Annual Review of Sociology

Bayesian Statistics in Sociology: Past, Present, and Future

Scott M. Lynch and Bryce Bartlett

Department of Sociology, Duke University, Durham, North Carolina 27708, USA;
email: scott.lynch@duke.edu

Annu. Rev. Sociol. 2019. 45:25.1–25.22

The *Annual Review of Sociology* is online at soc.annualreviews.org

https://doi.org/10.1146/annurev-soc-073018-022457

Keywords

Bayesian statistics, Markov chain Monte Carlo, missing data, multiple imputation, model selection, crisis in science

Abstract

Although Bayes' theorem has been around for more than 250 years, widespread application of the Bayesian approach only began in statistics in 1990. By 2000, Bayesian statistics had made considerable headway into social science, but even now its direct use is rare in articles in top sociology journals, perhaps because of a lack of knowledge about the topic. In this review, we provide an overview of the key ideas and terminology of Bayesian statistics, and we discuss articles in the top journals that have used or developed Bayesian methods over the last decade. In this process, we elucidate some of the advantages of the Bayesian approach. We highlight that many sociologists are, in fact, using Bayesian methods, even if they do not realize it, because techniques deployed by popular software packages often involve Bayesian logic and/or computation. Finally, we conclude by briefly discussing the future of Bayesian statistics in sociology.

INTRODUCTION

Although Bayes' paper on probability was published posthumously in 1763 (Bayes 1763), and his theorem was used extensively in nonacademic settings for two centuries, including in military intelligence (McGrayne 2011), Bayesian statistics did not emerge as a significant subfield of statistics until the 1950s, and its status was marginal in the discipline until 1990. Gelfand & Smith (1990) published a seminal paper showing how sampling methods can be used to facilitate Bayesian analyses. The methods described in that paper, coupled with tremendous increases in computing power, led to an explosion in the use of the Bayesian approach in statistics.

In the following decade, Bayesian statistics made headway into social science, but even now it has seen far less direct use in sociology than in other social sciences. A key reason for the dearth of Bayesian analyses in sociology may be that sociologists remain untrained in these methods and are therefore unfamiliar with their application and advantages. Thus, in the next sections, we review the basics of Bayesian statistics, with a focus on introducing its key concepts and terminology as well as highlighting criticisms against, and advantages of, the Bayesian approach. These sections are necessary to understand why the Bayesian approach has become increasingly popular.

Following this review, we discuss the rise of Bayesian statistics in social science in general and the development and application of Bayesian methods in the top journals in sociology over the past decade. In the concluding section, we discuss why the Bayesian approach may become more prominent in sociology in the coming years.

WHAT IS BAYESIAN STATISTICS?

Bayes' theorem shows how to reverse a conditional probability via a trivial application of basic rules of probability. Consider the following example. Suppose there are two jars on a countertop, each containing four marbles. In jar 1, one marble is black and the other three are white; in jar 2, two marbles are black and two are white. I blindfold you and have you draw a single marble from one of the jars, and the marble is black. What is the probability you selected from jar 2? A simple way to answer this question is to realize that there are three black marbles, two of which were in jar 2. So, the odds that the marble came from jar 2 are 2:1, meaning the probability you selected from jar 2 is 2/3. Bayes' theorem provides the formal recipe for this computation and is easily derived:

$$\mathrm{P}(B, A) = \mathrm{P}(A, B) \quad 1.$$

$$\mathrm{P}(B|A)p(A) = \mathrm{P}(A|B)\mathrm{P}(B) \quad 2.$$

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A|B)\mathrm{P}(B)}{\mathrm{P}(A)}. \quad 3.$$

Equation 1 simply states the fact that events A and B are permutable in their joint probability. Equation 2 substitutes in the familiar conditional probability rule for nonindependent joint events. Equation 3 is Bayes' theorem. The left-hand side is called the posterior probability for B, given A. In our example, it is the probability you selected from jar 2, after learning that the marble you selected was black. $\mathrm{P}(A|B)$ is the probability of obtaining a black marble, conditional on drawing from jar 2. In Bayesian lingo, $\mathrm{P}(B)$ is called the prior probability for B: It is the probability that you would presumably assign to drawing from jar 2 prior to having any knowledge of the color of the marble you selected (1/2). The denominator of the equation is the marginal probability for A,

here, the total probability that you would select a black marble. Thus,

$$\text{P(jar 2 | black marble)} = \frac{\text{P(black marble | jar 2)P(jar 2)}}{\text{P(black marble | jar 2)P(jar 2)} + \text{P(black marble | jar 1)P(jar 1)}} \quad 4.$$

$$= \frac{(2/4)(1/2)}{(2/4)(1/2) + (1/4)(1/2)} \quad 5.$$

$$= \frac{2}{3}. \quad 6.$$

Although our prior probability for selecting from jar 2 was 1/2, our posterior probability is much greater and stems from the knowledge that we obtained a black marble.

Although this example is trivial, Bayes' theorem is widely applicable and often produces counterintuitive results. Consider a more realistic (and classic) example. Suppose you test positive for a disease (+). What is the probability you actually have the disease (*D*), given this information? The posterior probability of interest is P(*D*|+), and the computation is

$$\text{P}(D|+) = \frac{\text{P}(+|D)\text{P}(D)}{\text{P}(+|D)\text{P}(D) + \text{P}(+|ND)\text{P}(ND)}. \quad 7.$$

The prior probability for having the disease, P(*D*), could be obtained, perhaps, from published prevalence rates. We may be able to obtain the probability of testing positive, given one has the disease (called the sensitivity of the test), from the maker of the test. The denominator provides the marginal probability for testing positive under the two possible states of the world—you either have the disease (*D*) or you do not (*ND*). The conditional probability on the right hand side of the denominator is the probability of a false positive [P(+|*ND*) = 1 − P(−|*ND*), where the true negative rate P(−|*ND*) is called the specificity of the test], and it may also be available from the maker of the test. To illustrate the implications of Bayes' theorem, suppose the sensitivity and specificity of the test are 0.9 and 0.8, respectively, and the prevalence of the disease in the population is 0.0001. Then, the posterior probability is

$$\text{P}(D|+) = \frac{(0.9)(0.0001)}{(0.9)(0.0001) + (1 - 0.8)(0.9999)} = 0.0004. \quad 8.$$

Thus, even with a positive test, the probability you have the disease is extremely small. To be sure, it is four times larger than the prior probability for having the disease, so the positive test is informative, but it is by no means convincing proof of having the disease.

In this case, the posterior probability is low because the test is poor and the disease prevalence is low. This result is why some tests, such as the prostate-specific antigen test for prostate cancer, are not recommended for men under age 50 who have no strong risk factors (see **https://www.cancer.gov/types/prostate/psa-fact-sheet**). With advancing age and other risk factors, the disease prevalence is higher, so that the posterior probability of disease given a positive test may be high enough to justify further, more invasive testing. Alternatively, one could develop a better test.

Although computations of posterior probabilities such as this one are enlightening and have changed our views of medical testing in the face of imperfect tests, we rarely work with point probabilities in sociological research. However, Bayes' theorem extends to probability distributions. In the example above, we treated the prior probability for having the disease as exact [$p \equiv \text{P}(D) = 0.0001$], but in reality we most likely do not know the exact value of p. Instead, it is surely only an estimate, even if the prevalence rate were derived from population data. Thus, we may opt to replace p with a prior distribution for it. That is, we may treat it as a random

variable that follows some distribution that represents our uncertainty regarding its exact value. An appropriate distribution for quantities that fall in the [0, 1] interval, such as probabilities, is the Beta(α, β) distribution, which has the following density function:

$$f(p|\alpha, \beta) = \left(\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) p^{\alpha-1}(1-p)^{\beta-1} \quad 9.$$

$$\propto p^{\alpha-1}(1-p)^{\beta-1}. \quad 10.$$

In this density, $\Gamma(.)$ is the gamma function, which is the continuous analog to the factorial function. The collection of terms in parentheses constitutes a normalizing constant: This collection merely scales the density function for p to make it a proper density function (i.e., so that it integrates to 1 and meets the requirement that the total of the probabilities of all events in the sample space is 1). Here, these gamma functions do not change the relative heights of the curve at different points on the x-axis, and so they are not informative about the relative likelihood of different values of p. The second line (Equation 10) drops this normalizing constant and shows that the beta density is proportional to ($\propto$) what is left. Proportionality is used quite often in Bayesian statistics to eliminate irrelevant constants.

