Scientific Research Publishing

Open Journal of Statistics, 2017, 7, 153-172
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Use of BayesSim and Smoothing to Enhance Simulation Studies

Jeffrey D. Hart

Department of Statistics, Texas A&M University, College Station, TX, USA
Email: hart@stat.tamu.edu

How to cite this paper: Hart, J.D. (2017) Use of BayesSim and Smoothing to Enhance Simulation Studies. *Open Journal of Statistics*, 7, 153-172.
https://doi.org/10.4236/ojs.2017.71012

Received: January 10, 2017
Accepted: February 25, 2017
Published: February 28, 2017

Abstract

The conventional form of statistical simulation proceeds by selecting a few models and generating hundreds or thousands of data sets from each model. This article investigates a different approach, called BayesSim, that generates hundreds or thousands of *models* from a prior distribution, but only one (or a few) data sets from each model. Suppose that the performance of estimators in a parametric model is of interest. Smoothing methods can be applied to BayesSim output to investigate how estimation error varies as a function of the parameters. In this way inferences about the relative merits of the estimators can be made over essentially the *entire parameter space*, as opposed to a few parameter configurations as in the conventional approach. Two examples illustrate the methodology: One involving the skew-normal distribution and the other nonparametric goodness-of-fit tests.

Keywords

Loss Function, Bayes Risk, Prior Distribution, Regression, Simulation, Skew-Normal Distribution, Goodness of Fit

1. Introduction

When investigating statistical methodology using simulation, the following strategy is almost always used. Choose several completely specified statistical models, generate hundreds or thousands of data sets from each model, apply the methodology to every data set, and then summarize the results. Usually some effort is made to choose models that represent a variety of model types, although this is often hard to do when only four or five models are considered. Having selected models, one hopes to show that one method is better than another for most if not all the models considered. However, even when Method A works better than Method B in all cases considered, there is a nagging doubt that one has failed to look at all the relevant models, and there might still be cases where B is better than A.

In this article a different simulation strategy is advocated. The idea is to generate one data set (or at most a few data sets) from each of hundreds or thousands of models that are randomly selected from a *prior*. We call such an approach *BayesSim*. Given an appropriate loss function, one may generate BayesSim output to estimate the Bayes risk of each method of interest and then use these estimates as at least part of a basis for choosing amongst methods. Such an approach was used in a simulation study of [1]. One may also use the output of BayesSim to estimate *local* risk of estimators. This is done by grouping output by model similarity, and then computing average loss within groups. An example of this approach may be found in [2].

The basic idea considered in this article is not new. It was proposed by [3], who suggested that BayesSim be applied to study the Bayesian coverage probability of interval estimates in the setting of parametric models. The article [4] proposes a comprehensive Bayesian approach in which parameters are generated from a prior and simulated data are analyzed using Bayesian methodology. One might then ask: "What is new in the current article, and/or what is the point of the current article?"

- A main purpose of the current article is to bring to the attention of statisticians methodology that seems to be unfamiliar to them, but which nonetheless could make their simulations more efficient and informative.
- The articles of [3] and [4] seem to be the exceptions proving the rule that BayesSim is greatly underutilized. The authors of [4] stated of their work: "We hope that it will stimulate readers to learn more about this important subject, and also encourage further research in this area." Unfortunately, their article seems not to have stimulated statisticians, and the current article is another effort to do so.
- The work of [3] and [4] focuses on parametric models. With the current interest in Bayesian nonparametrics, it seems that there is even greater scope for applying BayesSim. When the model is nonparametric, one can envision generating probability distributions from, say, a Dirichlet process, generating data from each distribution and applying some statistical method to each data set.
- The one idea that may be "new" in the current article is the idea of using nonparametric smoothing to analyze BayesSim output. If one opposes the notion of using Bayes principle (averaging all results with respect to a prior) to compare various methods, one may use smoothing in conjunction with BayesSim to compare methodology in the same "local" way that one does in a conventional simulation, but with the advantage that broader conclusions may be drawn.

There are two main considerations in the motivation for BayesSim: simulation efficiency and diversity of models considered. Generating thousands of data sets for each of two or three different models often seems like overkill for those few cases, and is unsatisfying from the standpoint of model diversity. In other words, one learns a lot about very little. In the author's opinion a much better understanding of the *general* performance of a method can be obtained with BayesSim, and this can be done without generating any more data sets than in the traditional approach. This notion can be easily quantified. Suppose one uses k parameter configurations in a conventional approach and generates n observations from each configuration. In BayesSim one may generate the same number of data sets by generating nk parameter values from a prior, and then one data set from each of the nk parameter values. Now, let $R(\theta)$ be the mean squared error of an estimator when the model parameter is θ. One may compare the conventional and BayesSim approaches by asking how efficiently each one can estimate $R(\cdot)$ using its information.

Another reason that BayesSim seems like a good idea is that it has the potential of actually simulating the average performance of a method that is used at a variety of times and places throughout the world. Arguably this *should* be the goal of a simulation study that is trying to make a case for a new method. If the prior used in BayesSim well-approximates the empirical distribution of models or parameter values encountered in practice, then BayesSim will achieve this goal. This suggests a new avenue of research: studying the empirical distributions of various types of parameters, perhaps by the use of meta analyses.

It is worth remarking that the proposed methodology is not Bayesian per se. Nonetheless, the name BayesSim seems appropriate on two counts. First of all, a good choice of prior distribution in BayesSim is based on the same considerations used in choosing an objective prior in a Bayesian analysis. Secondly, we advocate the use of Bayes principle, *i.e.*, averaging results with respect to the prior, as a means of comparing different methods.

A subtheme of the article is the contention that statisticians seldom use the sophisticated methods they develop to analyze simulation data. In the author's opinion the simulations that are an ever-present part of methodological papers would be more interesting and informative if they employed more of the vast array of new statistical methods. For example, in comparing estimation methods for multiparameter models, one could estimate performance measures as a function of the parameter vector by using additive nonparametric regression methods.

2. A Toy Example

Before giving a more detailed description of our method we present a toy example that illustrates the potential advantages of BayesSim. The example involves binomial experiments, whose statistical properties are, of course, well understood. Imagine for the moment, though, that nothing is known about those properties, and we wish to use simulation to learn about, say, the variability of sample proportions. Let $\hat{\theta}$ be the sample proportion in a binomial experiment with $n = 100$ and success probability θ.

Results of a "traditional" simulation are shown in Figure 1. Five settings for

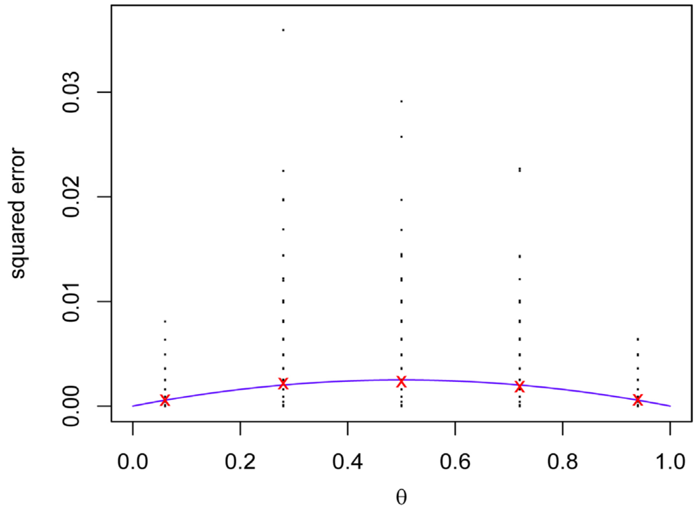

Figure 1. A traditional simulation involving binomial experiments. At each success probability θ, the 500 points are squared errors for sample proportions from simulated binomial experiments with $n=100$. Each red X is the sample mean of all 500 squared errors, and the blue line is the true mean squared error curve, $\theta(1-\theta)/100$.

