Received: 7 April 2020 | Revised: 22 April 2020 | Accepted: 4 May 2020 | Published on: 9 June 2020

DOI: 10.1002/sam.11464

WILEY

RESEARCH ARTICLE

MR plot: A big data tool for distinguishing distributions

Omid M. Ardakani[1] | **Majid Asadi**[2,3] | **Nader Ebrahimi**[4] | **Ehsan S. Soofi**[5]

[1]Department of Economics, Georgia Southern University, Savannah, Georgia

[2]Department of Statistics, University of Isfahan, Isfahan, Iran

[3]School of Mathematics, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

[4]Department of Statistics, Northern Illinois University, DeKalb, Illinois

[5]Lubar School of Business, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin

Correspondence
Ehsan S. Soofi, Lubar School of Business, University of Wisconsin-Milwaukee, P.O. Box 742, Milwaukee, WI 53201.
Email: esoofi@uwm.edu

Funding information
Institute for Research in Fundamental Sciences, Grant/Award Number: 99620213

Abstract

Big data enables reliable estimation of continuous probability density, cumulative distribution, survival, hazard rate, and mean residual functions (MRFs). We illustrate that plot of the MRF provides the best resolution for distinguishing between distributions. At each point, the MRF gives the mean excess of the data beyond the threshold. Graph of the empirical MRF, called here the *MR plot*, provides an effective visualization tool. A variety of theoretical and data driven examples illustrate that MR plots of big data preserve the shape of the MRF and complex models require bigger data. The MRF is an optimal predictor of the excess of the random variable. With a suitable prior, the expected MRF gives the Bayes risk in the form of the entropy functional of the survival function, called here the *survival entropy*. We show that the survival entropy is dominated by the standard deviation (*SD*) and the equality between the two measures characterizes the exponential distribution. The empirical survival entropy provides a data concentration statistic which is strongly consistent, easy to compute, and less sensitive than the *SD* to heavy tailed data. An application uses the New York City Taxi database with millions of trip times to illustrate the MR plot as a powerful tool for distinguishing distributions.

KEYWORDS

Bayes risk, concentration measures, distributional plots, mean residual plot, survival entropy, taxi trip time

1 | INTRODUCTION

Which distributional plot most clearly distinguishes between probability distributions? We address this question in light of big data where huge number of observations enables estimating probability distributions of a variable for different groups and levels of covariates. A probability distribution has various representations. The most familiar representations of the distribution of a continuous random variable T are the cumulative distribution function (CDF) $F(t) = P(T \le t)$, the survival function (SF) $\overline{F}(t) = P(T > t)$, and the probability density function (PDF) $f(t) = dF(t)/dt$. The CDF and SF are bounded in the unit interval and the area under PDF is one. These properties impose graphical limitations for providing clear visualizations of distributions that are different but close to each other. Probability distributions can be also represented by the hazard (failure) rate (HR) function and mean residual function (MRF). These representations are nonnegative functions without the area/range restrictions of the PDF, CDF, and SF, hence being capable of providing more pronounced separations of the distributions. The HR function of a continuous probability distribution is defined by $r(t) = f(t)/\overline{F}(t)$, which is interpreted in lifetime modeling as the

instantaneous failure of an item at age t for items that are older than t. The MRF is more readily interpretable in terms of the average. The main point of this paper is that, with big data, the plot of MRF of the empirical distribution is potentially a useful visualization technique for distinguishing distributions.

The residual or excess of a nonnegative random variable T with distribution function $F(t) = 0, t < 0$, given that it exceeds a threshold τ is defined by $T - \tau | T > \tau$. The PDF of the residual random variable is given by

$$f(t|\tau) = \frac{f(t)}{\overline{F}(\tau)}, \quad t > \tau, \ \overline{F}(\tau) > 0. \tag{1}$$

The MRF or mean excess of a random variable with a finite mean μ is defined as

$$m(\tau) = E_{T>\tau}(T - \tau | T > \tau), \quad \tau \geq 0,$$

where $E_{T>\tau}$ denotes the expectation with respect to the residual PDF (1). The MRF determines the distribution by

$$\overline{F}(t) = \frac{m(0)}{m(t)} \exp\left\{ - \int_0^t \frac{1}{m(\tau)} d\tau \right\}, \tag{2}$$

where $m(\tau)$ is a continuous function with the following properties:

$$0 \leq m(\tau) < \infty, \quad m(0) > 0, \quad m'(\tau) + 1 > 0 \tag{3}$$

$$\begin{cases} \text{If } m(\tau_0) = 0 \text{ for some } \tau_0, \ \text{ then } m(\tau) = 0 \text{ for all } \tau \geq \tau_0. \\ \text{If } m(\tau) > 0 \text{ for all } \tau > 0, \text{then } \ \int_0^\infty \frac{1}{m(\tau)} d\tau = \infty; \end{cases} \tag{4}$$

see Shaked and Shanthikumar [16] for details. Application of (2) provides probability models more complex than the well known families of distributions; see Asadi et al [4]. Properties (3) and (4) must be kept in mind when developing new models based on data.

The MRF originally was used in the form of life tables for centuries. Since the middle of the last centuries it is being used in many fields [2]. Guess and Proschan [10] point out applications for setting rates and benefits for life insurance, survival analysis, job mobility, length of wars, duration of strikes, landholding, optimal inventory policies for perishable items, optimal disposal of an asset, renewal theory, dynamic programming, branching processes, preventive maintenance, and "burn-in." Poynor [13] includes a comprehensive literature review of the MRF. More recently, Ardakani et al [1] proposed ranking forecast models by the mean excess error defined by the MRF of the absolute error of the forecast model, where τ is a tolerance threshold for the error such that only the errors with magnitudes larger than τ are penalized.

Guess and Proschan [10] echoed the important fact that "estimation of MR is more stable than estimation of the failure rate. Statistical properties of estimated means are better than those of estimated derivatives (which enter into failure rates)" [9]. This argument equally applies to the estimation of the PDF, which is the most widely used tool for visualization of distribution. It has been suggested in the lifetime analysis literature that the graph of empirical MRF provides a useful tool for data analysis [8,9,12]. However, graph of the empirical MRF seldom is used, which could be due to its data requirement. We will illustrate that the MR plot is a powerful visualization tool that should be routinely used for distinguishing distributions especially with big and complex data.

In Section 2 we present the known estimate of the MRF and the associated confidence band. The MRF of the empirical distribution provides the MR plot. An example compares visualizations provided by various representations of several neighboring distributions. A variety of data driven examples illustrate the merits of MR plots for applications with big data.

Section 3 presents a data concentration measure, called here *survival entropy statistic*, which is strongly consistent and less sensitive than the standard deviation (*SD*) for heavy tailed data. At each point τ, $m(\tau)$ is a local mean, hence it is the optimal predictor of the excess of the random variable under the quadratic loss. The variance of the residual distribution is the corresponding local risk measure. The global risk of $m(\tau)$ is given by the expected MRF which is the Bayes risk under a prior for the threshold. When the prior is identical to the PDF of T, the Bayes risk is a *survival entropy* (*SE*). We present a sharp upper bound for the survival entropy in terms of the *SD* of T. Attaining this new bound characterizes the exponential model and improves on an existing bound for the characterization. The empirical version of the survival entropy provides the *SE* statistic.

Section 4 illustrates the MR plot as a powerful visualization tool that separates distributions for which the SF and PDF fail to do so. Our empirical application uses the New York City (NYC) Taxi and Limousine Commission (TLC) database. This database provides millions of trip times per month since 2009 allowing estimation and comparison of distributions for different months and hours.

Section 5 summarizes the findings of the paper and gives some concluding remarks.

2 | MR PLOT

Big data provides reliable estimates of distributional representations, such as the kernel estimate of the PDF, the Kaplan-Meier estimate of SF, the HR estimate developed

by Rebora et al [15], and the estimate of the MRF developed by Yang [17].

