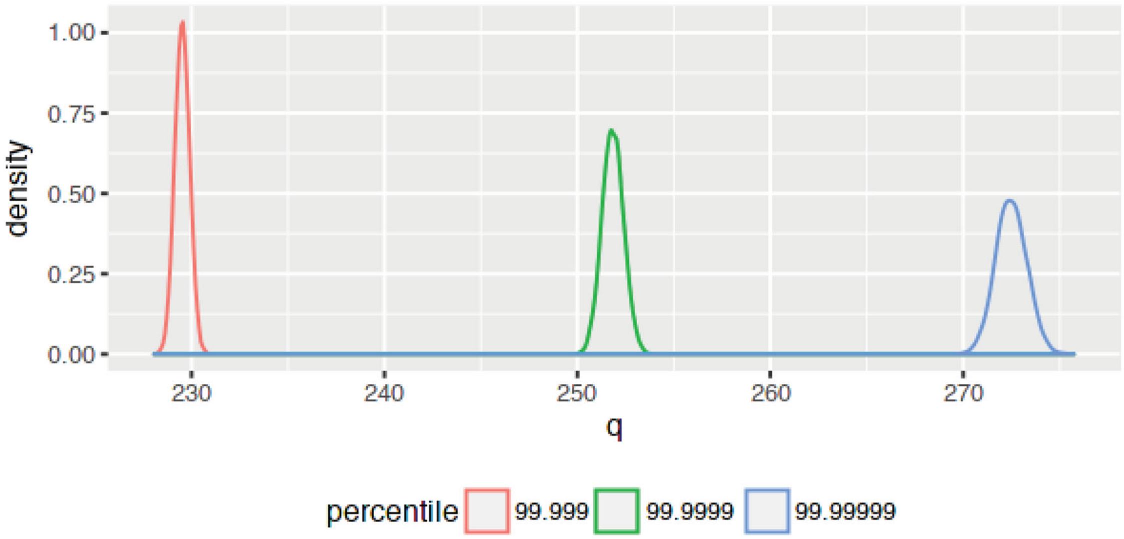

Figure 8. Kernel density estimates of the fiducial densities of the brightness of 99.999, 99.9999 and 99.99999 percentiles (q) solar flare events. One can see that an observed value of 253 roughly corresponds to 99.9999% (1 in a million). Values over 255 indicate events with brightness that is higher than the resolution of the instrument.

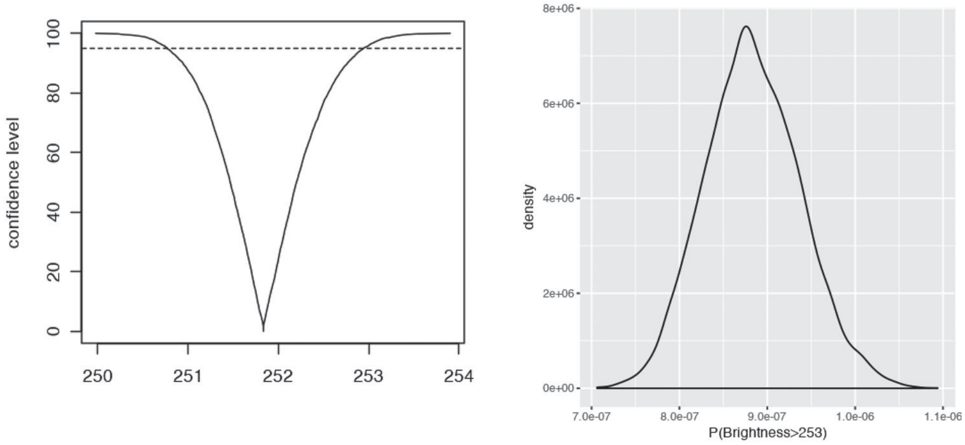

Figure 9. Left: Confidence curve for the 99.9999 percentile. Right: Fiducial probability of brightness greater than 253.