θ have been chosen: $(\theta_1,\cdots,\theta_5)=(0.06,0.28,0.50,0.72,0.94)$. Five hundred binomial experiments are performed at each θ_j, and **Figure 1** shows all 2500 values of $\left(\hat{\theta}-\theta\right)^2$. Obviously one obtains very good estimates of mean squared error at each θ_j, but is at least somewhat uncertain about the functional form of

$$R(\theta)=E\left[\left(\hat{\theta}-\theta\right)^2\right]=\theta(1-\theta)/100.$$

Figure 2 shows results from an application of BayesSim. Two thousand values of θ were generated from a beta (1/2, 1/2) distribution, which is the Jeffreys noninformative prior for a binomial experiment. A single binomial experiment was then conducted for each θ. A Gaussian kernel local linear smooth of the squared errors is quite close to the true mean squared error curve. The bandwidth of the smooth was chosen by one-sided cross-validation ([5]). Note that a total of 2000 points were generated in BayesSim while 2500 were generated in the traditional approach. **Figure 2** seems to accurately portray the facts that 1) generating 500 points at exactly the same θ is excessive, and 2) the entire mean squared curve is well estimated by the local linear smooth.

One *could* form an estimate of the mean squared error curve, $R(\theta)$, using the information from the traditional simulation. A possible approach would be to use a Fourier series estimate of the form

$$\hat{R}_T(\theta)=\hat{\phi}_0+2\sum_{j=1}^{4}\hat{\phi}_j\cos(\pi j\theta),\quad 0\le\theta\le 1, \tag{1}$$

where

$$\hat{\phi}_j=0.22\sum_{k=1}^{5}\bar{Y}_k\cos(\pi j\theta_k),\quad k=0,\cdots,4, \tag{2}$$

and $\bar{Y}_j$ is the average of the 500 mean squared errors at θ_j, $j=1,\cdots,5$.

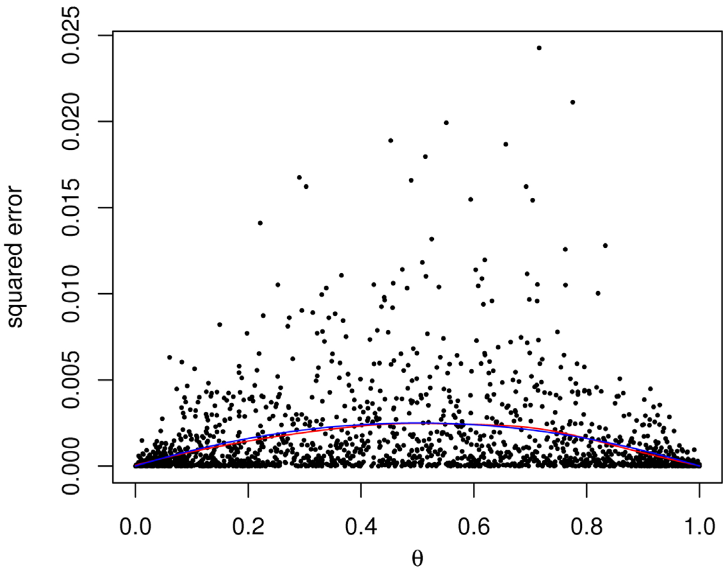

Figure 2. Using BayesSim in binomial experiments. Each of the two thousand points is of the form $\left(\theta,\left(y/100-\theta\right)^2\right)$, where y was generated from a binomial distribution with $n=100$ and success probability θ. The 2000 values of θ were generated from a beta distribution with both parameters equal to $1/2$. The red and blue lines are a local linear smooth and the true mean squared error curve, respectively.

It is reasonable to define the efficiency of an estimator $\hat{R}$ by its mean integrated square error:

$$\text{Eff}\left(\hat{R}\right)=E\left[\int_0^1\left(\hat{R}(\theta)-R(\theta)\right)^2 \mathrm{d}\theta\right]. \tag{3}$$

Doing so is analogous to the well-known practice of using information to measure the efficiency of an experimental design. Table 1 summarizes the efficiencies of the two approaches in our toy example. Each component of Eff, integrated variance and integrated squared bias, is smaller for BayesSim, this in spite of the fact that more data sets are generated in the traditional approach. It is worth noting that if the number of data sets generated in the traditional approach tended to infinity with the number, five, of parameter values fixed, the component IV would tend to 0 but IB would remain fixed. This well illustrates the diminishing returns aspect of generating very large numbers of data sets in the traditional approach. In contrast, both IV and IB would tend to 0 in the BayesSim approach if the number of replications tended to ∞.

We close this section with the following remarks.

- The results in Table 1 actually present the traditional approach in too favorable a light. The Fourier series estimate is a type of smoother and makes better use of the information in the simulated data than is typical. Usually, researchers would simply report what happened at the five parameter values.
- The Fourier series estimate in our toy example actually works quite well in spite of the fact that only five settings of θ were considered. This is due to the very smooth nature of $R(\theta)$ in this case. However, one should re-

Table 1. Efficiency of traditional and BayesSim approaches in estimating mean squared error curve. The quantities IV, IB and Eff are 10^9 (integrated variance), 10^9 (integrated squared bias) and 10^9 (mean integrated squared error), respectively.

Method	IV	IB	$MISE$
Traditional	15.03	8.40	23.43
BayesSim	14.58	5.95	20.53

member that this *is* a toy example. If the functional form of $R(\theta)$ were more complicated and/or there were multiple parameters, the advantage of BayesSim in conjunction with smoothing would be more pronounced.

3. BayesSim

It is important to clarify what will be meant by our use of the term "model." Occasionally we will use model in the sense of "parametric model," meaning a collection of probability density or mass functions indexed by a parameter. More often, however, model will mean a completely specified probability distribution from which data might be generated. In this sense, a standard normal distribution could be a model, if that were the density that generated independent observations. This use of model makes it easier to describe our most general methodology, wherein each "model" may be definable only in terms of infinitely many parameters, or where one might be considering several different parametric families all at once. In the latter case, one might employ BayesSim by first randomly choosing a parametric family, and then randomly generating a parameter value for the chosen family.