The estimate of MRF is based on the following representation:

$$m(\tau) = \frac{1}{\overline{F}(\tau)} \int_{\tau}^{\infty} \overline{F}(t)dt. \quad (5)$$

Let $T_1, \ldots, T_n$ be a random sample from a continuous CDF. The estimate of the MRF uses the empirical SF

$$\hat{\overline{F}}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(T_i > t),$$

where $\mathbf{1}(A)$ is the indicator function of the event A. The empirical MRF is found by using $\hat{\overline{F}}$ in (5):

$$\begin{aligned} \hat{m}(\tau) &= \frac{1}{\hat{\overline{F}}(\tau)} \int_{\tau}^{\infty} \hat{\overline{F}}(z)dz \cdot \mathbf{1}(\tau < \max_{1\le i\le n} t_i) \\ &= \frac{1}{n\hat{\overline{F}}(\tau)} \sum_{i=1}^{n} (t_{(i)} - \tau) \cdot \mathbf{1}(t_{(i)} > \tau) \\ &= \frac{1}{n_\tau} \sum_{i=1}^{n} (t_{(i)} - \tau) \cdot \mathbf{1}(t_{(i)} > \tau), \end{aligned}$$

where $0 = t_{(0)} < t(1) < \cdots < t_{(n)}$ are the ordered data points t_i's and n_τ is number of observations longer than τ. Asymptotic properties of $\hat{m}(\tau)$ are established by Yang [17] and Hall and Wellner [12]. The latter authors have developed an approximate confidence band for the $m(\tau)$ based on $\hat{m}(\tau)$, given by

$$P\left(\hat{m}(\tau) - \frac{aS_n}{\sqrt{n_\tau}} \le m(\tau) \le \hat{m}(\tau) + \frac{aS_n}{\sqrt{n_\tau}} \right) = Q(a), \quad (6)$$

where S_n is the sample SD and $Q(a)$ is given by

$$Q(a) = 1 - 4\sum_{k=1}^{\infty} (-1)^{k-1} \overline{\Phi}((2k-1)a),$$

where $\Phi(\cdot)$ is the standard normal CDF. Hall and Wellner [12] suggested further approximation and tabulated $Q(a) = 1 - 4\overline{\Phi}(a)$ for selected values of a; (eg, $Q(a) = 0.90$ and $a = 1.960$ provide the 90% band). R codes for computing $\hat{m}(\tau)$ with the band (6) are shown in the Appendix. (R Package "ismev" includes a command for computing an estimate of MRF and the band, which produced similar plots to ours). Zhao and Qin [18] have developed inference for the MRF based on the Chi-squared distribution and Poynor [13] has proposed Bayesian inference procedures for the MRF.

For examining the effect of sample size on the reliability and efficacy of the empirical MRF we use data generated from a few distributions and compare the MR plots with the plots of the MRFs of the data-generating distributions. Table 1 gives the SFs or PDFs and MRFs for the families of distributions used in this paper. The MRF formula with PDF is $m(\tau) = \int_{\tau}^{\infty} tf(t)dt/\overline{F}(\tau) - \tau$. We use the gamma family where the MRF has simple shapes (increasing and decreasing MRFs define classes of lifetime models for reliability analysis). We use the Lognormal family and the mixture of two Weibull models where the shapes of MRFs are more complex.

Figure 1 shows plots of the PDFs, SFs, HR functions, and MRFs of five neighboring distributions in the Gamma family with shape parameter $\alpha = 0.8, 0.9, 1, 1.1, 1.2$ and scale parameter $\lambda = 0.5$. The figure illustrates that the MRF distinguishes neighboring distributions better than the other three representations. The figure also displays some properties of the MRF of the gamma family which we wish to be preserved by the MR plots of the simulated data. The MRF of the gamma family is convex for $\alpha < 1$, constant for $\alpha = 1$ (the exponential case), and concave for $\alpha > 1$. The shape parameter α ranks the members of the gamma family increasingly by the MR order defined by $m_1(\tau) \le m_2(\tau)$ for all $\tau \ge 0$; (the MR order is unrelated with the well known stochastic order defined by $\overline{F}_1(t) \le \overline{F}_2(t)$ for all $t \ge 0$ and is implied by the HR order defined by $r_1(t) \ge r_2(t)$ for all $t \ge 0$; for details and more properties of the MRF see Shaked and Shanthikumar [16] and Gupta and Kirmani [11].)

Figure 2 shows MR plots of data simulated from the same Gamma distributions of Figure 1. The first row illustrates the sample size effect. The second row shows MR plots with 90% bands for the purpose of examining plausibility of the exponential model (constant MRF) for data from its neighboring gamma distributions. The following points are evident.

1. The first row plots reveal tradeoffs between the sample size and the threshold beyond which the MR plot can be accurate and distinguishes between distributions. For $n = 1000$ the monotonicity of the MR is distorted and the MR order between the neighboring distributions holds only for $\tau < 0.5$ and some distributions become indistinguishable beyond this threshold. For $n = 10\,000$ the monotonicity of the MR is distorted to a lesser extent and the MR order between the exponential and Gamma distributions with $\alpha = 1.1$ holds only for $\tau < 2.2$ and with $\alpha = 1.2$ holds only for $\tau < 3$. Beyond this threshold some distributions become indistinguishable. For $n = 50\,000$ the monotonicity of the MRF is preserved but the MR order between the Gamma distributions with $\alpha = 1.1$ and $\alpha = 1.2$ holds for $\tau < 4$; this issue persists for $n = 75\,000$. The panel for $n = 80\,000$ shows that MR plots match

TABLE 1 Examples of mean residual functions (MRFs)

Family	PDF or SF, $t \geq 0$	MRF[a], $\tau \geq 0$	Shape of MRF
Gamma	$f(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t}, \ \alpha, \lambda > 0$	$\frac{\overline{G}(\tau;\alpha+1,\lambda)}{\overline{G}(\tau;\alpha,\lambda)}\mu - \tau$	$\begin{cases} \text{Increasing} & \alpha < 1 \\ \text{Constant} & \alpha = 1 \\ \text{Decreasing} & \alpha > 1 \end{cases}$
Lognormal	$f(t) = \frac{1}{\sqrt{2\pi}\theta t} e^{-(\log t)^2/2\theta^2}, \ \theta > 0$	$\frac{1-\Phi(z_\tau-\theta)}{1-\Phi(z_\tau)}\mu - \tau, \ z_\tau = \frac{\log \tau}{\theta}$	Bathtub
Weibull	$\overline{F}(t) = e^{-t^\alpha}, \ \alpha > 0$	$\frac{\overline{G}(\tau^\alpha;1/\alpha,1)}{\overline{F}(\tau)}\mu$	$\begin{cases} \text{Increasing} & \alpha < 1 \\ \text{Constant} & \alpha = 1 \\ \text{Decreasing} & \alpha > 1 \end{cases}$
Weibull mixture	$\overline{F}(t) = w_1 e^{-t^\alpha} + w_2 e^{-t^2}$, $w_j \geq 0, w_1 + w_2 = 1, \alpha > 1$	$\frac{\overline{F}_1(\tau)}{\overline{F}(\tau)} w m_1(\tau) + \frac{\overline{F}_2(\tau)}{\overline{F}(\tau)}(1-w) m_2(\tau)$	$\begin{cases} \text{Bathtub}, & \tau \leq \tau_0 \\ \text{Upside down bathtub}, & \tau > \tau_0 \end{cases}$
Pareto	$\overline{F}(t) = \frac{1}{(1+t)^\alpha}, \ \alpha > 2$	$\frac{1}{\alpha-1}(\tau+1)$	Increasing

Abbreviations: PDF, probability density function; SF, survival function.

[a] $\mu = E(T), \ \overline{G}(\tau;\alpha,\lambda) = \int_\tau^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t} dt.$

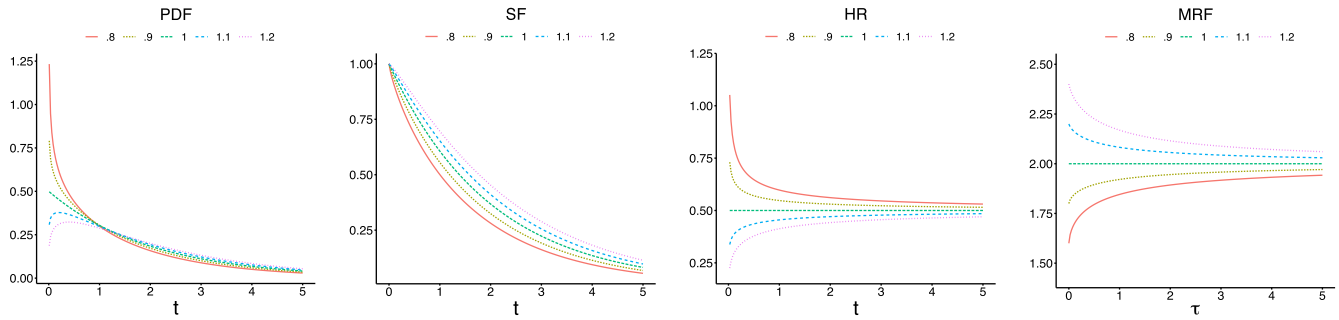

FIGURE 1 Plots of four representations of five neighboring Gamma distributions with $\alpha = 0.8, 0.9, 1.0, 1.1, 1.2$ and $\lambda = 0.5$

MRFs of the neighboring Gamma distributions in Figure 1.

