Submitted to *INFORMS Journal on Computing*
manuscript JOC-2017-01-OA-012

Reducing Simulation Input-Model Risk via Input Model Averaging

Barry L. Nelson
Northwestern University, nelsonb@northwestern.edu

Alan T. K. Wan
City University of Hong Kong, Alan.Wan@cityu.edu.hk

Guohua Zou
Capital Normal University, ghzou@amss.ac.cn

Xinyu Zhang
University of Science and Technology and Chinese Academy of Sciences, xinyu@amss.ac.cn, Corresponding Author

Xi Jiang
Northwestern University, xijiang2020@u.northwestern.edu

Input uncertainty is an aspect of simulation model risk that arises when the driving input distributions are derived or "fit" to real-world, historical data. While there has been significant progress on quantifying and hedging against input uncertainty, there has been no direct attempt to reduce it via better input modeling. The meaning of "better" depends on the context and the objective: our context is when (a) there are one or more families of parametric distributions that are plausible choices; (b) the real-world historical data are not expected to *perfectly* conform to any of them; and (c) our primary goal is to obtain higher-fidelity simulation *output* rather than to discover the "true" distribution. In this paper we show that frequentist model averaging can be an effective way to create input models that better represent the true, unknown input distribution, thereby reducing model risk. Input model averaging builds from standard input modeling practice, is not computationally burdensome, requires no change in how the simulation is executed nor any follow-up experiments, and is available on The Comprehensive R Archive Network (CRAN). We provide theoretical and empirical support for our approach.

Key words: Input Modeling, Stochastic Simulation, Input Uncertainty

1. Introduction

At a high level, stochastic simulations consist of inputs and logic: The inputs are the basic sources of uncertainty that defy further explanation; they are represented by fully specified probability models (e.g., exponential distribution with rate $\lambda = 7.2$). The logic consists of rules and algorithms that transform realizations of the inputs into sample paths of system performance (e.g., waiting times in queues); estimating system properties from these sample paths or "outputs" is the reason a simulation experiment is performed. The fidelity of the outputs in representing the performance of a real or conceptual system clearly depends—often in a very complex way—on the fidelity of the input models. In this paper we consider input models that are tuned or "fit" to samples of real-world, historical data. We refer to this activity as *input modeling*, and we propose a better way to do it.

Beyond the availability of good software, methods used for input modeling in the stochastic simulation practice community have not advanced much in the last 30 years.[1] Here is the recipe found in many textbooks (e.g., Banks et al. (2010), Law and Kelton (1991)) for fitting a marginal distribution F to describe an independent and identically distributed (i.i.d.) input process:

1. Given: $x_1, x_2, \ldots, x_N$ an i.i.d. sample from some unknown input distribution F^c, with "c" denoting "correct" or "true" distribution.

2. Fit $q \geq 1$ candidate parametric distributions $\mathcal{F} = \{F_1, F_2, \ldots, F_q\}$ using methods such as maximum likelihood estimation (MLE), least squares, or moment matching. This yields a set of fitted distributions, say $\widehat{\mathcal{F}} = \left\{\widehat{F}_1, \widehat{F}_2, \ldots, \widehat{F}_q\right\}$. The number of choices in current software ranges from $q = 10$ to 40.

[1] Many new input models have been invented, particularly for multivariate and nonstationary inputs; the lack of progress to which we refer is in methods for fitting these models to data.

3. Rank the choices using one or more summary measures of fit. Standard measures are hypothesis-test statistics such as chi-squared, Kolmogorov-Smirnov and Anderson-Darling, and likelihood-based statistics such as AIC and BIC.

4. From among the top choices, evaluate the fit, e.g., via p-values of the hypothesis tests or graphically via Q-Q plots or other tools.

5. Select $\widehat{F} = \textsf{Best Choice}\left\{\widehat{F}_1, \widehat{F}_2, \ldots, \widehat{F}_q\right\}$. Alternatively, use the (possibly smoothed) empirical distribution of $x_1, x_2, \ldots, x_N$ if nothing fits well.

Although this recipe can and should be made smarter, for instance by using the physical basis of the real input process to suggest an appropriate family of distributions, in practice Step 5 is often automated by selecting the distribution with, say, the minimum AIC statistic for each input process, bypassing Step 4. This approach is understandable because it is neither obvious to simulation practitioners how to do better, nor how much the choice actually matters. *Our proposal adopts Step 2, but rather than selecting one element of $\mathcal{F}$, it instead creates an "input model average" that often leads to a better input model.*

Our work is motivated by the current interest in simulation model risk due to input uncertainty, which is the uncertainty resulting from having only a finite sample N of real-world data from which to fit $\widehat{F}$. However, the input-uncertainty literature has emphasized either quantifying the variability in the simulation output due to the sampling variability in $\widehat{F}$, or selecting a defensive $\widehat{F}$, which means a distribution that is plausible with respect to the given data but leads to the worst-case (maximum or minimum) simulation output performance; see for instance Lam (2016). *Our objective is to reduce input uncertainty through our choice of $\widehat{F}$ via a rethinking about how the input models are created.* Our work is heavily motivated by recent advances in the statistics literature on using model averaging within the frequentist paradigm to improve parameter estimation efficiency and forecast accuracy (e.g., Hjort and Claeskens (2003), Hansen (2007), Wan et al. (2010) and Liang et al. (2011)).

What do we want in an input modeling solution?

- It should work within the framework of current input-modeling software, and in particular Step 2 above where we have a collection of candidate distributions, and impose only a modest additional computational burden.

- It should not require any change in how we actually execute the simulation, other than generating inputs from a different $\widehat{F}$. Thus, input modeling remains an upfront step in the simulation experiment (input uncertainty quantification, on the other hand, often requires additional follow-up experiments).

- It should improve simulation *output* fidelity when the true distribution is *not* in $\mathcal{F}$—so no single choice can be fully correct—but also tends to favor a single candidate model when it is close to F^c. This is consistent with the "view through the queue" criterion coined by Whitt (1981), which evaluates an input approximation by how well it reproduces the desired output, rather than whether it discovers the true input.

More pointedly, our model averaging approach is *not* a better way to discover the "true" real-world distribution when it is a member of the candidate set $\mathcal{F}$, either individually or as a mixture. In fact, our asymptotic analysis specifically assumes that $F^c \notin \mathcal{F}$, and shows that, under some assumptions on the candidate set of distributions $\mathcal{F}$, our model-average distribution gets as close as possible to the real-world distribution using only the component distributions in $\mathcal{F}$. Thus, model averaging is not generally consistent for F^c; however, if we include the empirical distribution (ED) as a candidate, then the model average places all weight on the ED as the sample size $N \to \infty$. Further, under very weak assumptions, the empirically optimal model-average distribution exists and is easily found.

In the end, we will recommend the following: Reduce the size of the candidate set $\mathcal{F}$ (if large) by using prior knowledge of the input process physics or by screening out poor choices using something like AIC or BIC; always include the ED in $\mathcal{F}$; and then do model averaging. However, model averaging can be applied in a completely automated fashion to a large candidate set, and the ED need not be included (say if a continuous $\widehat{F}$ is desired).

The paper is organized as follows. We describe the problem and context more fully in Section 2, and our input model averaging method in Section 3. An empirical evaluation follows in Section 4. Proofs of the some results are found in Appendices A–B of the online supplement.

2. Background and Examples

In this paper we focus on univariate input models, but the method extends naturally to random vectors. Generically, X and Y denote input and output random variables, respectively, and x and y are realized values. We use F_Y to refer to the (typically unknown) distribution of Y.

The following examples will be used to evaluate our methods; they were chosen because they mimic three distinct classes of problems found in simulation studies, and because we expect that different aspects of the inputs X (e.g., mean, variance, tail behavior) will affect the fidelity of their outputs Y. That is, even though the examples themselves are simple, they manifest complex and differing input-to-output behavior. The examples are important because we rely solely on empirical evaluation to establish the potential reduction in input uncertainty.

2.1. Stochastic Activity Network (SAN)

Stochastic activity networks are used in project planning when there is interest in the time to complete the project; an early paper on simulating such networks is Burt and Garman (1971). A realistic problem might have several hundred activities, constrained resources, etc., but as an illustration we consider one with five activities where the time to complete the project is

$$Y = \max\{X_1 + X_4, X_1 + X_3 + X_5, X_2 + X_5\}. \tag{1}$$

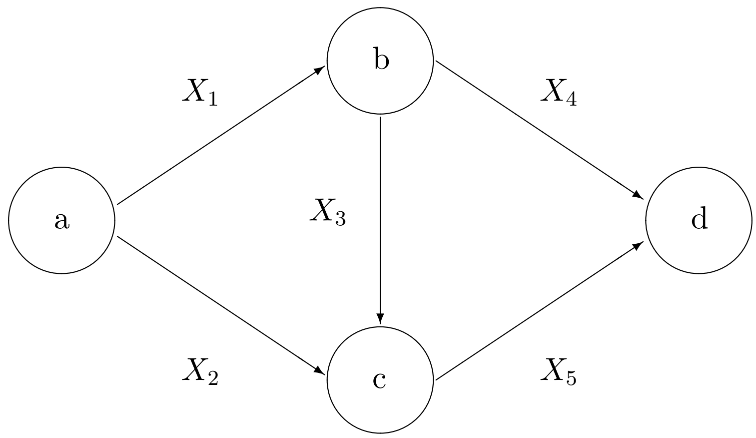

Figure 1 **A small stochastic activity network.**

See Figure 1. Thus, simulation of the SAN requires five input distributions for $X_1, X_2, \ldots, X_5$. Properties of Y that are of interest include $\mathrm{E}(Y)$, $\Pr\{Y > c\}$, $F_Y^{-1}(p)$ or the entire distribution. The natural experiment design for the simulation is to make R replications yielding i.i.d. outputs $Y_1, Y_2, \ldots, Y_R$. Each replication requires random activity-time inputs $X_1, X_2, \ldots, X_5$. Because activity times are summed, path durations tend to be normally distributed for realistically large projects, but for this small example the specific distributions of the X_i should matter.

2.2. $GI/G/1$ Queue

The $GI/G/1$ queue has a renewal arrival process of customers, some non-negative service-time distribution, and a single server; see for instance Gross et al. (2008). Let Y_i be the delay in queue of the ith arriving customer when the system starts empty. Then

$$Y_i = \max\{0, Y_{i-1} + X_{2,i-1} - X_{1,i}\},\ i = 1, 2, \ldots \tag{2}$$

with $Y_0 = X_{2,0} = 0$. There are two input distributions, the interarrival-time distribution of X_1 and the service-time distribution of X_2; successive interarrival times and service times are individually and jointly independent.

Under certain conditions it can be shown that $Y_i \xrightarrow{\mathcal{D}} Y$ as $i \to \infty$, and properties of Y, a specific Y_i, or the set $\{Y_i, i \leq T\}$ for some stopping time T, are of interest. Thus, the experiment design could be a single, long replication, or multiple shorter ones, and the number of inputs $X_{1,i}$ and $X_{2,i}$ needed may be fixed or random.

The $\mathrm{E}(Y)$ is tractable if the interarrival times are exponentially distributed and it depends only on the mean and variance of the service times X_2; the entire distribution of Y is tractable if the service time is also exponential. In general the distributions of Y_i and Y are not known. In our evaluation we focus on the distribution of Y_5, the wait of the 5th arriving customer, since the effect of the service-time distribution beyond its mean and variance should not yet have washed out.

2.3. Highly Reliable System (HRS)

Systems that are repairable or have signficant redundancy are designed to be highly reliable, meaning that system failure is rare. Let Y be the time of system failure. The following algorithm mimics a HRS for which a failure is avoided if a backup component is repaired (time to repair X_1) before the active component fails (time to failure X_2); it does not actually model such a system, but allows us to control the rarity of failure through the distributions of X_1 and X_2.

1. $Y = 0$; $i = 1$
2. until $X_{2,i} < X_{1,i}$ do

    $Y = Y + X_{2,i}$

    $i = i + 1$

loop

3. return $Y = Y + X_{i,1}$

If $\mathrm{E}(X_1) \ll \mathrm{E}(X_2)$ then the system will be highly reliable; just how reliable it is is characterized by properties of Y, such as its mean or a tail probability. In our example $X_{i,1}$ and $X_{i,2}$ are individually and jointly independent.

2.4. Input Uncertainty

To present our method we focus on a single, univariate input distribution F^c for which we have an i.i.d. sample $x_1, x_2, \ldots, x_N$ of real-world data. Because it is a real-world process, there is no expectation that F^c is a member of any standard family of parametric distributions, including those in our set $\mathcal{F}$, although some may be close.

The distribution of our generic output Y depends upon the choice of input distribution F; thus we write

$$Y \sim F_Y(y \mid F).$$

Based on the input data we fit a distribution denoted by $\widehat{F}$; thus, the simulation generates observations of

$$\widehat{Y} \sim F_Y(y \mid \widehat{F}).$$

Ideally $F_Y(y \mid \widehat{F}) = F_Y(y \mid F^c)$, but in practice we will be satisfied if the distributions are close in some relevant sense (e.g., have nearly the same mean). *Notice that what matters is the implied output distribution; the input distribution F^c itself is of less interest.* We let $Y^c \sim F_Y(y \mid F^c)$ denote the ideal output.

Research on *input uncertainty* addresses the problem that

$$F_Y(y \mid \widehat{F}) \neq F_Y(y \mid F^c)$$

often by focusing on the error in using $\widehat{Y}$ as an estimator of $\mathrm{E}(Y^c)$. See, for instance, the surveys in Barton (2012), Lam (2016) and Song et al. (2014). One reasonable objective is to form a confidence interval or a Bayesian credible interval for $\mathrm{E}(Y^c)$ that accounts for error in using $\widehat{F}$ as an estimator of F^c as well as the stochastic error from observing the simulated output $\widehat{Y}$ rather than $\mathrm{E}(\widehat{Y})$. There has been significant success in attacking this and related problems, including Cheng and Holland (1997), Cheng and Holland (1998), Chick (2001), Zouaoui and Wilson (2004), Ankenman and Nelson (2012), Barton et al. (2013), Corlu and Biller (2013), Fan et al. (2013), Xie et al. (2014), Song and Nelson (2015), Ghosh and Lam (2015), Zhou and Xie (2015), Song et al. (2015), Glynn and Lam (2018) and Lam and Qian (2018), to name a few. Notice that none of these papers attempt to *reduce* input uncertainty; instead they try to quantify it or hedge against it.

Unfortunately, experience has shown that the added error due to input uncertainty can be substantial, sometimes overwhelming, even when we have a significant quantity of real-world data. *Therefore, in this paper we look to reduce the input-uncertainty error by our choice of $\widehat{F}$, a reduction which would then be reflected in reduced measurements of it using the methods described in the papers cited above.* We are not attempting to create a defensive choice $\widehat{F}_{\mathrm{worst}}$, and in fact our approach would be an impediment to such robust analysis (see Lam (2016) for an excellent survey of these methods).

Reducing the effect of input-model uncertainty on the simulation output is challenging for obvious reasons: The standard families of distributions used in simulations are often supported by process physics, they are flexible meaning that they can assume a variety of shapes, and they are accompanied by provably efficient parameter-estimation methods, such as MLE. Improving upon the standard approach universally would be impossible, but we will demonstrate empirically that substantial improvements are possible in some situations, especially when the real-world input data do not perfectly conform to any known parametric distribution, as is frequently the case in practice. For completeness we also evaluate our model-average distribution $\widehat{F}$ against F^c, which we can do because we will create the input data.