In the kernel of the density that remains above after dropping the normalizing constant, p is the random variable, which has support—or can range—over the [0, 1] interval, and α and β are parameters that represent prior successes and failures. The parameters α and β can take any positive value, making the distribution extremely flexible in shape. The mean of the beta distribution is $\alpha/(\alpha + \beta)$, and the variance is $\alpha\beta/\left[(\alpha + \beta)^2(\alpha + \beta + 1)\right]$. α and β can be set to values that allow this prior distribution for p to be centered over the estimate (i.e., assumed population value) of 0.0001 and to be an appropriate width to represent the extent of our knowledge (or uncertainty) about p. For example, if $\alpha = \beta = 1$, the beta distribution is flat over the [0, 1] interval, indicating that all values of p are equally likely—a uniform distribution. This type of prior distribution is called a noninformative prior, because it indicates that we are ignorant about the parameter a priori. In contrast, if we are confident that $p = 0.0001$, we might set $\alpha = 10$ and $\beta = 99,990$ to reduce the variance of the prior and concentrate its mass over the mean of 0.0001.

In addition to posing a probability distribution for P(D), we may replace P(+|D) with a probability distribution as well. In our example, our only information was a single positive test for an individual, but the individual could take the test repeatedly. In each replication, the posterior probability from the previous test could be treated as the prior for the current test, and Bayes' theorem could be applied sequentially. The posterior probabilities arising from a sequence of 10 positive tests would be [0.0004, 0.002, 0.009, 0.039, 0.156, 0.454, 0.789, 0.944, 0.987, 0.997]. Thus, after 10 positive tests in a row, we would be very confident that the individual has the disease. This process of repeated application and replacement of the prior with the posterior from previous tests is called Bayesian updating, and it arguably mirrors the way humans learn about the world: We construct priors regarding how things work, and we continually update them with new information as we encounter it (see Andrew & Hauser 2011, Fallesen & Breen 2016 for recent discussions of Bayes and learning theory in sociology).

In sociological analyses, we may have thousands of tests (i.e., a sample), some of which are positive and some of which are negative, but we need not conduct Bayesian updating sequentially. Instead, we could represent all our data simultaneously with a joint probability distribution. To simplify this example, let us assume that the test is infallible—our positive (negative) cases are truly positive (negative)—and our goal is to update our knowledge of the prevalence rate, p, given a new sample of data consisting of x positive tests out of n total tests. The binomial distribution is

a distribution for counts in which each individual test can be considered a trial with an identical success probability p. The density/mass function for the binomial distribution is

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}. \quad 11.$$

Here, p is the probability of disease, x is the number of cases (positives), and n is the sample size. In classical statistical terminology, we would call this joint distribution for the data our likelihood function, and our analysis would focus exclusively on it [i.e., via maximum likelihood (ML) estimation; Eliason 1993].

With a prior distribution for p and a likelihood function for x, our posterior becomes a probability distribution as well:

$$f(p|x,n) = \frac{f(x|p)f(p)}{\int_S f(x|p)f(p)\mathrm{d}p} \quad 12.$$

$$\propto f(x|p)f(p) \quad 13.$$

$$\propto \left[p^x(1-p)^{n-x}\right]\left[p^{\alpha-1}(1-p)^{\beta-1}\right] \quad 14.$$

$$\propto p^{x+\alpha-1}(1-p)^{n-x+\beta-1} \quad 15.$$

$$\propto \mathrm{Beta}(\alpha^* = x+\alpha,\ \beta^* = n-x+\beta). \quad 16.$$

In Equation 12, the denominator has become an integral of the data over the entire sample space (S) for p, rather than a sum of the data over two states of the world (disease or no disease) as in Equation 7. This integral is sometimes called the marginal likelihood and is important in the computation of Bayes factors and other quantities (see subsequent sections). However, it is a normalizing constant here and can be dropped and replaced by the proportionality symbol. Furthermore, note that in Equation 14, the likelihood portion of the posterior no longer contains the leading combinatorial shown in Equation 11. This is because the posterior is a distribution for p, and not x, so the leading combinatorial is also a normalizing constant once the data are observed (and hence fixed). Given the ability to drop normalizing constants without affecting the relative likelihoods of different values of the random variable of interest, it is common in Bayesian statistics to see Bayes' theorem reduced to Posterior $\propto$ Likelihood $\times$ Prior.

Equation 16 shows that the posterior distribution for p is proportional to a beta distribution, with new parameters α^* and β^*. This highlights another term in the Bayesian lexicon: conjugacy. When the prior and likelihood are such that the posterior takes the same distributional form as the prior, the distributions for the data and parameter are said to be conjugate. Equation 16 also shows that the posterior is a compromise between the prior (what we knew before observing the new data) and the likelihood (the new data). Importantly, this new beta distribution is narrower than our prior (as can be seen if we compare variance calculations of the prior and posterior), reflecting the reduction in uncertainty about p with the addition of new information.

Markov Chain Monte Carlo Methods

A fundamental part of Bayesian analysis is summarizing the posterior distribution once it is obtained. In the example, the posterior distribution is a beta distribution, and so we can summarize what we know about the parameter using the posterior mean and posterior variance for p, which can be computed using the functions for the mean and variance shown earlier. Interval estimates for p can be constructed either using the posterior variance and a normality assumption, much as is done to produce ML-based confidence intervals, or directly, using quantile functions for

the beta distribution found in most software packages. For example, for a central 95% interval, we would find the 2.5th and 97.5th percentiles of the posterior to obtain the bounds using the `invibeta(a,b,prob)` function in Stata or the `qbeta(prob,a,b)` function in `R`, where `a` and `b` are the parameters of the posterior and `prob` is the cumulative probability. However, it is not always so easy to summarize posterior distributions. Indeed, difficulty with summarization has been a key factor limiting the use of Bayesian statistics in social science.

Most summary measures that are of interest, such as the mean, median, variance, and quantiles, involve integral calculus, and it is not always possible to perform integration analytically. Using conjugate priors can help resolve the difficulty because integrals for known distributions are often tractable. However, conjugate priors are not always reasonable, and nonconjugate priors often yield posterior distributions that are of unknown forms. Furthermore, even with conjugate priors, high-dimensional multivariate models—which are common in sociology—often yield unwieldy posterior distributions. Thus, Bayesians historically used normality assumptions and numerical methods to produce approximate summaries of posterior distributions, making conclusions subject to criticism for lack of accuracy. Moreover, such methods are generally not familiar to sociologists.

Everything changed, beginning in 1990, with the development of Markov chain Monte Carlo (MCMC) methods (Gelfand & Smith 1990). MCMC methods made conducting Bayesian analyses much easier, leading to an explosion in the use of Bayesian statistics throughout science. A full discussion of MCMC methods is beyond the scope of this article. However, a brief summary of them is warranted to facilitate reading subsequent sections and to provide insight into how these methods have helped increase the popularity of Bayesian statistics.