3.1. The Basic Method

Consider how a specific statistical methodology is often used in the real world. Over a certain period of time, N researchers working in different parts of the world and in different subject matter areas, apply this methodology. Let us assume that the true model from which researcher i obtains his or her data is M_i, $i=1,\cdots,N$. Now, let L_i be the loss experienced by researcher i upon applying the method in question. Then the efficacy of the method could be measured by the distribution of all L_is, or the average loss $N^{-1}\sum_{i=1}^{n} L_i$. This way of evaluating methodology is precisely what BayesSim tries to simulate. I submit that it would rarely happen that any of $M_1,\cdots,M_N$ would be the same, nor is it common that any researcher would obtain multiple data sets from the same model. For this reason, comparing the efficacy of methods by generating thousands of data sets from the same few models does not seem quite right. The point is that *variation due to differences in models is a real and prevalent factor in the performance of statistical methodology and is largely neglected in traditional simulations.*

We describe BayesSim using decision theoretic terminology. Let $\boldsymbol{Y}$ be a data vector and suppose that $\boldsymbol{Y}$ has distribution $f(\cdot|M)$ when the true model is M. Let $\delta(\boldsymbol{Y})$ be some statistical rule, perhaps an estimator or a test of hypothesis. The "action" prescribed by δ when the observed data are $\boldsymbol{y}$ is

$\delta(\boldsymbol{y})$. For example, if δ is an estimator, then $\delta(\boldsymbol{y})$ is the point estimate for data set $\boldsymbol{y}$. Now suppose that $L(M,a)$ is the loss when M is the true model and action a is taken. The frequentist risk, or simply risk, of δ when M is the true model is

$$R(M,\delta)=E_M L(M,\delta(\mathbf{Y})), \tag{4}$$

where E_M is expectation with respect to the distribution $f(\cdot|M)$.

The usual simulation strategy is to try to obtain a very good estimate of $R(M,\delta)$ for a few models M and each rule δ of interest. In estimation problems a typical choice for L is squared error, while in testing L is usually 0 - 1 loss, leading to power and Type I error probability as the criteria for comparing tests.

Now suppose that $\mathcal{M}$ is a collection of models, and let π be a probability measure over $\mathcal{M}$. Then the Bayes risk of δ with respect to π is

$$r(\delta,\pi)=\int_M R(M,\delta)\mathrm{d}\pi(M). \tag{5}$$

Obviously, $r(\delta,\pi)$ is a weighted average of the risks $R(M,\delta)$ that are usually considered individually in deciding on a good statistical rule. Bayes principle consists of choosing a rule to minimize Bayes risk, and is one of the main principles used in decision theory. Possibilities for $\mathcal{M}$ include parametric families, nonparametric, or infinite dimensional, families, and a collection of the form $\mathcal{M}=\mathcal{M}_1\cup\cdots\cup\mathcal{M}_m$, where each of $\mathcal{M}_i$, $i=1,\cdots,m$, is a parametric family.

In a simulation setting where $\mathcal{M}$ and π are chosen by the investigator, a consistent estimator of the Bayes risk of a rule δ may be constructed as follows. Let $M_1,\cdots,M_N$ be N independent draws from π. Generate a data vector $\boldsymbol{y}_i$ from M_i for each $i=1,\cdots,N$, and estimate the Bayes risk of rule δ by

$$\hat{r}(\delta,\pi)=\frac{1}{N}\sum_{i=1}^{N}L(M_i,\delta(\boldsymbol{y}_i)). \tag{6}$$

If $N\to\infty$, then under mild conditions $\hat{r}(\delta,\pi)$ converges in probability to $r(\delta,\pi)$. Consistency follows from the law of large numbers and the fact that

$$\begin{aligned} E\left[L(M_i,\delta(\mathbf{Y}_i))\right] &= E\left\{E\left[L(M_i,\delta(\mathbf{Y}_i))\middle|M_i\right]\right\} \\ &= E\left\{R(M_i,\delta)\right\} \\ &= r(\delta,\pi). \end{aligned} \tag{7}$$

Another factor that varies in the practical problem outlined above is sample size. A prior π_n for the sample size could also be incorporated into the procedure. On each replication of the simulation, one would first randomly select a model, then select a sample size n from π_n, and finally generate a data set of size n. However, mostly for the sake of simplicity, we will assume a fixed sample size throughout the remainder of the article.

3.2. More Efficient Estimation of Bayes Risk

A result of [3] makes use of a fundamental property to show that there exists a more efficient estimator of Bayes risk than the simple average of losses described in the previous section. Assume that $\mathcal{M}$ is a parametric family with (continuous) parameter space Θ and that $f(\boldsymbol{y}|\theta)$ is the distribution of $\boldsymbol{Y}$ given θ. For each $\boldsymbol{y}$ the posterior risk is

$$r(\boldsymbol{y})=\int_{\Theta}L(\theta,\delta(\boldsymbol{y}))\pi(\theta|\boldsymbol{y})\mathrm{d}\theta, \tag{8}$$

where

$$\pi(\theta|\boldsymbol{y})=\frac{f(\boldsymbol{y}|\theta)\pi(\theta)}{m(\boldsymbol{y})} \quad \text{and} \quad m(\boldsymbol{y})=\int_{\Theta}f(\boldsymbol{y}|\theta^*)\pi(\theta^*)\mathrm{d}\theta^*. \tag{9}$$

It is easy to see that the expected value of $r(\boldsymbol{Y})$ with respect to the marginal m is equal to $r(\delta,\pi)$.

A second strategy for estimating Bayes risk is therefore:

- Generate θ from π and $\boldsymbol{Y}$ from $f(\cdot|\theta)$.
- Repeat the previous step N times, producing $r(\boldsymbol{y}_1),\cdots,r(\boldsymbol{y}_N)$.
- Compute the estimate $\hat{r}_1=N^{-1}\sum_{i=1}^{N}r(\boldsymbol{y}_i)$.

Both $\hat{r}_1$ and $\hat{r}_2=\hat{r}(\delta,\pi)$ are unbiased estimators of $r(\delta,\pi)$. However, $\hat{r}_1$ has smaller variance, assuming the density π is not degenerate. This follows from the well-known iterated expectation rule for variance, which yields

$$\mathrm{Var}(\hat{r}_2)=\mathrm{Var}(\hat{r}_1)+\frac{1}{N}E\left[\left\{L(\theta,\delta(\boldsymbol{Y}))\middle|\boldsymbol{Y}\right\}\right]. \tag{10}$$

In the setting of [3], this result entails that, when applying BayesSim to evaluate interval estimates, the standard error of the average of posterior coverage probabilities is smaller than that of the proportion of intervals that contain their respective parameter values.

This efficiency result may be of limited usefulness in applications envisioned by the author. The main problem is that it will often be difficult to compute the posterior risk $r(\boldsymbol{y})$. The necessity of MCMC methods in most Bayesian data analyses is at the heart of this difficulty. However, if one knows how to generate independent samples from each posterior, then there is a fairly simple way to reduce the variability of $\hat{r}_2$ without having to compute posterior risk. For each $\boldsymbol{y}$ that is generated in BayesSim , generate a few, say M , independent values from $\pi(\cdot|\boldsymbol{y})$. Call these values $\theta_1(\boldsymbol{y}),\cdots,\theta_M(\boldsymbol{y})$. Now define the estimator

$$\hat{r}_3=\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}L(\theta_j(\mathbf{Y}_i),\mathbf{Y}_i). \tag{11}$$

This estimator is unbiased for $r(\delta,\pi)$ and has variance

$$\mathrm{Var}(\hat{r}_3)=\mathrm{Var}(\hat{r}_1)+\frac{1}{NM}E\left[\mathrm{Var}\left\{L(\theta,\delta(\boldsymbol{Y}))\middle|\boldsymbol{Y}\right\}\right], \tag{12}$$

which is smaller than that of $\hat{r}_2$ for any M larger than 1.

