Stat Comput (2015) 25:809–819
DOI 10.1007/s11222-015-9563-8

# Simulation-efficient shortest probability intervals

**Ying Liu[1,2] · Andrew Gelman[1] · Tian Zheng[1]**

Accepted: 8 March 2015 / Published online: 11 June 2015
© Springer Science+Business Media New York 2015

**Abstract** Bayesian highest posterior density (HPD) intervals can be estimated directly from simulations via empirical shortest intervals. Unfortunately, these can be noisy (that is, have a high Monte Carlo error). We derive an optimal weighting strategy using bootstrap and quadratic programming to obtain a more computationally stable HPD, or in general, shortest probability interval (Spin). We prove the consistency of our method. Simulation studies on a range of theoretical and real-data examples, some with symmetric and some with asymmetric posterior densities, show that intervals constructed using Spin have better coverage (relative to the posterior distribution) and lower Monte Carlo error than empirical shortest intervals. We implement the new method in an R package (`SPIn`) so it can be routinely used in post-processing of Bayesian simulations.

**Keywords** Bayesian computation · Highest posterior density · Bootstrap

## 1 Introduction

It is standard practice to summarize Bayesian inferences via posterior intervals of specified coverage (for example, 50 % and 95 %) for parameters and other quantities of interest. In the modern era in which posterior distributions are computed via simulation, we most commonly see central intervals: the $100(1-\alpha)\%$ central interval is defined by the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles. Highest posterior density (HPD) intervals [recommended, for example, in the classic book of Box and Tiao (1973)] are easily determined for models with closed-form distributions such as the normal and gamma but are more difficult to compute from simulations.

We would like to go back to the HPD, solving whatever computational problems necessary to get it to work. Why? Because for an asymmetric distribution, the HPD interval can be a more reasonable summary than the central probability interval. Figure 1 shows these two types of intervals for three distributions: for symmetric densities (as shown in the left panel in Fig. 1), central and HPD intervals coincide; whereas for the two examples of asymmetric densities (the middle and right panels in Fig. 1), HPDs are shorter than central intervals (in fact, the HPD is the shortest interval containing a specified probability).

In particular, when the highest density occurs at the boundary (the right panel in Fig. 1), we strongly prefer the shortest probability interval to the central interval; the HPD covers the highest density part of the distribution and also the mode. In such cases, central intervals can be much longer and have the awkward property at cutting off a narrow high-posterior slice that happens to be near the boundary, thus ruling out a part of the distribution that is actually strongly supported by the inference.

One concern with highest posterior density intervals is that they depend on parameterization. For example, the left endpoint of the HPD in the right panel of Fig. 1 is 0, but the interval on the logarithmic scale does not start at $-\infty$. Interval estimation is always conditional on the purposes to which the estimate will be used. Beyond this, univariate summaries cannot completely capture multivariate relation-

✉ Andrew Gelman
gelman@stat.columbia.edu

Ying Liu
liuying4490@gmail.com

Tian Zheng
tz33@columbia.edu

1 Department of Statistics, Columbia University, New York, USA

2 Present Address: Google Inc., New York, USA

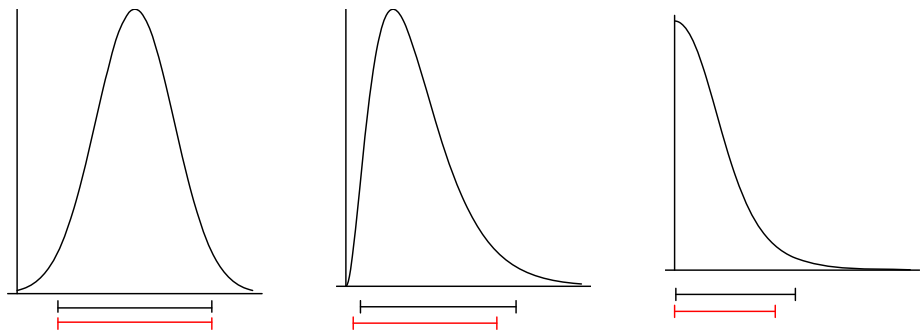

**Fig. 1** Simple examples of central (*black*) and highest probability density (*red*) intervals. The intervals coincide for a symmetric distribution, otherwise the HPD interval is shorter. The three examples are a normal distribution, a gamma with shape parameter 3, and the marginal posterior density for a variance parameter in a hierarchical model. (Color figure online)

ships. Thus, all this work is within the context of routine data analysis [e.g., Spiegelhalter et al. (2014), Gelman and Shirley (2011)] in which interval estimates are a useful way to summarize inferences about parameters and quantities of interest in a model in understandable parameterizations. We do not attempt a conclusive justification of HPD intervals here; we merely note that in the pre-simulation era such intervals were considered the standard, which suggests to us that the current preference for central intervals arises from computational reasons as much as anything else.

For the goal of computing an HPD interval from posterior simulations, the most direct approach is the *empirical shortest probability interval*, the shortest interval of specified probability coverage based on the simulations Chen and Shao (1999). For example, to obtain a 95 % interval from a posterior sample of size $n$, you can order the simulation draws and then take the shortest interval that contains $0.95n$ of the draws. This procedure is easy, fast, and simulation consistent (that is, as $n \to \infty$ it converges to the actual HPD interval assuming that the HPD interval exists and is unique). The only trouble with the empirical shortest probability interval is that it can be too noisy, with a high Monte Carlo error (compared to the central probability interval) when computed from the equivalent of a small number of simulation draws. This is a concern with current Bayesian methods that rely on Markov chain Monte Carlo (MCMC) techniques, where for some problems the effective sample size of the posterior draws can be low (for example, hundreds of thousands of steps might be needed to obtain an effective sample size of 500).

Figure 2 shows the lengths of the empirical shortest 95 % intervals based on several simulations for the three distributions shown in Fig. 1, starting from the $k$th order statistic. For each distribution and each specified number of independent simulation draws, we carried out 200 replications to get a sense of the typical size of the Monte Carlo error. The lengths of the 95 % intervals are highly variable when the number of simulation draws is small.

In this paper, we develop a quadratic programming strategy coupled with bootstrapping to estimate the endpoints of the shortest probability interval. Simulation studies show that our procedure, which we call Spin, results in more stable endpoint estimates compared to the empirical shortest interval (Fig. 3). Specifically, define the efficiency as

$$\text{Efficiency} = \frac{\text{MSE (empirical shortest interval)}}{\text{MSE (Spin)}},$$

so that an efficiency greater than 1 means that Spin is more efficient. We show in Fig. 3 that, in all cases that we experimented on, Spin is more efficient than the competition. We derive our method in Sect. 2, apply it to some theoretical examples in Sect. 3 and in two real-data Bayesian analysis problems and 60 BUGS examples in Sect. 4 and 5. We have implemented our algorithm as `SPIn`, a publicly available package in R Development Core Team (2011).

## 2 Methods

### 2.1 Problem setup

Let $X_1, \ldots, X_n \overset{iid}{\sim} F$, where $F$ is a continuous unimodal distribution. The goal is to estimate the $100(1-\alpha)\%$ shortest probability interval for $F$. Denote the true shortest probability interval by $(l(\alpha), u(\alpha))$. Define $G = 1 - F$, so that $F(l(\alpha)) + G(u(\alpha)) = \alpha$.