2. The second row plots show that the constant MRF of the exponential model remains outside of the 90% bands for data from gamma distributions with $\alpha = 0.8$, 1.2, but touches the bands for data from its immediate neighbors ($\alpha = 0.9, 1.1$) near $\tau = 5$.

Figure 3 illustrates MRFs of more complex distributions. The left panel shows the MRF of the Lognormal distribution with $\theta = 0.7$, which is bathtub shaped (decreasing to a point and then increasing). The right panel shows the MRF of the mixture of two Weibull distributions with $\alpha = 1.1$ and $w_1 = 1/3$. The MRF of this distribution is bathtub shaped to a point and then becomes decreasing to form an upside-down bathtub shaped. Big data allows developing accurate and reliable MR plots for complex distributions.

Figure 4 shows MR plots with 90% bands for data from the distributions in Figure 3. The following points are noteworthy.

1. The first row shows MR plots for the Lognormal distribution. It reveals that the empirical MRF with $n = 50\,000$ reverses the increasing feature for $\tau > 4.5$ and with $n = 100\,000$ matches the corresponding panel in Figure 3.
2. The second row shows the MR plots for the mixture of two Weibull distributions. These plots reveal that a huge sample size is needed for accurate estimation of this complex model. Sample sizes of $n = 600\,000$, $1\,200\,000$ are not sufficient to capture the upside down bathtub shape seen in corresponding panel in Figure 3. For $n = 2\,400\,000$ upside down bathtub shape is captured for $\tau < 4.5$.
3. Comparison of samples sizes for accurate estimation of the MRFs of these models and the Gamma family reveals that more complex distributions require larger sample sizes.

3 | SURVIVAL ENTROPY STATISTIC

At each threshold, the MRF is the mean of the excess distribution (1). This endows the MR plot with an intuitive

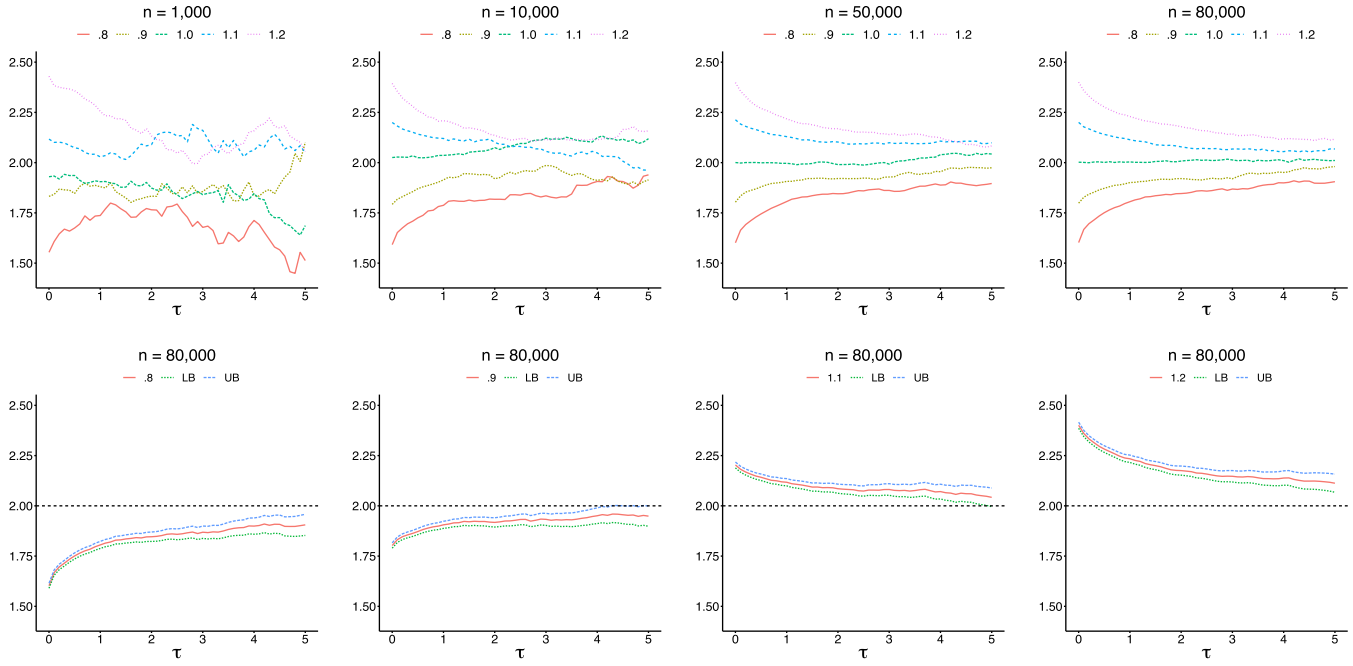

FIGURE 2 MR plots of data simulated from the five neighboring Gamma distributions of Figure 1 without band (first row) and with band (second row)

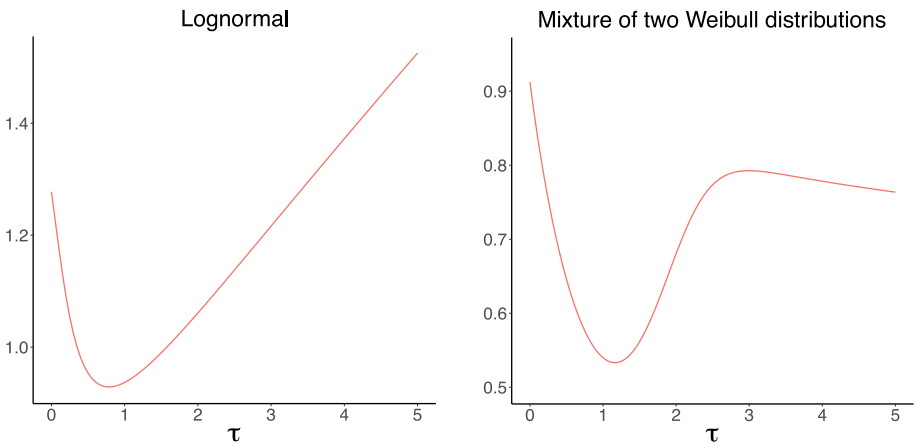

FIGURE 3 Examples of nonmonotone mean residual functions

and easily understood pointwise interpretation in the data analysis. Aggregation of these uncountably many local means over the continuous range $\tau \geq 0$ gives the global mean,

$$\mathcal{R}_w(m) = \int_0^\infty m(\tau)w(\tau)d\tau,$$

where $w(\cdot)$ is a suitably chosen weight function. Theoretically, at each threshold point, $m(\tau)$ is the optimal predictor of the excess under the quadratic loss and the variance of the excess distribution is its risk. But, $m(\tau)$ and its risk are local measures which depend on the threshold parameter τ. The global risk of $m(\tau)$ is its Bayes risk defined by the expected MRF under the prior $w(\tau)$.

Asadi and Zohrevand [6] showed that if $w(\tau)$ is identical to the PDF of the distribution of T, then

$$\mathcal{R}_f(m) = h(\overline{F}) = -\int_0^\infty \overline{F}(t)\log\overline{F}(t)dt \geq 0. \qquad (7)$$

Rao et al [14] introduced $h(\overline{F})$, referred to it as the cumulative residual entropy, and showed that $h(\overline{F}) \geq 0$ where the equality holds if and only if the distribution is degenerate. That is, $\mathcal{R}_f(m)$ is a measure of concentration of f. This measure has also been called the entropy functional of SF ([5]) due to its functional similarity with the Shannon entropy $H(f) = -\int_0^\infty f(t)\log f(t)dt$; for short, we call $\mathcal{R}_f(m)$ the *survival entropy* (*SE*), along the lines of