When one is more interested in using BayesSim to estimate frequentist risk $R(\theta,\delta)$ as a function of θ, then the efficiency issue raised in this section becomes moot. Estimation of $R(\theta,\delta)$ is the subject of Section 3.3.

3.3. Local Risk Estimation

One could argue that choosing Method A over Method B simply because A has smaller Bayes risk is not a good practice. After all, it is possible that A could have smaller Bayes risk than B, but the risk of B could be smaller than that of A for a majority of models. We have two responses to this point. Firstly, average performance is the default way of comparing methods on a single model. Researchers seem to be sold on the idea that method A is preferable to B for a given model if the risk of A is smaller than that of B, even though B may have smaller loss in a majority of data sets from that model. A second response is that when the prior in BayesSim is actually the empirical distribution of models, BayesSim output can be used to estimate the proportion of cases in practice where one method will have smaller loss than another. This is not something that can be done using the traditional simulation approach.

Even when there are doubts about the practical validity of the prior, valuable *local* information about methods can be obtained using the output from BayesSim. Suppose it is believed that the risk of a method or methods depends importantly on a certain d-dimensional functional of the model, say $g(M)$. When the model is parametric, $g(M)$ might simply be the vector of parameters. One may regress $L(M_i,\delta(\boldsymbol{y}_i))$ on $g(M_i)$ using BayesSim output in order to estimate the function $m(\boldsymbol{x})=\left[L(M_i,\delta(\boldsymbol{Y}_i))\middle|g(M_i)=\boldsymbol{x}\right]$. This will in part mitigate concerns that the Bayes risk involves too much averaging. If m varies with $\boldsymbol{x}$, then at least part of the variation in risk due to model differences is being explained by the regression. In the (perhaps rare) event that the mapping g is one to one, then $m(\boldsymbol{x})$ is tantamout to the risk $R(M,\delta)$.

Suppose the model is parametric and $g(M)$ is the vector of parameters θ. By definition in BayesSim, the distribution of $\boldsymbol{Y}$ given θ does not depend on the prior distribution of θ. For this reason, *one may obtain a consistent estimator of the risk* $m(\theta)$ *even though the prior is "wrong."* (By a "wrong" prior, we mean one that does not actually correspond to the empirical distribution of parameter values over all settings in which the parametric model is to be used.) This is an important point since it shows that if one is more interested in traditional risk than Bayes risk, then the particular prior used is not of utmost importance.

All the existing methods of estimating regression functions can be applied to inference of m from BayesSim output. For example, we may fit a parametric model to the data $\left(g(M_i),L(M_i,\delta(\boldsymbol{y}_i))\right)$, $i=1,\cdots,N$, which would be especially appealing when d is large.

If no parametric model for the BayesSim output is obvious, and d is large, then the curse of dimensionality comes into play. However, this should not be a real obstacle to the statistician, who simply needs to apply some of the vast array of new regression methodology that deals with this problem. For example, support vector machines ([6]) can be used to classify data in the presence of a large number of regressors. Given BayesSim output, data could be categorized according to values of the loss function, and then a support vector machine used to identify subsets of the regressor space that provide the best discrimination among these categories. Another approach in regression is to try to reduce the dimensionality of the regressor space. Additive models ([7]) and projection pursuit ([8]) are but two methods that could be used for this purpose.

When $\mathcal{M}$ is a class of functions, a possible way of summarizing elements of $\mathcal{M}$ is to use principal components for curves. We may regard a randomly chosen element f of $\mathcal{M}$ as a stochastic process. The process may be represented by a Karhunen-Loève expansion of the form

$$f(x)=\sum_{i=1}^{\infty}\theta_i b_k(x), \tag{13}$$

where $b_1, b_2, \cdots$ are eigenfunctions each having norm 1. Letting K be the covariance kernel of the process, b_1 has the property that it maximizes

$$\mathrm{Var}\left[\int b_1(x) f(x)\mathrm{d}x\right]=\int\int b_1(s) b_1(t) K(s,t)\mathrm{d}s\mathrm{d}t, \tag{14}$$

subject to the constraint that b_1 have norm 1. Each subsequent function b_i maximizes $\int\int b_i(s) b_i(t) K(s,t)\mathrm{d}s\mathrm{d}t$ subject to the constraints that b_i has norm 1 and is orthogonal to each of $b_1, \cdots, b_{i-1}$.

Now, suppose $f_1, \cdots, f_N$ are randomly chosen curves in a BayesSim analysis. Then we may compute the first few principal components of each curve, *i.e.*,

$$\theta_{ij}=\int b_i(x) f_j(x)\mathrm{d}x, \quad i=1,\cdots,I,\ j=1,\cdots,N, \tag{15}$$

and use these components as regressors, per the discussion at the beginning of this section.

4. Choosing Priors

Ideally the prior distribution π used in BayesSim is a good approximation to the frequency distribution of models encountered in practice. If this is the case, then the Bayes risk truly represents the average loss associated with a method over all cases encountered in practice. Considerations in the choice of π are more or less the same as those used in making an *objective* choice of prior in a Bayesian data analysis.

Suppose that the collection of models $\mathcal{M}$ is parametric. Then if one is investigating methodology to be used in a particular subject matter area, it would be sensible to choose π to approximate whatever is known about the distribution of parameter values in that area. If nothing is known about said distribution, then a noninformative prior could be used.

Likewise, a noninformative prior would seem to be a sensible choice when one is investigating methodology that will be used in a wide variety of subject matter areas. However, the question arises "Do commonly used noninformative priors well-approximate the frequency distribution of parameters over a variety of settings?" There are at least a couple of famous examples of when they apparently do. The Jeffreys noninformative prior for a proportion in a binomial experiment is the U-shaped beta distribution proportional to $\theta^{-1/2}(1-\theta)^{-1/2}$ for $\theta \in (0,1)$. The author of [9] collected data from a variety of settings that allowed him to estimate the frequency distribution of proportions. His data were more consistent with a U-shaped distribution than with the uniform distribution, the latter of which seems more intuitive. A similar phenomenon occurs with the distribution of leading digits of quantities coming from a variety of sources. The probability of digit j is $\log_{10}\left(1+j^{-1}\right)$ for $j=1,\cdots,9$ as opposed to $1/9$, a fact known as Benford's Law; see [10]. The author of [11] (p. 86) points out that the Jeffreys noninformative prior for a scale parameter produces this frequency distribution for leading digits, and thus can be expected to be the “correct” empirical distribution of scale parameters.

Whether the previous two examples are isolated, or there are numerous situations where, say, a Jeffreys prior coincides with the empirical distribution of parameters, is a question beyond the scope of this article. However, it seems to be a worthwhile question for further research, especially if BayesSim were to become a commonly used method of simulation. A means of addressing this question would be to perform meta analyses, or to study the results of existing meta analyses.

When the models of interest are nonparametric, construction of priors becomes more challenging. Here, recent proposals in the Bayesian literature are of interest. For example, a commonly used noninformative prior for probability densities is the Dirichlet process, [12]. In regression and other areas where models are curves, curves could be generated as sample paths of a Gaussian or some other stochastic process. Another possibility is to represent a curve using Fourier series, and to generate curves by generating sequences of Fourier coefficients.