The basic logic underlying the development and use of MCMC methods is that it is often easy to sample from distributions (e.g., posterior distributions), and samples from a distribution can be used to approximate integrals of it. For example, the mean, μ, of a distribution, $f(x)$, is defined as $\mu = \int_{-\infty}^{\infty} x\, f(x)\mathrm{d}x$. Although computing this integral may not be straightforward for a given distribution, computing the mean of a sample of size n from that distribution is straightforward: $\bar{x} = \sum x/n$. We can make the approximation of μ by $\bar{x}$ arbitrarily precise by simply drawing a larger sample, as indicated by the central limit theorem.

MCMC methods are recipes for sampling from distributions. The methods use simulation (Monte Carlo) to produce Markov chains of samples that have the posterior distribution as their stationary distribution (Gamerman & Lopes 2006). The most common such method is the Gibbs sampler, which produces samples from multivariate distributions by sequentially sampling from conditional distributions for subsets of model parameters. For example, in a linear model, where $y = X\beta + e$, with y an $n \times 1$ column vector, X an $n \times k$ design matrix, and e a residual assumed to be normally distributed with a mean of 0 and a variance of σ_e^2, the conditional posterior distributions for the model parameters (under noninformative priors for β and σ_e^2) are

$$\beta \mid \sigma_e^2, X, y \sim \mathrm{MVN}\left((X^TX)^{-1}(X^Ty),\ \sigma_e^2(X^TX)^{-1}\right) \quad 17.$$

$$\sigma_e^2 \mid \beta, X, y \sim \mathrm{IG}\left(\frac{n}{2},\ \frac{e^Te}{2}\right). \quad 18.$$

That is, if σ_e^2 is known (treated as fixed), the conditional posterior distribution for β is multivariate normal with a mean vector equal to the ordinary least squares (OLS) solution and a covariance matrix equal to a function of X and σ_e^2, also identical to the OLS/ML solution. If β is known (treated as fixed), then each residual, e, can be computed (i.e., $e = y - X\beta$), and the conditional posterior distribution for σ_e^2 is inverse gamma with parameters that are a function of the sample size and the sum of the squared errors (Gelman et al. 2013, Lynch 2007). Sampling from multivariate normal

and inverse gamma distributions is easy, thus reducing a more complex problem to a simpler one. Gibbs sampling, then, involves iteratively sampling from these conditional distributions until a large enough sample of β and σ_e^2 is obtained that their distributions can be summarized via basic descriptive calculations. The only trick to Gibbs sampling, in many cases, is to learn to discard normalizing constants and recognize kernels of distributions.

Not all MCMC sampling methods are as straightforward as the Gibbs sampler, and an important aspect of implementing MCMC methods is evaluating whether the algorithm converged to the right posterior distribution and sampled throughout it (i.e., mixed). If not, the algorithm may need to be altered and rerun before summarizing the results. This process can be tedious and technical and remains an impediment to greater use in social science. Nonetheless, with relatively little training, and with the development of software that implements MCMC sampling, including recent procedures in standard packages such as Stata (StataCorp 2017), the use of these methods is becoming increasingly within reach for the typical quantitative sociologist.

CRITICISMS AND ADVANTAGES OF THE BAYESIAN APPROACH

In addition to the practical difficulty with summarizing posterior distributions prior to the development of MCMC methods, two philosophical issues historically limited the use of the Bayesian approach. First, the use of priors introduces subjectivity into the scientific enterprise. From a classical statistical standpoint and a traditional view of science, subjectivity has no place in science. Yet, priors are subjective by their nature: The researcher chooses them. Because the posterior distribution is a weighted combination of the prior and likelihood, a strong prior can influence the conclusion of an analysis. Thus, a researcher can predetermine an outcome.

Second, in the previous section, we discussed how summaries of a posterior distribution can be made, including the construction of interval estimates. Although such interval estimates may be numerically similar to those obtained via classical methods, they do not have the same interpretation as ones derived classically. From a classical standpoint, parameters are fixed, while data are random, and so it is only appropriate to place probability distributions on data. In contrast, from a Bayesian standpoint, an unknown quantity may be fixed, but our lack of knowledge about it allows us to place probability distributions on it and make probabilistic statements about it. The posterior distribution is a probability distribution that represents our remaining uncertainty about a model parameter (or any unknown quantity) after observing new data. Thus, if we construct a 95% credible interval for p in the example above, we would say, "the parameter p is in the interval $[a, b]$ with probability 0.95." This interpretation is substantially different from that of a classical confidence interval, which is that "this interval either captures the parameter or it does not, but in repeated samples, 95% of such intervals constructed in this fashion will contain the parameter."

This interpretation highlights a fundamental philosophical difference between the Bayesian and classical approaches. Whereas classical Neyman/Pearson/Fisher statistical hypothesis/significance testing follows a deductive perspective that says that theories can only be falsified, the Bayesian approach is inherently inductive. Under the classical approach, we compute the probability of observing our data under the assumption that the (null) hypothesis is true. If that probability (the p-value) is small enough, we reject the (null) hypothesis and claim support for our (alternative) hypothesis of real interest. Although Bayesians generally do not test point hypotheses such as null hypotheses, instead focusing on interval estimates of parameters, the Bayesian approach does allow for probabilistic assessments of hypotheses. For example, if our hypothesis is that a parameter falls in some range, we may compute the probability that the hypothesis is true by computing the area contained in the posterior distribution in that range. This inversion of Popperian, deductive logic

is a key reason why the Bayesian approach was rejected in most scientific fields, and marginalized in statistics, until fairly recently.

These two reasons for rejecting Bayesian statistics have become less damning in recent years for at least five reasons. First, almost all Bayesian analyses in social science have involved the use of noninformative priors, yielding results nearly identical to those obtained via classical analyses, except for their interpretation. In fact, many introductory texts provide derivations and results to show the similarity between Bayesian and classical methods (e.g., Gelman et al. 2013, Lynch 2007). Some software documentation, such as that for `MPlus` for structural equation modeling, actually suggests using Bayesian methods as a more efficient method for obtaining approximate ML estimates in models with high numbers of latent variables (Asparouhov & Muthen 2012).

Second, although noninformative priors often contain some information about model parameters, sociological data sets tend to be large and overwhelm most priors, including informative ones. Thus, the choice of prior often has little effect on substantive conclusions.

Third, in recent years considerable attention has been devoted to prior elicitation, that is, the development of priors that incorporate existing knowledge into new analyses to formalize accumulation of knowledge (see Billari et al. 2014, Rendall et al. 2009 for recent examples). In most fields, we do not enter into analyses in ignorance, and yet we test the same null hypotheses again and again, and knowledge accumulates only informally via the literature one chooses to include in one's literature review. An arguably more coherent strategy for advancing science is to incorporate prior knowledge explicitly into our analyses so that knowledge can accumulate formally.

Fourth, priors can be used strategically to identify parameters that otherwise may not be identified. For example, in generalized linear models, some combination of means, intercepts/thresholds, and variances must be fixed to identify the coefficients (Johnson & Albert 1999). In a Bayesian analysis, we may place a strong prior on some parameters to identify the model while allowing flexibility to compensate for uncertainty that is not allowed if a parameter is treated as fixed (see An 2010, Cheng et al. 2008 for recent sociological examples).

Finally, there is growing recognition of a crisis emerging in science, and it is in no small part due to overreliance on *p*-values (Papineau 2018, Ziliak & McCloskey 2008). Scientific findings, especially in social science, are often not replicable, and contradictory findings are common (Freese & Peterson 2017). Yet this should be expected, given the classical statistical approach and the tendency in social science not to publish null findings (i.e., publication bias; see Lee & Conley 2016 for a recent sociological example). By design, true null hypotheses are rejected 5% of the time. If hundreds of scholars test the same true null hypothesis, statistically significant results will emerge and will be published, but they will be false (Ioannidis 2005). A solution to this problem is for science to shift toward the Bayesian paradigm, which is not focused on testing null hypotheses.