To estimate the interval, for $0 \le \Delta \le \alpha$, find $\Delta$ such that $G^{-1}(\alpha - \Delta) - F^{-1}(\Delta)$ is a minimum, i.e.,

$$\Delta^* = \text{argmin}_{\Delta \in [0,\alpha]} \{G^{-1}(\alpha - \Delta) - F^{-1}(\Delta)\}.$$

Taking the derivative,

$$\frac{\partial}{\partial \Delta}[(1-F)^{-1}(\alpha - \Delta) - F^{-1}(\Delta)] = 0,$$

we get

$$\frac{1}{f(G^{-1}(\alpha - \Delta))} - \frac{1}{f(F^{-1}(\Delta))} = 0, \tag{1}$$

where $f$ is the probability density function of $X$. The minimum can only be attained at solutions to (1), or $\Delta = 0$ or

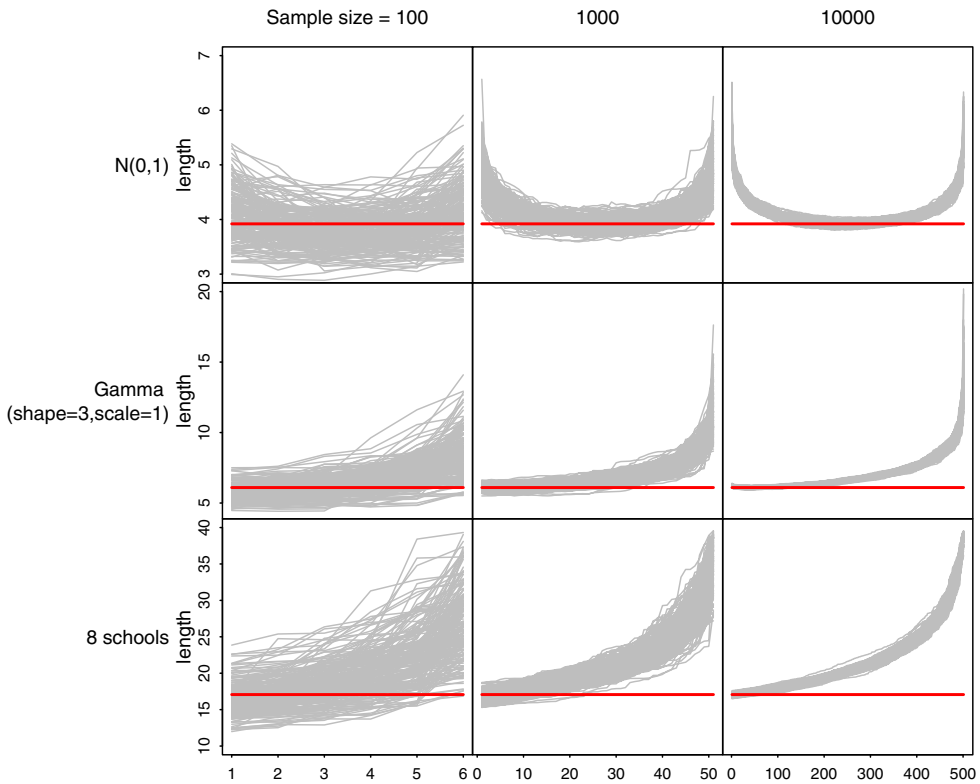


**Fig. 2** Lengths of 95 % empirical probability intervals from several simulations for each of three models. Each *gray curve* shows interval length as a function of the order statistic of the interval's lower endpoint; thus, the minimum value of the curve corresponds to the empirical shortest 95 % interval. For the (symmetric) normal example, the empirical shortest interval is typically close to the central interval (for example, with a sample of size 1000, it is typically near $(x_{(26)}, x_{(975)})$). The gamma and eight schools examples are skewed with a peak near the *left* of the distribution, hence the empirical shortest intervals are typically at the *leftend* of the scale. The *redlines* show the lengths of the true shortest 95 % probability interval for each distribution. The empirical shortest interval approaches the true value as the number of simulation draws increases but is noisy when the number of simulation draws is small, hence motivating a more elaborate estimator. (Color figure online)

$\alpha$ (Fig. 4). It can easily be shown that if $f'(x) \neq 0$ a.e., the solution to (1) exists and is unique. Then

$$l(\alpha) = F^{-1}(\Delta^*),$$
$$u(\alpha) = G^{-1}(\alpha - \Delta^*).$$

Taking the lower end for example, we are interested in a weighting strategy such that $\hat{l} = \sum_{i=1}^n w_i X_{(i)}$ (where $\sum w_i = 1$) has the minimum mean squared error (MSE), $\mathrm{E}\left(\left|\left|\sum_{i=1}^n w_i X_{(i)} - l(\alpha)\right|\right|^2\right)$. It can also be helpful to calculate $\mathrm{MSE}(X_{([n\Delta^*])}) = \mathrm{E}\left(||X_{([n\Delta^*])} - l(\alpha)||^2\right)$. In practice, we estimate $\Delta^*$ by $\hat{\Delta}$ such that

$$\hat{\Delta} = \mathrm{argmin}_{\Delta\in[0,\alpha]}\{\hat{G}^{-1}(\alpha - \Delta) - \hat{F}^{-1}(\Delta)\}, \tag{2}$$

where $\hat{F}$ represents the empirical distribution and $\hat{G} = 1 - \hat{F}$. This yields the widely used empirical shortest interval, which can have a high Monte Carlo error (as illustrated in Fig. 2). We will denote its endpoints by $l^*$ and $u^*$. The corresponding MSE for the lower endpoint is $\mathrm{E}(||X_{([n\hat{\Delta}])} - l(\alpha)||^2)$.

## 2.2 Quadratic programming

Let $\hat{l} = \sum_{i=1}^n w_i X_{(i)}$. Then

$$\begin{aligned}\mathrm{MSE}(\hat{l}) &= \mathrm{E}(\hat{l} - F^{-1}(\Delta^*))^2\\ &= \mathrm{E}\,(\hat{l} - \mathrm{E}\,\hat{l} + \mathrm{E}\,\hat{l} - F^{-1}(\Delta^*))^2\\ &= \mathrm{E}\,(\hat{l} - \mathrm{E}\,\hat{l})^2 + (\mathrm{E}\,\hat{l} - F^{-1}(\Delta^*))^2\\ &= \mathrm{Var} + \mathrm{Bias}^2,\end{aligned}$$

where $\mathrm{E}(\hat{l}) = \sum_{i=1}^n w_i \mathrm{E}X_{(i)}$ and $\mathrm{Var} = \sum_{i=1}^n w_i^2 \mathrm{Var}X_{(i)} + 2\sum_{i<j} w_i w_j \mathrm{cov}(X_{(i)}, X_{(j)})$. It has been shown [e.g., David and Nagaraja (2003)] that

$$\mathrm{E}(X_{(i)}) = Q_i + \frac{p_i q_i}{2(n+2)} Q_i'' + o(n^{-1}),$$

where $q_i = 1 - p_i$, $Q = F^{-1}$ is the quantile function, $Q_i = Q(p_i) = Q(\mathrm{E}U_{(i)}) = Q(\frac{i}{n+1})$, and $Q_i'' = \frac{Q_i}{f^2(Q_i)}$. Thus,