3. Input Model Averaging

To motivate the method that follows, recall that we have a set $\mathcal{F}$ of q candidate parametric distributions for F^c; for instance, $\mathcal{F}$ could contain

1. Exponential: $f_1(x) = \theta e^{-x\theta}, \qquad x \in [0, \infty)$,
2. Normal: $f_2(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-(x-\mu)^2/(2\sigma^2)}, \qquad x \in \mathbb{R}$,
3. Shifted Gamma: $f_3(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}(x-\xi)^{(\alpha-1)} e^{-(x-\xi)\beta}, \qquad x \in (\xi, \infty)$

where $f_m(x), m = 1, 2, 3$, are density functions for $\mathcal{F} = \{F_1, F_2, F_3\}$, and $\theta, \sigma, \beta, \alpha > 0$ and $\mu, \xi \in \mathbb{R}$ are unknown parameters.

Let $\widehat{F}_m(x)$ and $\widehat{f}_m(x)$ be the estimators of $F^c(x)$ and $f^c(x)$ under the m^{th} candidate distribution, and let $\mathbf{w} = (w_1, w_2, \ldots, w_q)^{\mathrm{T}}$ be a weight vector belonging to the set $\mathcal{W} = \{\mathbf{w} \in [0,1]^q : \sum_{m=1}^q w_m = 1\}$. The *model-average estimator* of $F^c(x)$ is

$$\widehat{F}(x, \mathbf{w}) = \sum_{m=1}^{q} w_m \widehat{F}_m(x) \tag{3}$$

and taking its derivative with respect to x we have the model average estimator of $f^c(x)$

$$\widehat{f}(x, \mathbf{w}) = \sum_{m=1}^{q} w_m \widehat{f}_m(x). \tag{4}$$

Clearly $\widehat{F}(x, \mathbf{w})$ includes each of the individual candidate distributions as a special case of $\mathbf{w}$, but it increases flexibility by allowing averages. A good choice of $\mathbf{w}$ is one that makes $\widehat{F}(x, \mathbf{w})$ close to $F^c(x)$ in a comprehensive way that we define precisely below. Of course, $F^c(x)$ is unknown, but the ED with cdf

$$\bar{F}(x) = N^{-1} \sum_{i=1}^{N} I(x_i \le x)$$

is unbiased and consistent for it, so we use $\bar{F}$ as a stand-in for F^c in fitting. Finally, to guard against overfitting we use cross-validation (CV) with $\bar{F}(x)$ to select $\mathbf{w}$; we describe the CV method in the next section. Given the CV-estimated weight

$\widehat{\mathbf{w}}$, variate generation is easy: Each time a value of X is needed, generate integer $M \sim \widehat{\mathbf{w}}$ to select the distribution, then generate $X \sim \widehat{F}_M$.

REMARK 1. Averaging distributions as dissimilar as exponential, normal and shifted-gamma may not seem sensible. However, practitioners often use software that fits a long list of distributions, and as we show later we can easily find the empirically optimal model average even for such heterogeneous cases.

REMARK 2. The model-average distribution $\widehat{F}(x, \mathbf{w})$ is clearly a mixture distribution, but it is different in spirit from mixing a finite number of distributions from a *common* family, which is a well-known distribution fitting approach (McLachlan and Peel 2004). We can, in fact, exploit finite mixtures of a common distribution by including such models in the candidate set $\mathcal{F}$ provided we have a method for fitting them.

3.1. Cross-Validation for Input Model Averaging

Let $\mathbf{x}_N = (x_1, x_2, \ldots, x_N)$ be the real-world data, which we model as i.i.d. copies of the random variable $X \sim F^c$. Here we develop a "frequentist model averaging" approach to estimate $F^c(x)$ by $F(x, \mathbf{w})$ using J-fold CV to tune $\mathbf{w}$ to $\mathbf{x}_N$; it is in the spirit of the Jackknife model average (JMA) of Hansen and Racine (2012), developed originally for improving the efficiency of estimators in a heteroscedastic linear regression model. Hansen and Racine (2012) proved that the JMA estimator of the regression coefficients has the smallest asymptotic expected squared errors among a large class of linear estimators including the least squares, ridge, Nadaraya-Watson and spline estimators. They also showed that the JMA estimator frequently outperforms the AIC and BIC model selection estimators, and Hansen (2007)'s Mallows model average estimators in finite samples. Zhang et al. (2013) showed that the merits of the JMA estimator carry over to models that admit a lagged dependent variable as a regressor and a non-diagonal error covariance structure.

To implement the JMA scheme for input modeling in stochastic simulation, we partition the data set $\mathbf{x}_N$ into J groups, such that for each group we have $S = N/J$

observations. For the j^{th} group, the observations are labelled as $x_{(j-1)S+1}, \ldots, x_{jS}$, where $j = 1, 2, \ldots, J$. Let $\widetilde{F}_m^{(-j)}(x)$ be the estimator (e.g., via MLE) of $F^c(x)$ with the observations of the j^{th} group *removed* from the data set for the m^{th} candidate distribution. Correspondingly, the model average estimator with the j^{th} group removed is

$$\widetilde{F}^{(-j)}(x, \mathbf{w}) = \sum_{m=1}^{q} w_m \widetilde{F}_m^{(-j)}(x).$$

The ED estimator of $F^c(x)$ using *only* the j^{th} group is

$$\bar{F}_{(j)}(x) = S^{-1} \sum_{s=1}^{S} I(x_{(j-1)S+s} \leq x) \tag{5}$$

and it is well known that $\mathrm{E}\left(\bar{F}_{(j)}(x)\right) = F^c(x)$. Our J-fold CV criterion is formulated to be

$$CV_J(\mathbf{w}) = \sum_{j=1}^{J} \sum_{s=1}^{S} \left\{ \widetilde{F}^{(-j)}(x_{(j-1)S+s}, \mathbf{w}) - \bar{F}_{(j)}(x_{(j-1)S+s}) \right\}^2.$$

In other words, we consider the squared difference between the model-average estimator constructed *without* the j^{th} group of real-world data, and the ED constructed from *only* the j^{th} group, summed across all groups. The empirically optimal weight vector resulting from this criterion is

$$\widehat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}} CV_J(\mathbf{w})$$

leading to the model average estimator $\widehat{F}(x, \widehat{\mathbf{w}})$ of $F^c(x)$.

The optimization problem we need to solve to find $\widehat{\mathbf{w}}$ can be formulated as a quadratic program (QP); see Jiang and Nelson (2018) for the formulation, and Nocedal and Wright (2006) for solving QPs. Specifically,

$$\begin{aligned} \text{minimize: } CV_J(\mathbf{w}) &= \sum_{j=1}^{J} \sum_{s=1}^{S} \left\{ \widetilde{F}^{(-j)}(x_{(j-1)S+s}, \mathbf{w}) - \bar{F}_{(j)}(x_{(j-1)S+s}) \right\}^2 \\ &= \sum_{j=1}^{J} \sum_{s=1}^{S} \left\{ \sum_{m=1}^{q} w_m \left(\widetilde{F}_m^{(-j)}(x_{(j-1)S+s}) - \bar{F}_{(j)}(x_{(j-1)S+s}) \right) \right\}^2 \end{aligned}$$

$$= \sum_{j=1}^{J} \sum_{s=1}^{S} \left\{ \sum_{m=1}^{q} w_m c_{mjs} \right\}^2$$

$$= \sum_{j=1}^{J} \sum_{s=1}^{S} \mathbf{w}^\top \mathbf{C}_{js} \ \mathbf{w} = \mathbf{w}^\top \mathbf{C} \, \mathbf{w}$$

$$\text{subject to:} \quad \sum_{m=1}^{q} w_m = 1$$

$$w_m \geq 0, \ m = 1, 2, \ldots, q$$

where the matrices $\mathbf{C}_{js}$ and $\mathbf{C}$ are defined in the obvious way. If the $q \times q$ matrix $\mathbf{C}$ is positive definite, then the objective function is strictly convex and the QP has a unique optimal solution; we refer to this as the *empirically optimal* model average. That each matrix $\mathbf{C}_{js}$ is positive semi-definite is clear from their construction. When X is continuous-valued and at least one of the candidate distributions is continuous, we show in Appendix C of the online supplement that the probability of there existing a $\mathbf{w}' \neq \mathbf{0}$ for which $(\mathbf{w}')^\top \mathbf{C} \, \mathbf{w}' = 0$ is zero; therefore, $\mathbf{C}$ is positive definite almost surely. The QP is easily solved via standard methods (Nocedal and Wright 2006). Notice that the construction of $\mathbf{C}$ is a one-time calculation, and that QPs of such small size ($q \leq 40$) can be solved very efficiently. The fitting algorithm is implemented in our R package FMAdist (`https://cran.r-project.org/package=FMAdist`).

REMARK 3. Model averaging is intuitively appealing, as it enlarges the space of input model choices beyond $\mathcal{F}$ while still including the individual candidate distributions in $\mathcal{F}$ as special cases; it employs cross-validation as a robust method for fitting; and the empirically optimal solution is easy to find under weak assumptions. Under more restrictive assumptions we show in Section 3.3 that this empirically optimal choice is in fact the best possible choice as $N \to \infty$.

3.2. Relationships to Other Input Modeling Methods

The greatest progress in input uncertainty quantification to date has been when the distribution family of F^c is assumed known (e.g., gamma(α, β)), so that the only uncertainty comes from the parameter estimates (e.g., $\widehat{\alpha}, \widehat{\beta}$). In our opinion, it is not universally the case that real-world data both conform perfectly to a parametric distribution and are measured to sufficient precision to be indistinguishable, even though parametric distributions are often good approximations.

There has been some work on input uncertainty quantification that allows for distribution family uncertainty, specifically Chick (2001) and Zouaoui and Wilson (2004). Both papers take a Bayesian perspective, placing a prior distribution on the correct model family (e.g., exponential, Weibull, gamma), as well as each distributions' parameters, and derive the posterior distributions given $\mathbf{x}_N$. While variate generation of inputs is identical to our method—first choosing the distribution family from the posterior, then generating variates—their goal is to fully represent input uncertainty in the posterior predictive distribution of the output Y, rather than trying to reduce it as we do; in fact we provide no estimate of input uncertainty.

Another appealing solution is to use a parametric function $\widehat{F}$ that has the flexibilty to get close to any F^c, and many distributions have been created for this purpose, including the generalized lambda distribution (Karian and Dudewicz 2000) to match moments or percentiles, and the Bézier distribution (Wagner and Wilson 1996) that can have an arbitrary number of parameters. However, these distributions were created to be flexible, rather than to conform to particular process physics, leading to the possibility of overfitting or manifesting unusual features that are not consistent with the data. There is, after all, a reason that the standard arsenal of normal, lognormal, logistic, Weibull, gamma, Pareto, etc. continue to be used: their existence is implied by theory that can hold approximately in practice.

As described in the previous section, the input model averaging approach allows us to exploit these tried-and-true families, but to extend their reach through averaging. To resist overfitting we use CV to select the weights; CV insures that the weights do not give an average that is inconsistent with the distribution of the data, and the empirically optimal weights are unique and easily found.

3.3. Asymptotic Properties of Input Model Averages

In this section we establish asymptotic properties of the empirically optimal model average under certain restrictions on the true distribution F^c, the individual candidate distributions in $\mathcal{F}$, and whether or not the ED $\bar{F}$ is in $\mathcal{F}$: As $N \to \infty$, (a) for certain classes of candidate distributions $\mathcal{F}$, when neither F^c nor $\bar{F}$ are in $\mathcal{F}$ individually, the empirically optimal model-average weights become the squared-error-optimal weights; and (b) when $\bar{F}$ is included in $\mathcal{F}$, then its weight converges to 1. The first result implies that under certain conditions cross validation provides the best-possible weights when no candidate distribution fits perfectly, while the second implies that it is consistent for F^c if we include the ED as a candidate.

We first establish the restrictions. Let $\boldsymbol{\beta}_m$ be the unknown parameter vector in the m^{th} candidate distribution, and let $\widehat{\boldsymbol{\beta}}_m$ be its MLE for $m = 1, 2, \ldots, q$, which we assume exists. It is worth noting that $\widehat{\boldsymbol{\beta}}_m$ $(1 \le m \le q)$ is determined from each candidate distribution individually, and not by the optimized linear combination of distributions. Further, let $\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{\beta}}_1^{\text{T}}, \ldots, \widehat{\boldsymbol{\beta}}_q^{\text{T}}\right)^{\text{T}}$ with dimension κ. We require that the size q of the candidate set is finite. Furthermore, we assume that the following conditions hold:

(i) For each $x \in \mathbb{R}$, the density function $f_m(x; \boldsymbol{\beta}_m)$ of the m^{th} candidate distribution is continuous at every $\boldsymbol{\beta}_m$ in the corresponding compact parameter space Θ_m.

(ii) $\mathrm{E}[\log f^c(x)]$ exists and $|\log f_m(x; \boldsymbol{\beta}_m)| < l(x)$, where $l(x)$ is integrable with respect to F^c.

(iii) There exists a vector $\boldsymbol{\beta}_m^*$ at which the Kullback-Leibler information $\int_{\mathbb{R}} \log[f^c(x)/f_m(x;\boldsymbol{\beta}_m)]]f^c(x)dx$ attains a unique minimum.

Under these conditions $\widehat{\boldsymbol{\beta}} \to \boldsymbol{\beta}^* = \left(\boldsymbol{\beta}_1^{*\mathrm{T}}, \ldots, \boldsymbol{\beta}_q^{*\mathrm{T}}\right)^{\mathrm{T}}$ almost surely as $N \to \infty$, i.e., the MLEs converge even when the distributions are misspecified. White (1982) further showed that

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = O_p(N^{-1/2}). \tag{6}$$

We assume that (6) is in force. The validity of (6) depends on Conditions (i)–(iii) as well as Assumptions A4, A5 and A6 of White (1982).

Remark 4. The canonical parameter space for many standard distributions is not compact, as assumed in (i); e.g., for the normal distribution $\sigma > 0$. However, as a practical matter assuming that there exists a large, but compact space in which each parameter lies, e.g., $\sigma \in [10^{-10}, 10^{10}]$, is reasonable since there is no requirement that the bound be known. In this sense all of the distributions that have a density in the examples in Section 4 below satisfy this condition.

We next define the notations needed to state our main results. Let $\mathbf{F}_0 = (F^c(x_1), \ldots, F^c(x_N))^{\mathrm{T}}$, the values of the true cdf evaluated at the data points, and $\widehat{\mathbf{F}}_m = (\widehat{F}_m(x_1), \ldots, \widehat{F}_m(x_N))^{\mathrm{T}}$, the corresponding quantity for the m^{th} candidate fitted distribution, for $m = 1, 2, \ldots, q$. We assume the ED is not one of the q candidates. For any fixed $\mathbf{w}$, define a corresponding vectors of values for the averaged distribution, with parameters fitted from data, $\widehat{\mathbf{F}}(\mathbf{w}) = (\widehat{F}(x_1, \mathbf{w}), \ldots, \widehat{F}(x_N, \mathbf{w}))^{\mathrm{T}}$, and with the limiting parameters $\mathbf{F}^*(\mathbf{w}) = \widehat{\mathbf{F}}(\mathbf{w}) \mid_{\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*}$.

Recall that CV leaves out sets of S consecutive data values in turn. In $\widetilde{\mathbf{F}}(\mathbf{w}) = (\widetilde{F}^{(-1)}(x_1, \mathbf{w}), \ldots, \widetilde{F}^{(-1)}(x_S, \mathbf{w}), \widetilde{F}^{(-2)}(x_{S+1}, \mathbf{w}), \ldots, \widetilde{F}^{(-J)}(x_N, \mathbf{w}))^{\mathrm{T}}$ we collect the cdf values for each data point based on the model average that excludes it; $\bar{\mathbf{F}} = (\bar{F}_{(1)}(x_1), \ldots, \bar{F}_{(1)}(x_S), \bar{F}_{(2)}(x_{S+1}), \ldots, \bar{F}_{(J)}(x_N))^{\mathrm{T}}$ is the corresponding vector using the ED. We assume J is fixed, so that $S \to \infty$ as $N \to \infty$.

Finally, define the discrepancy $L_N^*(\mathbf{w}) = \|\mathbf{F}^*(\mathbf{w}) - \mathbf{F}_0\|^2$, and let $\xi_N = \inf_{\mathbf{w}\in\mathcal{W}} L_N^*(\mathbf{w})$ (with all weights assigned to distributions other than the ED).