5. Examples

We consider two examples. One is parametric and the other nonparametric, involving the skew-normal distribution and goodness of fit, respectively.

5.1. A Parametric Model

Suppose we have a random sample $X_1,\cdots,X_n$ from the following skew-normal density:

$$f\left(x\middle|\xi,\omega,\delta\right)=\frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\frac{\delta}{\sqrt{1-\delta^2}}\left(\frac{x-\xi}{\omega}\right)\right),\quad -\infty<x<\infty, \tag{16}$$

where ϕ and Φ are the pdf and cdf, respectively, of the standard normal distribution, and the parameter space is $\left\{\left(\xi,\omega,\delta\right):-\infty<\xi<\infty,\omega>0,-1<\delta<1\right\}$. The quantity δ is the so-called *skew* parameter and controls the degree of skewness of the distribution. Obviously, when $\delta=0$ the density is $N\left(\xi,\omega^2\right)$. As δ tends to 1(−1), the density approaches a half-normal distribution with left (right) endpoint ξ.

We wish to compare the efficiency of maximum likelihood (ML) and method-of-moments (MOM) estimators of $\left(\xi,\omega,\delta\right)$ using BayesSim. This is a good example for use of BayesSim for a couple of reasons. Firstly, finding exact expressions for the mean squared error of estimators in finite samples appears to be hopeless. Secondly, the information matrix of the skew-normal model has a well-known singularity at $\delta = 0$; see [13]. This means that, even for large samples, the usual infomation matrix should not be used for approximating standard errors of MLEs when δ is near 0.

We will consider three priors, which differ only with respect to the distribution of δ. Each prior is such that (ξ, ω) is independent of δ, δ has a beta distribution, ω has a gamma distribution with shape and rate parameters each $1/2$, and, given ω, ξ is normal with mean 0 and variance ω^2. For $a > 0$ and $b > 0$, the beta density is proportional to $u^{a-1}(1-u)^{b-1}$ for $0 < u < 1$. The three beta densities employed have (a,b) equal to $(1/2,1)$, $(1,1)$ and $(1,1/2)$. The first and third of these emphasize values of δ near 0 and 1, respectively, while the second is uniform. Since this example is for the sake of illustration, we consider only the sample size $n = 100$. We generate 10,000 independent and identically distributed three-vectors from each of the three priors. For each given three-vector (ξ, ω, δ), we generate one sample of size 100 from a skew-normal distribution having those parameters.

MOM estimates are computed from sample moments using the following equations:

$$\mu = E(X_i) = \xi + \omega\delta\sqrt{\frac{2}{\pi}} \tag{17}$$

$$\sigma^2 = \mathrm{Var}(X_i) = \omega^2\left(1 - \frac{2\delta^2}{\pi}\right) \tag{18}$$

$$\gamma = \frac{E\left[(X_i - \mu)^3\right]}{\sigma^3} = \left(\frac{4-\pi}{2}\right)\frac{\delta^3}{\left(\pi/2 - \delta^2\right)^{3/2}} \tag{19}$$

Maximum likelihood estimates were approximated using the nonlinear optimization routine optim in R.

Basic results are summarized in Table 2 and Table 3. The Bayes risks in Table 2 correspond to squared error loss. For example, the entry 0.099 in the first line is $\sum_{i=1}^{10,000}\left(\hat{\delta}_{i,MLE} - \delta_i\right)^2 / 10,000$, where δ_i and $\hat{\delta}_{i,MLE}$ are the value of δ and the ML estimate of δ, respectively, generated on the ith replication of the simulation that used the Beta $(1/2,1)$ prior. One interesting conclusion is that the performance of both ML and MOM estimators degrades as the proportion of δs near 1 becomes less. Also, not surprisingly, the ML estimators are substantially more efficient than the MOM estimators. Table 2 provides more insight by giving bounds for the probability that ML error is less than MOM error.

More detailed information can be obtained by studying the relationship between estimation error and the parameters (ξ, ω, δ). Consider Figure 3 in which ML estimates of δ are plotted against δ for the case where the δ prior was uniform. There is an intriguing pattern in which most of the points are

Table 2. Estimates of Bayes risk for maximum likelihood and method-of-moments estimators. CI denotes a 95% confidence interval for $r\left(\hat{\theta}_{MOM},\pi\right)-r\left(\hat{\theta}_{MLE},\pi\right)$.

		Parameter		
Prior		ξ	ω	δ
Beta $(1/2,1)$	MLE	0.404	0.092	0.099
	MOM	1.426	0.155	0.522
	CI	(0.924, 1.121)	(0.052, 0.073)	(0.413, 0.432)
Uniform	MLE	0.349	0.077	0.109
	MOM	1.137	0.102	0.499
	CI	(0.700, 0.877)	(0.016, 0.032)	(0.379, 0.401)
Beta $(1,1/2)$	MLE	0.290	0.073	0.083
	MOM	0.903	0.092	0.361
	CI	(0.525, 0.699)	(0.010, 0.027)	(0.268, 0.289)

Table 3. Ninety-five percent confidence intervals for probability that MLE error is smaller than MOM error.

	Parameter		
Prior	ξ	ω	δ
Beta $(1/2,1)$	(0.762, 0.779)	(0.627, 0.646)	(0.777, 0.793)
Uniform	(0.695, 0.713)	(0.548, 0.569)	(0.717, 0.734)
Beta $(1,1/2)$	(0.672, 0.690)	(0.548, 0.568)	(0.713, 0.731)

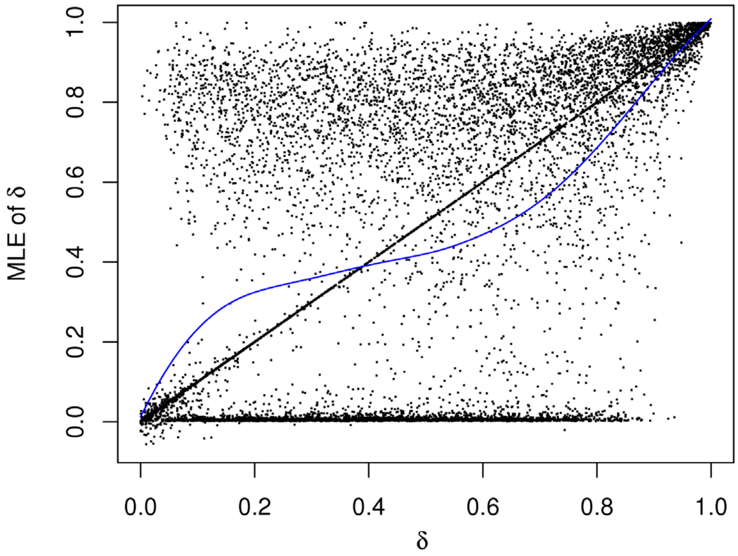

Figure 3. Maximum likelihood estimates of δ plotted against δ. The solid (blue) line is a local linear smooth.

scattered widely throughout the interval $(0,1)$, but a small proportion of points lie very near the 45 degree line, as one would hope most of them would. The solid line is a local linear estimate, indicating considerable bias in the ML estimates for most values of δ. Further investigation shows that the parameter

ω explains the peculiar scatterplot. In Figure 4 one sees that almost all the points on the 45 degree line occur when $\omega < 0.2$. Also, the local linear smooths in Figure 4 show that the ML estimates are more nearly unbiased when ω is small.