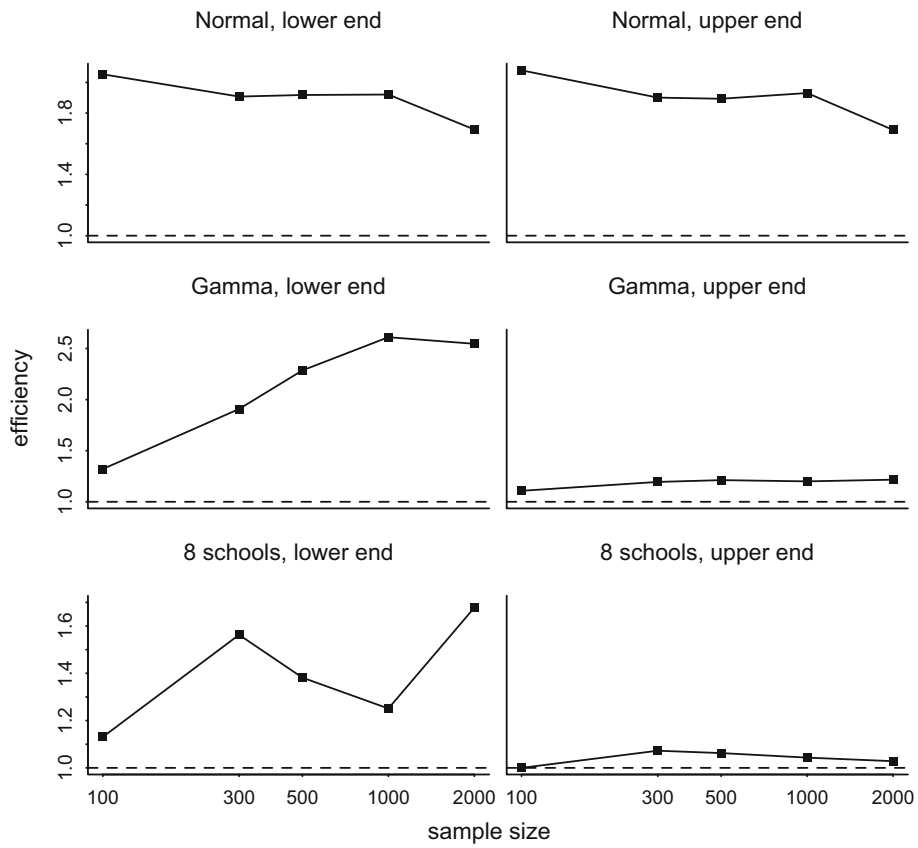


**Fig. 3** Efficiency of Spin for 95 % shortest intervals for the three distributions shown in Fig. 1. For the eight schools example, Spin is compared to a modified empirical HPD that includes the zero point in the simulations. The efficiency is always greater than 1, indicating that Spin always outperforms the empirical HPD. The jagged appearance of some of the lines may arise from discreteness in the order statistics for the 95 % interval

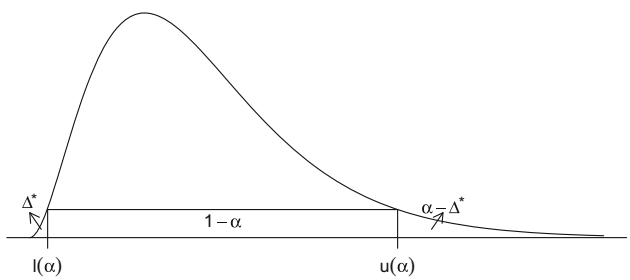


**Fig. 4** Notation for shortest probability intervals

$$\mathrm{E}(\hat{l}) \doteq \sum_{i=1}^{n} w_i \left( Q_i + \frac{p_i q_i}{2(n+2)} Q_i'' \right). \tag{3}$$

It has also been shown [e.g., David and Nagaraja (2003)] that

$$\mathrm{Var}\, X_{(i)} = \frac{p_i q_i}{n+2} Q_i'^2 + o(n^{-1})$$
$$\mathrm{cov}(X_{(i)}, X_{(j)}) = \frac{p_i q_j}{n+2} Q_i' Q_j' + o(n^{-1}), \text{ for } i < j,$$

where $Q_i' = \frac{1}{dp_i/dQ_i} = \frac{1}{f(Q_i)}$ ($f(Q_i)$ is called the density quantile function). Thus,

$$\mathrm{Var}(\hat{l}) = \sum_{i=1}^{n} w_i^2 \frac{p_i q_i}{n+2} Q_i'^2 + 2 \sum_{i<j} w_i w_j \frac{p_i q_j}{n+2} Q_i' Q_j' + o(n^{-1}). \tag{4}$$

Putting (3) and (4) together yields,

$$\mathrm{MSE}(\hat{l}) = \sum_{i=1}^{n} w_i^2 \frac{p_i q_i}{n+2} Q_i'^2 + 2 \sum_{i<j} w_i w_j \frac{p_i q_j}{n+2} Q_i' Q_j' + \left[ \sum_{i=1}^{n} w_i \left( Q_i + \frac{p_i q_i}{2(n+2)} Q_i'' \right) - Q(\Delta^*) \right]^2 + o(n^{-1}). \tag{5}$$

Finding the optimal weights that minimize MSE as defined in (5) is then approximately a quadratic programming problem.

In this study, we impose triangle kernels centered around the endpoints of the empirical shortest interval on the weights for computational stability. Specifically, the estimate of the lower endpoint has the form,

$$\hat{l} = \sum\nolimits_{i=i^*-b/2}^{i^*+b/2} w_i X_{(i)},$$

where $i^*$ is the index of the endpoint of the empirical shortest interval, $b$ is the bandwidth in terms of data points, and $w_i$ decreases linearly when $i$ moves away from $i^*$. We choose $b$ to be of order $\sqrt{n}$ in this study. This optimization problem is equivalent to minimizing MSE with the following constraints:

$$\begin{aligned}
&\sum_{i=i^*-b/2}^{i^*+b/2} w_i = 1 \\
&\frac{w_i - w_{i-1}}{X_{(i)} - X_{(i-1)}} = \frac{w_{i-1} - w_{i-2}}{X_{(i-1)} - X_{(i-2)}} \\
&\quad \text{for } i = i^* - b/2+2, \ldots, i^*, i^*+2, \ldots, i^*+b/2 \\
&\frac{w_{i^*} - w_{i^*-1}}{X_{(i^*)} - X_{(i^*-1)}} = \frac{w_{i^*} - w_{i^*+1}}{X_{(i^*+1)} - X_{(i^*)}} \\
&w_{i^*-b/2} \geq 0 \\
&w_{i^*+b/2} \geq 0 \\
&w_{i^*} - w_{i^*+1} \geq 0. \qquad (6)
\end{aligned}$$

The above constraints reflect the piecewise linear and symmetric pattern of the kernel. In practice, $Q$, $f$, and $\Delta^*$ can be substituted by the corresponding sample estimates $\hat{Q}$, $\hat{f}$, and $\hat{\Delta}$.