For proving the results, we need the following regularity conditions:

Condition 1 *There exists a neighborhood $\mathcal{N}$ of $\boldsymbol{\beta}^*$ such that*

$$\sup_{\widetilde{\boldsymbol{\beta}}\in\mathcal{N}} \left\| \frac{\partial \widehat{F}(x_i,\mathbf{w})}{\partial \widehat{\boldsymbol{\beta}}} \Big|_{\widehat{\boldsymbol{\beta}}=\widetilde{\boldsymbol{\beta}}} \right\| = O_p(1)$$

uniformly for $i = 1, 2, \ldots, N$ and $\mathbf{w} \in \mathcal{W}$.

Condition 2 *For all $\mathbf{w}\in\mathcal{W}$, $N^{-1/2}\|\widehat{\mathbf{F}}(\mathbf{w}) - \widetilde{\mathbf{F}}(\mathbf{w})\|^2 = O_p(1)$, and $N^{-1/2}\{\widehat{\mathbf{F}}(\mathbf{w}) - \widetilde{\mathbf{F}}(\mathbf{w})\}^{\mathrm{T}}\{\widehat{\mathbf{F}}(\mathbf{w}) - \bar{\mathbf{F}}\} = O_p(1)$.*

Condition 3 *When $N\to\infty$, there exists a sequence $c_N \to 0$ such that $\xi_N^2 \ge N/c_N$ almost surely.*

Condition 3 is well-defined even if F^c is a nontrivial mixture of two or more elements of the candidate set $\mathcal{F}$. It can be seen that

$$\widehat{\mathbf{F}}(\mathbf{w}) - \widetilde{\mathbf{F}}(\mathbf{w}) = \Bigg(\sum_{m=1}^{q} w_m \left\{ \widehat{F}_m(x_1) - \widetilde{F}_m^{(-1)}(x_1) \right\}, \ldots, \sum_{m=1}^{q} w_m \left\{ \widehat{F}_m(x_S) - \widetilde{F}_m^{(-1)}(x_S) \right\}, \\ \sum_{m=1}^{q} w_m \left\{ \widehat{F}_m(x_{S+1}) - \widetilde{F}_m^{(-2)}(x_{S+1}) \right\}, \ldots, \sum_{m=1}^{q} w_m \left\{ \widehat{F}_m(x_N) - \widetilde{F}_m^{(-J)}(x_N) \right\} \Bigg)^{\mathrm{T}}. \quad (7)$$

Hence Condition 2 requires the difference between the regular and leave-S out estimators to decrease sufficiently quickly as N increases. On the other hand, Condition 3 requires that ξ_N grows at a rate no slower than $N^{1/2}$. This in turn implies that the correct input distribution F^c must not be among the candidate distributions in the model average.

THEOREM 1. *If $F^c \notin \mathcal{F}$, $\bar{F} \notin \mathcal{F}$, and Conditions 1–3 hold, then as the real-world sample size $N \to \infty$*

$$\frac{\sum_{i=1}^{N} \left[\widehat{F}(x_i, \widehat{\mathbf{w}}) - F^c(x_i) \right]^2}{\inf_{\mathbf{w}\in\mathcal{W}} \sum_{i=1}^{N} \left[\widehat{F}(x_i, \mathbf{w}) - F^c(x_i) \right]^2} \xrightarrow{P} 1. \quad (8)$$

The proof is in Appendix A of the online supplement. Notice that the numerator and denominator are the sum of squared deviations of the model average estimator from the true distribution, a comprehensive measure of fit. In the numerator the weight $\widehat{\mathbf{w}}$ is obtained via J-fold CV with the empirical distribution; while in the denominator the minimum possible squared deviation weight is chosen. The result shows that as the sample size increases, J-fold CV yields error *no larger than the minimum possible error* with the given set of candidate distributions, which we would expect to be smaller than choosing any *one* distribution from $\mathcal{F}$ when $F^c \notin \mathcal{F}$. Notice that the condition "$F^c \notin \mathcal{F}$" does not prohibit F^c from being a nontrivial mixture of two or more elements of $\mathcal{F}$.

Theorem 1 does not establish that $\widehat{F}(x_i, \widehat{\mathbf{w}})$ is asymptotically consistent for F^c. However, as noted earlier, the ED is unbiased and consistent for F^c, Therefore, we also consider including $\bar{F}$ in the candidate set $\mathcal{F}$ for model averaging. While (8) no longer holds, model averaging becomes consistent in the sense that in the limit all of the weight is on the ED.

THEOREM 2. *If $F^c \notin \mathcal{F}$ but $\bar{F} \in \mathcal{F}$, and if Conditions 1–3 hold, then $\widehat{w}_{ED} \xrightarrow{P} 1$ as $N \to \infty$.*

The proof is in Appendix B of the online supplement. The effect of including $\bar{F}$ is that the other parametric distributions smooth the ED and provide better tail behavior. This is important because the ED being unbiased does not imply that the output $\bar{Y} \sim F_Y(y \mid \bar{F})$ has the same distribution, or even the same mean, as the ideal output Y^c because the simulation is in general a highly nonlinear transformation of inputs to outputs. See the commentary in Song et al. (2015). Nevertheless, we will show empirically that the ED of the entire data set, $\bar{F}$, is often a good choice when the criterion is recovering the distribution of Y^c, and model averaging with the ED can be superior to either the ED alone or model averaging of parametric distributions.

REMARK 5. Among the list of possible candidates $\mathcal{F}$ could be kernel density estimators and the more-recent log-concave density estimators (Kim et al. 2016, Cule et al. 2010). These semi-parametric methods do not leverage process physics—an advantage of our approach—but do have excellent convergence rates to the true distribution. However, we would expect our fits to be smoother for small to moderate N. That said, these methods are not natural candidates for our model averaging because they directly estimate the density, while we require the cdf.

4. Experiments

In this section we evaluate our proposal empirically. Recall that our interest is in how properties of the simulation output $\widehat{Y} \sim F_Y(y \mid \widehat{F})$ compare to the ideal output $Y^c \sim F_Y(y \mid F^c)$ (Section 4.1), and also how our fitted model-average distribution $\widehat{F}(x; \widehat{\mathbf{w}})$ compares to the distribution that generated the input data F^c (Section 4.2). The assessments in this section are quantitative; see Jiang and Nelson (2018) for some graphical illustrations of the attained fits and the online data supplement for additional documentation

We reach the following broad conclusions: Model averaging, especially including the ED as a candidate, is often substantially superior to any *single* choice from $\mathcal{F}$, and typically no worse. Given a large number q of candidate distributions it is best to screen out obviously poor choices first and do model averaging over a smaller subset of $\mathcal{F}$. When either the size of the real-world data sample N is large, or none of the candidate distributions has the capability to fit well (e.g., data are multi-modal but choices in $\mathcal{F}$ are all unimodal), then the weight on the ED, $\widehat{w}_{ED}$, tends to be large. Thus, the ED provides protection against a badly chosen candidate set (which can occur when using the built-in set in a software package) as well as providing the consistency established by Theorem 2. Although not specifically targeted in our experiments, it seems clear that employing a candidate set with common, appropriate support, and containing candidates that are well justified by process physics when available, is helpful. Finally, we found no systematic difference from using $J = 5$ or 10 folds for CV; we would never use more than 10 folds, and recommend $J = 5$ when N is small.

4.1. Evaluation of the Output Distribution

To evaluate the various methods with respect to the output distribution, we use the relative distribution method of Handcock and Morris (2006). A brief explanation of the method follows: Let the distributions of $\widehat{Y}$ and Y^c be denoted by $F_{\widehat{Y}}(y)$ and $F_{Y^c}(y)$, respectively. Define the grade transformation of $\widehat{Y}$ to Y^c as

$$U = F_{Y^c}(\widehat{Y}), \tag{9}$$

obtained by transforming $\widehat{Y}$ by F_{Y^c}. The CDF of U can be expressed as

$$G(u) = F_{\widehat{Y}}(F_{Y^c}^{-1}(u)) \tag{10}$$

for $0 \le u \le 1$, where $F_{Y^c}^{-1}(u) = \inf\{y \mid F_{Y^c}(y) \ge u\}$ is the quantile function of F_{Y^c}.

It is easily seen that if $\widehat{Y} \stackrel{\mathcal{D}}{=} Y^c$, then the CDF of U is a 45° line. When $\widehat{Y} \stackrel{\mathcal{D}}{\neq} Y^c$, then the closer $G(u)$ is to the 45° line, the better the fit provided by $\widehat{Y}$. In our analysis, we use the unsigned area between $G(u)$ and the 45° line over $0 \le u \le 1$ to evaluate the effectiveness of the method. We denote the gap by $A(u) = |G(u) - u|$, so that the area is $A = \int_0^1 A(u)\, du$. Clearly we could compare a list of individual properties, such as the mean and variance, but A provides a comprehensive measure of performance. The relative distribution method is in the same spirit as the tail-probability plot method (see, for instance, Heyde and Kou (2004)). When the true distribution of Y^c is not available, as in most of our examples, then it is represented by a very large sample from F_{Y^c}; this is possible for us because the distributions of the inputs F^c are known. To implement the relative distribution analysis we used the codes at `https://csde.washington.edu/~handcock/RelDist/Software/R/`.

The following are features of our empirical evaluation:

- We apply input model averaging to cases of the SAN, $GI/G/1$ and HRS simulations as described in Sections 2.1–2.3, for different quantities of real-world data, N, used to fit the input models. We generate the "real-world data" from fully specified distributions.
- We consider instances in which the candidate set $\mathcal{F}$ does, or does not, contain the true input distributions F^c. The single "best fit" distribution, which represents common input-modeling practice, is selected from this set both by minimum AIC and minimum BIC.
- We refer to our frequentist model averaging method as JFCV (for J-fold cross-validation). The ED is considered both as an individual input distribution method, and a candidate within the JFCV model average. We refer to the JFCV method that includes the ED as a default candidate as the JFCV(ED) method. Thus, our five competing methods are AIC, BIC, ED, JFCV and JFCV(ED).
- When evaluating the performance of the methods we consider both the area A discussed above and the tail area $A_{\text{tail}} = \int_{0.9}^{1} A(u)\,du$. We are interested in A_{tail} because there is a common belief that the ED, which does not model the tail of the distribution beyond the largest data point in the real-world sample, may be an inferior method when interest centers on the tail of the simulation output Y. Each experiment is repeated for 100 macroreplications and the results reported are averages of A or A_{tail} across these 100 macroreplications. When presenting the results we usually normalize the average area (of A or A_{tail} depending on the focus of interest) generated by the JFCV method to 1 although in some cases we also present the raw average area. Hence, if the relative average area produced by a method is larger than 1 then it is inferior to JFCV, and vice versa, based on this metric. All results are displayed to statistically meaningful digits of precision. We also examine how the JFCV weight $\widehat{\mathbf{w}}$ changes as N increases.
- For the SAN experiment we also very precisely estimate the mean squared error (MSE) of a point estimate for the probability of late project completion.

4.1.1. SAN Experiment We begin with the SAN described in Section 2.1, for which there are five input distributions for the five activity times, $X_1, X_2, \ldots, X_5$. For Cases I–III the true distributions are made up of mixture distributions that are *not* contained in any of the candidate sets; whereas Cases IV–V includes distributions that are contained in the candidate sets. Results for Cases I–V are reported in Appendix D of the online supplement. Here we report Case VI, which uses a candidate set $\mathcal{F}$ that is common to all commercial distribution fitting products: $\mathcal{F}_4 = \{\text{normal}, \text{lognormal}, \text{beta}, \text{exponential}, \text{gamma}, \text{Weibull}\}$ plus possibly the ED. In addition to the candidate set $\mathcal{F}_4$, we also consider a smaller subset within it containing the "best" three based on minimum AIC and BIC selections. We refer to this subset as $\mathcal{F}_4^{(3)}$ and apply the JFCV methods under this subset as well as the full set $\mathcal{F}_4$. In the event that AIC and BIC do not lead to the same set of best distributions, averaging under $\mathcal{F}_4^{(3)}$ will involve more than three distributions.

The true activity-time distributions are Pareto, Rayleigh and loglogistic, as shown in Table 1. In each case the mean activity time is approximately 1. None of these are contained in the candidate set.

The results are displayed in Table 2. JFCV is superior to any single choice made via best AIC or BIC, and JFCV(ED), which includes the ED in the candidate set, is substantially better than JFCV alone. In this example selecting a subset of the top 3 distributions before modeling averaging has little or no effect; however, in Appendix D of the online supplement we show it can be useful in other scenarios for the SAN, as well as in our distribution-to-distribution comparisons in Section 4.2.

Although A provides a comprehensive measure of output-distribution performance, we also display some results for the MSE of a point estimate of $\Pr\{Y^c > 6.65\}$, since the probablity of completing beyond a due date is often important in project planning; 6.65 is the 0.9 quantile (based on a side experiment with 1 million replications). This case is to give a sense of the effect of modeling averaging on

Table 1 True activity-time distributions for Case VI SAN example.

Activity	Distribution	Parameters	CDF
X_1	Rayleigh	$\sigma = \pi/2$	$1-\exp(x^2/(2\sigma^2)), x \geq 0$
X_2	Pareto	$\mu = 1/4, \sigma = 3/16, \xi = 3/4$	$1-(1+\xi(x-\mu)/\sigma)^{-1/\xi}, x \geq \mu$
X_3	Pareto	$\mu = 1/2, \sigma = 1/4, \xi = 1/2$	$1-(1+\xi(x-\mu)/\sigma)^{-1/\xi}, x \geq \mu$
X_4	loglogistic	$\alpha = 0.23, \beta = 1.21$	$\frac{1}{1+(x/\alpha)^{-\beta}}, x \geq 0$
X_5	loglogistic	$\alpha = 2/\pi, \beta = 2$	$\frac{1}{1+(x/\alpha)^{-\beta}}, x \geq 0$

Table 2 Numerical results for SAN experiment Case VI.

	Actual Average A					Relative Average A				
Scenario	JFCV	AIC	BIC	ED	JFCV(ED)	JFCV	AIC	BIC	ED	JFCV(ED)
$\mathcal{F}_4$	0.05	0.06	0.06	0.06	0.04	1.00	1.07	1.07	1.06	0.77
$\mathcal{F}_4^{(3)}$	0.06	0.06	0.06	0.06	0.04	1.00	1.05	1.05	1.05	0.76

Table 3 MSE results for SAN experiment, Case VI, for estimating $\Pr\{Y^c > 6.65\}$.

N	R	Candidates	MSE	SE(MSE)
100	1000	ED	0.00254	4.8E-05
100	1000	Best AIC	0.00116	1.6E-05
100	1000	$\mathcal{F}_4 + \text{ED}$	0.00079	1.8E-05
1000	1000	ED	0.00016	2.9E-06
1000	1000	Best AIC	0.00093	8.2E-06
1000	1000	$\mathcal{F}_4 + \text{ED}$	0.00015	3.1E-06

point-estimator performance. Table 3 displays results for real-world sample sizes $N = 100, 1000$ and $R = 1000$ replications of the SAN; a large number of replications are required so that point estimator variance does not overwhelm the bias reduction we hope to be revealed. The MSE is estimated from 5000 macroreplications of the entire experiment, and the standard error of the estimate is also displayed. We see that when N is small, model averaging yields substantial improvement over the ED or best AIC choices; when N is larger the ED and model average are indistinguishable.