Figure 5 reveals how profoundly poor moment estimates of δ are unless δ is near 1. It's as if the moment estimate "thinks" δ is zero unless δ is larger than 0.6. In a sense this is not too surprising as plots of the skew normal density reveal that skewness is almost unnoticeable until δ gets close to 1.

Although it is possible that conclusions as in the last two paragraphs would arise from a carefully done conventional simulation, it seems that the conclusions are much clearer and more easily reached when BayesSim is used.

5.2. Goodness-of-Fit Tests

Here we compare the performance of two nonparametric goodness-of-fit tests of the null hypothesis that a data set is drawn from a normal distribution. Let $X_1, \cdots, X_n$ be a random sample from an unknown density f. We wish to test the null hypothesis that f is normally distributed with mean μ and variance

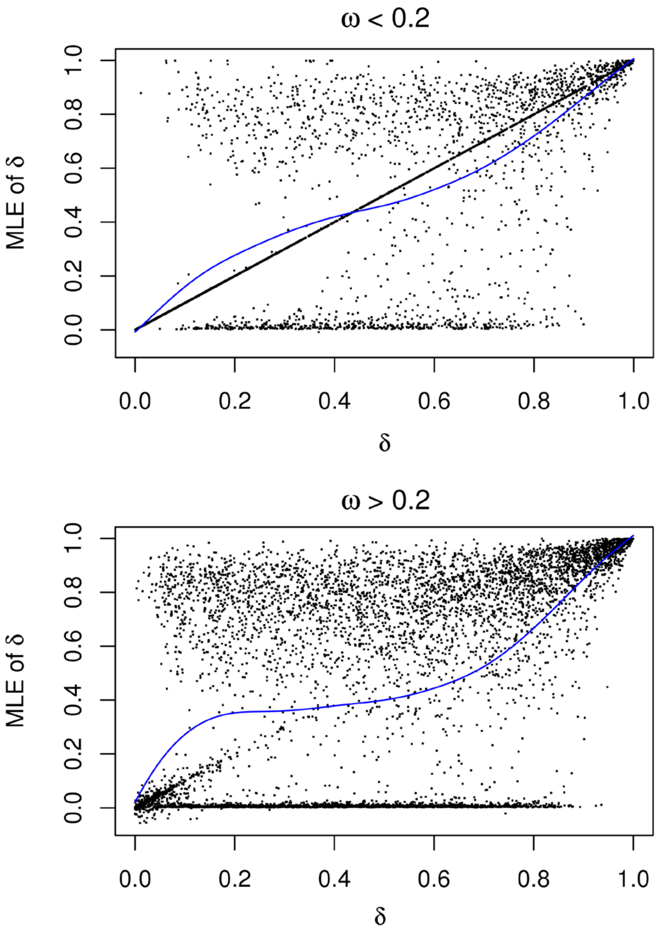

Figure 4. Maximum likelihood estimates of δ plotted against δ for $\omega < 0.2$ and $\omega > 0.2$. In each plot the solid (blue) line is a local linear smooth.

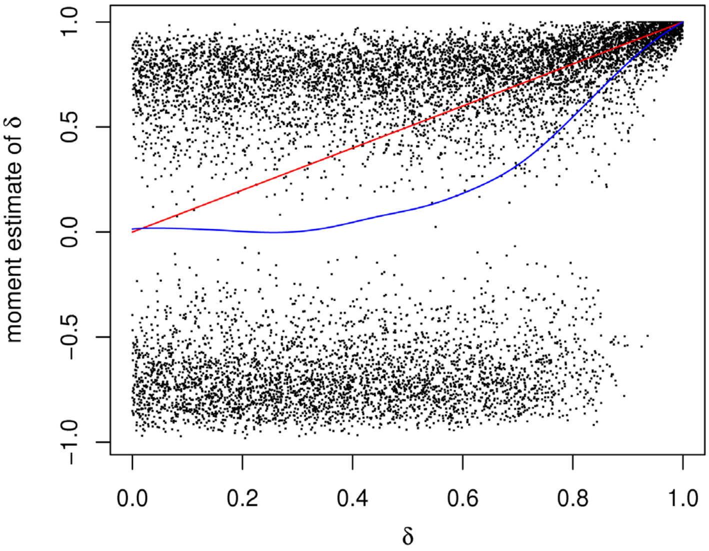

Figure 5. Method of moments estimates of δ plotted against δ. The straight (red) line is the 45 degree line and the other is a local linear smooth.

σ^2, where neither μ nor σ^2 is specified. The tests considered are the popular Shapiro-Wilk test ([14]) and a test based on a kernel density estimator. The latter test employs a nonparametric likelihood ratio of the form

$$\Lambda(h)=\frac{\prod_{i=1}^{n}\hat{f}_i\left(X_i\middle|h\right)}{\left[2\pi^2\right]^{-n/2}\exp(-n/2)},\tag{20}$$

where $\hat{\sigma}^2$ is the maximum likelihood estimator of σ^2 on the assumption of normality, and $\hat{f}_i\left(\cdot\middle|h\right)$ is a "leave-one-out" kernel density estimator:

$$\hat{f}_i\left(x\middle|h\right)=\frac{1}{(n-1)h}\sum_{j\neq i}K\left(\frac{x-X_j}{h}\right).\tag{21}$$

The numerator of $\Lambda(h)$ is the likelihood cross-validation criterion, as studied by [15]. To deal with the bandwidth selection problem, our test statistic is $\hat{\Lambda}=\sup_{0<h\le 2}\Lambda(h)$, and for the kernel, we use

$$K(z)=\left\{(8\pi e)^{1/2}\,\Phi(1)\right\}^{-1}\exp\left[-\frac{1}{2}\left\{\ln\left(1+|z|\right)\right\}^2\right].\tag{22}$$

Results in [15] show that this kernel generally performs well when used in $\Lambda(h)$, whereas "traditional," lighter-tailed kernels such as the Gaussian do not when the underlying density is sufficiently heavy-tailed.

Our interest is in comparing the power of the two tests, and so in using BayesSim we generate densities that are not normal. Each density generated is a mixture of two or more normals, having the form

$$f(x)=\sum_{i=1}^{M}w_i\frac{1}{\sigma_i}\phi\left(\frac{x-\mu_i}{\sigma_i}\right).\tag{23}$$

Densities are generated as follows:

- A value of M between 2 and 20 is selected from a distribution such that the probability of m is proportional to m^{-1}, $m = 2, \cdots, 20$.
- Given m, values $w_1, \cdots, w_m$ are selected from the Dirichlet distribution with all m parameters equal to $1/2$.
- Given m and independent of $w_1, \cdots, w_m$, $1/\sigma_1^2, \cdots, 1/\sigma_m^2$ are independent and identically distributed as gamma with shape and rate each $1/2$, and, conditional on $\sigma_1, \cdots, \sigma_m$, $\mu_1, \cdots, \mu_m$ are independent with μ_j distributed $N\left(0, \sigma_j^2\right)$, $j = 1, \cdots, m$.