The above quadratic programming problem can be rewritten in the conventional matrix form,

$$\mathrm{MSE}(\hat{l}) \doteq \frac{1}{2} w^T \mathbf{D} w - d^T w,$$

where

$$w = (w_{i^*-b/2}, \ldots, w_{i^*+b/2})^T,$$

and $\mathbf{D} = (d_{ij})$ is a symmetric matrix with

$$d_{ij} = \begin{cases} 2(Q_i^2 + \frac{p_i q_i}{n+2} Q_i'^2), \ i = j \\ 2(\frac{Q_i' Q_j'}{n+2} p_i q_j + Q_i Q_j), \ i < j, \end{cases}$$

$$d^T = 2Q(\Delta^*) Q_i,$$

subject to

$$\mathbf{A}^T w \geq w_0,$$

with appropriate $\mathbf{A}$ and $w_0$ derived from the linear constraints in (6).

### 2.3 Proof of simulation consistency of the estimated HPD

The following result ensures the simulation consistency of our endpoint estimators when we use the empirical distribution and kernel density estimate.

Under regularity conditions, with probability 1,

$$\lim_{n\to\infty} \min_w \left( \frac{1}{2} w^T \hat{\mathbf{D}}_n w - \hat{d}_n^T w \right) = \min_w \left( \frac{1}{2} w^T \mathbf{D} w - d^T w \right),$$

where $\hat{\mathbf{D}}_n$ and $\hat{d}_n$ are empirical estimates of $\mathbf{D}$ and $d$ based on empirical distribution function and kernel density estimates.

To see this, we first show that $\hat{\mathbf{D}}_n \to \mathbf{D}$ and $\hat{d}_n \to d$ uniformly as $n \to \infty$ almost surely. By the Glivenko–Cantelli theorem, $||\hat{F} - F||_\infty \overset{a.s.}{\to} 0$, which implies $\hat{Q} \rightsquigarrow Q$ almost surely, where $\rightsquigarrow$ denotes weak convergence, i.e., $\hat{Q}(t) \to Q(t)$ at every $t$ where $Q$ is continuous [e.g., Vaart (1998)]. It has also been shown that $\int \mathrm{E}_f(\hat{f}(x) - f(x))^2 dx = O(n^{-4/5})$ under regularity conditions [see, e.g., Vaart (1998)], which implies that $\hat{f}(x) \to f(x)$ almost surely for all $x$. The endpoints of the empirical shortest interval are simulation consistent Chen and Shao (1999).

The elements in matrix $\hat{\mathbf{D}}_n$ result from simple arithmetic manipulations of $\hat{Q}$ and $\hat{f}$, so $\hat{d}_{ij} \to d_{ij}$ with probability 1, which implies,

$$\hat{\mathbf{D}}_n \to \mathbf{D} \text{ uniformly and almost surely,}$$

given $\mathbf{D}$ is of finite dimension. We can prove the almost sure uniform convergence of $\hat{d}_n$ to $d$ in a similar manner.

The optimization problem $\min_w(\frac{1}{2} w^T \hat{\mathbf{D}}_n w - \hat{d}_n^T w)$ corresponds to calculating the smallest eigenvalue of an augmented matrix constructed from $\hat{\mathbf{D}}_n$ and $\hat{d}_n$. The above uniform convergence then implies,

$$\lim_{n\to\infty} \min_w (w^T \hat{\mathbf{D}}_n w - \hat{d}_n^T w) = \min_w (w^T \mathbf{D} w - d^T w).$$

The same proof works for the upper endpoint.

### 2.4 Bootstrapping the procedure to get a smoother estimate

Results from quadratic programming as described above show that, as expected, Spin has a much reduced bias than the empirical shortest intervals. This is because the above procedure takes the shape of the empirical distribution into account. However, the variance remains at the same magnitude as that of the empirical shortest interval (as we shall see in the left panel in Fig. 10), because the optimal weights derived from the empirical distribution are also subject to the same level of variability as the empirical shortest intervals. We can use the bootstrap Efron (1979) to smooth away some of this noise and thus further reduce the variance in the interval. Specifically, we bootstrap the original posterior draws $B$ times (in this study we set $B = 50$) and calculate the Spin optimal weights for each of the bootstrapped samples. Here we treat the weights as general functions of the posterior distribution under study rather than the endpoints of HPD interval of the posterior samples. We then compute the final weights as the average of the $B$ sets of weights obtained from the above procedure (Fig. 5).

$$
X_1, X_2, \cdots, X_n \xrightarrow{\text{bootstrap}} \left\{ \begin{array}{lcl}
X_1^{(1)}, X_2^{(1)}, \cdots, X_n^{(1)} & \xrightarrow{Spin} & w_1^{(1)}, w_2^{(1)}, \cdots, w_n^{(1)} \\
X_1^{(2)}, X_2^{(2)}, \cdots, X_n^{(2)} & \xrightarrow{Spin} & w_1^{(2)}, w_2^{(2)}, \cdots, w_n^{(2)} \\
\cdots & & \\
X_1^{(B)}, X_2^{(B)}, \cdots, X_n^{(B)} & \xrightarrow{Spin} & w_1^{(B)}, w_2^{(B)}, \cdots, w_n^{(B)}
\end{array} \right.
$$

$$
\downarrow \text{average}
$$

$$
w_1, w_2, \cdots, w_n
$$

**Fig. 5** Bootstrapping procedure to get more stable weights

### 2.5 Bounded distributions

As defined so far, our procedure necessarily yields an interval within the range of the simulations. This is undesirable if the distribution is bounded with the boundary included in the HPD interval (as in the right graph in Fig. 1). To allow boundary estimates, we augment our simulations with a pseudo-datapoint (or two, if the distribution is known to be bounded on both sides). For example, if a distribution is defined on $(0, \infty)$, then we insert another datapoint at 0; if the probability space is (0, 1), we insert additional points at 0 and 1.

### 2.6 Discrete and multimodal distributions

If a distribution is continuous and unimodal, the highest posterior density region and shortest probability interval coincide, at least for parameters or quantities of interest with unimodal or approximately unimodal posteriors, so that a single interval is itself a reasonable inferential summary. More generally, the highest posterior density region can be formed from disjoint intervals. For distributions with known boundary of disjoint parts, Spin can be applied to different regions separately and a HPD region can be assembled using the derived disjoint intervals. When the nature of the underlying true distribution is unknown and the sample size is small, the inference of unimodality can be difficult. Therefore, in this paper, we have focused on estimating the shortest probability interval, recognizing that, as with interval estimates in general, our procedure is less relevant for multimodal distributions.

## 3 Results for simple theoretical examples

We conduct simulation studies to evaluate the performance of our methods. We generate independent samples from the normal, $t(5)$, and gamma(3) distributions and construct 95 % intervals using these samples. We consider sample sizes of 100, 300, 500, 1000, and 2000. For each setup, we generate 20,000 independent replicates and use these to compute root mean squared errors (RMSEs) for upper and lower endpoints. We also construct empirical shortest intervals as defined in (2), parametric intervals and central intervals for comparison. For parametric intervals, we calculate the sample mean and standard deviation. For the normal distribution, the interval takes the form of mean $\pm$ 1.96 sd (for the $t$ distribution, we also implement the same form as "Gaussian approximation" for comparison); for the gamma, we use the mean and standard deviation to estimate its parameters first, and then numerically obtain the HPD interval using the resulted density with the two estimates plugged in. The empirical 95 % central interval is defined as the 2.5th and 97.5th % percentiles of the sampled data points. We also use our methods to construct optimal central intervals (see Sect. 6) for the two symmetric distributions.