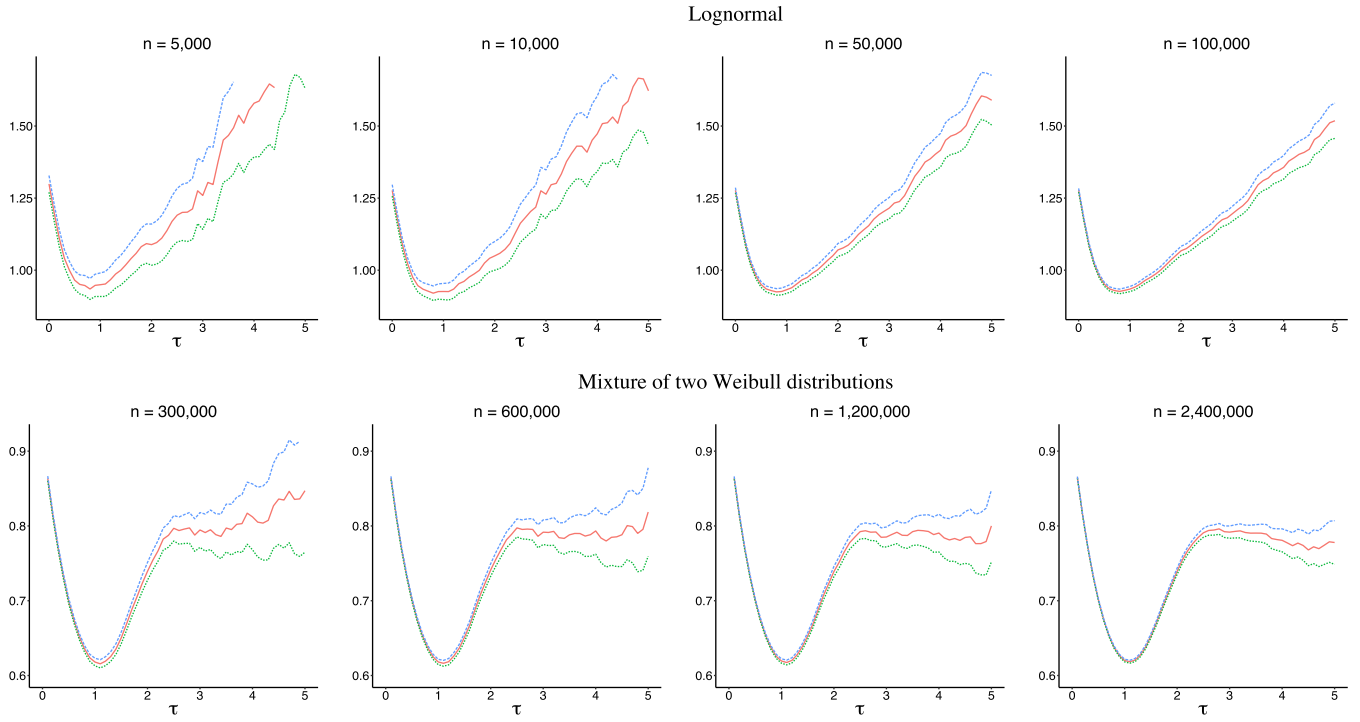

FIGURE 4 MR plots with 90% bands using data simulated from Lognormal and mixture of Weibull distributions of Figure 3

Zografos and Nadarajah [19]. The survival entropy can be represented in terms of the Shannon entropy of a related distribution as follows:

$$h(\overline{F}) = \mu H(f_e) - \log \mu,$$

where $f_e(t) = \overline{F}(t)/\mu$, $\mu = E(T) < \infty$ is known as the equilibrium distribution of $\overline{F}$ [5, 14, Proposition 4]. Rao et al [14] also showed that under the usual sampling assumption, $h(\hat{\overline{F}})$ converges to $h(\overline{F})$, almost surely.

Recently, Ardakani et al [1] used $\hat{\overline{F}}$ in (8) and provided the following statistic:

$$\widehat{\mathcal{R}}(m) = h(\hat{\overline{F}}) = -\sum_{i=1}^{n} a_i(t_{(i)} - t_{(i-1)}), \tag{8}$$

where $t_{(0)} = 0$ and

$$a_i = \frac{n+1-i}{n} \log \frac{n+1-i}{n}.$$

This *SE statistic* is a nonnegative information concentration measure which can be easily calculated for data from continuous distributions. R codes for computing this measure are given in the Appendix.

As a strongly consistent estimator, $\widehat{\mathcal{R}}(m)$ provides accurate estimates of $\mathcal{R}_f(m)$ with big data. Figure 5 illustrates the reliability and efficacy of the sample survival entropy (8) using data generated from two Weibull subfamilies and the Pareto family (parameterized as shown in Table 1). The Weibull family with shape parameter $\alpha \geq 1$ is a light (nonheavy) tailed distribution. The Weibull with shape parameter $\alpha < 1$ and Pareto families are heavy tailed distributions. Plots of the *SD* are also included for the purpose of comparison with the *SE* and sample *SE* as alternative measures of concentration. The left column shows plots of $\mathcal{R}_f(m)$ and σ. For the Weibull family $\mathcal{R}_f(m) = \mu/\alpha$; for the Pareto family the variance is defined for $\alpha > 2$ and $\mathcal{R}_f(m) = \alpha\mu/(\alpha - 1)$, $\alpha > 1$. The middle and right columns show plots of $\widehat{\mathcal{R}}_f(m)$ and the *SD* of the empirical CDF, $\hat{\sigma}$, which is equal to the maximum likelihood estimate of σ. The following points are evident.

1. For the light tailed Weibull subfamily, $\widehat{\mathcal{R}}(m)$ with 10 000 observations provides a rather accurate estimate and with 50 000 gives a very accurate estimate of $\mathcal{R}_f(m)$.
2. For the heavy tailed Weibull subfamily, $\widehat{\mathcal{R}}(m)$ based on 10 000 observations is rather volatile, but with 100 000 gives a very accurate estimate of $\mathcal{R}_f(m)$. For the Pareto family, $\widehat{\mathcal{R}}(m)$ based on 10 000 observations is very volatile, but with 200 000 gives a very accurate estimate of $\mathcal{R}_f(m)$.
3. In all cases $\mathcal{R}_f(m)$ and $\widehat{\mathcal{R}}(m)$ are dominated by σ and $\hat{\sigma}$, respectively. For the light tailed Weibull subfamily the difference between the two concentration measures is negligible. But for the heavy tailed distributions the difference between the two measures is substantial indicating that the *SE* is less sensitive against the tail thickness than the *SD*. The estimated *SD* for the Pareto

with $2 < \alpha < 3$ based on 300 000 data points is highly volatile, in contrast with the nearly smooth sample *SE*.

The following theorem gives the general relationship between the survival entropy and *SD* of the distribution.

Theorem 3.1. *Let T be a nonnegative continuous random variable with the SF $\overline{F}$ and variance $\sigma^2 < \infty$. Then*

$$\mathcal{R}_f(m) \le \sigma, \tag{9}$$

where the equality holds if and only if the distribution is exponential.

Proof Asadi [3] has shown the following representation: $h(\overline{F}) = \mathrm{Cov}(T, -\log \overline{F}(T))$.

Using this representation and the Cauchy-Schwartz inequality we obtain

$$\begin{aligned}\mathcal{R}_f^2(m) &= \mathrm{Cov}^2(-\log \overline{F}(T), T) \le \mathrm{Var}(-\log \overline{F}(T))\mathrm{Var}(T)\\ &= \mathrm{Var}(T) = \sigma^2,\end{aligned}$$

where the last equality follows from the fact that $\mathrm{Var}(-\log \overline{F}(T)) = 1$, because $-\log \overline{F}(T)$ has exponential distribution with mean one. The characterization of exponential distribution in the case of equality $\mathcal{R}_f(m) = \sigma$ is found as follows. The correlation between T and $-\log \overline{F}(T)$ is

$$\kappa = \mathrm{Cor}(T, -\log \overline{F}(T)) = \frac{\mathcal{R}_f(m)}{\sigma}.$$

From $0 \le \kappa \le 1$ we have $\kappa = 0$ if and only if the distribution of T is degenerate $P(T = a) = 1$ and $\kappa = 1$ if and only if there exist two constants a and b such that $P[-\log \overline{F}(T) = a + bT] = 1$. That is, $\overline{F}(t) = e^{-a-bt}$ for almost all t in the support of $\overline{F}$. From the initial condition $\overline{F}(0) = 1$ we have $a = 0$ and since $\overline{F}(t)$ is decreasing we have $b > 0$.

The variance and *SD* are commonly used as risk function. The variance is the risk of mean under the quadratic loss, however, we have not found a formal decision theoretic derivation of the *SD* as a measure or risk (expected loss). By Theorem 3.1, the *SD* is an upper bound for the Bayes risk of the MRF. The inequality (9) shows that among the distributions with given *SD*, the exponential distribution has maximum MR Bayes risk under the prior $w(\tau) = f(\tau)$.

From Theorem 3.1 and the consistency of $\widehat{\mathcal{R}}(m)$ we have the following corollary.

Corollary 3.1. *Let $\widehat{\sigma}_n$ be a consistent estimator of σ. Then*

$$\Pr[\widehat{\mathcal{R}}_n(m) \le \widehat{\sigma}_n] \to 1, \qquad \text{almost surely,}$$

and at the limit the two concentration statistics are identical if and only if the data generating distribution is exponential.