likelihood and Jacobian (3) are

$$f(\mathbf{y};\sigma,\gamma)=\sigma^{-n}\prod_{i=1}^{n}\left(1+\frac{\gamma}{\sigma}(x_i-a)\right)^{-1-1/\gamma}I_{(a,\infty)}(x_i),$$

$$J(y,\sigma,\gamma)=\gamma^{-2}D\begin{pmatrix} x_1-a & (1+\frac{\gamma}{\sigma}(x_1-a))\log(1+\frac{\gamma}{\sigma}(x_1-a)) \\ \vdots & \vdots \\ x_n-a & (1+\frac{\gamma}{\sigma}(x_n-a))\log(1+\frac{\gamma}{\sigma}(x_n-a)) \end{pmatrix}.$$

Wandler and Hannig (2012) provided an MCMC algorithm for sampling from the fiducial distribution of the generalized Pareto parameters for small datasets. A sample from fiducial distribution for a β-percentile is then obtained by plugging the fiducial samples (σ^*,γ^*) into the inverse distribution function

$$q^*=a+\frac{\sigma^*}{\gamma^*}\left((1-\beta)^{-\gamma^*}-1\right).$$

For this solar dataset, we used the averaged pixel values (summary statistics P2) computed from Schuh et al. (2013) and the proposed method G to parallelize the computations. The number of images was 7697. Method G combined sample of (σ^*,γ^*) to generate fiducial distribution for the $\beta = 99.999\%$, 99.9999%, and 99.99999% percentiles of the brightness. Figure 8 reports the Gaussian kernel-based estimates for the fiducial densities of these the brightness. These densities are shown in can help astronomers determine the frequency, predict the occurrences of the solar flares, and understand the limitations of their instruments. Figure 9 displays the confidence curve for the 99.9999 percentile for the solar flare brightness. The 95% confidence interval is (250.8, 253.0) and a solar flare of brightness in this range is likely to happen with 1 in a million chance. The fiducial probability of brightness greater than 253 is also computed and displayed in Figure 9.

The simulations were run on UCDavis Department of Statistics computer cluster, each node of the cluster is equipped with a 32-core AMD Opteron(TM) Processor 6272. The program took about 15 sec to finish the fiducial sample generation process when 32 workers are in work and it took about 80 sec when only 4 workers are in place. The number of MCMC runs for each workers was 50,000.

6. Conclusion and Discussion

In this article, generalized fiducial inference is paired with importance sampling to develop a method for the distributed analysis of massive datasets. In addition to point estimates, the resulting method is also capable of producing uncertainty measures for such quantities. Another attractive feature of the method is that it only requires minimal communications amongst workers. Via mathematical calculations and numerical experiments, the method is shown to enjoy excellent theoretical and empirical properties.

The proposed method assumes the sub-sample in each worker is a random sample from the original dataset. Therefore, a useful extension of the current work is to relax this assumption. Another important extension is to allow for heterogeneity that is a common feature of massive datasets that are obtained from potentially disparate sources. One possible computationally efficient approach for handling this issue was proposed in the "small data" inter-laboratory comparison context by Hannig et al. (2018). Their idea seems especially promising in the massive data context if one could ensure that the within each subset is relatively homogeneous, while the data between subsets is potentially heterogeneous.

Appendix A. Technical Details

This appendix provides technical details.

Assumptions

We begin with a set of assumptions which allow us to work on the theories.

We start by listing the standard assumptions sufficient to prove that the maximum likelihood estimators are asymptotically normal (Lehmann and Casella 1998).

(A0) The distributions $P_{\boldsymbol{\theta}}$ are distinct.

(A1) The set $\{y: f(y|\boldsymbol{\theta})>0\}$ is independent of the choice of $\boldsymbol{\theta}$.

(A2) The data $\mathbf{Y}=\{Y_1,\dots,Y_n\}$ are iid with probability density $f(\cdot|\boldsymbol{\theta})$.

(A3) There exists an open neighborhood about the true parameter value $\boldsymbol{\theta}_0$ such that all third partial derivatives $(\partial^3/\partial\theta_i\partial\theta_j\partial\theta_k)f(\mathbf{y}|\boldsymbol{\theta})$ exist in the neighborhood, denoted by $B(\boldsymbol{\theta}_0,\delta)$.