Figure 6 shows the intervals constructed for the standard normal distribution and the $t(5)$ distribution based on 500 simulation draws. The empirical shortest intervals tend to be too short in both cases, while Spins have better endpoint estimates. Empirical central intervals are more stable than empirical shortest intervals, and Spins have comparable RMSE for N(0, 1) and smaller RMSE for $t(5)$. Our methods can further improve RMSE based on the empirical central intervals as shown in the "central (QP)" row in Fig. 6. The RMSE is the smallest if one specifies the correct parametric distribution and uses that information to construct interval estimates, while in practice the underlying distribution is usually not totally known, and misspecifying it can result in far-off intervals (the right bottom panel in Fig. 6).

Figure 7 shows the empirical shortest, Spin, and parametric intervals constructed from 500 samples of the gamma distribution with shape parameter 3. Spin gets more accurate endpoint estimates than empirical shortest intervals. Specifically, for the lower end where the density is relatively high, Spin estimates are less variable, and for the upper end at the tail of the density, Spin shows a smaller bias. Again, the lowest RMSE comes from taking advantage of the parametric form of the posterior distribution, which is rarely practical in real MCMC applications. Hence, the RMSE using the parametric form represents a rough lower bound on the Monte Carlo error in any HPD computed from simulations.

Figure 8 shows the intervals constructed for MCMC normal samples. Specifically, the Gibbs sampler is used to draw samples from a standard bivariate normal distribution with correlation 0.9. We use this example to explore how Spin works on simulations with high autocorrelation. Two chains each with 1000 samples are drawn with Gibbs sampling. For one variable, every ten draws are recorded for Spin construction, resulting in 200 samples, which is roughly the level of the effective sample size in this case. This is a typical scenario in practice when MCMC techniques are adopted for multivariate distributions. Again Spin greatly outperforms the empirical shortest interval in case of highly correlated draws.

We further investigate coverage probabilities of the different intervals constructed (Fig. 9). Empirical shortest intervals

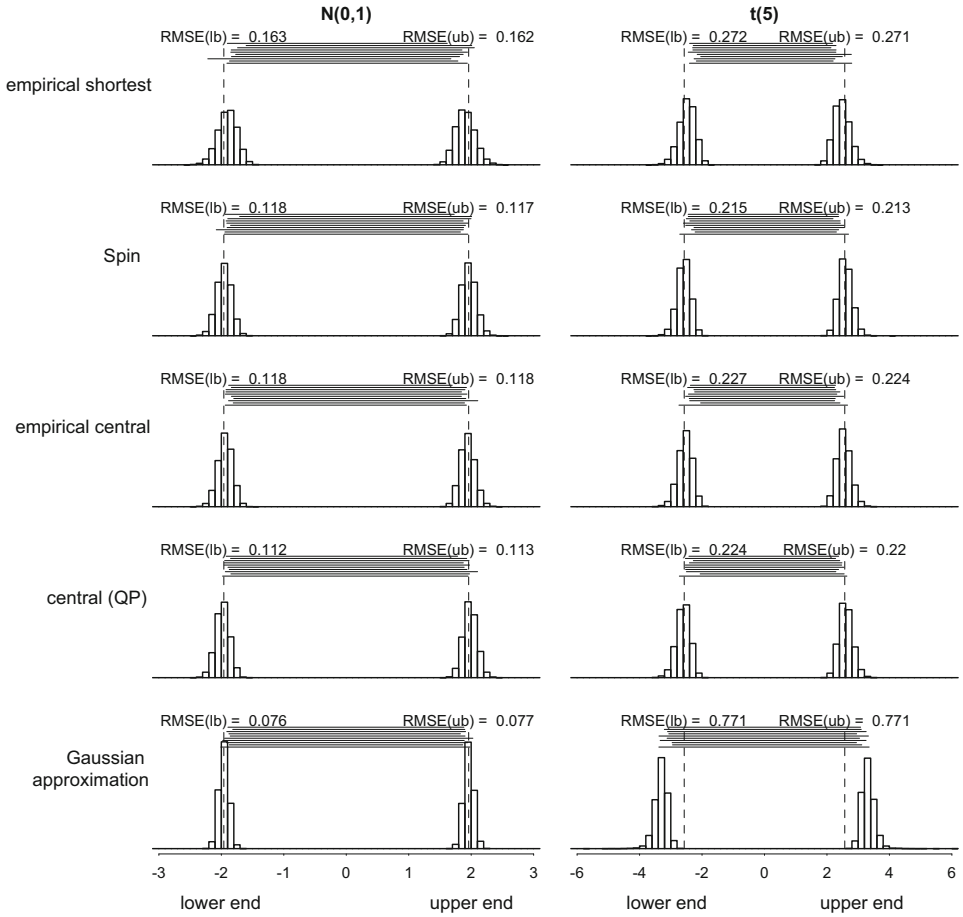


**Fig. 6** Spin for symmetric distributions: 95 % intervals for the normal and $t(5)$ distributions, in each case based on 500 independent draws. Each *horizontal bar* represents an interval from one simulation. The histograms of the lower ends and the upper ends are based on results from 20,000 simulations. The *dotted vertical lines* represent the true endpoints of the HPD intervals. Spin greatly outperforms the raw empirical shortest interval. The central interval (and its quadratic programming improvement) does even better for the Gaussian but is worse for the $t(5)$ and in any case does not generalize to asymmetric distributions. The intervals estimated by fitting a Gaussian distribution do the best for the normal model but are disastrous when the model is wrong

have the lowest coverage probability, which is as expected since they are biased toward shorter intervals (see Figs. 6, 7). Coverage probabilities of Spin are closer to the nominal coverage (95 %) for both normal and gamma distributions. Comparable coverage is observed for central intervals. As expected, parametric intervals represent a gold standard and have the most accurate coverage.

Figure 10 shows the bias–variance decomposition of different interval estimates for normal and gamma distributions under sample sizes 100, 300, 500, 1000, and 2000. We average lower and upper ends for the normal case due to symmetry. For both distributions, Spin has both well-reduced variance and bias compared to the empirical shortest intervals. The upper end estimates of empirical central intervals for the gamma have a large variance since the corresponding density is low so the observed simulations in this region are more variable. It is worth pointing out that the computational time for Spin is negligible compared to sampling, thus it is a more efficient way to obtain improved interval estimates. In the normal example shown in the left panel in Fig. 10, rather than increasing the sample size from 300 to 500 to reduce error, one can spend less time to compute Spin with the 300 samples and get a even better interval.

We also carried out experiments with even bigger samples and intervals of other coverages (90 and 50 %), and got similar results. Spin beats the empirical shortest interval in RMSE (which makes sense, given that Spin is optimizing over a class of estimators that includes the empirical shortest as a special case).

## 4 Results for two real-data examples

In this section, we apply our methods to two applied examples of hierarchical Bayesian models, one from education and one from sociology. In the first example, we show the advantages of Spin over central and empirical shortest intervals; in the second example, we demonstrate the routine use of Spin to summarize posterior inferences.