Generating densities in this way provides a variety of different distributional types, including ones that are skewed, lepto- or platykurtic, and/or multimodal. Furthermore, functionals of interest, such as moments and norms are readily calculated for normal mixtures, as pointed out by [16].

Ten thousand densities were generated independently as described above, and then one data set of size $n = 200$ was generated from each of these densities. A P-value for the Shapiro-Wilk test was determined using the R function shapiro. test. The critical value of a size 0.05 test based on $\hat{\Lambda}$ was approximated by generating 10,000 values from a standard normal density.

Since we are interested in power, a 0 - 1 loss function is used, in which case the Bayes risk of each test is estimated by the proportion of rejections among all ten thousand replications. These two proportions were 0.8294 and 0.8003 for the Shapiro-Wilk and kernel-based tests, respectively. A 95% confidence interval for the difference in Bayes risks is $(0.0245, 0.0334)$.

To gain more insight into the difference between the tests, we considered the relationship of rejection probabilities to various characteristics of the selected densities. Letting μ and σ denote the mean and standard deviation of a selected density f, we computed the skewness coefficient $E(X-\mu)^3/\sigma^3$, the excess kurtosis $E(X-\mu)^4/\sigma^4 - 3$ and the Kullback-Leibler divergence

$$KL(f) = -\ln\sigma - \int_{-\infty}^{\infty} \phi(z) \ln f(\sigma z + \mu) \mathrm{d}z \tag{24}$$

where ϕ is the standard normal density. The quantity $KL(f)$ measures the discrepancy between ϕ and a standardized version of f. It represents the consistency parameter for the kernel-based test, and hence a strong dependence is expected between $KL(f)$ and the power of that test. This is illustrated in **Figure 6**, where 0, 1 data (representing do not reject and reject H_0, respectively) have been jittered to make the results clearer.

The same plot as in **Figure 6** for the Shapiro-Wilk test is very similar to **Figure 6**, but a more detailed analysis reveals a key difference between the tests. It turns out that the slight power advantage of Shapiro-Wilk is explained almost entirely by cases where $KL(f)$ is relatively small. This difference between the two tests is in part confirmed by **Figure 7**. For the Shapiro-Wilk and kernel-based tests, respectively, let $\pi_L(x; SW)$ and $\pi_L(x; K)$ be the probability of rejecting H_0 given that $\ln(KL(f)) \le x$. An estimate of $\pi_L(x; SW) - \pi_L(x; K)$ is shown in **Figure 7**. In only 9 of the 4694 cases where $\ln(KL(f))$ exceeded 0.5 did the two tests reach different conclusions. This fact and **Figure 7** strongly

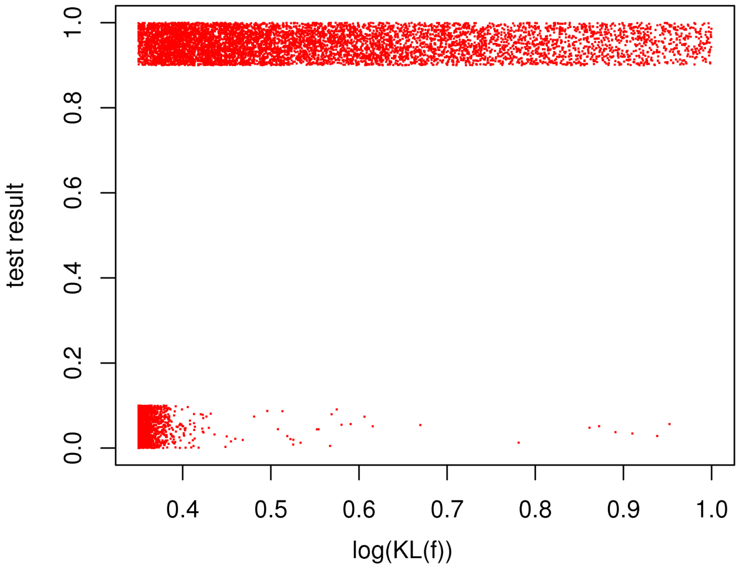

Figure 6. *Results for kernel-based test as a function of Kullback-Leibler divergence.* The data points have been jittered. Those between 0.9 and 1 on the vertical scale represent rejections of H_0 and those between 0 and 0.1 are failures to reject.

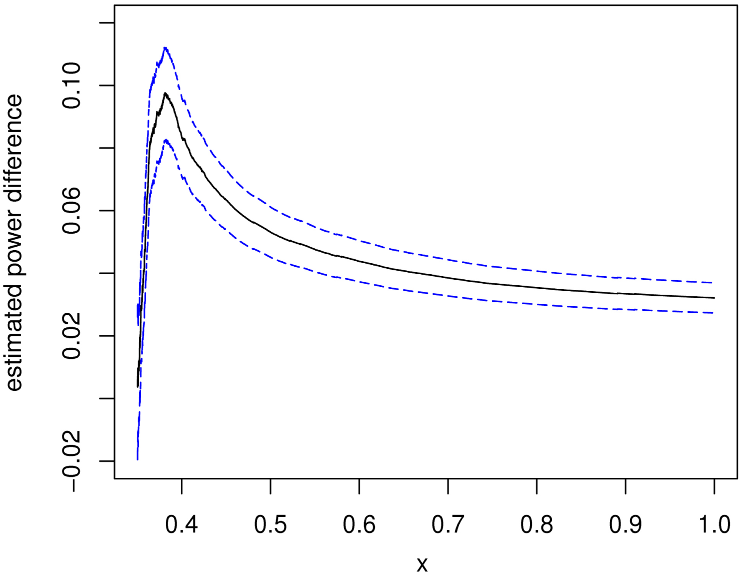

Figure 7. *Estimated difference in powers for Shapiro-Wilk and kernel-based tests given that* $\ln(KL(f)) < x$. The black (solid) line is the difference between the two estimates and the blue (dashed) lines are pointwise confidence limits.

suggest that the only substantial difference in the power of the two tests occurs at small values of $\ln(KL(f))$.

Finally, we present heat plots showing how the power of the two tests are related to skewness and kurtosis. Define the transformation t by $t(x) = \text{sgn}(x)|x|^{1/5}$, for any real x. Letting s and k denote values of the skewness and kurtosis coefficents defined previously, define independent variables $t(s)$ and $t(k)$. Also define y_{1i} to be 1 if the Shapiro-Wilk test rejected H_0 for replication i of the simulation and 0 otherwise. The variables y_{2i}, $i = 1,\cdots,10,000$ are defined similarly for the kernel-based test. We then computed Nadaraya-Watson kernel estimates from the data sets $(t(s_i),t(k_i),y_{1i})$, $i = 1,\cdots,10,000$, and $(t(s_i),t(k_i),y_{2i})$, $i = 1,\cdots,10,000$. These estimates are shown in Figure 8 and Figure 9 in the form of heat plots. The plots are very similar, but there is a subtle difference. The blue/green region is slightly shorter in the direction of skewness for the Shapiro-Wilk test than it is for the kernel-based test. This indicates a slight superiority of the former test in detecting skewness. It is arguable that this sort of subtle effect is more easily detected with BayesSim than with traditional simulation methodology.