(A4) The first and second derivatives of $L(\boldsymbol{\theta},y)=\log f(y|\boldsymbol{\theta})$ satisfy

$$E_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial\theta_j}L(\boldsymbol{\theta},y)\right]=0$$

and

$$\begin{aligned}I_{j,k}(\boldsymbol{\theta})&=E_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial\theta_j}L(\boldsymbol{\theta},y)\cdot\frac{\partial}{\partial\theta_k}L(\boldsymbol{\theta},y)\right]\\&=-E_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}L(\boldsymbol{\theta},y)\right].\end{aligned}$$

(A5) The information matrix $I(\boldsymbol{\theta})$ is positive-definite for all $\boldsymbol{\theta}\in B(\boldsymbol{\theta}_0,\delta)$

(A6) There exists functions $M_{jkl}(\mathbf{y})$ such that

$$\sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_0,\delta)}\left|\frac{\partial^3}{\partial\theta_j\partial\theta_k\partial\theta_l}L(\boldsymbol{\theta},y)\right|\le M_{j,k,l}(y)\quad\text{and}$$

$$E_{\boldsymbol{\theta}_0}M_{j,k,l}(Y)<\infty.$$

Next, we state conditions sufficient for the Bayesian posterior distribution to be close to that of the MLE (van der Vaart 1998; Ghosh and Ramamoorthi 2003). The prior used is the limiting fiducial prior Let $\pi(\boldsymbol{\theta})=E_{\boldsymbol{\theta}_0}J_0(Y_0,\boldsymbol{\theta})$ and $L_n(\boldsymbol{\theta})=\sum L(\boldsymbol{\theta},Y_i)$

(B1) For any $\delta>0$, there exists $\epsilon>0$ such that

$$P_{\boldsymbol{\theta}_0}\left\{\sup_{\boldsymbol{\theta}\notin B(\boldsymbol{\theta}_0,\delta)}\frac{1}{n}\left(L_n(\boldsymbol{\theta})-L_n(\boldsymbol{\theta}_0)\right)\le-\epsilon\right\}\to 1$$

(B2) $\pi\,(\boldsymbol{\theta})$ is positive at $\boldsymbol{\theta}_0$

Finally, we state assumptions on the Jacobian function. Recall $\pi(\boldsymbol{\theta})=E_{\boldsymbol{\theta}_0}J_0(X_0,\boldsymbol{\theta})$.

(C1) For any $\delta>0$

$$\inf_{\boldsymbol{\theta}\notin B(\boldsymbol{\theta}_0,\delta)}\frac{\min_{i=1\dots n}L(\boldsymbol{\theta},Y_i)}{|L_n(\boldsymbol{\theta})-L_n(\boldsymbol{\theta}_0)|}\xrightarrow{P_{\boldsymbol{\theta}_0}}0,$$

where $L_n\,(\boldsymbol{\theta})=\sum_{i=1}^{n}\log f\left(y_i;\boldsymbol{\theta}\right)$ and $B\,(\boldsymbol{\theta}_0,\delta)$ is a neighborhood of diameter δ centered at $\boldsymbol{\theta}_0$.

(C2) The Jacobian function $\binom{n}{p}^{-1} J(\mathbf{Y}, \boldsymbol{\theta}) \overset{\text{a.s.}}{\to} \pi(\boldsymbol{\theta})$ uniformly on compacts in $\boldsymbol{\theta}$.

(D1) The MCMC chain $\{\boldsymbol{\theta}_{k,t}\}$ is an ergodic Markov chain with marginal density π_k defined in (8) and satisfying at least one of the followings:

(a) geometrically ergodic, or
(b) uniformly ergodic, or
(c) polynomially ergodic of order $m > 1$ with $E_{\pi_k} M(\boldsymbol{\theta}) < \infty$, where M is defined in Equation (3) of Jones (2004).