Our first example is a Bayesian analysis from Rubin (1981) of a hierarchical model of data from a set of experiments performed on eight schools. The group-level scale parameter (which corresponds to the between-school standard deviation of the underlying treatment effects) has a posterior distribution that is asymmetric with a mode at zero (as shown in the right panel of Fig. 1). Central probability intervals for this scale parameter (as presented, for example,

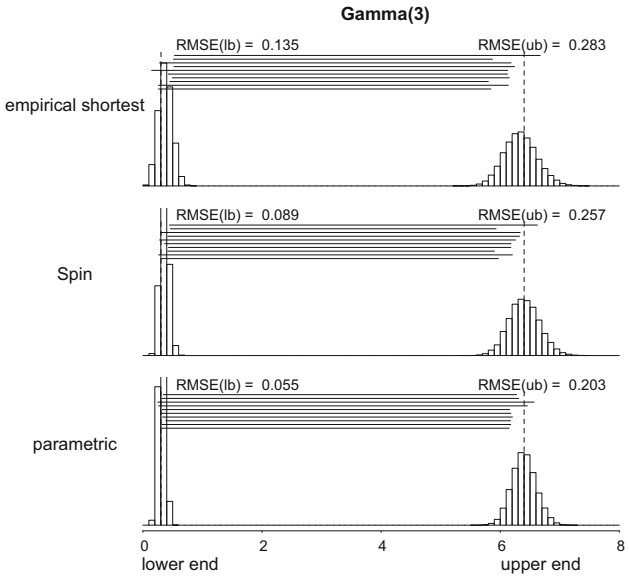


**Fig. 7** Spin for an asymmetric distribution. 95 % intervals for the gamma distributions with shape parameter 3, as estimated from 500 independent draws. Each *horizontal bar* represents an interval from one simulation. The histograms are based on results from 20,000 simulations. The *dotted vertical lines* represent the true endpoints of the HPD interval. Spin outperforms the empirical shortest interval. The interval obtained from a parametric fit is even better but this approach cannot be applied in general, rather, it represents an optimality bound for any method

in the analysis of these data by Gelman et al. (1995)) are unsatisfying in that they exclude a small segment near zero where the posterior distribution is in fact largest. Figure 11 shows the 95 % empirical shortest intervals and Spin constructed from 500 draws. The results of empirical shortest intervals for 8 schools are from including the zero point in the simulations. Spin has smaller RMSE than both empirical shortest and central intervals (Figs. 11, 12).

For our second example, we fit the social network model of Zheng et al. (2006) using MCMC and construct 95 % Spins for the overdispersion parameters based on 200 posterior draws. The posterior is asymmetric and bounded below at 1. Figure 13 is a partial replot of Fig. 4 from Zheng et al. (2006) with Spins added. For this type of asymmetric posterior, we prefer the estimated HPDs to the corresponding central intervals as HPDs more precisely capture the values of the parameter that are supported by the posterior distribution.

## 5 Results for BUGS examples

In this section, we apply our methods to 60 examples from BUGS Spiegelhalter et al. (2014). The 60 examples include 5398 parameters. For each parameter, 1000 MCMC samples

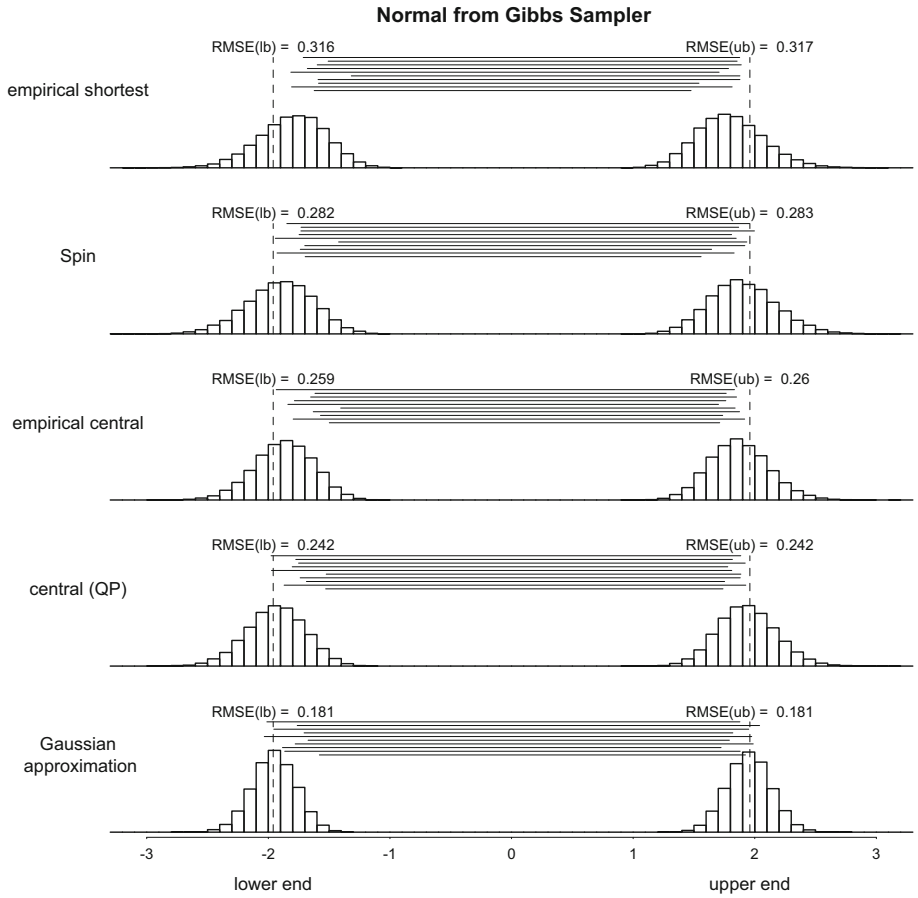


**Fig. 8** Spin for MCMC samples. 95 % intervals for normal samples from Gibbs sampler, in each case based on 200 draws. Each *horizontal bar* represents an interval from one simulation. The histograms are based on results from 20,000 simulations. The *dotted vertical lines* represent the true endpoints of the HPD intervals. Spin greatly outperforms the raw empirical shortest interval. The central interval (and its quadratic programming improvement) does even better. Again the intervals estimated by fitting a Gaussian distribution do the best

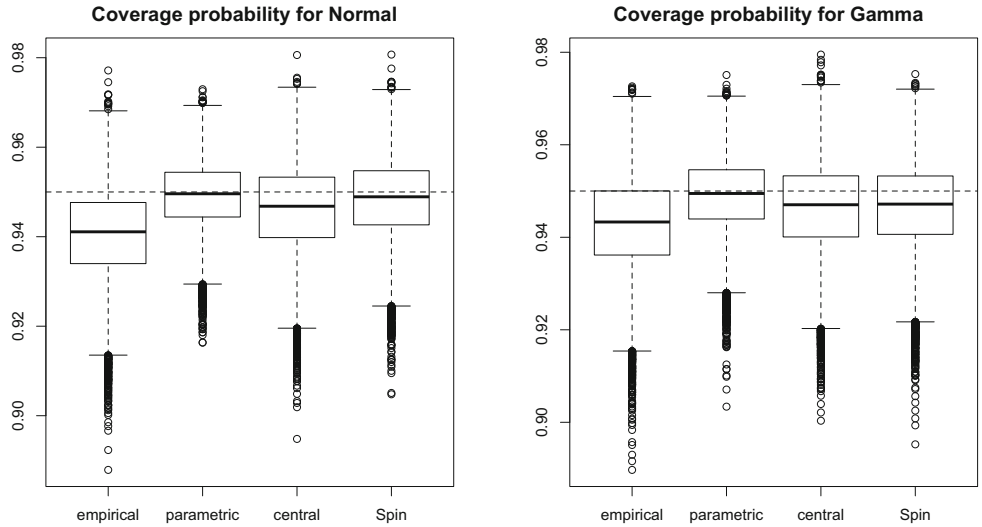


**Fig. 9** Distribution of coverage probabilities for Spin and other 95 % intervals calculated based on 500 simulations for the normal and gamma(3) distributions