4.1.2. $GI/G/1$ **Experiment** Next we examine results for two cases of the $GI/G/1$ queue described in Section 2.2: An $M/M/1$ queue (Case VII), meaning exponential interarrival and service times, and a $GI/G/1$ queue with balanced hyperexponential interarrival times

$$X_1 \sim \begin{cases} \text{exponential}(1) & \text{with probability } 1/2 \\ \text{exponential}(20) & \text{with probability } 1/2 \end{cases}$$

and service times that follow the mixture distribution,

$$X_2 \sim \begin{cases} \text{unif}(10, 20) & \text{with probability } 2/5 \\ \text{gamma}(2.875, 1/2) & \text{with probability } 3/5 \end{cases}$$

which we label as Case VIII. In both cases the implied traffic intensity is $\mathrm{E}(X_2)/\mathrm{E}(X_1) = 0.9$, and the output we consider is the waiting time of the 5th arrival Y_5. We consider candidate sets

$$\mathcal{F}_1 = \{\text{truncated normal, beta, gamma}\}$$
$$\mathcal{F}_2 = \mathcal{F}_1 \cup \{\text{lognormal, Weibull}\}$$
$$\mathcal{F}_3 = \mathcal{F}_2 \cup \{\text{negative binomial}, \text{discrete uniform}, \text{Poisson}, \text{continuous uniform}, \text{loglogistic}, \text{inverse Gaussian}, \text{Pareto}, \text{binomial}\}$$

Tables 4 and 5 contain the $M/M/1$ results for relative average A and A_{tail}, respectively. For capturing the entire output distribution of Y_5 as measured by A, JFCV and JFCV(ED) tend to be better than AIC, BIC and ED, even though the true exponential distribution is in all candidate sets $\mathcal{F}_1$–$\mathcal{F}_3$ in the form of the gamma distribution, and again in sets $\mathcal{F}_2$–$\mathcal{F}_3$ in the form of the Weibull distribution. AIC and BIC improve substantially in capturing the tail behavior as measured by A_{tail}, but do not beat JFCV(ED).

Tables 6 and 7 present corresponding results for the $GI/G/1$. When interest centers on A, the JFCV(ED) method is the clear favorite, followed by the ED, which has an edge over the JFCV that in turn delivers better performance than

Table 4 Numerical results for $M/M/1$ queue with $N = 100$.

		Relative Average A				
Case	Scenario	JFCV	AIC	BIC	ED	JFCV(ED)
VII	$\mathcal{F}_1$	1.00	1.04	1.04	1.00	0.98
	$\mathcal{F}_2$	1.00	1.04	1.04	1.01	0.98
	$\mathcal{F}_3$	1.00	1.00	0.99	0.96	0.94
	$\mathcal{F}_3^{(3)}$	1.00	1.09	1.08	1.05	1.00
	$\mathcal{F}_3^{(6)}$	1.00	1.07	1.06	1.02	0.98

Table 5 Numerical results for $M/M/1$ queue, tail estimation, with $N = 100$.

		Relative Average A_{tail}				
Case	Scenario	JFCV	AIC	BIC	ED	JFCV(ED)
VII	$\mathcal{F}_1$	1.00	0.74	0.74	1.24	0.84
	$\mathcal{F}_2$	1.00	0.71	0.71	1.17	0.71
	$\mathcal{F}_3$	1.00	0.43	0.42	0.65	0.47
	$\mathcal{F}_3^{(3)}$	1.00	0.82	0.81	1.26	0.77
	$\mathcal{F}_3^{(6)}$	1.00	0.53	0.53	0.82	0.51

Table 6 Numerical results for $GI/G/1$ queue with $N = 100$.

		Relative Average A				
Case	Scenario	JFCV	AIC	BIC	ED	JFCV(ED)
VIII	$\mathcal{F}_1$	1.00	0.94	0.94	0.90	0.83
	$\mathcal{F}_2$	1.00	1.05	1.05	0.96	0.88
	$\mathcal{F}_3$	1.00	1.15	1.16	0.81	0.78
	$\mathcal{F}_3^{(3)}$	1.00	1.30	1.31	0.92	0.84
	$\mathcal{F}_3^{(6)}$	1.00	1.24	1.25	0.88	0.82

AIC and BIC in the majority of cases. When interest centers on A_{tail}, JFCV(ED) remains the best, the ED can yield worse performance than the JFCV, and AIC and BIC selections can be particularly bad when there is a large set of candidate distributions.

4.1.3. HRS Experiment Finally, we consider the HRS example of Section 2.3. We consider the following setup, labelled as Case IX in our subsequent presentation of results. Let the inputs be

$$X_1 \sim \begin{cases} \text{unif}(0,1), & \text{with probability } 0.5 \\ \text{exponential}(1), & \text{with probability } 0.5 \end{cases}$$

Table 7 Numerical results for $GI/G/1$ queue, tail estimation, with $N = 100$.

		Relative Average A_{tail}				
Case	Scenario	JFCV	AIC	BIC	ED	JFCV(ED)
VIII	$\mathcal{F}_1$	1.00	0.64	0.64	0.99	0.58
	$\mathcal{F}_2$	1.00	0.82	0.82	1.35	0.67
	$\mathcal{F}_3$	1.00	1.40	1.42	0.68	0.37
	$\mathcal{F}_3^{(3)}$	1.00	2.49	2.54	1.22	0.59
	$\mathcal{F}_3^{(6)}$	1.00	1.66	1.69	0.81	0.38

Table 8 Numerical results for HRS with $N = 100$.

		Relative Average A				
Case	Scenario	JFCV	AIC	BIC	ED	JFCV(ED)
IX	$\mathcal{F}_1$	1.00	1.00	1.00	0.58	0.37
	$\mathcal{F}_2$	1.00	0.73	0.73	0.64	0.41
	$\mathcal{F}_3$	1.00	1.26	1.27	1.27	0.75
	$\mathcal{F}_3^{(3)}$	1.00	1.17	1.18	1.18	0.70
	$\mathcal{F}_3^{(6)}$	1.00	1.29	1.31	1.30	0.79

$$X_2 \sim \begin{cases} \mathrm{N}(100, 100), & \text{with probability } 0.5 \\ \mathrm{gamma}(20, 0.2), & \text{with probability } 0.5. \end{cases}$$

Notice that $\mathrm{E}(X_1) = 1$ and $\mathrm{E}(X_2) = 100$. Recall that Y is thought of as the time to failure.

Tables 8 and 9 present the results. This is a very difficult problem for which the distribution of Y is very sensitive to the input distributions. This makes the performance of JFCV(ED) impressive as it is clearly the best across all cases. The performance of the JFCV, AIC, BIC and ED methods is rather diverse. AIC's performance is either on a par with, or slightly better than, BIC. None of the JFCV, AIC, BIC and ED can strictly dominate the others, although JFCV tends to be the winner when considering A.

4.2. Evaluation of the Input Distribution

In this section we present results that directly assess the quality of the model average fit $\widehat{F}(x; \widehat{\mathbf{w}})$ with respect to the true distribution F^c. In the unlikely event

Table 9 **Numerical results for HRS, tail estimation, with $N = 100$.**

		Relative Average A_{tail}				
Case	Scenario	JFCV	AIC	BIC	ED	JFCV(ED)
IX	$\mathcal{F}_1$	1.00	1.00	1.00	0.66	0.55
	$\mathcal{F}_2$	1.00	0.85	0.85	0.70	0.57
	$\mathcal{F}_3$	1.00	0.74	0.75	0.70	0.51
	$\mathcal{F}_3^{(3)}$	1.00	1.20	1.22	1.13	0.82
	$\mathcal{F}_3^{(6)}$	1.00	0.83	0.85	0.79	0.61

that $F^c \in \mathcal{F}$, one should not expect model averaging to do better since an empirical weight of precisely 1 assigned to any particular distribution, including F^c, is a probability 0 outcome. Therefore, we focus on cases in which $F^c \notin \mathcal{F}$.

Specifically, our candidate set is all or part of

$$\mathcal{F} = \{\text{normal}, \text{lognormal}, \text{exponential}, \text{Weibull}, \text{gamma}, \text{ED}\}$$

while F^c is Rayleigh, Pareto, generalized lambda (Karian and Dudewicz 2000), hyperexponential, or mixtures of these. For measures of fit, we compared the mean and variance of the fitted distributions to those of F^c (as a sanity check), but more importantly compute

Kolmogorov-Smirnov distance (K-S): $\max_x \left|\widehat{F}(x;\widehat{\mathbf{w}}) - F^c(x)\right|$

Cramér von-Mises distance (Cv-M): $\int \left[\widehat{F}(x;\widehat{\mathbf{w}}) - F^c(x)\right]^2 dF^c(x)$

Anderson-Darling distance (A-D): $\int \frac{\left[\widehat{F}(x;\widehat{\mathbf{w}}) - F^c(x)\right]^2}{F^c(x)(1-F^c(x))} dF^c(x).$

K-S examines the largest absolute gap between the cdfs; Cv-M and A-D are likelihood weighted squared areas between them, with A-D further emphasizing differences in the tails. We also recorded the weights assigned to each distribution in the model average. Real-world sample sizes of $N = 100, 1000$ were employed, and all results were averaged over 1000 macroreplications of the experiment.

We present results that represent the more-favorable and less-favorable performance of model averaging from this large number of cases. Not surprisingly, no approach dominates on all instances and all measures, so "favorable" is somewhat subjective. Overall, we found the following:

- Model averaging can improve over any single choice from $\mathcal{F}$, and the best model average tends never to be worse.
- Including the ED in $\mathcal{F}$ is almost always valuable for measures other than K-S; ED alone often has the poorest K-S performance, which makes sense as it is a discrete approximation to a continuous F^c.
- Reducing the size of $\mathcal{F}$ to the top AIC/BIC choices before model averaging improves fit; often model averaging the single best fit and the ED is the concensus best choice.
- The more distinct F^c is from any other choice in $\mathcal{F}$, the more weight is applied to the ED; for instance, this occurred when we created a bimodel true distribution F^c via a mixture (all of the candidates in $\mathcal{F}$ are unimodal).
- In a targeted test to study the effect of nested distributions, we found that using $\mathcal{F} = \{\text{exponential}, \text{Weibull}, \text{gamma}\}$ for model averaging when F^c is exponential leads to a noticably poorer fit than choosing any one of the candidates. A tentative recommendation is to avoid nesting, such as including exponential and Erlang in a set that already includes Weibull and gamma.

4.2.1. More-favorable Performance Here F^c is Rayleigh with parameter 0.5 from which we have $N = 1000$ observations, with full candidate set $\mathcal{F} = \{\text{normal}, \text{lognormal}, \text{exponential}, \text{gamma}, \text{ED}\}$, and we use $J = 5$ folds for fitting the weights. The gamma distribution is always the best AIC fit. Results are shown in Table 10. Either gamma + ED or using all of $\mathcal{F}$ provide arguably the best fits based on our three performance measures. For the same experiment with only $N = 100$ "real-world" observations, gamma + ED was the best choice, and better than model averaging larger sets. This suggests that when the quantity of input data is small it is even more important to first screen the larger set $\mathcal{F}$ before model averaging.

Table 10 Results from 1000 macroreplications for $N = 1000$ observations from a Rayleigh distribution F^c.

Distributions	$\mathbf{w}$	K-S	Cv-M	A-D
gamma	1	0.034	0.501	3.137
ED	1	0.028	0.172	1.021
gamma+ED	(0.257, 0.743)	0.026	0.182	1.080
$\mathcal{F}$	(0.387, 0.052, 0.017, 0.479, 0.065)	0.018	0.148	1.941

Table 11 Results from 1000 macroreplications for $N = 100$ observations from a Pareto distribution F^c.

Distributions	$\mathbf{w}$	K-S	Cv-M	A-D
lognormal	1	0.065	0.203	1.256
gamma	1	0.064	0.206	1.259
Weibull	1	0.054	0.144	0.931
ED	1	0.085	0.164	0.990
logn+ED	(0.748, 0.252)	0.064	0.164	1.002
Weibull+ED	(0.678 0.322)	0.060	0.137	0.849
gamma+ED	(0.407, 0.593)	0.071	0.152	0.914
logn+Weibull+ED	(0.571, 0.312, 0.117)	0.056	0.143	0.875
logn+gamma+Weibull+ED	(0.575, 0.217, 0.098, 0.110)	0.056	0.142	0.872
$\mathcal{F}$	(0.073, 0.668, 0.039, 0.118, 0.103)	0.060	0.154	1.261

For a second favorable example, F^c is Pareto with location parameter 1 and shape parameter 3 from which we have $N = 100$ observations, with full candidate set $\mathcal{F} = \{\text{normal}, \text{lognormal}, \text{gamma}, \text{Weibull}, \text{ED}\}$, and we use $J = 5$ folds for fitting the weights. Either the gamma, lognormal or Weibull distribution was chosen as the best AIC fit in some macroreplication, so we included them all as individual choices. Results are shown in Table 11. Individually the Weibull provides a good fit in this case, yet improvement is still possible by model averaging a smaller set of distributions than the full set.

Although not shown here because the result is obvious, model averaging including the ED had very favorable performance relative to any single choice when F^c was obtained by a mixture (e.g., of two Rayleigh's with different parameters) so as to create a bimodel distribution; in such cases the ED received a weight of around 0.9. This illustrates that model averaging with the ED provides protection against a poorly chosen candidate set, which might occur if distribution fitting

Table 12 Results from 1000 macroreplications for $N = 100$ observations from a generalized lambda distribution F^c.

Distributions	**w**	K-S	Cv-M	A-D
logn	1	0.090	0.266	4.119
logn+ED	(0.170, 0.830)	0.080	0.162	1.079
logn+gamma+normal	(0.810, 0, 0.190)	0.096	0.269	4.164
logn+gamma+normal+ED	(0.091, 0, 0.011, 0.8)	0.079	0.163	1.091
$\mathcal{F}$	(0.014, 0.381, 0.130, 0.008, 0.008, 0.459)	0.111	0.231	9.118

was automated. Of course, bimodal and mixture distributions can be included as candidates.

4.2.2. Less-favorable Performance In all of our experiments there was *some* model average distribution that did as well or better than any single choice, but in a few cases this was very sensitive to the distributions chosen as candidates; the most extreme case follows.

In this example F^c is a generalized lambda distribution with $\lambda_1 = 3$, $\lambda_2 = 2$, $\lambda_3 = 1.5$ and $\lambda_4 = 0.5$. With these choices the density has a bathtub shape. On each of 1000 macroreplications, we obtained $N = 100$ observations, with full candidate set $\mathcal{F} = \{\text{normal}, \text{lognormal}, \text{exponential}, \text{gamma}, \text{Weibull}, \text{ED}\}$, and used $J = 5$ folds for fitting the weights. The lognormal was chosen as the best AIC fit, but we explored other combinations as well. Results are shown in Table 12. Notice that averages of lognormal + ED and lognormal + gamma + normal + ED offer significant improvement on all measures over the single lognormal choice, but lognormal + gamma + normal and the full set $\mathcal{F}$ have inferior A-D statistics.

5. Conclusions

Model risk due to input uncertainty arises because the fitted input distribution $\widehat{F}$ deviates from the true distribution of the input data F^c. When F^c is known to belong to a certain parametric family, then it makes sense to use statistically efficient parameter estimates, which would often be the MLEs. Many methods for quantifying the impact of input parameter uncertainty on simulation performance estimates for this case have been proposed.

However, at best we should expect a standard parametric family to be a good approximation for F^c, which means that there is error that does not disappear even as the real-world input sample size $N \to \infty$. When the input data are also used to select the family, as is common practice, then the possible error is compounded.

In this paper we proposed using frequentist model averaging as an innovative way to construct better input models, meaning input models that yield more faithful output performance. Since the optimal weights are unknown, we estimated them using J-fold cross-validation. We showed that under mild conditions the empirically optimal model average is unique and easily obtained; and under more restrictive conditions the empirically optimal weights yield the best possible weighted average distribution as the sample size increases. This method augments current input modeling practice, and requires no alternation of the simulation model or additional simulation runs.

We also observed that the empirical distribution (ED) is frequently a very good input-modeling choice when the objective is to get close to the ideal output distribution, F_{Y^c}; this seems not to be very well known. Including the ED in the candidate set $\mathcal{F}$ for model averaging hedges against possible inadequacy of the ED, as occurred in some of our examples, especially when tail behavior of Y is of interest. The JFCV(ED) input models were often the best by a significant margin, were always very good performers in our experiments and seem to be a powerful addition to the standard input modeling pallet.