The simple sharp upper bound (9) improves on an existing bound. Rao et al [14, Theorems 8 and 10] give an entropy power lower bound and a moment ratio upper bound for $h(\overline{F})$, which can been represented in a more interpretable form [1] as follows. For a nonnegative continuous random variable T with the SF $\overline{F}$ and $E(T^2) < \infty$,

$$e^{H(f)-H(f_\gamma)} \le h(\overline{F}) \le \frac{E(T^2)}{2E(T)}, \tag{10}$$

where $H(f)$ is the Shannon entropy of f and $H(f_\gamma)$ is the entropy of exponential distribution with mean e^γ, and $\gamma = 0.5772\ldots$ is the Euler constant. The following corollary gives the improvement provided by Theorem 3.1 over the upper bound in (10).

Corollary 3.2. *Let T be a nonnegative continuous random variable with the SF $\overline{F}$ and variance $\sigma^2 < \infty$. Then*

$$\sigma \le \frac{E(T^2)}{2E(T)},$$

where the equality holds if and only if the coefficient of variation $c = \sigma/E(T) = 1$. The difference between the two sides gives the improvement as $.5E(T)(c-1)^2$.

Figure 6 illustrates the entropy power lower bounds and the *SD* upper bounds for survival entropies of the Weibull and Pareto families. For the Gamma and Lognormal families $\mathcal{R}_f(m)$ is not available in closed form. For these families the figure shows plots of $\widehat{\mathcal{R}}(m), \widehat{\sigma}$, and plug-in estimate $\widehat{H}(f)$. The following points are noteworthy.

1. In all four cases concentration measures are closer to the upper bounds than the lower bounds.
2. The Gamma family is light tailed for all $\alpha > 0$ and the plot of $\widehat{\mathcal{R}}(m)$ based on 50 000 data points is nearly smooth.
3. The Lognormal family is heavy tailed and 100 000 data points are needed to produce a smooth plot of $\widehat{\mathcal{R}}(m)$.

4 | MR PLOTS OF TAXI TRIP TIMES

The TLC database includes many millions of records of pick-up and drop-off date/times and several other variables for two taxi companies, Yellow and Green. We consider the data for trip time of the Yellow Company recorded for January 2009 to June 2019. Heterogeneity of trip time distributions for hours and days has been noted before, for example, by Ata et al [7]. Our analysis will include examining heterogeneity for months of the year and pick up hours

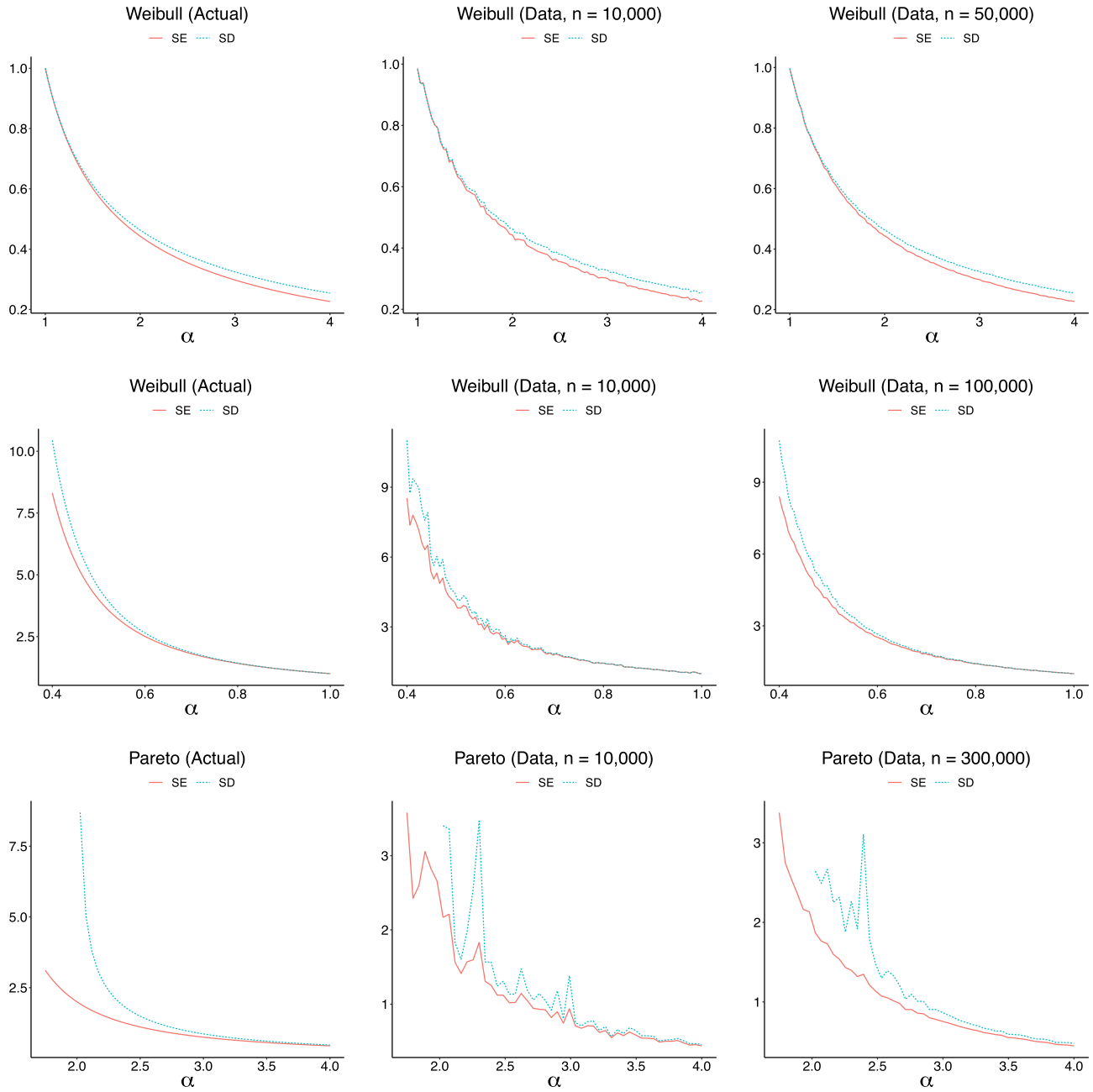

FIGURE 5 Plots of $\mathcal{R}_f(m)$, $\widehat{\mathcal{R}}(m)$, and the standard deviations of a light tailed family (Weibull with shape parameter $\alpha \geq 1$) and two heavy tailed families (Weibull with shape parameter $\alpha < 1$ and Pareto) distributions

of trip time distributions for full 24 hours and for months of a year.

Table 2 gives summary measures (traditional and *SE* statistic) for the Yellow Taxi data for June of each year since 2009 (the number of trips has been decreasing since 2013 partially due to increasing share of the Green Company since 2014). Data screening indicated two single extremely unusual cases for 2 months which were excluded before computing the summaries shown in the table. Still the difference between the *SD*s and *SE*s are substantial, indicating the tails of the data are very thick. As seen in the table 99.5th percentiles of data do not exceed 120 minutes. The last two columns show the *SD* and *SE* for trimmed data at 120. For the trimmed data, the *SD* provides very tight upper bound for *SE*.

4.1 | Comparison of months

Table 3 gives the summary measures for the months of 2018; (the latest year that the data is available for every month; data for March is excluded because the database

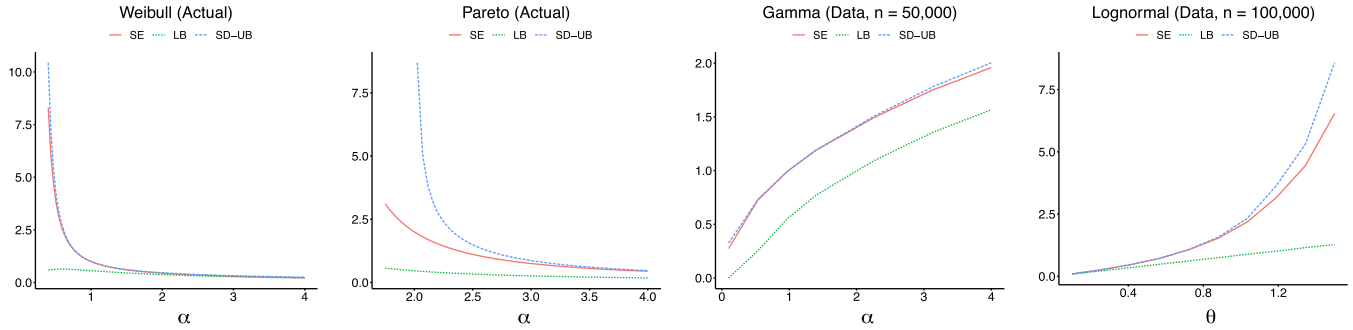

FIGURE 6 Plots of $\mathcal{R}_f(m)$ or $\widehat{\mathcal{R}}(m)$, its standard deviation upper bound, and the entropy power lower bound for four families of distributions