Moreover, if $k \neq k'$, chains from different workers, say $\{\boldsymbol{\theta}_{k,t}\}$ and $\{\boldsymbol{\theta}_{k'}^{(t)}\}$, are independent given $\mathbf{y}$.

(E1) Let $u_k(\mathbf{y}, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \frac{J(\mathbf{y}_k, \boldsymbol{\theta})}{J(\mathbf{y}, \boldsymbol{\theta})}$, there exists $U(\mathbf{y})$ s.t. $u_k(\mathbf{y}, \boldsymbol{\theta}) \leq U(\mathbf{y})$ for all $\boldsymbol{\theta}$ with probability tending to 1.

(E2) $\int_{\mathbb{R}^p} |t| f_{\sqrt{n}(\boldsymbol{\vartheta} - \hat{\boldsymbol{\theta}})}(t) \mathrm{d}t \xrightarrow{P} \int_{\mathbb{R}^p} |t| \phi(t; 0, I^{-1}(\boldsymbol{\theta}_0)) \mathrm{d}t$, where $f_{\sqrt{n}(\boldsymbol{\vartheta} - \hat{\boldsymbol{\theta}})}$ is the scaled generalized fiducial density and ϕ is the multivariate normal density function.

Proofs

Proof of Proposition 3. We first consider

$$
\begin{aligned}
&\sqrt{n} E\left[\left|\frac{J(\mathbf{y}_k, \boldsymbol{\vartheta})}{J(\mathbf{y}, \boldsymbol{\vartheta})} - \frac{J(\mathbf{y}_k, \hat{\boldsymbol{\theta}}_n)}{J(\mathbf{y}, \hat{\boldsymbol{\theta}}_n)}\right| \Big| \mathbf{y}\right] \\
&= \int_{\Xi} \sqrt{n} \left|\frac{J(\mathbf{y}_k, \boldsymbol{\theta})}{J(\mathbf{y}, \boldsymbol{\theta})} - \frac{J(\mathbf{y}_k, \hat{\boldsymbol{\theta}}_n)}{J(\mathbf{y}, \hat{\boldsymbol{\theta}}_n)}\right| r(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\
&= \int_{\mathbb{R}^p} \sqrt{n} \left|\frac{J(\mathbf{y}_k, \hat{\boldsymbol{\theta}}_n + \frac{t}{\sqrt{n}})}{J(\mathbf{y}, \hat{\boldsymbol{\theta}}_n + \frac{t}{\sqrt{n}})} - \frac{J(\mathbf{y}_k, \hat{\boldsymbol{\theta}}_n)}{J(\mathbf{y}, \hat{\boldsymbol{\theta}}_n)}\right| f_{\sqrt{n}(\boldsymbol{\vartheta} - \hat{\boldsymbol{\theta}}_n)}(t) \mathrm{d}t \\
&= \int_{\mathbb{R}^p} |u_k(\mathbf{y}, \hat{\boldsymbol{\theta}}_n + \lambda_t t/\sqrt{n})| |t| f_{\sqrt{n}(\boldsymbol{\vartheta} - \hat{\boldsymbol{\theta}}_n)}(t) dt \quad \text{where } 0 \leq \lambda_t \leq 1 \\
&= \int_{\mathbb{R}^p} |u_k(\mathbf{y}, \hat{\boldsymbol{\theta}}_n + \lambda_t t/\sqrt{n})| |t| \phi(t; 0, I^{-1}(\boldsymbol{\theta}_0)) \mathrm{d}t \\
&\quad + \int_{\mathbb{R}^p} |u_k(\mathbf{y}, \hat{\boldsymbol{\theta}}_n + \lambda_t t/\sqrt{n})| |t| \left[f_{\sqrt{n}(\boldsymbol{\vartheta} - \hat{\boldsymbol{\theta}}_n)}(t) - \phi(t; 0, I^{-1}(\boldsymbol{\theta}_0))\right] \mathrm{d}t
\end{aligned}
$$