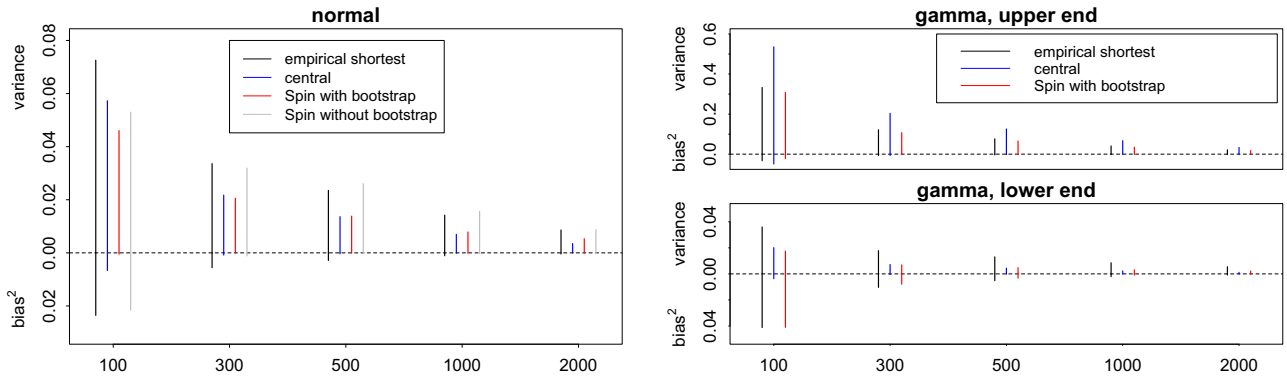


**Fig. 10** Bias–variance decomposition for 95 % intervals for normal and gamma(3) examples, as a function of the number of simulation draws. Because of the symmetry of the normal distribution, we averaged its errors for upper and lower endpoints. Results from Spin without bootstrap are shown for normal for description purpose. (Color figure online)

are drawn using Stan Stan Development Team: RStan (2014), and upper and lower endpoints of empirical HPD and Spin intervals are estimated. The above procedure is conducted for 100 times and the Monte Carlo variance is calculated. Since we do not know the true endpoints of the intervals, we define the efficiency only based on variance as

$$\text{Efficiency} = \frac{\text{Var (empirical shortest interval)}}{\text{Var (Spin)}},$$

We compute the average computational efficiency for all the parameters in each of the 60 models. Figure 14 shows the efficiency of Spin against HPD intervals versus the average computation time (in seconds per effective sample size). It can be seen that almost all the examples result in efficiency greater than 1. We investigate the example corresponding to the lowest point. It turns out that many of the parameters in this specific example are not from unimodal distributions, under which cases, HPD is actually not reasonable.

## 6 Discussion

We have presented a novel optimal approach for constructing reduced error shortest probability intervals (Spin). Simulation studies and real-data examples show the advantage of Spin over the empirical shortest interval. Another commonly used interval estimate in Bayesian inference is the central interval. For symmetric distributions, central intervals and HPDs are the same; otherwise we agree with Box and Tiao (1973) that the HPD is generally preferable to the central interval as an inferential summary (Fig. 1). In our examples, we have found that for symmetric distributions Spin and empirical central intervals have comparable RMSEs and coverage probabilities (Figs. 6, 9, 10). Therefore, we recommend Spin as a default procedure for computing HPD intervals from simulations, as it is as computationally stable as the central intervals which are currently standard in practice.

We set the bandwidth parameter $b$ in (6) to $\sqrt{n}$, which seems to work well for a variety of distributions. We also

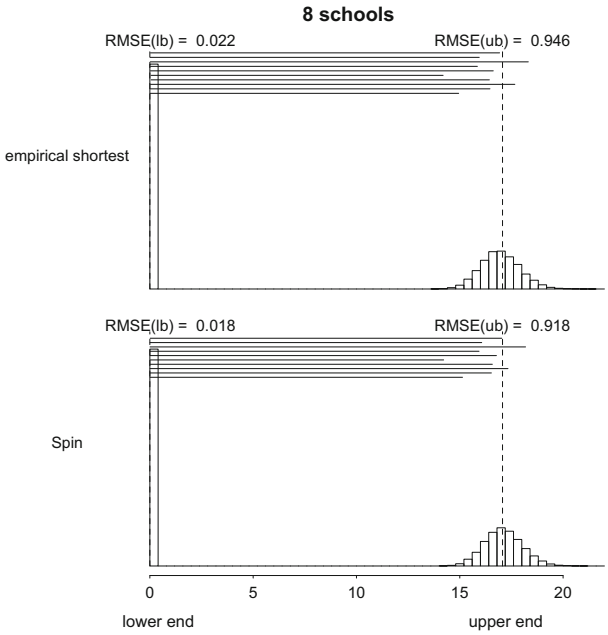


**Fig. 11** Spin for the group-level standard deviation parameter in the eight schools example, as estimated from 500 independent draws from the posterior distribution (which is the *right density curve* in Fig. 1, a distribution that is constrained to be nonnegative and has a minimum at zero). The histograms in this figure are based on results from 20,000 simulations. The *dotted vertical lines* represent the true endpoints of the HPD interval as calculated numerically from the posterior density. Spin does better than the empirical shortest interval, especially at the *left end*, where its smoothing tends to (correctly) pull the lower bound of the interval all the way to the boundary at 0

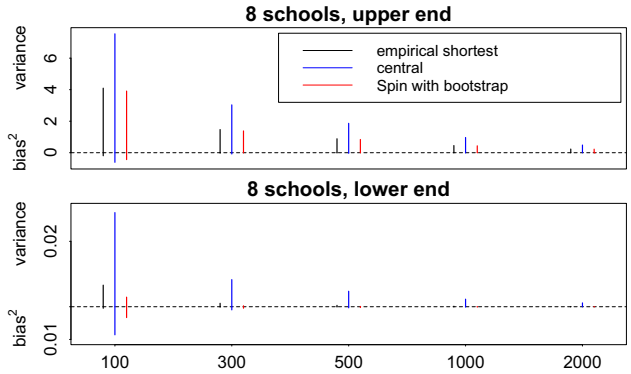


**Fig. 12** Bias–variance decomposition for 95 % intervals for the eight school example, as a function of the number of simulation draws