TABLE 2 Summary measures for the trip times of Yellow Taxi Company in New York City for the month of June

		Trip time (two unusual cases excluded)							≤120 minutes	
Year	**Number of trips**	**25%**	**Median**	**75%**	**Mean**	***SD***	***SE***	**99.5%**	***SD***	***SE***
2009	14 140 661	6.00	10.00	15.52	12.19	10.61	9.47	54.93	9.00	8.96
2010	14 790 816	6.00	10.00	15.52	12.24	10.75	9.53	55.00	9.04	9.00
2011	15 067 434	6.28	10.13	16.00	12.67	9.44	9.38	56.00	9.33	9.30
2012	15 061 955	6.35	10.38	16.37	12.88	10.03	9.80	58.00	9.63	9.59
2013	14 329 345	6.42	10.68	17.00	13.09	9.91	9.85	58.08	9.81	9.79
2014	13 767 977	6.85	11.00	17.77	13.73	10.54	10.47	61.00	10.38	10.35
2015	12 308 993	6.80	11.40	18.55	15.48	38.17	18.49	70.35	11.48	11.45
2016	11 123 909	6.87	11.57	19.02	16.88	54.56	24.94	74.97	11.91	11.88
2017	9 647 850	6.70	11.28	18.67	16.88	55.14	25.60	78.15	12.40	12.36
2018	8 706 550	6.68	11.30	18.70	17.64	65.10	29.85	78.40	12.09	12.06
2019	6 932 286	6.88	11.60	19.17	18.74	72.09	33.63	83.57	12.46	12.43

includes only two values 0 and 1440). The table indicates heterogeneity of the trip time distributions according to the months. The summary measures show that the distributions of trip time for different months are different. We examine distributional plots for distinguishing between these heterogeneous distributions.

Figure 7 shows the PDF, SF, HR, and MR plots of NYC Yellow Taxi trip time for months of 2018. The PDF plots show the distributions for all months have the same unimodal shape but cannot distinguish between most of the months. The SF plots show stochastic dominance between the distributions for some months, but cannot distinguish between most of the months. The HR plots improve over the PDF and SF plots for distinguishing between the months. The MR plots further improve over the HR plots for distinguishing between the months. The patterns of MR plots are bathtub shaped up to about 30 minutes and then turning to upside down bathtub shaped, somewhat similar to the pattern for the mixture of two Weibull distributions but more pronounced. With the huge numbers of observations at hand, the MR plots provide highly precise estimates of the MRFs the monthly trip times. This can be seen from the margins of error for the approximate confidence bands (6). We computed the margins of error for confidence bands for the 95% confidence band at the 25%-99th percentiles of the monthly trip time empirical distributions (using $a = 2.241$ for $Q(a) = 0.95$ tabulated in Hall and Wellner [12]. The maximum 95% of margin of error is ± .0195 for the 25th percentile of the trip time in November. With margins of error less than ± .0195, each MR plot in Figure 7 provides estimate of the respective MRF with extremely high precision.

4.2 | Comparison of pickup hours

Table 4 gives the summary measures for the taxi trip times by the pickup hour for June 2019. The hourly number of

TABLE 3 Summary measures for the trip times of Yellow Taxi company in New York City for months of 2018

		Trip time (Two unusual cases excluded)							≤120 minutes	
Year	**Number of trips**	**25%**	**Median**	**75%**	**Mean**	***SD***	***SE***	**99.5%**	***SD***	***SE***
January	8 752 251	6.25	10.38	16.85	15.67	60.27	25.99	66.52	10.24	10.22
February	8 484 877	6.32	10.47	16.95	15.76	60.57	26.08	64.52	10.12	10.11
March	Not available; the database includes only two values 0 and 1440									
April	9 297 982	6.58	11.08	18.25	17.06	63.30	28.50	74.53	11.53	11.51
May	9 216 416	6.78	11.47	19.03	17.71	62.98	29.13	81.18	12.42	12.38
June	8 706 550	6.68	11.30	18.70	17.64	65.10	29.85	78.40	12.09	12.06
July	7 843 324	6.55	11.05	18.18	17.44	68.54	30.81	71.87	11.21	11.19
August	7 842 557	6.57	11.05	18.15	17.39	68.15	30.59	71.32	11.18	11.16
September	8 033 713	6.80	11.62	19.40	18.27	68.13	31.38	78.65	12.15	12.13
October	8 810 579	6.97	11.87	19.70	18.58	69.01	31.86	80.00	12.21	12.20
November	8 137 451	6.88	11.78	19.65	19.12	74.32	34.81	87.45	12.38	12.35
December	8 166 400	6.63	11.48	19.23	18.76	75.50	34.83	80.00	11.84	11.82

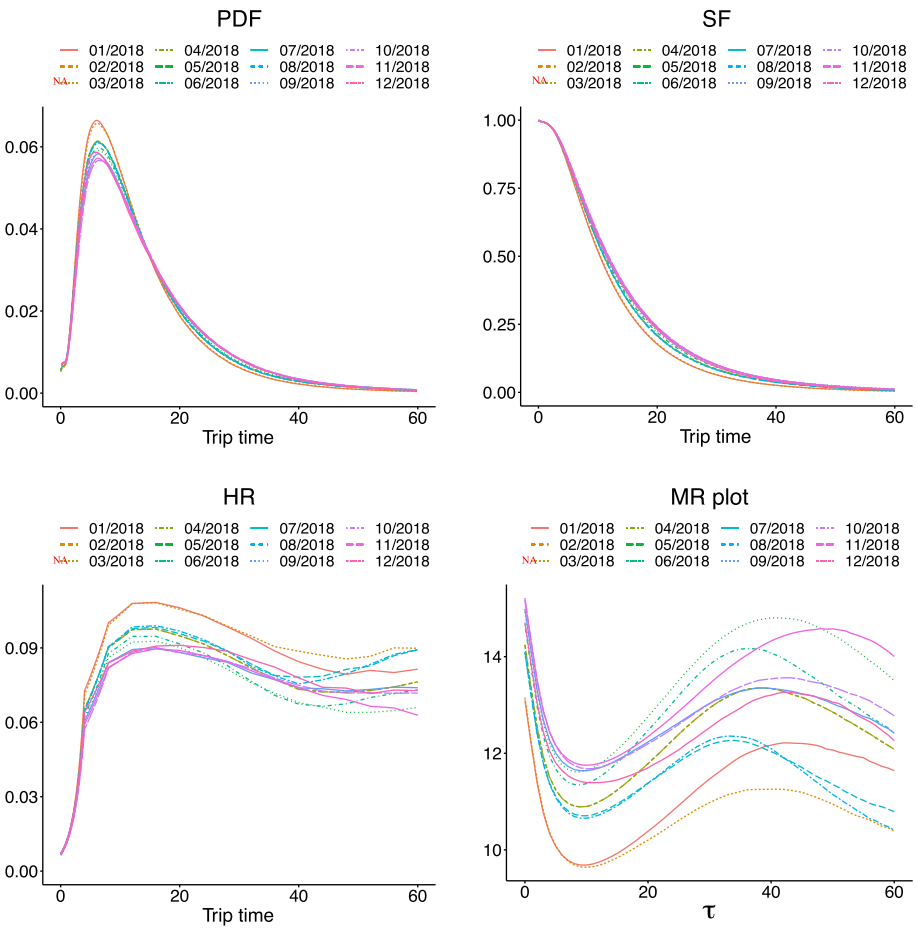

FIGURE 7 Probability density function (PDF), survival function (SF), hazard rate (HR), and MR plots of New York City (NYC) Yellow Taxi trip time for months of 2018 (NA, Not available)

trips varies from 58 435 in 4 AM to 436 149 in 6 PM, however, the trip time statistics are similar to the monthly data.