Acknowledgments

Nelson's work was partially supported by the National Science Foundation of the United States under Grant Number CMMI-1634982. Wan's work was supported by the City University of Hong Kong (Grant No. 7004985), Hong Kong Research Grants Council (Grant No. CityU 11500419) and the National Natural Science Foundation of China (Grant No. 71973116). Zou's work was supported by the Ministry of Science and Technology of China (Grant No. 2016YFB0502301) and the National Natural Science Foundation of China (Grant Nos. 11971323 and 11529101). The authors thank Eunhye Song for insights on averaging with the empirical distribution, and the area editor, associate editor and referees for enlightening reviews.

References

Ankenman BE, Nelson BL (2012) A quick assessment of input uncertainty. Laroque C, Himmelspach J, Pasupathy R, Rose O, Uhrmacher AM, eds., *Proceedings of the 2012 Winter Simulation Conference*, 21:1–21:10 (Piscataway, NJ, USA: IEEE Press).

Banks J, Carson J, Nelson B, Nicol D (2010) *Discrete-Event System Simulation* (Upper Saddle River, New Jersey: Prentice Hall).

Barton RR (2012) Tutorial: Input uncertainty in output analysis. Laroque C, Himmelspach J, Pasupathy R, Rose O, Uhrmacher AM, eds., *Proceedings of the 2012 Winter Simulation Conference*, 1–12 (Piscataway, NJ, USA: IEEE Press), ISSN 0891-7736.

Barton RR, Nelson BL, Xie W (2013) Quantifying input uncertainty via simulation confidence intervals. *INFORMS Journal on Computing* 26(1):74–87.

Burt JM, Garman MB (1971) Conditional Monte Carlo: A simulation technique for stochastic network analysis. *Management Science* 18(3):207–217.

Cheng RCH, Holland W (1997) Sensitivity of computer simulation experiments to errors in input data. *Journal of Statistical Computation and Simulation* 57(1-4):219–241.

Cheng RCH, Holland W (1998) Two-point methods for assessing variability in simulation output. *Journal of Statistical Computation and Simulation* 60(3):183–205.

Chick SE (2001) Input distribution selection for simulation experiments: Accounting for input uncertainty. *Operations Research* 49(5):744–758.

Corlu C, Biller B (2013) A subset selection procedure under input parameter uncertainty. Pasupathy R, Kim SH, Tolk A, Hill R, Kuhl ME, eds., *Proceedings of the 2013 Winter Simulation Conference*, 463–473 (Piscataway, NJ, USA: IEEE Press).

Cule M, Samworth R, Stewart M (2010) Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72(5):545–607.

Fan W, Hong LJ, Zhang X (2013) Robust selection of the best. Pasupathy R, Kim SH, Tolk A, Hill R, Kuhl ME, eds., *Proceedings of the 2013 Winter Simulation Conference*, 868–876 (Piscataway, NJ, USA: IEEE Press), URL `http://dl.acm.org/citation.cfm?id=2675983.2676094`.

Ghosh S, Lam H (2015) Mirror descent stochastic approximation for computing worst-case stochastic input models. Yilmaz L, Chan WKV, Moon I, Roeder TMK, Macal C, Rossetti MD, eds., *Proceedings of the 2015 Winter Simulation Conference*, 425–436 (Piscataway, NJ, USA: IEEE Press).

Glynn PW, Lam H (2018) Constructing simulation output intervals under input uncertainty via data sectioning. Rabe M, Juan A, Mustafee N, Skoogh A, Jain S, Johansson B, eds., *Proceedings of the 2018 Winter Simulation Conference*, 1551–1562 (IEEE).

Gross D, Shortle JF, Thompson JM, Harris CM (2008) *Fundamentals of Queueing Theory* (New York: Wiley), 4th edition.

Handcock MS, Morris M (2006) *Relative Distribution Methods in the Social Sciences* (New York: Springer).

Hansen BE (2007) Least squares model averaging. *Econometrica* 75:1175–1189.

Hansen BE, Racine J (2012) Jackknife model averaging. *Journal of Econometrics* 167(1):38–46.

Heyde C, Kou S (2004) On the controversy over tailweight distributions. *Operations Research Letters* 32:399–408.

Hjort NL, Claeskens G (2003) Frequentist model average estimators. *Journal of the American Statistical Association* 98:879–899.

Jiang WX, Nelson BL (2018) Better input modeling via model averaging. Rabe M, Juan A, Mustafee N, Skoogh A, Jain S, Johansson B, eds., *Proceedings of the 2018 Winter Simulation Conference*, 1575–1586 (IEEE).

Karian ZA, Dudewicz EJ (2000) *Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Methods* (New York: CRC Press).

Kim AK, Samworth RJ, et al. (2016) Global rates of convergence in log-concave density estimation. *The Annals of Statistics* 44(6):2756–2779.

Lam H (2016) Input uncertainty and robust analysis in stochastic simulation: Advanced tutorial. Roeder TMK, Frazier PI, Szechtman R, Zhou E, Huschka T, Chick SE, eds., *Proceedings of the 2016 Winter Simulation Conference*, in press (Piscataway, NJ, USA: IEEE Press).

Lam H, Qian H (2018) Subsampling variance for input uncertainty quantification. Rabe M, Juan A, Mustafee N, Skoogh A, Jain S, Johansson B, eds., *Proceedings of the 2018 Winter Simulation Conference*, 1611–1622 (IEEE).

Law AM, Kelton WD (1991) *Simulation Modeling and Analysis* (New York: McGraw-Hill), 2nd edition.

Liang H, Zou G, Wan ATK, Zhang X (2011) On optimal weight choice in a frequentist model average estimator. *Journal of the American Statistical Association* 106:1053–1066.

McLachlan G, Peel D (2004) *Finite Mixture Models* (John Wiley & Sons).

Nocedal J, Wright SJ (2006) *Numerical Optimization* (New York: Springer), 2nd edition.

Song E, Nelson BL (2015) Quickly assessing contributions to input uncertainty. *IIE Transactions* 47:1–17.

Song E, Nelson BL, Hong LJ (2015) Input uncertainty and indifference-zone ranking & selection. Yilmaz L, Chan WKV, Moon I, Roeder TMK, Macal C, Rossetti MD, eds., *Proceedings of the 2015 Winter Simulation Conference*, 414–424 (Piscataway, NJ, USA: IEEE Press), ISBN 978-1-4673-9741-4, URL `http://dl.acm.org/citation.cfm?id=2888619.2888665`.

Song E, Nelson BL, Pegden CD (2014) Advanced tutorial: Input uncertainty quantification. Tolk A, Diallo S, Ryzhov I, Yilmaz L, Buckley S, Miller J, eds., *Proceedings of the 2014 Winter Simulation Conference*, 162–176 (Piscataway, NJ, USA: IEEE Press).

Wagner M, Wilson J (1996) Using univariate Bézier distributions to model simulation input processes. *IIE Transactions* 28(9):699–711.

Wan ATK, Zhang X, Zou G (2010) Least squares model averaging by Mallows criterion. *Journal of Econometrics* 156:277–283.

White H (1982) Maximum likelihood estimation of misspecified models. *Econometrica* 50:1–25.

Whitt W (1981) Approximating a point process by a renewal process: The view through a queue, an indirect approach. *Management Science* 27(6):619–636.

Xie W, Nelson BL, Barton RR (2014) A Bayesian framework for quantifying uncertainty in stochastic simulation. *Operations Research* 62(6):1439–1452.

Zhang X, Wan ATK, Zou G (2013) Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics* 174(2):82–94.

Zhou E, Xie W (2015) Simulation optimization when facing input uncertainty. Yilmaz L, Chan WKV, Moon I, Roeder TMK, Macal C, Rossetti MD, eds., *Proceedings of the 2015 Winter Simulation Conference*, 3714–3724 (Piscataway, NJ, USA: IEEE Press), ISBN 978-1-4673-9741-4, URL `http://dl.acm.org/citation.cfm?id=2888619.2889116`.

Zouaoui F, Wilson JR (2004) Accounting for input-model and input-parameter uncertainties in simulation. *IIE Transactions* 36(11):1135–1151.

Online Supplement

Appendix A: Proof of Optimality

We first provide the lemma below that is useful for proving the main results that follow. This lemma is based on the idea of Li (1987).

LEMMA 1. *Assume that the random functions $L_n(\mathbf{w})$ and $a_n(\mathbf{w})$ are continuous in $\mathbf{w}$, and $R_n(\mathbf{w})$ is a real-valued function, where $\mathbf{w} \in \mathcal{W}$ and $\mathcal{W}$ is a compact set in $\mathbb{R}^q$. Let*

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}\in\mathcal{W}}{argmin}\{L_n(\mathbf{w}) + a_n(\mathbf{w})\}.$$

If $L_n(\mathbf{w}) > 0$ and $R_n(\mathbf{w}) > 0$ almost surely for large n, and

$$\sup_{\mathbf{w}\in\mathcal{W}} \frac{|a_n(\mathbf{w})|}{R_n(\mathbf{w})} \xrightarrow{P} 0 \tag{11}$$

and

$$\sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{L_n(\mathbf{w})}{R_n(\mathbf{w})} - 1 \right| \xrightarrow{P} 0, \tag{12}$$

then

$$\frac{L_n(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{W}} L_n(\mathbf{w})} \xrightarrow{P} 1.$$

Proof. Our proof is a modification of Gao et al. (2019). From the definition of infimum, for any fixed n, there exist a non-negative sequence $\vartheta_{n,m}$ and a weight vector sequence $\mathbf{w}_{n,m} \in \mathcal{W}$ such that as $m \to \infty$, $\vartheta_{n,m} \to 0$ and

$$\inf_{\mathbf{w}\in\mathcal{W}} L_n(\mathbf{w}) = L_n(\mathbf{w}_{n,m}) - \vartheta_{n,m} \inf_{\mathbf{w}\in\mathcal{W}} L_n(\mathbf{w}).$$

So we have,

$$\begin{aligned}
0 \le 1 - \frac{\inf_{\mathbf{w}\in\mathcal{W}} L_n(\mathbf{w})}{L_n(\widehat{\mathbf{w}})} &= \frac{L_n(\widehat{\mathbf{w}}) + a_n(\widehat{\mathbf{w}}) - a_n(\widehat{\mathbf{w}}) - L_n(\mathbf{w}_{n,m}) + \vartheta_{n,m} \inf_{\mathbf{w}\in\mathcal{W}} L_n(\mathbf{w})}{L_n(\widehat{\mathbf{w}})} \\
&= \frac{\min_{\mathbf{w}\in\mathcal{W}}[L_n(\mathbf{w}) + a_n(\mathbf{w})] - a_n(\widehat{\mathbf{w}}) - L_n(\mathbf{w}_{n,m}) + \vartheta_{n,m} \inf_{\mathbf{w}\in\mathcal{W}} L_n(\mathbf{w})}{L_n(\widehat{\mathbf{w}})} \\
&\le \frac{L_n(\mathbf{w}_{n,m}) + a_n(\mathbf{w}_{n,m}) - a_n(\widehat{\mathbf{w}}) - L_n(\mathbf{w}_{n,m}) + \vartheta_{n,m} \inf_{\mathbf{w}\in\mathcal{W}} L_n(\mathbf{w})}{L_n(\widehat{\mathbf{w}})} \\
&\le \frac{|a_n(\mathbf{w}_{n,m})|}{L_n(\widehat{\mathbf{w}})} + \frac{|a_n(\widehat{\mathbf{w}})|}{L_n(\widehat{\mathbf{w}})} + \vartheta_{n,m} \\
&\le \frac{|a_n(\mathbf{w}_{n,m})|}{\inf_{\mathbf{w}\in\mathcal{W}} L_n(\mathbf{w})} + \frac{|a_n(\widehat{\mathbf{w}})|}{L_n(\widehat{\mathbf{w}})} + \vartheta_{n,m}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{|a_n(\mathbf{w}_{n,m})|/L_n(\mathbf{w}_{n,m})}{1-\vartheta_{n,m}\inf_{\mathbf{w}\in\mathcal{W}} L_n(\mathbf{w})/L_n(\mathbf{w}_{n,m})} + \frac{|a_n(\widehat{\mathbf{w}})|}{L_n(\widehat{\mathbf{w}})} + \vartheta_{n,m} \\
&\le \frac{\sup_{\mathbf{w}\in\mathcal{W}}\{|a_n(\mathbf{w})|/L_n(\mathbf{w})\}}{1-\vartheta_{n,m}\inf_{\mathbf{w}\in\mathcal{W}} L_n(\mathbf{w})/L_n(\mathbf{w}_{n,m})} + \sup_{\mathbf{w}\in\mathcal{W}}\frac{|a_n(\mathbf{w})|}{L_n(\mathbf{w})} + \vartheta_{n,m}.
\end{aligned}
$$

Letting $m\to\infty$ on both sides of the equation above we obtain

$$
0 \le 1 - \frac{\inf_{\mathbf{w}\in\mathcal{W}} L_n(\mathbf{w})}{L_n(\widehat{\mathbf{w}})} \le 2\times \sup_{\mathbf{w}\in\mathcal{W}} \frac{|a_n(\mathbf{w})|}{L_n(\mathbf{w})}. \tag{13}
$$

In the following, we will show that

$$
\sup_{\mathbf{w}\in\mathcal{W}} \frac{|a_n(\mathbf{w})|}{L_n(\mathbf{w})} \xrightarrow{P} 0. \tag{14}
$$

Together with (13), this proves Lemma 1.

Now, it is clear that

$$
\sup_{\mathbf{w}\in\mathcal{W}} \frac{|a_n(\mathbf{w})|}{L_n(\mathbf{w})} \le \sup_{\mathbf{w}\in\mathcal{W}} \frac{|a_n(\mathbf{w})|}{R_n(\mathbf{w})} \times \sup_{\mathbf{w}\in\mathcal{W}} \frac{R_n(\mathbf{w})}{L_n(\mathbf{w})}.
$$

By (11), it suffices to show that

$$
\sup_{\mathbf{w}\in\mathcal{W}} \frac{R_n(\mathbf{w})}{L_n(\mathbf{w})} = O_p(1). \tag{15}
$$

It can be verified that

$$
-\sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{L_n(\mathbf{w})}{R_n(\mathbf{w})} - 1\right| \le \inf_{\mathbf{w}\in\mathcal{W}} \frac{L_n(\mathbf{w})}{R_n(\mathbf{w})} - 1 \le \sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{L_n(\mathbf{w})}{R_n(\mathbf{w})} - 1\right|,
$$

which, together with (12), implies that

$$
\left|\inf_{\mathbf{w}\in\mathcal{W}} \frac{L_n(\mathbf{w})}{R_n(\mathbf{w})} - 1\right| \le \sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{L_n(\mathbf{w})}{R_n(\mathbf{w})} - 1\right| = o_p(1).
$$

Thus, we have

$$
\inf_{\mathbf{w}\in\mathcal{W}} \frac{L_n(\mathbf{w})}{R_n(\mathbf{w})} = 1 + o_p(1),
$$

which shows that (15) is true. This completes the proof of the lemma. □

Proof of Theorem 1. Recall from Section 3.3 that $\boldsymbol{\beta}_m$ is the unknown parameter vector in the m^{th} candidate distribution, $m = 1, 2, \ldots, q$, $\widehat{\boldsymbol{\beta}}_m$ its MLE that we assume exists, and $\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\beta}}_q^{\mathrm{T}}\right)^{\mathrm{T}}$ with dimension κ and q being finite. As stated in Section 3.3, from White (1982), under some regularity conditions, there exists a vector $\boldsymbol{\beta}^*$ such that

$$
\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = O_p(N^{-1/2}), \tag{16}
$$

as $N\to\infty$. In other words, the MLEs converge even when the distributions are misspecified. Our asymptotic results hold when (16) and Conditions (i)–(iii) are in force.