At the same time that these criticisms of the Bayesian approach have diminished, the advantages of Bayesian statistics have become increasingly recognized. As already stated, the ability to incorporate prior knowledge and the ability to identify otherwise unidentifiable parameters are important advantages of the Bayesian approach over the classical approach. However, there are numerous other advantages that are relevant for sociological research, including advantages for model evaluation and selection, for estimation of tertiary parameters and uncertainty therein that is not easily directly estimated otherwise, for handling missing data, for hierarchical and related modeling, and for applications in which taking a classical approach is difficult or impossible. In each of these cases, the Bayesian approach tends to be superior to classical approaches in terms of ease of implementation and adjustment to handle data idiosyncrasies. In the past decade, each

of these advantages has been exploited in sociological literature, as we discuss in greater detail in subsequent sections.

THE RISE OF BAYESIAN STATISTICS IN SOCIAL SCIENCE

With the development of MCMC methods and concurrent increases in computing power, sampling from posterior distributions has often become fast and easy. Although the revolution in statistics began in 1990, it took another 15 years before Bayesian statistics began to make significant inroads into the social sciences, as **Figure 1** illustrates using publication counts for articles involving Bayesian statistics in statistics, probability, and key social science disciplines. **Figure 1*a*** shows that over the past 40 years, not surprisingly, the most papers involving Bayesian statistics were published in journals focused on statistics and probability (17,595 papers). 4,600 papers were

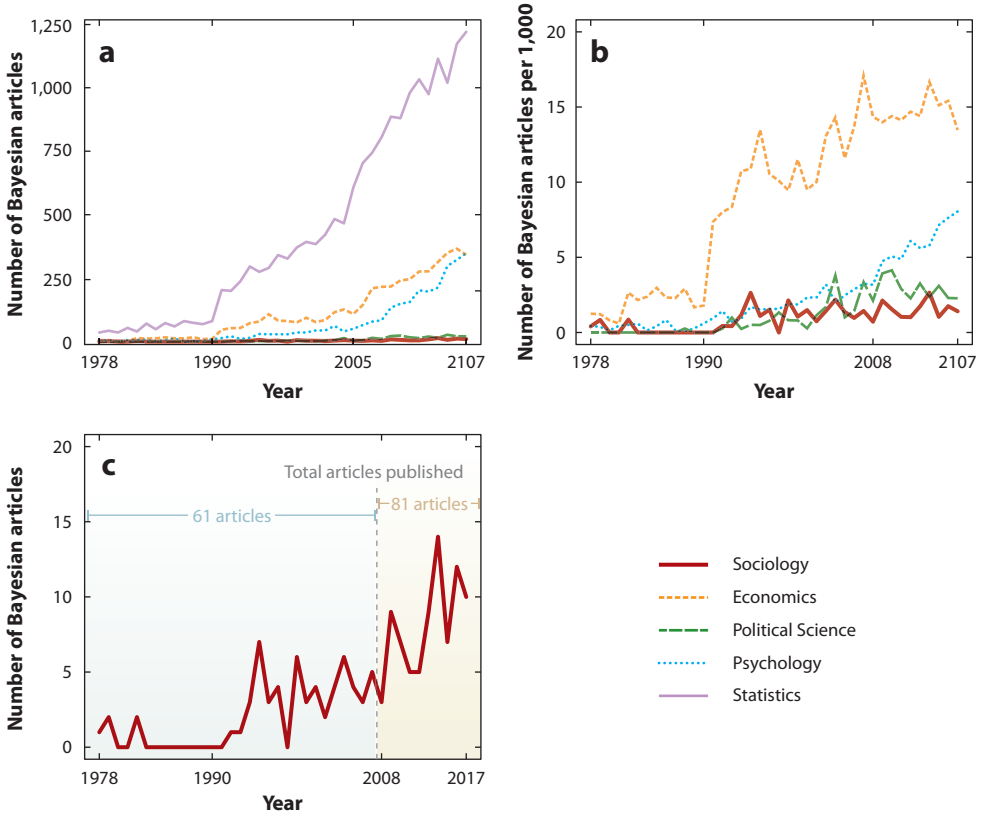

Figure 1

Publication counts and rates for articles involving Bayesian statistics in statistics and probability and key social science disciplines from 1978–2017. (*a*) Number of articles published involving Bayesian statistics in statistics and key social science disciplines over the past 40 years. (*b*) Number of Bayesian articles per 1,000 articles published. (*c*) Number of Bayesian articles in sociology by year. Data were obtained from Web of Science searches with the topic "TI=bayes*" and Web of Science category set to "WC=Statistics and Probability," "WC=Sociology," "WC=Economics," "WC=Political Science," "WC=Psychology." Denominators for rates were obtained by eliminating "TI=bayes*" to obtain counts of all published papers in each disciplinary category.

published in economics; 2,808 were published in psychology; 267 were published in political science; and 142 were published in sociology. As **Figure 1*a*** shows, the number of papers published in statistics grew rapidly after 1990, while the social sciences did not see much increase until around 2005. Even then, neither sociology nor political science seemed to see much of an increase in publication counts; the increase was concentrated in economics and psychology. Although no database search is perfect, the patterns shown in the figure follow the same general shape as those shown in a recent review of Bayesian statistics in psychology for both psychology and other disciplines (van de Schoot et al. 2017).

An increase in Bayesian statistics articles may simply result from an increase in the total number of publications. Our Web of Science search revealed that all of the disciplines shown in **Figure 1** experienced substantial increases in total publication counts over the period. For example, statistics publications increased from 1,425 in 1978 to 11,226 in 2017! **Figure 1*b*** shows the number of Bayesian statistics articles per 1,000 articles published, by year, for the social sciences. This panel excludes the rates for the category of statistics and probability because its rates are well above those for the social sciences. For example, in 1978, roughly 25 of every 1,000 papers published in statistics were Bayesian. That rate remained relatively stable until 1990, when it jumped to about 66 papers per 1,000 and increased steadily to about 110 per 1,000 by 2005. The rate has remained relatively stable since then at about 10–12% of all statistics publications.

The rate of Bayesian publications in the social sciences is much lower. Prior to 1990, about 3 papers per 1,000 published in economics were Bayesian, but the rate was close to 0 for political science, psychology, and sociology. In 1990, the publication rate increased substantially in economics, surpassing 10 per 1,000 by the late 1990s. The rate stabilized by 2008 at about 15 per 1,000. In the other disciplines, the rate increased only slightly. The publication rate in sociology hovers around 1–2 papers per 1,000, just below the rate for political science. Interestingly, the publication rate in psychology grew steadily beginning around 2005 so that it was nearly 9 per 1,000 by 2017.

Despite the small magnitude of the Bayesian publication rate in sociology, the annual count of publications in sociology is growing and is now well above 0. **Figure 1*c*** shows these counts for the past 40 years. In many years prior to 1990, there were no Bayesian publications in sociology. As with the other disciplines, the count of Bayesian publications began to rise after 1990 and remained relatively stable at about four papers per year from the mid-1990s until around 2008. Since 2008, the count has increased so that in the most recent years the publication count has averaged about 10 per year. This recent increase since 2008 may be due to the publication of at least three books on Bayesian statistics that are aimed specifically at social scientists (Gill 2002, 2007, 2014; Jackman 2009; Lynch 2007), as well as a book on hierarchical modeling written at a very readable level for social scientists (Gelman & Hill 2007).