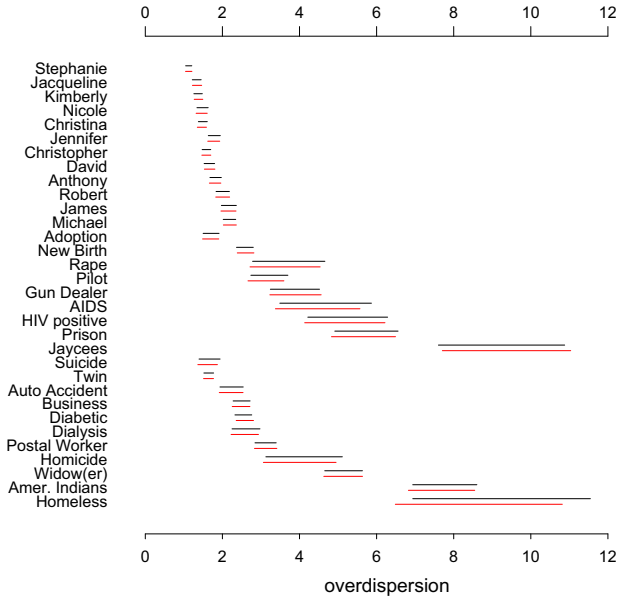


**Fig. 13** 95 % central intervals (*black lines*) and Spins (*red lines*) for the overdispersion parameters in the "How many X's do you know?" study. The parameter in each row is a measure of the social clustering of a certain group in the general population: groups of people identified by first names have low overdispersion and are close to randomly distributed in the social network, whereas categories such as airline pilots or American Indians are more overdispersed (that is, non-randomly distributed). We prefer the Spins as providing better summaries of these highly skewed posterior distributions. However, the differences between central intervals and Spins are not large; our real point here is not that the Spins are much better but that they will work just fine in routine applied Bayesian practice, satisfying the same needs as were served by central intervals but without that annoying behavior when distributions are highly asymmetric. (Color figure online)

carried out sensitivity analysis by varying $b$ and found that large $b$ tends to result in more stable endpoint estimates where the density is relatively high but can lead to noisy estimates where the density is low. This makes sense: in low-density regions, adding more points to the weighted average may introduce noise instead of true signals. Based on our experiments, we believe the default value $b = \sqrt{n}$ is a safe general choice.

Our approach can be considered more generally as a method of using weighted averages of order statistics to construct optimal interval estimates. One can replace $Q(\Delta^*)$ in (5) by the endpoints of any reasonable empirical interval estimates, and obtain improved intervals by using our quadratic programming strategy (such as the improved central intervals shown in Fig. 6).

One concern that arises is the computational cost of performing Spin itself. Our simulations show Spin intervals to have better simulation coverage and appreciably lower mean squared error compared to the empirical HPD, but for simple problems in which one can quickly draw direct posterior simulations, it could be simpler to forget Spin and instead just double the size of the posterior sample. Many times, though, we find ourselves computing Bayesian models using elaborate Markov chain simulations for which it can take many steps of the algorithm, or for which each step is computationally expensive (for example, in models with differential equation solvers), so that hours or even days of computing time are required to obtain an effective sample size of a few hundred posterior simulation draws. In such cases, the computational cost of Spin is relatively small. Thus, we think

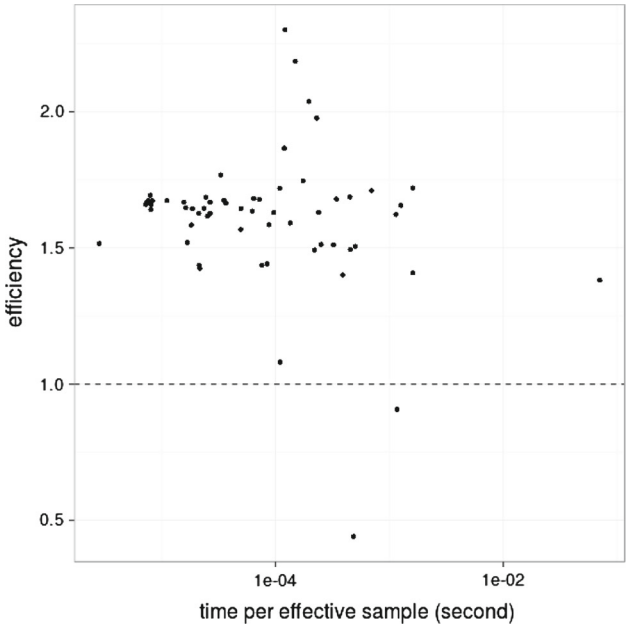


**Fig. 14** Computational efficiency (ratio of Monte Carlo variances in repeated simulations) of Spin compared to empirical HPD intervals, plotted versus average computation time (in seconds per effective sample size), for each of 60 BUGS examples. Spin outperforms empirical HPD intervals in almost all the cases, typically with computational efficiency around 1.7. The one point at the bottom of the graph comes from a model which has many parameters with bimodal posterior distributions, in which case the highest posterior density interval can be difficult to interpret in any case

Spin makes sense as a default option for posterior summaries, especially with simulations that are costly.

We have demonstrated that our Spin procedure works well in a range of theoretical and applied problems, that it is simulation consistent, computationally feasible, addresses the boundary problem, and is optimal within a certain class of procedures that include the empirical shortest interval as a special case. We do not claim, however, that the procedure is optimal in any universal sense. We see the key contribution of the present paper as developing a practical procedure to compute shortest probability intervals from simulation in a way that is superior to the naive approach and is competitive (in terms of simulation variability) with central probability intervals. Now that Spin can be computed routinely, we anticipate further research improvements on posterior summaries.

**Acknowledgments** We thank Chia-Hui Huang, Daniel Lee, and Matt Hoffman for research assistance and the National Science Foundation (Grant CNS-1205516), Institute of Education Sciences (Grant DE R305D140059), and Department of Energy (Grant DE-SC0002099) for partial support of this work.

## References

Box, G.E.P., Tiao, G.C.: Bayesian Inference in Statistical Analysis. Wiley, New York (1973)

Chen, M.H., Shao, Q.M.: Monte Carlo estimation of Bayesian credible and HPD intervals. J. Comput. Gr. Stat. **8**, 69–92 (1999)

David, H.A., Nagaraja, H.N.: Order Statistics, 3rd edn. Wiley, New York (2003)

Efron, B.: Bootstrap methods: another look at the jackknife. Ann. Stat. **7**, 1–26 (1979)

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis. CRC Press, London (1995)

Gelman, A., Shirley, K.: Inference from simulations and monitoring convergence. In: Brooks, S., Gelman, A., Jones, G., Meng, X.L. (eds.) Handbook of Markov Chain Monte Carlo, pp. 163–174. CRC Press, London (2011)

R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2011)

Rubin, D.B.: Estimation in parallel randomized experiments. J. Educ. Stat. **6**, 377–401 (1981)

Spiegelhalter, D., Thomas, A., Best, N., Gilks, W., Lunn, D.: BUGS: Bayesian Inference Using Gibbs Sampling. MRC Biostatistics Unit, Cambridge. http://www.mrc-bsu.cam.ac.uk/bugs (1994, 2002). Accessed 15 Nov 2014

Stan Development Team: RStan: the R interface to Stan, Version 2.5. http://mc-stan.org/rstan.html (2014). Accessed 15 Nov 2014

van der Vaart, A.W.: Asymptotic Statistics. Cambridge University Press, Cambridge (1998)

Zheng, T., Salganik, M.J., Gelman, A.: How many people do you know in prison?: using overdispersion in count data to estimate social structure in networks. J. Am. Stat. Assoc. **101**, 409–423 (2006)