In addition to the notation in Section 3.3, define the $\kappa\times 1$ vector $\boldsymbol{\nu}_i(\mathbf{w},\widetilde{\boldsymbol{\beta}}_i)=\partial\widehat{F}(x_i,\mathbf{w})/\partial\widehat{\boldsymbol{\beta}}|_{\widehat{\boldsymbol{\beta}}=\widetilde{\boldsymbol{\beta}}_i}$, where $\widetilde{\boldsymbol{\beta}}_i$ is a $\kappa-$dimensional vector, $i=1,2,\ldots,N$. Further, write $\mathbf{Q}(\mathbf{w})=\mathbf{Q}(\mathbf{w},\widetilde{\boldsymbol{\beta}}_1,\ldots,\widetilde{\boldsymbol{\beta}}_N)=\Big(\boldsymbol{\nu}_1(\mathbf{w},\widetilde{\boldsymbol{\beta}}_1),\ldots,\boldsymbol{\nu}_N(\mathbf{w},\widetilde{\boldsymbol{\beta}}_N)\Big)^{\mathrm{T}}$, and denote $L_N(w)=||\widehat{\mathbf{F}}(w)-\mathbf{F}_0||^2$.

First, using the notations in Theorem 1, the conditions of $L_n(\mathbf{w})>0$ and $R_n(\mathbf{w})>0$ almost surely given in Lemma 1 are implied by

$$\inf_{\mathbf{w}\in\mathcal{W}} L_N(\mathbf{w})>0 \quad (a.s.) \quad \text{and} \quad \xi_N>0 \quad (a.s.) \tag{17}$$

for large N. The second part of (17) holds automatically for large N (say $N\geq N_0$) under Condition 3, which implies that ξ_N has a non-zero lower bound (a.s.) for large N. To explain the first part of (17), recognizing that $L_N(\mathbf{w})$ is continuous and $\mathcal{W}$ is compact, we can write $\inf_{\mathbf{w}\in\mathcal{W}} L_N(\mathbf{w})=L_N(\mathbf{w}^{(N)})$, where $\mathbf{w}^{(N)}=(w_1^{(N)},\cdots,w_q^{(N)})^T\in\mathcal{W}$ and the superscript $^{(N)}$ means that the corresponding sample size is N. If

$$\inf_{\mathbf{w}\in\mathcal{W}} L_N(\mathbf{w})=0 \quad \text{a.s.} \quad \text{(infinitely often)}$$

then there is a subsequence of $\{N\}$, say $\{N_l\}$, such that $L_{N_l}(\mathbf{w}^{(N_l)})=0$ (a.s.), $l=1,2,\cdots$. That is, for N_l $(l=1,2,\cdots)$, the following holds almost surely:

$$\begin{aligned}
&||\widehat{\mathbf{F}}(\mathbf{w}^{(N_l)})-\mathbf{F}_0||^2=0\\
\Leftrightarrow&\sum_{i=1}^{N_l}\Big[\widehat{F}(x_i,\mathbf{w}^{(N_l)})-F^c(x_i)\Big]^2=0\\
\Leftrightarrow&\sum_{i=1}^{N_l}\Big[\sum_{m=1}^{q} w_m^{(N_l)}F_m(x_i,\widehat{\boldsymbol{\beta}}_m^{(N_l)})-F^c(x_i)\Big]^2=0\\
\Leftrightarrow&\sum_{m=1}^{q} w_m^{(N_l)}F_m(x_i,\widehat{\boldsymbol{\beta}}_m^{(N_l)})=F^c(x_i), \quad i=1,\cdots,N_l.
\end{aligned} \tag{18}$$

Because $\mathbf{w}^{(N_l)}=(w_1^{(N_l)},\cdots,w_q^{(N_l)})^T$ is a bounded sequence in $\mathbb{R}^q$ and q is fixed, there is a subsequence $\Big\{\mathbf{w}^{N_{l,k}}\Big\}(k=1,2,\cdots)$, which is convergent. Without loss of generality, we let $\mathbf{w}^{(N_l)}\to\mathbf{w}^*=(w_1^*,\cdots,w_q^*)^T$. It is clear that $\mathbf{w}*\in\mathcal{W}$.

Now, from (16) and (18) and the continuity of F_m in $\boldsymbol{\beta}_m$, we have

$$\sum_{m=1}^{q} w_m^* F_m(x_i,\boldsymbol{\beta}_m^*)=F^c(x_i), \quad i=1,\cdots,N_0. \quad (a.s.)$$

Thus,

$$\begin{aligned}
\xi_{N_0}&\leq||\mathbf{F}^*(\mathbf{w}^*)-\mathbf{F}_0||^2\\
&=\Big[\sum_{m=1}^{q} w_m^* F_m(x_1,\boldsymbol{\beta}_m^*)-F^c(x_1)\Big]^2+\cdots+\Big[\sum_{m=1}^{q} w_m^* F_m(x_{N_0},\boldsymbol{\beta}_m^*)-F^c(x_{N_0})\Big]^2
\end{aligned}$$

$$= 0, \quad (a.s.)$$

which contradicts the fact that $\xi_{N_0} > 0$ (a.s.). Therefore, $\inf_{\mathbf{w}\in\mathcal{W}} L_N(\mathbf{w}) = 0$ (a.s.) holds only for finitely many sample sizes, and hence the first part of (17) is satisfied.

Next, we prove (8). Let us decompose $L_N(\mathbf{w})$ and $CV_J(\mathbf{w})$ as follows.

$$\begin{aligned} L_N(\mathbf{w}) &= \left\|\widehat{\mathbf{F}}(\mathbf{w}) - \mathbf{F}_0\right\|^2 \\ &= L_N^*(\mathbf{w}) + \left\|\widehat{\mathbf{F}}(\mathbf{w}) - \mathbf{F}^*(\mathbf{w})\right\|^2 + 2\left\{\mathbf{F}^*(\mathbf{w}) - \mathbf{F}_0\right\}^{\mathrm{T}}\left\{\widehat{\mathbf{F}}(\mathbf{w}) - \mathbf{F}^*(\mathbf{w})\right\} \\ &\equiv L_N^*(\mathbf{w}) + \Pi_N(\mathbf{w}) \end{aligned}$$

and

$$\begin{aligned} CV_J(\mathbf{w}) &= \left\|\widetilde{\mathbf{F}}(\mathbf{w}) - \bar{\mathbf{F}}\right\|^2 \\ &= L_N(\mathbf{w}) + \left\|\widehat{\mathbf{F}}(\mathbf{w}) - \widetilde{\mathbf{F}}(\mathbf{w})\right\|^2 - 2\left\{\widehat{\mathbf{F}}(\mathbf{w}) - \widetilde{\mathbf{F}}(\mathbf{w})\right\}^{\mathrm{T}}\left\{\widehat{\mathbf{F}}(\mathbf{w}) - \bar{\mathbf{F}}\right\} \\ &\quad +2\left\{\widehat{\mathbf{F}}(\mathbf{w}) - \widetilde{\mathbf{F}}(\mathbf{w})\right\}^{\mathrm{T}}(\mathbf{F}_0 - \bar{\mathbf{F}}) + 2\widetilde{\mathbf{F}}(\mathbf{w})^{\mathrm{T}}(\mathbf{F}_0 - \bar{\mathbf{F}}) - (\mathbf{F}_0 + \bar{\mathbf{F}})^{\mathrm{T}}(\mathbf{F}_0 - \bar{\mathbf{F}}) \\ &\equiv L_N(\mathbf{w}) + \Xi_N(\mathbf{w}) - (\mathbf{F}_0 + \bar{\mathbf{F}})^{\mathrm{T}}(\mathbf{F}_0 - \bar{\mathbf{F}}). \end{aligned}$$

Note that the last term in the expression above is independent of the weight vector $\mathbf{w}$, so

$$\widehat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}}\left\{L_N(\mathbf{w}) + \Xi_N(\mathbf{w})\right\}.$$

Hence, from Lemma 1, to prove (8), it suffices to prove

$$\sup_{\mathbf{w}\in\mathcal{W}} \frac{|\Xi_N(\mathbf{w})|}{L_N^*(\mathbf{w})} = o_p(1) \tag{19}$$

and

$$\sup_{\mathbf{w}\in\mathcal{W}} \frac{|\Pi_N(\mathbf{w})|}{L_N^*(\mathbf{w})} = o_p(1). \tag{20}$$

From (16) and assuming that Condition 1 holds, we have, uniformly for all $\mathbf{w}\in\mathcal{W}$,

$$\begin{aligned} &N^{-1/2}\left\|\mathbf{Q}(\mathbf{w})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|^2 \\ \le\ & N^{-1/2}\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|^2 \lambda_{\max}\left\{\mathbf{Q}^{\mathrm{T}}(\mathbf{w})\mathbf{Q}(\mathbf{w})\right\} \\ =\ & N^{-1/2}O_p(N^{-1})O_p(N) \\ =\ & O_p(N^{-1/2}), \end{aligned} \tag{21}$$

where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue of a matrix; recognizing also that the elements of vector $|\mathbf{F}^*(\mathbf{w}) - \mathbf{F}_0|$ are uniformly upper bounded by 2. Then

$$N^{-1/2}\left\{\mathbf{F}^*(\mathbf{w}) - \mathbf{F}_0\right\}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = N^{-1/2}O_p(N^{1/2}) = O_p(1). \tag{22}$$

By the Taylor-series expansion,

$$\widehat{\mathbf{F}}(\mathbf{w}) - \mathbf{F}^*(\mathbf{w}) = \mathbf{Q}(\mathbf{w})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*). \tag{23}$$

From (21)–(23), Condition 3, and recognizing that any element of the vector $|\mathbf{F}^*(\mathbf{w}) - \mathbf{F}_0|$ is upper bounded by 2, we have

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{W}} \frac{|\Pi_N(\mathbf{w})|}{L_N^*(\mathbf{w})} \\
&\le \xi_N^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left\|\widehat{\mathbf{F}}(\mathbf{w}) - \mathbf{F}^*(\mathbf{w})\right\|^2 + 2\xi_N^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \{\mathbf{F}^*(\mathbf{w}) - \mathbf{F}_0\}^{\mathrm{T}} \left\{\widehat{\mathbf{F}}(\mathbf{w}) - \mathbf{F}^*(\mathbf{w})\right\}\right| \\
&= \frac{N^{1/2}}{\xi_N} N^{-1/2} \sup_{\mathbf{w}\in\mathcal{W}} \left\|\mathbf{Q}(\mathbf{w})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|^2 + 2\frac{N^{1/2}}{\xi_N} N^{-1/2} \sup_{\mathbf{w}\in\mathcal{W}} \left| \{\mathbf{F}^*(\mathbf{w}) - \mathbf{F}_0\}^{\mathrm{T}} \mathbf{Q}(\mathbf{w})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right| \\
&= o_p(1),
\end{aligned}$$

which is (20). Similarly, from Conditions 2 and 3, we have

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{W}} \frac{\left| \left\|\widehat{\mathbf{F}}(\mathbf{w}) - \widetilde{\mathbf{F}}(\mathbf{w})\right\|^2 - 2\left\{\widehat{\mathbf{F}}(\mathbf{w}) - \widetilde{\mathbf{F}}(\mathbf{w})\right\}^{\mathrm{T}} \left\{\widehat{\mathbf{F}}(\mathbf{w}) - \bar{\mathbf{F}}\right\} + 2\left\{\widehat{\mathbf{F}}(\mathbf{w}) - \widetilde{\mathbf{F}}(\mathbf{w})\right\}^{\mathrm{T}} (\mathbf{F}_0 - \bar{\mathbf{F}})\right|}{L_N^*(\mathbf{w})} \\
&= \quad o_p(1).
\end{aligned} \tag{24}$$

Now, consider $F_0(x) - \bar{F}_{(j)}(x)$. Note that the Kolmogorov-Smirnov statistic (Billingsley 1968),

$$\sqrt{S} \sup_x |F_0(x) - \bar{F}_{(j)}(x)|$$

is convergent in distribution. Hence, for any x, from (5), we have

$$F_0(x) - \bar{F}_{(j)}(x) = F_0(x) - S^{-1} \sum_{s=1}^{S} I(x_{(j-1)S+s} \le x) = O_p(S^{-1/2}),$$

and thus,

$$\max_{i\in\{1,2,\dots,N\}} |F_0(x_i) - \bar{F}_{(j)}(x_i)| \le \sup_x |F_0(x) - \bar{F}_{(j)}(x)| = O_p(S^{-1/2}),$$

which, along with the fact that $x_1, \dots, x_N$ are i.i.d., implies that, uniformly for $i \in \{1, 2, \dots, N\}$,

$$F_0(x_i) - \bar{F}_{(j)}(x_i) = O_p(S^{-1/2}). \tag{25}$$

From $c_N \to 0$, (25), and recognizing that any element of $|\widehat{\mathbf{F}}_m|$ is upper bounded by 1, we obtain, for any $j \in \{1, 2, \dots, J\}$,

$$\begin{aligned}
&c_N^{1/2} S^{-1/2} \left| \sum_{s=1}^{S} \widetilde{F}_m^{(-j)}(x_{(j-1)S+s}) \left\{F_0(x_{(j-1)S+s}) - \bar{F}_{(j)}(x_{(j-1)S+s})\right\}\right| \\
&\le c_N^{1/2} S^{-1/2} \sum_{s=1}^{S} \left| \widetilde{F}_m^{(-j)}(x_{(j-1)S+s}) \left\{F_0(x_{(j-1)S+s}) - \bar{F}_{(j)}(x_{(j-1)S+s})\right\}\right|
\end{aligned}$$

$$\begin{aligned}
&\le c_N^{1/2} S^{-1/2} \sum_{s=1}^{S} \left| F_0(x_{(j-1)S+s}) - \bar{F}_{(j)}(x_{(j-1)S+s}) \right| \\
&= o_p(1),
\end{aligned}$$

which, along with Condition 3 and the assumption that J is finite, implies that

$$\begin{aligned}
&\xi_N^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \widetilde{\mathbf{F}}(\mathbf{w})^{\mathrm{T}} (\mathbf{F}_0 - \bar{\mathbf{F}}) \right| \\
&\le \sum_{m=1}^{q} \frac{N^{1/2}}{\xi_N} N^{-1/2} \left| \widetilde{\mathbf{F}}_m^{\mathrm{T}} (\mathbf{F}_0 - \bar{\mathbf{F}}) \right| \\
&= \sum_{m=1}^{q} J^{-1/2} \frac{N^{1/2}}{\xi_N} \left| \sum_{j=1}^{J} S^{-1/2} \sum_{s=1}^{S} \widetilde{F}_m^{(-j)}(x_{(j-1)S+s}) \left\{ F_0(x_{(j-1)S+s}) - \bar{F}_{(j)}(x_{(j-1)S+s}) \right\} \right| \\
&\le \sum_{m=1}^{q} J^{-1/2} c_N^{1/2} \left| \sum_{j=1}^{J} S^{-1/2} \sum_{s=1}^{S} \widetilde{F}_m^{(-j)}(x_{(j-1)S+s}) \left\{ F_0(x_{(j-1)S+s}) - \bar{F}_{(j)}(x_{(j-1)S+s}) \right\} \right| \\
&= \sum_{m=1}^{q} J^{-1/2} \left| \sum_{j=1}^{J} c_N^{1/2} S^{-1/2} \sum_{s=1}^{S} \widetilde{F}_m^{(-j)}(x_{(j-1)S+s}) \left\{ F_0(x_{(j-1)S+s}) - \bar{F}_{(j)}(x_{(j-1)S+s}) \right\} \right| \\
&\le \sum_{m=1}^{q} J^{-1/2} \sum_{j=1}^{J} c_N^{1/2} S^{-1/2} \sum_{s=1}^{S} \left| \widetilde{F}_m^{(-j)}(x_{(j-1)S+s}) \left\{ F_0(x_{(j-1)S+s}) - \bar{F}_{(j)}(x_{(j-1)S+s}) \right\} \right| \\
&\le \sum_{m=1}^{q} J^{-1/2} \sum_{j=1}^{J} c_N^{1/2} S^{-1/2} \sum_{s=1}^{S} \left| F_0(x_{(j-1)S+s}) - \bar{F}_{(j)}(x_{(j-1)S+s}) \right| \\
&\le q J^{-1/2} \sum_{j=1}^{J} c_N^{1/2} S^{-1/2} \cdot S \max_{i\in\{1,2,\dots,N\}} \left| F_0(x_i) - \bar{F}_{(j)}(x_i) \right| \\
&= c_N^{1/2} \cdot O_p(1) \qquad (26) \\
&= o_p(1). \qquad (27)
\end{aligned}$$

From (24) and (27), we obtain (19) and this completes the proof. □

Appendix B: Convergence of the Weight on ED

Proof of Theorem 2. Here we show that when we employ JFCV(ED) the weight on the ED converges to 1 as the real-world sample size $N \to \infty$. This is important because the ED is unbiased and asymptotically consistent for F^c, so as the sample size increases the incorrect choices in $\mathcal{F}$ should not enter into the average. The following result shows that this is indeed the case when the weights are estimated via J-fold CV.