For the first integral, since $u_k(\mathbf{y}, \hat{\boldsymbol{\theta}}_n + \lambda_t t/\sqrt{n}) \xrightarrow{P} 0$ as $n \to \infty$ and the integrand is dominated, by dominated coverage theorem, it goes to 0 in probability. For the second integral, since u_k is bounded and $\int_{\mathbb{R}^p} |t| f_{\sqrt{n}(\boldsymbol{\vartheta} - \hat{\boldsymbol{\theta}})}(t) \mathrm{d}t \xrightarrow{P} \int_{\mathbb{R}^p} |t| \phi(t; 0, I^{-1}(\boldsymbol{\theta}_0)) \mathrm{d}t$, it also goes to 0 in probability. Finally, Equation (15) follows directly from the definition of $R_k^*(A)$ and Equation (14).

The proposition can be immediately relaxed for the case $u_k(\mathbf{y}, \cdot)$ is bounded with some polynomial in $\boldsymbol{\theta}$ with probability tending to 1. To do this, we have to replace assumption (E2) by a similar condition with higher moment. □

Proof of Proposition 4. First, for $\varepsilon > 0$, consider

$$
\begin{aligned}
&P\left\{\left|\sqrt{n}\frac{d_k}{T}\left[\sum_{t=0}^{T} \tilde{w}_k(\boldsymbol{\theta}_{k,t}) - \frac{J(\mathbf{y}_k, \hat{\boldsymbol{\theta}}_n)}{J(\mathbf{y}, \hat{\boldsymbol{\theta}}_n)} \sum_{t=0}^{T} w_k(\boldsymbol{\theta}_{k,t})\right]\right| > \varepsilon \Big| \mathbf{y}\right\} \\
&= P\left\{\left|\frac{d_k}{T} \sum_{t=0}^{T} w_k(\boldsymbol{\theta}_{k,t}) \sqrt{n} \left[\frac{J(\mathbf{y}_k, \boldsymbol{\theta}_{k,t})}{J(\mathbf{y}, \boldsymbol{\theta}_{k,t})} - \frac{J(\mathbf{y}_k, \hat{\boldsymbol{\theta}}_n)}{J(\mathbf{y}, \hat{\boldsymbol{\theta}}_n)}\right]\right| > \varepsilon \Big| \mathbf{y}\right\} \\
&\leq \frac{1}{\varepsilon} E\left[\left|\frac{d_k}{T} \sum_{t=0}^{T} w_k(\boldsymbol{\theta}_{k,t}) \sqrt{n} \left[\frac{J(\mathbf{y}_k, \boldsymbol{\theta}_{k,t})}{J(\mathbf{y}, \boldsymbol{\theta}_{k,t})} - \frac{J(\mathbf{y}_k, \hat{\boldsymbol{\theta}}_n)}{J(\mathbf{y}, \hat{\boldsymbol{\theta}}_n)}\right]\right| \Big| \mathbf{y}\right] \\
&\leq \frac{d_k}{\varepsilon} \frac{1}{T} \sum_{t=0}^{T} E\left[w_k(\boldsymbol{\theta}_{k,t}) \sqrt{n} \left|\frac{J(\mathbf{y}_k, \boldsymbol{\theta}_{k,t})}{J(\mathbf{y}, \boldsymbol{\theta}_{k,t})} - \frac{J(\mathbf{y}_k, \hat{\boldsymbol{\theta}}_n)}{J(\mathbf{y}, \hat{\boldsymbol{\theta}}_n)}\right| \Big| \mathbf{y}\right] \\
&= \frac{1}{\varepsilon} E\left[1\{\boldsymbol{\vartheta} \in A\} \sqrt{n} \left|\frac{J(\mathbf{y}_k, \boldsymbol{\vartheta})}{J(\mathbf{y}, \boldsymbol{\vartheta})} - \frac{J(\mathbf{y}_k, \hat{\boldsymbol{\theta}}_n)}{J(\mathbf{y}, \hat{\boldsymbol{\theta}}_n)}\right| \Big| \mathbf{y}\right] \\
&\leq \frac{1}{\varepsilon} E\left[\sqrt{n} \left|\frac{J(\mathbf{y}_k, \boldsymbol{\vartheta})}{J(\mathbf{y}, \boldsymbol{\vartheta})} - \frac{J(\mathbf{y}_k, \hat{\boldsymbol{\theta}}_n)}{J(\mathbf{y}, \hat{\boldsymbol{\theta}}_n)}\right| \Big| \mathbf{y}\right] \xrightarrow{P} 0,
\end{aligned} \tag{17}
$$