KEY AREAS IN BAYESIAN STATISTICS IN SOCIOLOGY: 2008–2017

To obtain a more refined perspective on the use of Bayesian statistics in sociology, we conducted a Web of Science search of papers published in the field's top journals over the past decade (2008–2017). These journals include all American Sociological Association journals that publish original work [*American Sociological Review* (*ASR*), *Sociological Methodology* (*SM*), *Social Psychology Quarterly* (*SPQ*), *Sociological Theory* (*ST*), *Journal of Health and Social Behavior* (*JHSB*), and *Sociology of Education* (*SE*)], as well as key outlets for demographers and methodologists in sociology [*Demography* (*DEM*) and *Sociological Methods & Research* (*SMR*)], and two journals that are often viewed as part of the big three [*American Journal of Sociology* (*AJS*) and *Social Forces* (*SF*)].

We found 35 papers, although we acknowledge that that may be an undercount, given the limitations of database searches. *ST* and *JHSB* published no papers involving Bayesian statistics; *AJS*, *ASR*, *SPQ*, and *SE* each published one; and *SF* published two. *DEM*, *SMR*, and *SM* led the field, with 11, 10, and 8 papers, respectively. We were disappointed by the small number of papers: The rate of publication is 3.5 papers per year across 10 journals that, together, publish several hundred papers annually. At best, then, less than 1% of papers published in the discipline's top journals directly involved Bayesian statistics. We extended our search backward, from 2007 to 1990, and found that these journals published a total of 43 papers that *at least* had the word "Bayesian" in their abstracts over that period, for a rate of 2.4 papers per year. Most of those publications were concentrated in the two methodology journals: *SM* published 12 papers involving Bayesian methods and *SMR* published 24. *SPQ*, *SE*, *ST*, and *SF* published none; *DEM* published only one (Assuncao et al. 2005); and *ASR*, *AJS*, and *JHSB* each published two (Berk et al. 1992, Matsueda et al. 2006; Western 1994, 2001; McKinlay et al. 2002; Vanlandingham et al. 1995). Thus, the publication rate has increased in recent years, especially given that the majority of papers from the early period were concentrated in special issues or clusters in the methods journals. *SMR* had three special issues devoted to Bayesian methods, including a special issue introducing Bayesian statistics (Western 1999), a special issue on the Bayesian Information Criterion (BIC) (Winship 1999), and a special issue devoted to model selection (Weakliem 2004); and *SM* had a cluster of papers devoted to Bayesian statistics in 1995 on model selection (Raftery 1995a) with discussion (Gelman & Rubin 1995, Hauser 1995, Raftery 1995b) and significance testing in population data (Berk et al. 1995a) with discussion (Berk et al. 1995b, Bollen 1995, Firebaugh 1995, Rubin 1995). These areas—introduction of Bayesian statistics to the discipline, model and variable selection, and significance testing—remain important, but over the past decade, the use and development of Bayesian approaches have been concentrated in the following areas: (*a*) model/variable selection and tertiary analyses, (*b*) handling of missing data, (*c*) hierarchical and related models, and (*d*) other applications for which the classical approach is ill-suited. We discuss the recent sociological literature in these areas in the following sections.

Model/Variable Selection and Tertiary Analyses

Model and variable selection are key issues in sociology. When models are nested—meaning that one model is a generalization of another—classical methods allow for formal comparison (Bollen 1989). However, there are no formal tests available to compare nonnested models in a classical setting, despite increasing interest in doing so. For example, latent class analysis (LCA) is increasingly common. LCA is a type of finite mixture model in which individuals are assumed to be members of unobserved classes that differ from one another in their response patterns to a collection of variables (Lynch & Taylor 2016). A key question in such analyses is, How many latent classes exist in the population? This is a question of model selection, and models with different numbers of classes are fundamentally nonnested.

Variable selection typically refers to choosing regressors that are important in predicting an outcome. Although variable selection has not been overly common historically in sociology—because we generally seek effects of causes, rather than causes of effects—interest in variable selection has increased in recent decades. Variable selection can be viewed as a form of model selection that often involves nonnested models. For example, consider hypothetical model 1, with predictors $\{A, B, C\}$; model 2, with predictors $\{A, B\}$; and model 3, with predictors $\{B, C\}$. Models 2 and 3 are nested in model 1, but model 3 is not nested in Model 2 (nor vice versa). A comparison of models 2 and 3, then, cannot rely on classical methods.

Computation of Bayes factors facilitates comparison of nonnested models. The posterior odds favoring one model (M_1) over another (M_2), given the data (X) (Raftery 1995a), is computed as

$$\frac{\mathrm{P}(M_1|X)}{\mathrm{P}(M_2|X)} = \left[\frac{\mathrm{P}(X|M_1)}{\mathrm{P}(X|M_2)}\right]\left[\frac{\mathrm{P}(M_1)}{\mathrm{P}(M_2)}\right]. \quad 19.$$

The latter term on the right-hand side of Equation 19 is the prior odds for the two models and is often assumed to be 1 (both models are equally likely a priori). The first term on the right-hand side is the Bayes factor: the ratio of the marginal likelihoods of the data under the two models. As discussed in the first section of the article, the marginal likelihood for a model is the integral of the data under all values of the parameter, which can be extended to explicitly consider that values for the parameter, θ, stem from the model:

$$f(X|M) = \int f(X|\theta_M, M) f(\theta|M) \mathrm{d}\theta. \quad 20.$$

This integral is difficult to compute and is one reason that MCMC methods became popular: They do not rely on this computation. This quantity, however, can be approximated using minimal output from most canned software routines: the model log-likelihood, the model degrees of freedom, and the sample size. The BIC is computed as $-2\ln(L) + d\ln(n)$, where $\ln(L)$ is the log-likelihood function computed at the ML estimation, d is the number of parameters estimated in the model, and n is the sample size (Raftery 1995a). Thus, the BIC for a given model is a function of the model chi-square penalized by the number of parameters in the model. Model selection proceeds by choosing the model with the more favorable posterior odds, which means the model with the smaller BIC.

As an alternative to selecting a single model, in some cases one may prefer to combine results of multiple models rather than select one in particular. Bayesian model averaging (BMA) allows for the combination of results from K models via $\mathrm{P}(\theta|X) = \sum_{k=1}^{K} \mathrm{P}(\theta|D, M_k)\mathrm{P}(M_k|D)$, where θ is the parameter of interest (Hoeting et al. 1999, Raftery 1995a). The former term on the right-hand side of the equation is the posterior distribution for the parameter under each model; this distribution is weighted by the latter term on the right-hand side, which is the posterior model probability and can be approximated using the BIC. A key difficulty with BMA is that, in cases in which variable selection is the goal, there are 2^K possible models, which may be an overwhelming number of models to evaluate. Madigan & Raftery (1994) describe a procedure called Occam's window to reduce the number of models to consider prior to averaging over them.

In our review, we identified several papers in sociology concerned with developing or testing variable selection and averaging methods. Bollen et al. (2012) developed two new Bayes factor approximations and compared them to the BIC. They found their new measures and the BIC all work well with large samples and high model R^2, but none work well in other cases. They found evidence that one of their new measures may perform slightly better than the BIC when n is small. Feng et al. (2017) developed a Bayesian adaptive lasso procedure to estimate parameters of structural equation models with ordinal data. The method uses a Gibbs sampler to obtain samples from the posterior distribution, with variable selection built into algorithm rather than left as a secondary step. Morgan et al. (2016) compared four approaches to modeling individual impressions of social events under affect control theory: stepwise regression, ANOVA, BMA, and Bayesian model sampling, an alternative to BMA. Using simulations, they found that the Bayesian approaches work better at selecting the correct covariates and tend to produce estimates with less bias and variance than the traditional methods. Sutton (2013) investigated changes in incarceration rates across 15 countries from 1960–2000 using a Bayesian change-point model. The change-point is unknown and so it is incorporated as an additional set of parameters in the model,

but the addition of the change-point implies interactions between the time of change and all other covariates in the model. Thus, Sutton used BMA to help reduce the model space. He found that key changes in prison regimes happened in the 1980s among countries with unregulated labor markets, decentralized polities, or weak labor unions. Kuiper et al. (2013) showed how to combine evidence for competing hypotheses (i.e., $\theta < 0$, $\theta = 0$, or $\theta > 0$) from studies in which the measures and designs differ, by computing probabilities for the hypotheses within studies and updating the Bayes factor sequentially across studies. Finally, Warren et al. (2008) used an extension of the logic underlying the BIC to investigate the effect of high school exit exams on employment and earnings using data from the US Census and the Current Population Survey. Given the size of the data, the authors used an adjusted criterion for the t statistic to assess significance of coefficients: Rather than using $t = 1.96$, they used a value of $\sqrt{\ln(n)}$, a value derived from the BIC.