Because the proof is easier to follow, we begin with the case of $q = 2$ candidate distributions, the first being a misspecified parametric distribution and the second being the ED. Denote their weights chosen by the proposed cross-validation criterion as $\widehat{w}_1$ and $\widehat{w}_{ED}$ respectively and write

$\widehat{\mathbf{w}} = (\widehat{w}_1, \widehat{w}_{ED})^{\mathrm{T}}$. We will prove that $\widehat{w}_{ED} \to 1$ in probability as $N \to \infty$. To facilitate the proof, consider the weight vector $\widetilde{\mathbf{w}} = (0,1)^{\mathrm{T}}$. From the proof in Appendix A and Condition 2, we know that

$$CV_J(\widetilde{\mathbf{w}}) = O_p(1) \tag{28}$$

and

$$CV_J(\widehat{\mathbf{w}}) = O_p(N^{1/2}) + \widehat{w}_1^2 \left\|\widehat{\mathbf{F}}_1 - \bar{\mathbf{F}}\right\|^2$$

as $N \to \infty$. These yield

$$\widehat{w}_1^2 = \left\|\widehat{\mathbf{F}}_1 - \bar{\mathbf{F}}\right\|^{-2} CV_J(\widehat{\mathbf{w}}) - \left\|\widehat{\mathbf{F}}_1 - \bar{\mathbf{F}}\right\|^{-2} O_p(N^{1/2}). \tag{29}$$

Since $\widehat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}} CV_J(\mathbf{w})$, we have $CV_J(\widetilde{\mathbf{w}}) \geq CV_J(\widehat{\mathbf{w}})$, which, along with (28), implies

$$CV_J(\widehat{\mathbf{w}}) = O_p(1). \tag{30}$$

By (29) and (30), we have

$$\widehat{w}_1^2 = O_p(N^{1/2}) \left\|\widehat{\mathbf{F}}_1 - \bar{\mathbf{F}}\right\|^{-2},$$

We now establish the relationship between $||\widehat{\mathbf{F}}_1 - \bar{\mathbf{F}}||$ and ξ_N. Let $\mathbf{w}^{(1)} = (1,0)^T$. Then $\widehat{\mathbf{F}}_1$ can be written as $\widehat{\mathbf{F}}(\mathbf{w}^{(1)})$. Hence we have

$$\begin{aligned} ||\widehat{\mathbf{F}}_1 - \bar{\mathbf{F}}|| &= ||\widehat{\mathbf{F}}(\mathbf{w}^{(1)}) - \bar{\mathbf{F}}|| \\ &\geq ||\mathbf{F}^*(\mathbf{w}^{(1)}) - \mathbf{F}_0|| - ||\widehat{\mathbf{F}}(\mathbf{w}^{(1)}) - \mathbf{F}^*(\mathbf{w}^{(1)})|| - ||\mathbf{F}_0 - \bar{\mathbf{F}}|| \\ &\geq \sqrt{\xi_N} - ||\widehat{\mathbf{F}}(\mathbf{w}^{(1)}) - \mathbf{F}^*(\mathbf{w}^{(1)})|| - ||\mathbf{F}_0 - \bar{\mathbf{F}}||. \end{aligned}$$

From equations (21) and (23), it is seen that

$$||\widehat{\mathbf{F}}(\mathbf{w}^{(1)}) - \mathbf{F}^*(\mathbf{w}^{(1)})|| = O_p(1),$$

which, together with $||\mathbf{F}_0 - \bar{\mathbf{F}}|| = O_p(1)$, implies that

$$||\widehat{\mathbf{F}}_1 - \bar{\mathbf{F}}|| \geq \sqrt{\xi_N} + O_p(1).$$

Recognising this relationship between $||\widehat{\mathbf{F}}_1 - \bar{\mathbf{F}}||$ and $\sqrt{\xi_N}$, and using Condition 3, we obtain

$$N^{-1/4}||\widehat{\mathbf{F}}_1 - \bar{\mathbf{F}}|| \geq (\xi_N^2/N)^{1/4} + o_p(1) \geq c_N^{-1/4} + o_p(1).$$

As $c_N \to 0$, we have $N^{1/4}||\widehat{\mathbf{F}}_1 - \bar{\mathbf{F}}||^{-1} = o_p(1)$. Hence, $\widehat{w}_1 = o_p(1)$ and thus

$$\widehat{w}_{ED} \to 1$$

in probability, as $N \to \infty$.

Next we consider the general case of $q \geq 2$ candidate distributions with the first $q-1$ being misspecified parametric distributions and the q^{th} being the ED. Denote their weights chosen by the proposed cross-validation criterion by $\widehat{w}_1, \widehat{w}_2, \cdots, \widehat{w}_{q-1}, \widehat{w}_{ED}$ respectively and write $\widehat{\mathbf{w}} = (\widehat{w}_1, \widehat{w}_2, \cdots, \widehat{w}_{q-1}, \widehat{w}_{ED})^{\mathrm{T}}$. We will prove that $\widehat{w}_{ED} \to 1$ in probability as $N \to \infty$. Consider the weight vector $\widetilde{\mathbf{w}} = (0, 0, \cdots, 0, 1)^{\mathrm{T}}$, $\bar{\mathbf{w}} = (w_1/(1-w_{ED}), w_2/(1-w_{ED}), \cdots, w_{q-1}/(1-w_{ED}), 0)^{\mathrm{T}}$ and $\widehat{\bar{\mathbf{w}}} = (\widehat{w}_1/(1-\widehat{w}_{ED}), \widehat{w}_2/(1-\widehat{w}_{ED}), \cdots, \widehat{w}_{q-1}/(1-\widehat{w}_{ED}), 0)^{\mathrm{T}}$. Let $\bar{\mathbf{F}}_{ED} = (\bar{F}(x_1), \ldots, \bar{F}(x_N))^{\mathrm{T}}$. Note that

$$\begin{aligned} L_N(\mathbf{w}) &= \left\| \widehat{\mathbf{F}}(\mathbf{w}) - \mathbf{F}_0 \right\|^2 \\ &= \left\| (1-w_{ED})\widehat{\mathbf{F}}(\bar{\mathbf{w}}) + w_{ED}\bar{\mathbf{F}}_{ED} - \mathbf{F}_0 \right\|^2 \\ &= \left\| (1-w_{ED})(\widehat{\mathbf{F}}(\bar{\mathbf{w}}) - \mathbf{F}_0) + w_{ED}(\bar{\mathbf{F}}_{ED} - \mathbf{F}_0) \right\|^2 \\ &= (1-w_{ED})^2 \left\| \widehat{\mathbf{F}}(\bar{\mathbf{w}}) - \mathbf{F}_0 \right\|^2 + w_{ED}^2 \left\| \bar{\mathbf{F}}_{ED} - \mathbf{F}_0 \right\|^2 \\ &\quad +2(1-w_{ED})w_{ED}(\widehat{\mathbf{F}}(\bar{\mathbf{w}}) - \mathbf{F}_0)^{\mathrm{T}}(\bar{\mathbf{F}}_{ED} - \mathbf{F}_0). \end{aligned}$$

Together with the proof in Appendix A and Condition 2, and recognizing that $||\mathbf{F}_0 - \bar{\mathbf{F}}|| = O_p(1)$ and $||\mathbf{F}_0 - \bar{\mathbf{F}}_{ED}|| = O_p(1)$, we have

$$CV_J(\widetilde{\mathbf{w}}) = O_p(1)$$

and

$$CV_J(\widehat{\mathbf{w}}) = O_p(N^{1/2}) + (1-\widehat{w}_{ED})^2 \left\| \widehat{\mathbf{F}}(\widehat{\bar{\mathbf{w}}}) - \mathbf{F}_0 \right\|^2$$

as $N \to \infty$. As $\widehat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w} \in \mathcal{W}} CV_J(\mathbf{w})$, we have $CV_J(\widetilde{\mathbf{w}}) \geq CV_J(\widehat{\mathbf{w}})$, which, along with the two results given above, implies that

$$(1-\widehat{w}_{ED})^2 = O_p(N^{1/2}) \left\| \widehat{\mathbf{F}}(\widehat{\bar{\mathbf{w}}}) - \mathbf{F}_0 \right\|^{-2}. \tag{31}$$

By the definition of ξ_N, we know that

$$\begin{aligned} \left\| \widehat{\mathbf{F}}(\widehat{\bar{\mathbf{w}}}) - \mathbf{F}_0 \right\| &\geq \left\| \mathbf{F}^*(\widehat{\bar{\mathbf{w}}}) - \mathbf{F}_0 \right\| - \left\| \widehat{\mathbf{F}}(\widehat{\bar{\mathbf{w}}}) - \mathbf{F}^*(\widehat{\bar{\mathbf{w}}}) \right\| \\ &\geq \sqrt{\xi_N} - \left\| \widehat{\mathbf{F}}(\widehat{\bar{\mathbf{w}}}) - \mathbf{F}^*(\widehat{\bar{\mathbf{w}}}) \right\| \end{aligned} \tag{32}$$

almost surely. From equations (19) and (21), it is seen that

$$||\widehat{\mathbf{F}}(\widehat{\bar{\mathbf{w}}}) - \mathbf{F}^*(\widehat{\bar{\mathbf{w}}})|| = O_p(1),$$

which, together with (32), implies that

$$||\widehat{\mathbf{F}}(\widehat{\mathbf{w}}) - \mathbf{F}_0|| \geq \sqrt{\xi_N} + O_p(1).$$

Together with (31) and Condition 3, this implies

$$1 - \widehat{w}_{ED} = o_p(1),$$

and thus

$$\widehat{w}_{ED} \to 1$$

in probability, as $N \to \infty$. $\square$

Appendix C: Uniqueness of the Model-Average Weight

Recall that

$$\begin{aligned} CV_J(\mathbf{w}) &= \sum_{j=1}^{J}\sum_{s=1}^{S}\left\{\widetilde{F}^{(-j)}(x_{(j-1)S+s},\mathbf{w}) - \bar{F}_{(j)}(x_{(j-1)S+s})\right\}^2 \\ &= \sum_{j=1}^{J}\sum_{s=1}^{S}\left\{\sum_{m=1}^{q} w_m\left(\widetilde{F}_m^{(-j)}(x_{(j-1)S+s}) - \bar{F}_{(j)}(x_{(j-1)S+s})\right)\right\}^2 \\ &= \sum_{j=1}^{J}\sum_{s=1}^{S}\left\{\sum_{m=1}^{q} w_m c_{mjs}\right\}^2 \\ &= \sum_{j=1}^{J}\sum_{s=1}^{S}\mathbf{w}^\top\mathbf{C}_{js}\ \mathbf{w} = \mathbf{w}^\top\mathbf{C}\,\mathbf{w} \end{aligned}$$

and we need to establish that $\mathbf{C}$ is positive definite (it is clearly at least positive semi-definite). Without loss of generality, assume that within each fold j, the data are sorted so that $x_{(j-1)S+1} < x_{(j-1)S+2} < \cdots < x_{jS}$. Then for $\mathbf{C}$ to be only positive semi-definite, there must be a weight vector $\mathbf{w}' \neq \mathbf{0}$ such that

$$\sum_{m=1}^{q} w'_m c_{mjs} = \sum_{m=1}^{q} w'_m\left(\widetilde{F}_m^{(-j)}(x_{(j-1)S+s}) - \frac{s}{S}\right) = 0.$$

Therefore,

$$\sum_{m=1}^{q} w'_m \widetilde{F}_m^{(-j)}(x_{(j-1)S+s}) = \frac{s}{S}$$

for each $j = 1, 2, \ldots, J$ and $s = 1, 2, \ldots, S$. If X has a density, then $x_{(j-1)S+s}$ and $x_{(j'-1)S+s}$ are distinct with probabilty 1 when $j \neq j'$, and we would expect $\widetilde{F}_m^{(-j)}(\cdot)$ to be distinct for each j, m (although not required in our argument). Since any continuous $\widetilde{F}_m^{(-j)}$ is increasing, the existence of $\mathbf{w}'$ has probability 0.

Appendix D: Additional SAN Empirical Evaluation

This appendix summarizes a large collection of empirical results for the SAN problem that further explore the use of model averaging. We employed the following candidate sets of distributions for this evaluation:

$$\begin{aligned}
\mathcal{F}_1 &= \{\text{truncated normal, beta, gamma}\} \\
\mathcal{F}_2 &= \mathcal{F}_1 \cup \{\text{lognormal, Weibull}\} \\
\mathcal{F}_3 &= \mathcal{F}_2 \cup \{\text{negative binomial, discrete uniform, Poisson,} \\
&\qquad \text{continuous uniform, loglogistic, inverse Gaussian, Pareto, binomial}\}
\end{aligned}$$

See Table 13 for the true distribution cases of $X_1, X_2, \ldots, X_5$.

We start with Cases I–III, for which F^c is *not* in the candidate set. We can see from Table 14 that for these cases, applying model averaging without including the ED in the candidate set does not always lead to an improvement in efficiency relative to the AIC and BIC choices. The JFCV method results in improved efficiency over the AIC and BIC choices only in a small number of situations. Although there are exceptions, AIC and BIC typically yield the same performance. The ED method invariably results in the most accurate outcomes among the JFCV, AIC, BIC and ED methods. Significantly, when the ED is included as a candidate distribution in the model average, then the resultant JFCV(ED) estimator has a very clear advantage over all of the methods considered and outperforms the ED convincingly. The JFCV(ED) estimator may be viewed as a smoothed variant of the ED. The superior performance of the JFCV(ED) method indicates that there is an advantage to smoothing the ED using the other candidate distributions as "smoothers."

For Cases IV–V the true activity-time distributions are included in the candidate sets, except under $\mathcal{F}_1$. The JFCV(ED) method is not found to be superior to the AIC and BIC model selections, which generally result in the best estimation outcomes; however, the performance of JFCV(ED) is close to AIC and BIC. The ED usually performs worse than AIC, BIC and JFCV, and the JFCV(ED) remains superior to the JFCV for the majority of situations. The less compelling performance of JFCV and JFCV(ED) for Cases IV–V is attributed to the fact that the true input distributions, namely the lognormal and Weibull distributions, are contained in the candidate sets $\mathcal{F}_2$ and $\mathcal{F}_3$. These are distinctive distributions and it comes as no surprise that the AIC and BIC methods successfully identify them as the true distributions. On the other hand, under $\mathcal{F}_1$, the candidate set does not contain the true input distributions, and the ranking of the various estimators is very similar to that observed for Cases I–III.