Figure 8 shows the distributional plots of the trip times by the pickup hours for June 2019. As before, the PDF plots show the distributions for all hours have the same unimodal shape but cannot distinguish between most of the pickup hours. The SF plots show stochastic dominance between the distributions for some hours but cannot

TABLE 4 Summary measures for the trip times of Yellow taxi company in New York by the pickup hour in June 2019

		Trip time (two unusual cases excluded)							≤120 minutes	
Year	Number of trips	25%	Median	75%	Mean	*SD*	*SE*	99.5%	*SD*	*SE*
0	227 611	6.58	10.92	17.77	17.56	78.03	33.65	53.60	9.12	9.04
1	159 695	6.15	10.12	16.20	16.10	75.09	31.23	47.87	8.23	8.16
2	112 930	5.90	9.67	15.20	16.15	81.44	34.28	50.18	7.70	7.64
3	76 754	5.72	9.58	15.38	16.21	81.76	34.68	50.24	7.89	7.83
4	58 435	5.55	10.08	17.28	16.52	77.44	33.30	50.19	8.62	8.49
5	67 088	4.87	8.57	17.00	16.86	80.86	36.62	58.08	9.97	9.90
6	147 389	5.02	8.13	14.08	18.38	94.78	45.45	103.48	11.54	11.41
7	244 724	5.98	9.70	15.60	18.38	87.08	40.93	94.90	11.82	11.62
8	313 431	6.82	11.35	18.32	18.23	72.79	33.38	79.27	11.71	11.67
9	321 213	7.03	11.93	19.48	18.27	68.69	31.14	69.63	11.29	11.28
10	326 653	7.17	12.12	19.77	18.53	69.56	31.51	69.93	11.33	11.31
11	344 333	7.38	12.48	20.37	19.03	69.93	31.95	70.62	11.67	11.65
12	363 974	7.43	12.62	20.82	19.26	68.72	31.79	71.32	12.11	12.09
13	370 407	7.38	12.60	20.88	19.65	68.80	32.80	80.18	13.20	13.17
14	383 846	7.47	12.82	21.85	21.19	72.43	36.56	94.16	15.35	15.28
15	376 372	7.38	12.67	21.78	21.83	75.28	38.85	102.05	16.46	16.34
16	354 675	7.23	12.63	22.13	22.03	74.41	38.95	104.33	17.14	17.01
17	399 472	7.32	12.68	21.68	20.97	70.59	35.76	93.13	15.72	15.62
18	434 149	7.07	11.95	19.55	18.84	69.74	32.56	77.52	12.58	12.54
19	414 303	6.80	11.17	17.78	17.08	66.95	29.33	64.62	10.57	10.55
20	376 041	6.78	11.13	17.87	16.82	64.46	27.99	60.58	10.37	10.34
21	388 904	6.97	11.38	18.27	17.20	66.69	28.90	60.57	10.24	10.22
22	369 552	7.10	11.57	18.50	17.32	67.30	28.93	59.73	10.08	10.06
23	300 335	6.88	11.37	18.47	17.50	71.58	30.77	50.18	9.78	9.73

distinguish between most of the pickup hours. The HR plots improve over the PDF and SF plots for distinguishing between the pickup hours. The MR plots further improve over the HR plots for distinguishing between the pickup hours. In most cases, the patterns of MR plots are bathtub shaped up to about 30 minutes and then turning to upside down bathtub shaped, somewhat similar to the pattern for the mixture of two Weibull distributions but more pronounced. The patterns of MR plots for late evening and early morning pickup times are bathtub shaped, similar to the pattern for the Lognormal distribution but less pronounced.

Figure 9 shows the MR plots of the trip times by the pickup hour in four groups according to the *SE* statistic. The late evening-early morning hours (11 PM to 5 AM) form the lowest, followed by the evening hours (7 PM to 10 PM), then by morning-noontime hours (6 AM to 12 PM), and then by the afternoon hours (1 PM to 6 PM). The shapes of the MR plots vary from the bathtub shaped for pickup hours (11 PM to 5 AM) to bathtub-upside down bathtub shaped the daytime pickup hours (6 AM to 6 PM).

5 | DISCUSSION AND CONCLUSIONS

Nowadays availability of huge number of observations enables estimating various representations of probability distributions accurately. Like the CDF and SF, a simple estimation procedure for the MRF is available via the empirical distribution. The graph of this empirical MRF provides the MR plot. The main point of this paper is

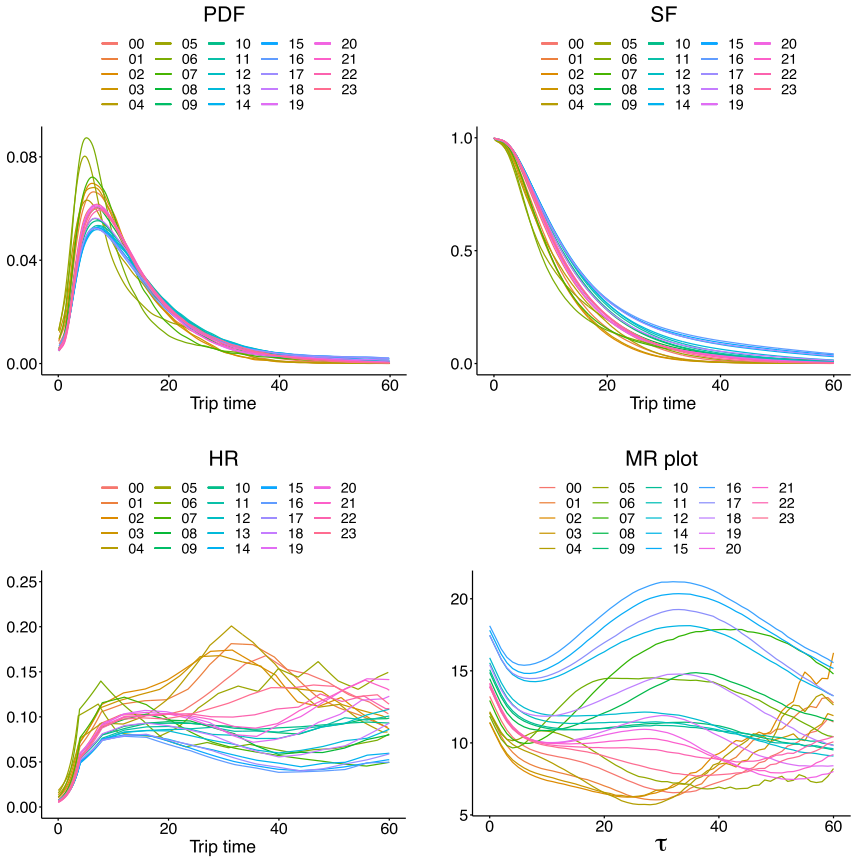

FIGURE 8 Probability density function (PDF), survival function (SF), hazard rate (HR), and MR plots of New York City (NYC) Yellow Taxi trip time for pickup hours in June 2019

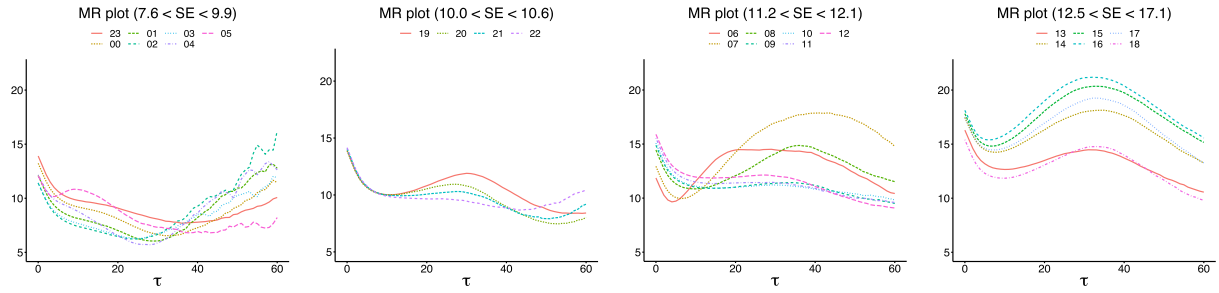

FIGURE 9 MR plots of New York City (NYC) Yellow Taxi trip time for pickup hours in June 2019

that the MR plot provides a visualization tool for big data. Because the MRF is not confined in the unit interval, the MR plot can distinguish between different distributions more clearly than the widely used plots of the PDF, SF, and HR function. We showed that very large number of observations are needed for preserving the properties of the MRF and for distinguishing neighboring distributions in the Gamma family. Producing correct images of the MRFs of complex distributions, like a mixture, may require millions of observations.

At each threshold point, the MRF gives the mean of the excess above the threshold, so under the quadratic loss it is the optimal predictor of the excess of a random variable. Using the weight function as a prior for the threshold parameter, the overall mean of the MRF gives its Bayes risk. When the weight function is identical to the distribution of the variable, the expected MRF is in the form of the survival entropy. At the theoretical level, while it is well known that the variance is the risk of the mean under the quadratic loss, we have not seen an existing formalization for the common use of the *SD* as a measure of risk. Our result provides a formalization for this common practice.

The *SE* statistic is a data concentration measure. We showed that this statistic is asymptotically dominated by the *SD*, and the equality of the two measures characterizes the exponential data. The *SD* bound is more efficient than an existing moment ratio bound. MR plots of simulated big data from light tailed distribution families (Weibull with shape parameter greater than and equal to one and Gamma) and heavy tailed distributions families (Weibull

with shape parameter less than one and Pareto) in terms their parameters provided insightful results. Plots of the *SE* statistic based on 50 000 simulated data points provided very accurate estimate for the light tailed Weibull subfamily. For the heavy tailed Weibull subfamily 100 000 data points and for the Pareto family 200 000 data points were required. These plots indicated that the *SE* statistic is less sensitive than the *SD* to the tail thickness of data.