Each of these studies either developed extensions, creatively used, or evaluated Bayesian model selection methods. Moreover, the application of Bayesian methods in this area has become routine in sociology: A SocIndex search found that, since 1990, 1,848 articles in sociology and demography included "BIC" in the title, abstract, or text, with nearly half (898) occurring since 2008. The key reason for the rise in popularity of the BIC is that model comparison is an important part of social science research, and yet the classical approach offers no method for comparing nonnested models.

Despite the increasing popularity of the BIC, some have criticized its use (Gelman & Rubin 1995, Gelman et al. 2013, Weakliem 1999). These criticisms include—among other concerns—that the choice of the best model is highly sensitive to priors, that the BIC strongly prefers parsimonious models, and that it performs poorly in models with high collinearity between predictor variables.

An alternative to using a single measure, such as the BIC, for model selection, is a less formalistic, more flexible approach involving the posterior predictive distribution (PPD). The PPD can be defined as

$$f(y^*|y) = \int f(y^*|\theta) f(y|\theta) \mathrm{P}(\theta) \mathrm{d}\theta, \qquad 21.$$

where y^* is future (out of sample) data, y is the observed data, and θ is the model parameter (vector). The middle two terms under the integral are proportional to the posterior distribution, and the parameter is integrated out. Thus, the PPD is the distribution of future data conditional on the observed data after accounting for parametric uncertainty. If a model fits the observed data well, then y^* should look much like y, and unlimited tests can be constructed to evaluate model fit using simulated PPD data, thereby facilitating the comparison of multiple models. Implementing PPD simulation is straightforward. For example, suppose one constructed a linear model for $y|x$ and used Gibbs sampling as shown in Equations 17 and 18 to draw $g = 1 \ldots G$ samples of parameters β and σ_e^2. For each sampled parameter, $\theta^{(g)}$ and $\sigma_e^{2(g)}$, one could simulate a new sample of data $y^{*(g)} \sim N(X\beta^{(g)}, \sigma_e^{2(g)})$ to obtain G new data samples. Individuals' values for y^* could be compared with the observed values of y to detect outliers, or distributions of sample statistics $T(y^*)$ such as the sample mean, median, or variance could be computed and compared with the observed values, $T(y)$. Indeed, one can compute a Bayesian p-value as the probability of obtaining future data, given the observed data: p-value $= \mathrm{P}(T(y^*) > T(y)|y)$ (Lynch 2007, Lynch & Western 2004, Rubin 1984). An extremely large or small p-value implies that the model does not fit the data well, because it does not generate new data that look like the original data.

Despite the recommendations of various authors that PPD simulation be used more often (Gelman et al. 2013, Lynch 2007, Lynch & Western 2004), we found only one paper in sociology over the past decade that employed it. Oravecz et al. (2014) developed a cultural consensus

model of true/false items with "don't know" responses allowed in examining the extent of agreement among respondents regarding basic science questions and questions about aging. They used MCMC methods because their modeling problem is intractable with ML methods, and they used PPD simulation to evaluate the fit of their models.

Samples from the posterior distribution for parameters can be used not only for PPD sampling, but also for tertiary analyses, that is, analyses that involve parameters that were not directly sampled but are functions (or functionals) of parameters (and variable values) that were sampled. For example, if we used a Gibbs sampler to sample parameters from a linear model as described above, but our real interest lay in the distribution for the product of β_1 and β_2, we could produce a posterior distribution for $\omega = \beta_1\beta_2$ by simply computing ω for each Gibbs sample. The resulting distribution for ω would have a probabilistic interpretation, just as the original parameters do.

These functions can be simple, such as the example of a product of parameters, or they can be complex, involving linear and nonlinear functions of both parameters and variable values (functionals). For example, Lynch & Brown (2010) developed a method for estimating the influence of covariates on intervals for healthy life expectancy from cross-sectional survey data coupled with population-level mortality data using a two-stage Gibbs sampler. This approach built on an earlier approach the authors developed for producing multistate life tables from panel data using Gibbs sampling (Lynch & Brown 2005). Under both approaches, (*a*) Gibbs sampling is used to produce samples from the posterior distribution for the parameters of a bivariate probit model predicting health status and mortality risk. (*b*) Each Gibbs sample is then combined with a given covariate profile to produce age-specific transition probability matrices. (*c*) Finally, standard multistate life table calculations are used to produce life tables. The application of these calculations to all of the Gibbs samples produces a posterior distribution for the life table quantities, enabling straightforward summarization of uncertainty in them (via the simple methods discussed earlier). Comparisons of subpopulations can be made by repeating the process but changing the covariate values at step *b* to change the subpopulation the life table represents.

Both the Bayesian cross-sectional and panel methods resolve problems with traditional approaches that use approximation methods—such as the delta method—or bootstrapping to attempt to measure uncertainty in estimates. The former approach is (by definition) only approximate, while the latter approach can be problematic when cell sizes are small, and it also has no direct probabilistic interpretation. A Bayesian approach resolves both of these issues.

Missing Data

Missing data is common in sociology, and methods to handle it have varied considerably over recent decades. The easiest methods for handling missingness include listwise deletion and simple (e.g., mean/mode) imputation, but both strategies are problematic because they bias standard errors and potentially parameter estimates as well (Little & Rubin 2002). The Bayesian approach is to treat missing data as an unknown and include it in the posterior distribution. Suppose θ is the parameter (vector) of interest, y_{obs} are the observed data, and y_{miss} are the missing data. Then, generically, the posterior is

$$f(\theta, y_{\text{miss}}|y_{\text{obs}}) \propto f(y_{\text{miss}}|\theta, y_{\text{obs}})f(y_{\text{obs}}|\theta)\text{P}(\theta). \quad 22.$$

We are generally not interested in the missing data itself, only θ. Thus, we can integrate out the missing data:

$$f(\theta|y_{\text{obs}}) = \int f(\theta, y_{\text{miss}}|y_{\text{obs}})\partial y_{\text{miss}}. \quad 23.$$

This integration can be done via MCMC methods (often via Gibbs sampling), by (*a*) simulating y_{miss} from its conditional posterior distribution (an imputation step conditional on θ and y_{obs}), (*b*) simulating θ from its conditional posterior distribution (given y_{obs} and the newly imputed y_{miss}), and (*c*) iterating. The resulting values of θ constitute a sample from θ's marginal posterior distribution. Importantly, so long as the missing data are missing at random (Little & Rubin 2002), neither the parameter nor its posterior variance is biased.

Our Web of Science search identified only one paper over the past decade that discussed this approach to handling missing data in sociology, although it did so in the context of multiple imputation (MI). Von Hippel (2013) compared the performance of direct ML estimation to two MI methods: imputation based on draws from the posterior PPD and imputation based on a single ML estimate. He found that MI using PPD draws performed less well than the other methods, although all three methods work well in a relatively simple model.