We also investigate the impact of varying the sample size on the performance of the various methods. We focus on Case I, and vary N from 20 to 1000. Table 15 presents the results. To evaluate the effect of a change in N on the actual difference between $G(u)$ produced by a given method and the 45° line, we also report the actual average area A in addition to the relative average area for each method. The results show that with few exceptions, other things being equal, an increase in N has the effect of bringing the distribution of $\widehat{Y}$ closer to the distribution of Y^c. For the comparisons in terms of relative area, the pattern of results observed for N=100 in Table 14 also apply in broad terms for other values of N. Specifically, for all values of N considered, JFCV(ED) is always the best of all methods; with few exceptions, ED yields more efficient estimates than the JFCV, which can have an advantage over AIC and BIC selection although the converse can also be true.

In addition, we examine the change in the empirical weight $\widehat{w}_{\mathrm{ED}}$ for the ED in the JFCV(ED) estimator as N varies, using Case I as an example. The results, also reported in Table 15, show that other things being equal, an increase in N is invariably accompanied by an increase in $\widehat{w}_{\mathrm{ED}}$. When N=1000, $\widehat{w}_{\mathrm{ED}}$ is very close to unity under all candidate sets. This finding corroborates the theoretical result in Appendix B which shows that $\widehat{w}_{\mathrm{ED}}$ converges in probability to unity as N approaches infinity.

In experiments not reported here, we also varied the number of CV folds $J = 5, 10$ for sample sizes $N = 100, 1000$ and found the conclusions above to be robust to the choice of J.

The results above are for the average area, A. Table 16 reports the results when interest centers on A_{tail}. At a high level, notice that the ED can perform quite badly when tail behavior of the output is of interest, and that AIC and BIC perform better when the true distribution is in the candidate set, but JFCV(ED) is quite close. Under Cases I–III, the JFCV(ED) estimator usually delivers the best outcome; exceptions occur under $\mathcal{F}_1$ for Cases I and II, where it is found that the JFCV(ED) estimator can have slightly worse performance than the JFCV estimator. Again, the AIC and BIC selection methods are not always inferior strategies; in fact, they yield better estimates than the JFCV method, although worse estimates than the JFCV(ED) for most situations in Cases I–III. Similar to the situation when interest focuses on A, for Cases IV–V AIC and BIC generally yield the best estimates except under $\mathcal{F}_1$. Comparing Tables 14 and 16, there is an apparent deterioration in the relative performance of the ED method when interest is shifted to A_{tail}. As noted previously, one deficiency of the ED method is that it takes no account of information beyond the largest data point in the real-world sample, and as such it may fail to accurately generate the tail of the simulation output. The worse performance of the ED observed in Table 16 for Cases IV–V is likely a reflection of this deficiency.

Table 13 **True input distributions for SAN experiment, where $B_i \sim \text{Bernoulli}(p_i)$, $i = 1, 2, \ldots, 5$, with $(p_1, p_2, \ldots, p_5) = (0.5, 0.5, 0.5, 0.5, 0.5)$, and $B_iG_i(x) + (1 - B_i)H_i(x)$, meaning that X_i follows $G_i(x)$ when $B_i = 1$ and $H_i(x)$ when $B_i = 0$.**

Case	X_1	X_2	X_3	X_4	X_5
I	B_1exponential(1/1.5) $+(1-B_1)$exponential(10)	B_1exponential(1/1.8) $+(1-B_1)$exponential(15)	B_1exponential(1/1.2) $+(1-B_1)$exponential(1/0.08)	B_1exponential(1/1.8) $+(1-B_1)$exponential(5)	B_1exponential(1/1.5) $+(1-B_1)$exponential(20)
II	B_1lognormal$(10, \sqrt{10})$ $+(1-B_1)$lognormal$(5, \sqrt{10})$	B_1exponential(1/1.8) $+(1-B_1)$exponential(15)	B_1Weibull(2, 6) $+(1-B_1)$exponential(1/0.08)	B_1exponential(1/1.8) $+(1-B_1)$exponential(5)	B_1Weibull(1, 6) $+(1-B_1)$exponential(20)
III	B_1unif(0, 2) $+(1-B_1)$exponential(15)	B_2HalfNor(0, 0, 4) $+(1-B_2)$unif(10, 20)	B_3exponential(2) $+(1-B_3)$exponential(12)	B_4gamma(15, 1) $+(1-B_4)$lognormal$(10, \sqrt{10})$	B_5gamma(15, 1) $+(1-B_5)$beta(0.5, 0.5)
IV	lognormal$(10, \sqrt{10})$	lognormal$(10, \sqrt{10})$	lognormal$(5, \sqrt{10})$	lognormal$(10, \sqrt{10})$	lognormal$(10, \sqrt{10})$
V	Weibull(2, 6)	Weibull(2, 6)	Weibull(1, 6)	Weibull(2, 6)	Weibull(2, 6)

Table 14 **Numerical results for SAN Experiment with $N = 100$.**

		Relative Average A				
Case	Scenario	JFCV	AIC	BIC	ED	JFCV(ED)
I	$\mathcal{F}_1$	1.0000	1.0001	1.0001	0.9624	0.7999
	$\mathcal{F}_2$	1.0000	1.2277	1.2277	0.8090	0.7236
	$\mathcal{F}_3$	1.0000	0.5916	0.5916	0.3880	0.3526
	$\mathcal{F}_3^{(3)}$	1.0000	0.9653	0.9653	0.6332	0.5435
	$\mathcal{F}_3^{(6)}$	1.0000	0.9101	0.9101	0.5969	0.5073
II	$\mathcal{F}_1$	1.0000	0.9384	0.9384	0.7343	0.5675
	$\mathcal{F}_2$	1.0000	0.6593	0.6593	0.4628	0.3743
	$\mathcal{F}_3$	1.0000	0.7619	0.7619	0.5538	0.4653
	$\mathcal{F}_3^{(3)}$	1.0000	0.7843	0.7843	0.5700	0.4515
	$\mathcal{F}_3^{(6)}$	1.0000	0.9185	0.9185	0.6676	0.5185
III	$\mathcal{F}_1$	1.0000	0.6655	0.6655	0.8816	0.6590
	$\mathcal{F}_2$	1.0000	1.0312	1.0312	0.7109	0.5333
	$\mathcal{F}_3$	1.0000	0.6552	0.6588	0.4309	0.3556
	$\mathcal{F}_3^{(3)}$	1.0000	0.7628	0.7670	0.5016	0.3845
	$\mathcal{F}_3^{(6)}$	1.0000	0.7619	0.7661	0.5010	0.3883
IV	$\mathcal{F}_1$	1.0000	0.8947	0.8947	1.1735	0.8293
	$\mathcal{F}_2$	1.0000	0.6306	0.6306	0.9502	0.6783
	$\mathcal{F}_3$	1.0000	0.9363	0.9363	1.4419	0.9736
	$\mathcal{F}_3^{(3)}$	1.0000	0.9836	0.9836	1.5148	1.0275
	$\mathcal{F}_3^{(6)}$	1.0000	0.9850	0.9850	1.5168	1.0096
V	$\mathcal{F}_1$	1.0000	0.6587	0.6587	0.8005	0.5806
	$\mathcal{F}_2$	1.0000	0.5111	0.5111	0.7192	0.6207
	$\mathcal{F}_3$	1.0000	0.7800	0.7800	1.0957	0.8384
	$\mathcal{F}_3^{(3)}$	1.0000	0.8921	0.8921	1.2531	0.9462
	$\mathcal{F}_3^{(6)}$	1.0000	0.9126	0.9126	1.2819	0.9691

Table 15 Numerical results for SAN experiment Case I with varying N.

		Actual Average A					Relative Average A					Weight
Sample Size	Scenario	JFCV	AIC	BIC	ED	JFCV (ED)	JFCV	AIC	BIC	ED	JFCV (ED)	on the ED in JFCV(ED)
20	$\mathcal{F}_1$	0.0602	0.0576	0.0576	0.0756	0.0593	1.0000	0.9559	0.9559	1.2561	0.9856	0.9564
	$\mathcal{F}_2$	0.0666	0.0719	0.0719	0.0756	0.0598	1.0000	1.0807	1.0807	1.1365	0.8980	0.9290
	$\mathcal{F}_3$	0.0863	0.0806	0.0806	0.0756	0.0629	1.0000	0.9337	0.9337	0.8760	0.7279	0.9123
	$\mathcal{F}_3^{(3)}$	0.0775	0.0806	0.0806	0.0756	0.0615	1.0000	1.0402	1.0402	0.9759	0.7928	0.9276
	$\mathcal{F}_3^{(6)}$	0.0784	0.0806	0.0806	0.0756	0.0610	1.0000	1.0284	1.0284	0.9649	0.7780	0.9154
50	$\mathcal{F}_1$	0.0357	0.0357	0.0357	0.0469	0.0344	1.0000	1.0000	1.0000	1.3151	0.9650	0.9741
	$\mathcal{F}_2$	0.0401	0.0517	0.0517	0.0469	0.0362	1.0000	1.2894	1.2894	1.1712	0.9020	0.9583
	$\mathcal{F}_3$	0.0841	0.0530	0.0530	0.0469	0.0421	1.0000	0.6298	0.6298	0.5579	0.5005	0.9347
	$\mathcal{F}_3^{(3)}$	0.0564	0.0530	0.0530	0.0469	0.0351	1.0000	0.9395	0.9395	0.8323	0.6220	0.9393
	$\mathcal{F}_3^{(6)}$	0.0603	0.0530	0.0530	0.0469	0.0352	1.0000	0.8785	0.8785	0.7782	0.5836	0.9374
100	$\mathcal{F}_1$	0.0338	0.0338	0.0338	0.0325	0.0270	1.0000	1.0001	1.0001	0.9624	0.7999	0.9796
	$\mathcal{F}_2$	0.0402	0.0493	0.0493	0.0325	0.0291	1.0000	1.2277	1.2277	0.8090	0.7236	0.9699
	$\mathcal{F}_3$	0.0838	0.0496	0.0496	0.0325	0.0295	1.0000	0.5916	0.5916	0.3880	0.3526	0.9474
	$\mathcal{F}_3^{(3)}$	0.0514	0.0496	0.0496	0.0325	0.0279	1.0000	0.9653	0.9653	0.6332	0.5435	0.9575
	$\mathcal{F}_3^{(6)}$	0.0545	0.0496	0.0496	0.0325	0.0276	1.0000	0.9101	0.9101	0.5969	0.5073	0.9541
200	$\mathcal{F}_1$	0.0252	0.0252	0.0252	0.0230	0.0187	1.0000	1.0000	1.0000	0.9105	0.7428	0.9904
	$\mathcal{F}_2$	0.0274	0.0489	0.0489	0.0230	0.0194	1.0000	1.7831	1.7831	0.8375	0.7062	0.9785
	$\mathcal{F}_3$	0.0916	0.0487	0.0485	0.0230	0.0222	1.0000	0.5315	0.5298	0.2505	0.2423	0.9585
	$\mathcal{F}_3^{(3)}$	0.0521	0.0487	0.0485	0.0230	0.0185	1.0000	0.9339	0.9310	0.4402	0.3550	0.9718
	$\mathcal{F}_3^{(6)}$	0.0497	0.0487	0.0485	0.0230	0.0181	1.0000	0.9791	0.9761	0.4616	0.3646	0.9680
500	$\mathcal{F}_1$	0.0213	0.0213	0.0213	0.0123	0.0097	1.0000	1.0000	1.0000	0.5778	0.4564	0.9937
	$\mathcal{F}_2$	0.0234	0.0491	0.0491	0.0123	0.0098	1.0000	2.1043	2.1043	0.5280	0.4202	0.9885
	$\mathcal{F}_3$	0.0853	0.0485	0.0486	0.0123	0.0121	1.0000	0.5686	0.5696	0.1445	0.1413	0.9752
	$\mathcal{F}_3^{(3)}$	0.0506	0.0485	0.0486	0.0123	0.0100	1.0000	0.9588	0.9606	0.2437	0.1981	0.9838
	$\mathcal{F}_3^{(6)}$	0.0465	0.0485	0.0486	0.0123	0.0107	1.0000	1.0440	1.0459	0.2653	0.2308	0.9809
1000	$\mathcal{F}_1$	0.0205	0.0205	0.0205	0.0105	0.0069	1.0000	1.0000	1.0000	0.5129	0.3346	0.9969
	$\mathcal{F}_2$	0.0240	0.0503	0.0503	0.0105	0.0070	1.0000	2.0927	2.0927	0.4376	0.2923	0.9929
	$\mathcal{F}_3$	0.0904	0.0507	0.0508	0.0105	0.0082	1.0000	0.5611	0.5618	0.1163	0.0906	0.9827
	$\mathcal{F}_3^{(3)}$	0.0521	0.0507	0.0508	0.0105	0.0077	1.0000	0.9736	0.9748	0.2018	0.1478	0.9913
	$\mathcal{F}_3^{(6)}$	0.0478	0.0507	0.0508	0.0105	0.0077	1.0000	1.0622	1.0636	0.2201	0.1618	0.9894

Table 16 Numerical results for SAN experiment, tail estimation, with $N = 100$.

		Relative Average A_{tail}				
Case	Scenario	JFCV	AIC	BIC	ED	JFCV(ED)
I	$\mathcal{F}_1$	1.0000	1.0055	1.0055	1.3160	1.0107
	$\mathcal{F}_2$	1.0000	2.3594	2.3594	1.2719	0.9784
	$\mathcal{F}_3$	1.0000	0.8978	0.8978	0.4734	0.4049
	$\mathcal{F}_3^{(3)}$	1.0000	0.7232	0.7232	0.3813	0.3253
	$\mathcal{F}_3^{(6)}$	1.0000	0.7113	0.7113	0.3751	0.3168
II	$\mathcal{F}_1$	1.0000	1.0055	1.0055	1.3160	1.0107
	$\mathcal{F}_2$	1.0000	2.3594	2.3594	1.2719	0.9784
	$\mathcal{F}_3$	1.0000	0.8978	0.8978	0.4734	0.4049
	$\mathcal{F}_3^{(3)}$	1.0000	0.7232	0.7232	0.3813	0.3253
	$\mathcal{F}_3^{(6)}$	1.0000	0.7113	0.7113	0.3751	0.3168
III	$\mathcal{F}_1$	1.0000	0.9838	0.9838	1.0550	0.8540
	$\mathcal{F}_2$	1.0000	1.0795	1.0795	1.0658	0.8150
	$\mathcal{F}_3$	1.0000	0.8768	0.8909	0.7875	0.6275
	$\mathcal{F}_3^{(3)}$	1.0000	0.7930	0.8058	0.7123	0.5508
	$\mathcal{F}_3^{(6)}$	1.0000	0.8536	0.8673	0.7667	0.6033
IV	$\mathcal{F}_1$	1.0000	0.9543	0.9543	1.3362	0.6634
	$\mathcal{F}_2$	1.0000	0.6131	0.6131	1.4322	0.6836
	$\mathcal{F}_3$	1.0000	0.7875	0.7875	1.9576	0.9596
	$\mathcal{F}_3^{(3)}$	1.0000	0.9548	0.9548	2.3735	1.1548
	$\mathcal{F}_3^{(6)}$	1.0000	0.9460	0.9460	2.3515	1.1396
V	$\mathcal{F}_1$	1.0000	0.9250	0.9250	0.8011	0.4883
	$\mathcal{F}_2$	1.0000	0.7296	0.7296	1.3324	0.8154
	$\mathcal{F}_3$	1.0000	0.6188	0.6188	1.1271	0.7072
	$\mathcal{F}_3^{(3)}$	1.0000	0.6621	0.6621	1.2059	0.6789
	$\mathcal{F}_3^{(6)}$	1.0000	0.7137	0.7137	1.2999	0.7856

References

Billingsley, P. 1968. *Convergence of Probability Measures*. New York: John Wiley and Sons.

Gao, Y., Zhang, X., Wang, S., Chong, T. T. L. and Zou, G. 2019. Frequentist model averaging for threshold models. *Annals of the Institute of Statistical Mathematics* **71**, 275–306.

Li, K. C. 1987. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics* **15**, 958–975.

White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.