The assumption of identical distribution is fundamental for statistics and other data analysis methods. Big data allows visualization of distributions of a variable for various categories. We examined distributions of the taxi trip time using millions of observations extracted from the TLC database. Distribution plots of the trip time for different months and pickup hours confirmed that the MR plot distinguishes heterogeneous distributions much more clearly than plots of the PDF, SF, and HR function. The *SE* statistics for different years, months, and pickup hours are substantially below the respective *SD*s indicating heavy tailed data. For data trimmed at 120 minutes (above 99.5 percentile) just notches below the *SD* of the respective data. Finally the *SE* statistic suggests sensible grouping of distributions of the taxi trip times according to the pickup time.

ACKNOWLEDGMENTS

The authors are grateful to Professors Jia Li and Bertrand Clarke for the invite to present this topic at the *Statistical Analysis and Data Mining* Invited Session of 2019 Symposium on Data Science & Statistics (SDSS) and the subsequent submission to the journal. Asadi's research was carried out at the IPM Isfahan branch and was in part supported by a grant from IPM (No. 99620213). Soofi's research was partially supported by the 2019 Business Advisory Council Faculty Scholar Award.

ORCID

Omid M. Ardakani https://orcid.org/0000-0003-0486-3359
Majid Asadi https://orcid.org/0000-0002-9259-0524
Nader Ebrahimi https://orcid.org/0000-0003-1687-2412
Ehsan S. Soofi https://orcid.org/0000-0003-1311-3509

REFERENCES

1. O. M. Ardakani, N. Ebrahimi, and E. S. Soofi, *Ranking forecasts by stochastic error distance, information and reliability measures*, Int. Stat. Rev. 86(3) (2018), 442–468.
2. K. J. Arrow, S. Karlin, and H. Scarf, *Studies in the mathematical theory of inventory and production*, Stanford University Press, Redwood City, 1958.
3. M. Asadi, *A new measure of association between random variables*, Metrika 80 (2017), 649–661.
4. M. Asadi et al., *Dynamic minimum discrimination information models*, J. Appl. Probab. 42 (2005), 643–660.
5. M. Asadi et al., *New maximum entropy methods for modeling lifetime distributions*, Naval Res. Logist. 61 (2014), 427–434.
6. M. Asadi and Y. Zohrevand, *On the dynamic cumulative residual entropy*, J. Statist. Plann. Inference 137 (2007), 1931–1941.
7. B. Ata, N. Barjestehy, and S. Kumar, *Spatial pricing: An empirical analysis of taxi rides in New York City*, Working Paper, 2019
8. M. C. Bryson and M. M. Siddiqui, *Soma criteria for aging*, J. Amer. Statist. Assoc. 64 (1969), 1472–1483.
9. I. P. Gertsbakh and K. B. Kordonskiy, *Models of failure*, Springer-Verlag, New York, 1969.
10. F. Guess and F. Proschan, *Mean residual life: Theory and applications*, Handbook of Statist. 7 (1988), 215–224.
11. R. Gupta and S. N. U. A. Kirmani, *On order relations between reliability measures*, Commun. Stat.: Stoch.Model. 3 (1987), 149–156.
12. W. J. Hall and J. A. Wellner, Estimation of mean residual life, Ph.D. Thesis, University of Rochester, Department of Statistics Technical Report, University of Rochester, Department of Statistics, 1979.
13. Poynor, V. A., Bayesian inference for mean residual life functions in survival analysis, Ph.D. Thesis, University of California Santa Cruz, 2010.
14. M. Rao et al., *Cumulative residual entropy: A new measure of information*, IEEE Trans. Inform. Theory 50 (2004), 1220–1228.
15. P. Rebora, A. Salim, and M. Reilly, *bshazard: A flexible tool for nonparametric smoothing of the hazard function*, R Journal 6 (2014), 114–122.
16. M. Shaked and J. G. Shanthikumar, *Stochastic orders*, Springer, New York, 2007.
17. G. L. Yang, *Estimation of a biometric function*, Ann. Statist. 6 (1978), 112–116.
18. Y. Zhao and G. Qin, *Inference for the mean residual life function via empirical likelihood*, Commun. Stat. Theory Methods 35 (2006), 1025–1036.
19. K. Zografos and S. Nadarajah, *Survival exponential entropies*, IEEE Trans. Inform. Theory 51 (2005), 1239–1246.

AUTHOR BIOGRAPHIES

Omid Ardakani is an Associate Professor of Economics in the Parker College of Business at Georgia Southern University. He is the Shirley and Philip Solomons, Sr. Research Fellow at the Department of Economics. He received his PhD in Economics from the University of Wisconsin-Milwaukee. His research interests lie at the interface of econometrics, computational statistics, and monetary and financial economics.

Majid Asadi is a Professor of Statistics at the Department of Statistics, University of Isfahan, Iran and a Senior Associate Researcher, School of Mathematics, IPM. He is an elected member of International Statistical Institute. He received BS and MS in Statistics from Ferdowsi University of Mashhad, Iran and PhD in Statistics from the University of

Sheffield, UK. His research interests include reliability theory, reliability modeling of systems and networks, information-theoretic probability modeling, and ordered random variables. He served as the Dean of the Faculty of Sciences, University of Isfahan (2012-2016) and Head of the Department of Statistics, University of Isfahan (2008-2009). He received Isfahan Province Governor Research Award in (2010) and was recognized as the Distinguished Lecturer at the University of Isfahan (2014) and as the Exemplary Professor of the Country in Iran (2015).

Nader Ebrahimi is Emeritus Professor of Statistics at the Department of Statistics at Northern Illinois University. He is Fellow of the American Statistical Association and elected member of International Statistical Institute. He received BS and MS from Shiraz University, Iran and PhD from Iowa State University. His research interests are in the areas of Reliability theory and Survival analysis, information theoretic statistics and modeling and applications in hardware and Software engineering, nano, and medical sciences.

Ehsan Soofi is a University of Wisconsin-Milwaukee Distinguished Professor at the Lubar School of Business. He is Fellow of the American Statistical Association and elected member of International Statistical Institute. He received BA in mathematics from UCLA, MA in statistics from the University of California, Berkeley, and PhD in Applied Statistics from the University of California, Riverside. His research interests are in information-theoretic and Bayesian approaches to distribution theory and statistics, and their applications in reliability, economics, and management sciences. Currently, he is an Associate Editor of Econometric Reviews. His professional activities include serving as the Guest Editor of Econometric Reviews, Special Issue on Bayesian Inference and Information: In Memory of Arnold Zellner (2014); Associate Editor of the Journal of the American Statistical Association (1990-2005); Chair of the Leonard J. Savage Thesis Award Committee (1992-2002); Chair of the International Society for Bayesian Analysis-Industrial Statistics Section; and the Vice President (1999-2001) of International Association for Statistical Computing.

How to cite this article: Ardakani OM, Asadi M, Ebrahimi N, Soofi ES. MR plot: A big data tool for distinguishing distributions. *Stat Anal Data Min: The ASA Data Sci Journal*. 2020;13:405–418. https://doi.org/10.1002/sam.11464

APPENDIX A

Empirical MRF with band and the survival entropy statistic are computed by the **mrf** and **survent** functions in R. The R codes are shown below.

The **mrf** function computes $\hat{m}(\tau)$ using two arguments (requires two inputs) as follows:
x: A vector of nonnegative data.
tau: A sequence of threshold.

```
mrf <- function(t, tau) {
  z <- t - tau
  value <- z[z>0]
  mrf <- mean(value, na.rm = TRUE)
  a = 1.960 ## Q(a) = .90
  ## a = 0.871, Q(a) = .25; a = 1.149,
  Q(a) = .50; a = 1.534, Q(a) = .75;
  ## a = 2.241, Q(a) = .95; a = 2.807,
  Q(a) = .99;
  upper = mrf + a * sd(t)/
  sqrt(length(value))
  lower = mrf - a * sd(t)/
  sqrt(length(value))
  print(c(lower, mrf, upper))
  }
```

The **survent** function computes $\widehat{\mathcal{R}}(m)$

```
survent <- function(t) {
  value <- diff(sort(t))
  no <- c(1:length(t))
  surv <- 1 - ((no-1)/length(t))
  surv <- surv[1:(length(t) - 1)]
  lsurv <- log(surv)
  survent <- -sum(value*surv*lsurv)
  return(survent)
  }
```