Despite the lack of recent Bayesian literature on missing data in sociology, we include a section on the topic in this review because missing data handling is an area in which Bayesian methods have indirectly entered into the mainstream. A SocIndex search found that "multiple imputation" appeared in 490 articles in sociology between 1990 and 2008 and in more than 2,000 since 2008. MI arguably has become the most commonly used method for handling missing data, and it is fundamentally Bayesian (Rubin 1977, 1976, 1987; Little & Rubin 2002). Few sociologists may recognize that MI is fundamentally Bayesian, however, and the most-cited articles that discuss MI do not make explicit its Bayesian roots (Acock 2005).

The reason many may not recognize that MI is a Bayesian approach is that MI separates the imputation step from the modeling step, allowing individuals to use canned procedures to conduct MI prior to invoking traditional modeling procedures (Johnson & Young 2011, Acock 2005). Whereas a fully Bayesian approach would use a Gibbs sampler like we discussed above, and alternate between imputation and parameter sampling steps, MI involves imputing multiple data sets before estimating the desired model on each data set and combining the results using Rubin's rules (Rubin 1987). The imputation step typically involves a distinct model or strategy for imputing the missing data, most commonly assuming the data arise from a multivariate (normal) distribution or from a series of chained equations. Either way, missing data are simulated from the PPD implied by the imputation model. The imputed data sets are then used subsequently in a classical modeling procedure, and in many cases, coefficients and standard errors obtained from the models are combined automatically by the software, leaving the entire Bayesian enterprise out of view.

Hierarchical and Related Methods and Applications

Although the use of Bayesian methods for comparing or combining models and handling missing data has exploded in recent years, other applications of Bayesian methods have been relatively limited, with most studies employing hierarchical methods. The Bayesian approach, coupled with Gibbs sampling, is well suited for modeling data that have a hierarchical structure. Indeed, the Bayes theorem itself has a natural hierarchical structure: Data depend on a parameter, and the parameter depends on a prior that itself contains higher-level parameters that may be fixed or may themselves be represented by hyperprior distributions. For example, suppose our data structure involves students nested within classes, and suppose we believe there are unique class effects, α, on student outcomes, y. We could construct a hierarchical probability model as

$$y_{ij} \sim \mathrm{N}(\alpha_j, \sigma_y^2), \quad 24.$$

$$\alpha_j \sim \mathrm{N}(\mu_\alpha, \tau_\alpha), \quad 25.$$

where at level 1, y_{ij} is a normally distributed outcome for student i in class j, and α_j is a class-specific mean, with student variation around it within-classes captured by σ_y^2. At level 2, the random effects, α, are assumed normally distributed with a grand mean of μ_α and variance τ_α. With a prior distribution on σ_y^2, μ_α, and τ_α parameters, a posterior distribution can be constructed as

$$f(\alpha, \mu_\alpha, \sigma_y^2, \tau_\alpha | y) \propto f(y|\alpha, \mu_\alpha, \sigma_y^2, \tau_\alpha) f(\alpha|\mu_\alpha, \tau_\alpha) f(\mu_\alpha) f(\tau_\alpha) f(\sigma_y^2) \quad 26.$$

$$\propto \left(\prod_{j=1}^{J} \left(\prod_{i=1}^{n_j} \mathrm{N}(\alpha_j, \sigma_y^2) \right) \mathrm{N}(\mu_\alpha, \tau_\alpha) \right) f(\mu_\alpha) f(\tau_\alpha) f(\sigma_y^2) \quad 27.$$

$$\propto \left(\prod_{j=1}^{J} \prod_{i=1}^{n_j} \mathrm{N}(\alpha_j, \sigma_y^2) \right) \left(\prod_{j=1}^{J} \mathrm{N}(\mu_\alpha, \tau_\alpha) \right) f(\mu_\alpha) f(\tau_\alpha) f(\sigma_y^2) \quad 28.$$

The first equation represents the posterior distribution generically, while the second shows the full posterior with products of normal distributions inserted to reflect that the data consist of individual student outcomes nested within J classes, with n_j representing the number of students in class j. The prior distributions remain generic but are typically normal for μ, with fixed hyperparameters for its mean and variance, and inverse gamma for τ and σ^2, also with fixed hyperparameters. As shown here, the parameters μ_α, τ_α, and σ_y^2 have independent priors (i.e., the joint prior is simply a product of priors). The third equation rearranges the second to show that the latter four terms on the right-hand side combine with the likelihood term to produce a joint distribution for all known and unknown quantities by the basic joint distribution rule for nonindependent probabilities. For fixed values of given unknown quantities, this joint distribution is proportional to the posterior distribution. Gibbs sampling can be employed to sample from the posterior distribution for the four parameters by fixing all components on the right-hand side but one, discarding proportionality constants, recognizing the resulting conditional distributions, and then iteratively sampling from them as discussed earlier. The hierarchical structure in this model can be extended indefinitely, in theory, via a conditional probability chain rule. For example, suppose classes are nested in schools, which themselves are nested in communities. Such a structure would just necessitate extending the conditional distributions on the right hand side to reflect the additional levels. Further, dependencies between parameters at different levels can be readily incorporated. Thus, it is often easier to estimate parameters of complex hierarchical models using a Bayesian approach than an ML approach (Raudenbush 2002).

A number of papers over the past decade in sociology and demography have used a Bayesian approach to conduct hierarchical modeling. Raftery and colleagues published several papers over the past decade in which they developed Bayesian population projection methods. In one paper, they developed a hierarchical model for the total fertility rate for all countries; the method combines data across countries to produce stable estimates of trends from which projections can be made (Alkema et al. 2011). Raftery et al. (2013) developed a similar approach for projecting total life expectancy for all countries, and Azose & Raftery (2015) introduced a similar approach for estimating immigration. The purpose underlying the development of these methods was to move beyond the longstanding United Nations approach to projection that has no probabilistic interpretation. Alexander et al. (2017) estimated subnational mortality rates using a Bayesian hierarchical model to help compensate for the problems that locales with sparse data pose. Richardson et al. (2013) developed a hierarchical Poisson regression model to estimate mortality for immigrants relative to native Pacific residents in New Zealand. Wisniowski et al. (2015) developed a

Bayesian extension of the Lee-Carter method for forecasting to allow for probabilistic forecasts that include migration and fertility and split the projections by sex.

Data with spatial dependencies are also hierarchically structured, making the Bayesian approach useful for such analyses. Xu (2014) noted that it is very common for spatial models to employ Bayesian methods because the complex hierarchical formulations require a high degree of integration and often cannot be estimated with ML methods. In his paper, he conducted simulations to study the relative advantages of traditional hierarchical versus spatial modeling, and he illustrated these approaches by examining neighborhood effects on child mortality in New Jersey prior to 1900. Savitz & Raudenbush (2009) used two Bayesian approaches to estimate collective efficacy. Spatial dependence was included in one but not the other. They found that the method that incorporated spatial dependence worked best. Finally, Jones et al. (2015) investigated ethnic residential segregation in London using a hierarchical Poisson model. They used a Bayesian approach because other methods overestimate higher-level variances. Importantly, their method allowed for the modeling of four levels of segregation simultaneously (ranging from the output level to the borough level). Beyond spatial analyses, other studies relying on complex data and parameter hierarchies have employed Bayesian approaches. Tighe et al. (2010) used Bayesian hierarchical models to pool survey data on small religious minority groups. They demonstrated how multiple surveys, each involving different sampling strategies, can be combined for meta-analysis by applying Bayesian hierarchical models. Finally, Zhou (2015) developed Bayesian estimators for modeling mobility tables across countries and found that a Bayesian hierarchical model captures mobility better than the traditional ML approach.

Other Applications

A number of papers in sociology employing Bayesian strategies, either conceptually or methodologically, are difficult to classify. Two papers used a Bayesian approach in order to incorporate prior information into analyses. Billari et al. (2014) focused on eliciting expert opinions for making population forecasts. Their approach treated expert opinion as data and involved MCMC methods for estimation. Rendall et al. (2009) used population-level information as priors in regression models of survey data to estimate Hispanic fertility.