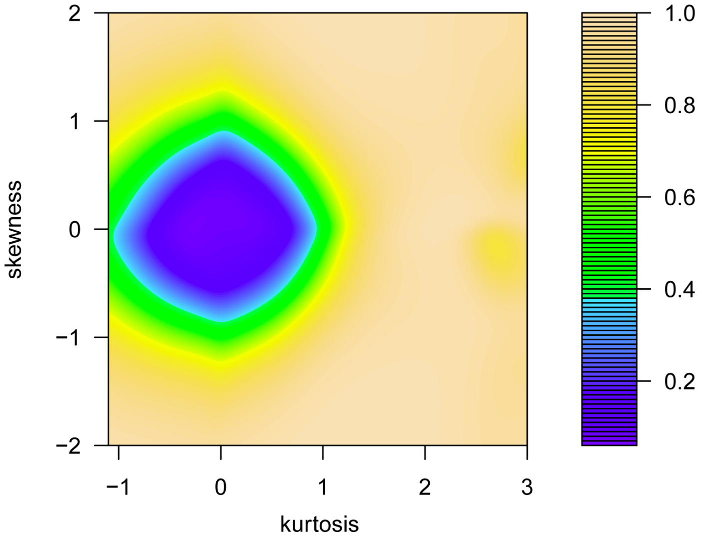

Figure 8. Power as a function of transformed skewness and kurtosis for the Shapiro-Wilk test.

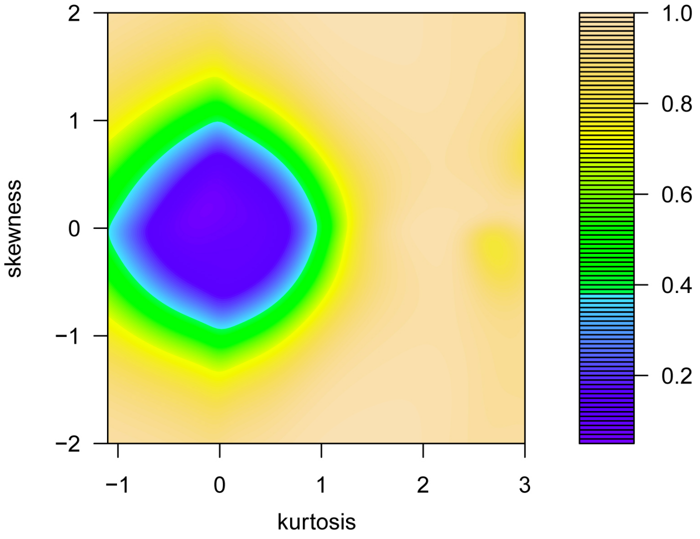

Figure 9. Power as a function of transformed skewness and kurtosis for the kernel-based test.

6. Discussion

An overlooked method of simulation called BayesSim has been propounded. The method provides the possibility of learning more about the statistical methodology being considered while at the same time generating no more (or even less) data than in the conventional method. Does the author think that BayesSim should supplant the conventional approach? Of course not. There are undoubtedly many situations where the latter method will be preferable. However, in the author's opinion there are equally many cases where BayesSim can provide valuable information that is not attainable by generating data from just a few models. Furthermore, conclusions that might be reached with the conventional approach are often much more evident and compelling when data from thousands of different models are considered.

Some will argue that the increasingly complicated methodology of modern statistics would make it difficult to choose satisfactory priors in BayesSim. That may be true but the process of choosing a prior in BayesSim cannot be more difficult than choosing a prior for a complex model in a Bayesian data analysis. And the difficulty of choosing priors has certainly not prevented an explosion in the use of Bayesian statistics.

Another reason the author finds BayesSim appealing is that it provides a rich source of data to which statistical researchers may apply the sophisticated methods that they are constantly developing. Kernel methods, nonparametric additive modeling, projection pursuit and support vector machines are but a few of the methods that could be used to investigate how, for example, estimator performance is related to parameters. In short, I believe that statisticians will find that BayesSim can make their simulation studies more informative, and more interesting.

Acknowledgements

This research was supported by National Science Foundation Grant DMS-0604801.

References

[1] Savchuk, O., Hart, J.D. and Sheather, S.J. (2010) Indirect Cross-Validation for Density Estimation. *Journal of the American Statistical Association*, **105**, 415-423. https://doi.org/10.1198/jasa.2010.tm08532

[2] Hart, J.D. (2009) Frequentist-Bayes Lack-of-Fit Tests Based on Laplace Approximations. *Journal of Statistical Theory and Practice*, **3**, 681-704. https://doi.org/10.1080/15598608.2009.10411954

[3] Rubin, D.B. and Schenker, N. (1986) Efficiently Simulating the Coverage Properties of Interval Estimates. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, **35**, 159-167. https://doi.org/10.2307/2347266

[4] Andradóttir, S. and Bier, V.M. (2000) Applying Bayesian Ideas in Simulation. *Simulation Practice and Theory*, **8**, 253-280. https://doi.org/10.1016/S0928-4869(00)00025-2

[5] Hart, J.D. and Yi, S. (1998) One-Sided Cross-Validation. *Journal of the American*

Statistical Association, **93**, 620-631. https://doi.org/10.1080/01621459.1998.10473715

[6] Lee, Y., Lin, Y. and Wahba, G. (2004) Multicategory Support Vector Machines, Theory, and Application to the Classification of Microarray Data and Satellite Radiance Data. *Journal of the American Statistical Association*, **99**, 67-81. https://doi.org/10.1198/016214504000000098

[7] Buja, A., Hastie, T. and Tibshirani, R. (1989) Linear Smoothers and Additive Models. *Annals of Statistics*, **17**, 453-555. https://doi.org/10.1214/aos/1176347115

[8] Friedman, J.H. and Stuetzle, W. (1981) Projection Pursuit Regression. *Journal of the American Statistical Association*, **76**, 817-823. https://doi.org/10.1080/01621459.1981.10477729

[9] Pearson, E. (1925) Bayes Theorem, Examined in the Light of Experimental Sampling. *Biometrika*, **17**, 388-442. https://doi.org/10.1093/biomet/17.3-4.388

[10] Hill, T.P. (1995) A Statistical Derivation of the Significant-Digit Law. *Statistical Science*, **10**, 354-363.

[11] Berger, J.O. (1980) Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York. https://doi.org/10.1007/978-1-4757-1727-3

[12] Ferguson, T.S. (1973) A Bayesian Analysis of Some Nonparametric Problems. *Annals of Statistics*, **1**, 209-230. https://doi.org/10.1214/aos/1176342360

[13] Hallin, M. and Ley, C. (2012) Skew-Symmetric Distributions and Fisher Information—A Tale of Two Densities. *Bernoulli*, **18**, 747-763. https://doi.org/10.3150/12-BEJ346

[14] Shapiro, S.S. and Wilk, M.B. (1965) An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, **52**, 591-611. https://doi.org/10.1093/biomet/52.3-4.591

[15] Hall, P. (1987) On Kullback-Leibler Loss and Density Estimation. *Annals of Statistics*, **15**, 1491-1519. https://doi.org/10.1214/aos/1176350606

[16] Marron, J.S. and Wand, M.P. (1992) Exact Mean Integrated Squared Error. *Annals of Statistics*, **20**, 712-736. https://doi.org/10.1214/aos/1176348653