as $n \to \infty$, by Proposition 3. Similarly,

$$
\begin{aligned}
&P\left\{\left|\sqrt{n}\frac{d_k}{T}\left[\sum_{t=0}^{T} 1\{\boldsymbol{\theta}_{k,t} \in A\} \tilde{w}_k(\boldsymbol{\theta}_{k,t})\right.\right.\right. \\
&\quad \left.\left.\left. - \frac{J(\mathbf{y}_k, \hat{\boldsymbol{\theta}}_n)}{J(\mathbf{y}, \hat{\boldsymbol{\theta}}_n)} \sum_{t=0}^{T} 1\{\boldsymbol{\theta}_{k,t} \in A\} w_k(\boldsymbol{\theta}_{k,t})\right]\right| > \varepsilon \Big| \mathbf{y}\right\} \xrightarrow{P} 0.
\end{aligned} \tag{18}
$$

Recall that

$$
\tilde{R}_k(A) = \frac{T^{-1} \sum_{t=0}^{T} 1\{\boldsymbol{\theta}_{k,t} \in A\} \tilde{w}_k(\boldsymbol{\theta}_k^{(t)})}{T^{-1} \sum_{t=0}^{T} \tilde{w}_k(\boldsymbol{\theta}_{k,t})}.
$$

Equations (17) and (18) imply that the numerator and denominator of $\tilde{R}_k(A)$ could well approximate, up to a constant, the numerator and denominator of $\hat{R}_k(A)$ in Equation (11). By properties of convergence in probabilities, we have for large enough T and any ε, as $n \to \infty$,

$$
P\left[\sqrt{n}\left|\tilde{R}_k(A) - \hat{R}_k(A)\right| > \varepsilon \Big| \mathbf{y}\right] \xrightarrow{P} 0.
$$

Second, by Proposition 1, we have, $\sqrt{T}(\hat{R}_k(A) - R(A))$ given $\mathbf{y}$ is stochastically bounded. Finally,

$$
\begin{aligned}
&P\left[\sqrt{n}\left|\tilde{R}_k(A) - R(A)|\mathbf{y}\right| > \varepsilon \Big| \mathbf{y}\right] \\
&\leq P\left[\sqrt{n}\left|\tilde{R}_k(A) - \hat{R}_k(A)\right| + \sqrt{n}\left|\hat{R}_k(A) - R(A)|\mathbf{y}\right| > \varepsilon \Big| \mathbf{y}\right] \\
&= P\left[\sqrt{n}\left|\tilde{R}_k(A) - \hat{R}_k(A)\right| + \sqrt{\frac{n}{T}}\sqrt{T}\left|\hat{R}_k(A) - R(A)|\mathbf{y}\right| > \varepsilon \Big| \mathbf{y}\right] \xrightarrow{P} 0,
\end{aligned}
$$

$T \to \infty, n \to \infty, n/T \to 0$. □

Supplementary Material

An R implementation of this algorithm can be available as a supplementary material.

Acknowledgments

The authors are most grateful to the reviewers and the associate editor for their most constructive comments which led to a much improved version of the article.

Funding

Hannig's research was supported by the National Science Foundation under grant nos. IIS-1633074 and DMS-1916115. Lee's research was supported by the National Science Foundation under grant nos. DMS-1811405, DMS-1811661 and DMS-1916125.

ORCID

Jan Hannig http://orcid.org/0000-0002-4164-0173
Thomas C. M. Lee http://orcid.org/0000-0001-7067-405X

References

Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. (2015), "Distributed Estimation and Inference With Statistical Guaranteesm," arXiv: 1509.05457. [1]

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

Cheng, G., and Shang, Z. (2015), "Computational Limits of Divide-and-Conquer Method," arXiv:1512.09226. [8]

Entezari, R., Craiu, R. V., and Rosenthal, J. S. (2018), "Likelihood Inflating Sampling Algorithm," *Canadian Journal of Statistics*, 46, 147–175. [1,2]

Fisher, R. A. (1930), "Inverse Probability," *Proceedings of the Cambridge Philosophical Society*, pp. 528–535. [2]

Geweke, J. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica: Journal of the Econometric Society*, 57, 1317–1339. [3]

Ghosh, J. K., and Ramamoorthi, R. V. (2003), *Bayesian Nonparametrics*, New York: Springer. [10]

Hannig, J., Feng, Q., Iyer, H., Wang, C., and Liu, X. (2018), "Fusion learning for Inter-Laboratory Comparisons," *Journal of Statistical Planning and Inference*, 195, 64–79. [10]

Hannig, J., Iyer, H., Lai, R. C. S., and Lee, T. C. M. (2016), "Generalized Fiducial Inference: A Review and New Results," *Journal of the American Statistical Association*, 111, 1346–1361. [1,2]

Huang, Z., and Gelman, A. (2005), "Sampling for Bayesian Computation with Large Datasets," available at SSRN 1010107. [1]

Jones, G. L. (2004), "On the Markov Chain Central Limit Theorem," *Probability Surveys*, 1, 299–320. [3,11]

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]

Kong, A. (1992), "A Note on Importance Sampling Using Renormalized Weights," Tech. Rep., Department of Statistics. Chicago, IL: University of Chicago . [4]

Lehmann, E. L., and Casella, G. (1998), *Theory of Point Estimation*, New York: Springer. [10]

Leisen, F., Craiu, R., and Casarin, R. (2016), "Embarrassingly Parallel Sequential Markov-Chain Monte Carlo for Large Sets of Time Series," *Statistics and Its Interface*, 9, 497–508. [1]

Liu, J. S. (1996), "Metropolized Independent Sampling with Comparisons to Rejection Sampling and Importance Sampling," *Statistics and Computing*, 6, 113–119. [4]

Liu, Q. (2016), "Importance Weighted Consensus Monte Carlo for Distributed Bayesian Inference," in *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, UAI'16*. pp. 497–506. Arlington, VA: AUAI Press. [1]

Neiswanger, W., Wang, C., and Xing, E. (2014), "Asymptotically Exact, Embarrassingly Parallel MCMC," in *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pp.623–632. [1]

Pickands III, J. (1975), "Statistical Inference Using Extreme Order Statistics," *The Annals of Statistics*, 3, 119 – 131. [8]

Schuh, M. A., Angryk, R. A., Pillai, K. G., Banda, J. M., and Martens, P. C. (2013), "A Large-Scale Solar Image Dataset With Labeled Event Regions," in *ICIP*, Melbourne, Australia, 4349–4353. [8,10]

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. (2016), "Bayes and Big Data: The Consensus Monte Carlo Algorithm," *International Journal of Management Science and Engineering Management*, 11, 78–88. [1]

Srivastava, S., Cevher, V., Tran-Dinh, Q., and Dunson, D. B. (2015), "Wasp: Scalable Bayes via Barycenters of Subset Posteriors," in *Artificial Intelligence and Statistics*, 912–920. [1,2]

van der Vaart, A. W. (1998), *Asymptotic Statistics*. Vol. 3 of Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge: Cambridge University Press. [10]

Wandler, D. V., and Hannig, J. (2012), "Generalized Fiducial Confidence Intervals for Extremes," *Extremes*, 15, 67–87. [8,10]

Wang, X., Guo, F., Heller, K. A., and Dunson, D. B. (2015), "Parallelizing Mcmc with Random Partition Trees," in *Advances in Neural Information Processing Systems*, Montreal, Canada: MIT Press, pp. 451–459. [1,2]