Other papers adopted a Bayesian approach to address topics that are difficult to address using classical methods. In network analysis, there have been 20 papers published involving Bayesian methods over the past decade in *SN* alone. One reason that Bayesian methods are so common in network analysis, as Butts (2011) notes, is that quantities of interest in network analysis are often impossible to derive analytically or to obtain via other methods. In the one paper we identified in the 10 journals listed earlier, Butts (2011) exemplified this feature of Bayesian analysis in developing a method to estimate characteristics of superpopulations from which one has complete sample network data or simply summary measures from it. Cheng et al. (2008) used a Bayesian mixture model to handle missclassification of individuals into small groups, with an application to identifying whether children in the Early Childhood Early Longitudinal Study are from adoptive, biological, or mixed families. They used a moderately strong prior to identify model parameters that are not otherwise identified, and they compared models in which more versus less information was incorporated to reduce the potential for misclassification in assessing whether children with biological versus adoptive parents are more likely to be disabled. They found that models that did not address the potential misclassification of children yielded smaller effects of family type on disability propensity.

Stamey et al. (2017) developed a Bayesian approach to address misclassification of outcome variables. They assumed there are two response variables, both binary, with one variable subject

to missclassification. They compared two adolescent groups; one received a comprehensive sex education program, while the other received a basic program. The outcomes included a measure of sexual experience and a measure of intentions. They found that ignoring missclassification tended to bias the effect of the treatment toward 0, although all methods produced nonsignificant results from a classical perspective.

Another paper used a Bayesian approach to handle uncertainty in propensity scores (An 2010). Traditional propensity score methods ignore uncertainty in the propensity score: One model is estimated in order to obtain propensity score estimates, but no compensation is made for the fact that these scores are only estimates. As a result, traditional propensity score methods (somewhat paradoxically) overstate the variance in the treatment effect. The model she develops samples the propensity score and the structural model parameters in one step, and she shows that the Bayesian approach works better than the traditional two-step approach.

Whalen & Boeri (2014) investigated trajectories of binary sequences using a Bayesian approach. Their goal was to develop visualization and other techniques for identifying discontinuity (and remission) in sequences of measures of drug use.

Lee & Conley (2016) examined the influence of having daughters versus sons on changes in political orientation. In the United Kingdom, it has been found that having daughters tends to lead parents to shift leftward in their political orientation, but in the United States, the opposite pattern has been found: Having daughters leads to a rightward shift. The authors examined data from 36 countries using additive regression tree models. These models split the data into subgroups to obtain heterogeneous treatment effects—here, the effect of having a daughter/son on political views. They found no significant effects and suggested that previous findings may be due to publication bias.

Finally, over the past decade, a small collection of papers did not employ Bayesian methods but simply utilized the underlying logic of Bayesian statistics. Abell (2009) developed a Bayesian approach to thinking about causality using narrative (small *n*), historical data. The method he developed was not itself Bayesian, but the logic grounding the method was. Andrew & Hauser (2011) investigated the evolution of educational expectations. Status attainment theory suggests that educational expectations are fixed early in life, but learning theory—which is fundamentally Bayesian in its logic—suggests that expectations evolve as individuals gain more experience. Although the conceptual background in that paper was Bayesian, they used a classical structural equation modeling approach in their analysis. Using a similar conceptual model, Fallesen & Breen (2016) developed a Bayesian updating model for divorce. They argued that having a child with colic is stressful, so that experience increases the pace with which individuals learn whether they are compatible with their spouse. Their conclusion is pessimistic: Having a child with colic increases the rapidity with which marriages end, but only among those who would ultimately divorce. Finally, Schroeder et al. (2016) developed a Bayesian generalization of affect control theory in social psychology. Whereas traditional affect control theory analyses model identities as a single quantity, the authors represent identities via probability distributions to reflect uncertainty. The logic is fundamentally Bayesian: Uncertainty about another's identity decreases as one gets to know the other person. The authors used simulation to show the emergence of increasing stability in identity, much as posterior distributions—and subsequent priors—become narrower as new information is incorporated into a corpus of knowledge.

CONCLUSIONS

This article has reviewed the key concepts and methods that underlie contemporary Bayesian statistics and has described the rise of Bayesian statistics in statistics, the social sciences, and

sociology. As we have discussed, the development of MCMC methods simplified the process of conducting Bayesian analyses and produced a rapid increase in the application of Bayesian statistics in statistics. The use of the Bayesian approach was delayed in social science and remains relatively rare in sociology even today. Nonetheless, the number of Bayesian publications in sociology is increasing, and some of the tools developed from the Bayesian approach have become ubiquitous. We expect this trend to continue and even accelerate for several reasons. First, the crisis in science is not going away. Recognition of the problems with the classical statistical approach of rejecting null hypotheses is growing, and the problem is crystallizing as the number of scholars and publications in social science—and therefore the number of false positives—grows, given the inherent bias to publish studies that have significant results while eschewing studies with null results. The Bayesian approach rejects such either/or reasoning, with new studies building on prior studies via the use of priors. We believe the decline in traditional criticisms against priors, and the growing emphasis on the benefits of Bayesian logic evidenced in many of the papers we reviewed, bodes well for science, including sociology. Second, the data we use and the questions we ask in sociology continue to become increasingly large and complicated, respectively, thus requiring more sophisticated models that pose difficulty for classical methods. As highlighted in our review, the Bayesian approach using MCMC methods holds considerable promise in this regard: A key reason the literature we discussed employed Bayesian methods was to overcome limitations of traditional methods. Third, generic software packages such as Stata have begun to introduce procedures that perform general MCMC simulation (StataCorp 2017), facilitating implementation in a host of models. As we showed, we have already seen the impact of this increased availability of user-friendly software in the growth of the use of MI and model selection measures such as the BIC. For these reasons, as much as we were disappointed that only a tiny fraction of contemporary sociology articles directly employs Bayesian methods, we are optimistic that Bayesian analysis will become more common in sociology over the coming decades.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Abell P. 2009. A case for cases: comparative narratives in sociological explanation. *Sociol. Methods Res.* 38:38–70

Acock AC. 2005. Working with missing values. *J. Marriage Fam.* 67:1012–28

Alexander M, Zagheni E, Barbieri M. 2017. A flexible Bayesian model for estimating subnational mortality. *Demography* 54:2025–41

Alkema L, Raftery AE, Gerland P, Clark SJ, Pelletier F, et al. 2011. Probabilistic projections of the total fertility rate for all countries. *Demography* 48:815–39

An W. 2010. Bayesian propensity score estimators: incorporating uncertainties in propensity scores into causal inference. *Sociol. Methodol.* 40:151–89

Andrew M, Hauser RM. 2011. Adoption? Adaptation? Evaluating the formation of educational expectations. *Soc. Forces* 90:497–520

Asparouhov T, Muthen B. 2012. *Comparison of computational methods for high dimensional item factor analysis.* Tech. Rep., Mplus version 7, Muthén & Muthén, Los Angeles, CA. **https://www.statmodel.com/download/HighDimension11.pdf**

Assuncao R, Schmertmann C, Potter J, Cavenaghi S. 2005. Empirical Bayes estimation of demographic schedules for small areas. *Demography* 42:537–58

Azose JJ, Raftery AE. 2015. Bayesian probabilistic projection of international migration. *Demography* 52:1627–50

Bayes T. 1763. An essay toward solving a problem in the doctrine of chances. *Philos. Trans. R. Soc.* 53:370–418

Berk R, Campbell A, Klap R, Western B. 1992. The deterrent effect of arrest in incidents of domestic violence—a Bayesian analysis of 4 field experiments. *Am. Sociol. Rev.* 57:698–708

Berk R, Western B, Weiss R. 1995a. Statistical inference for apparent populations. *Sociol. Methodol.* 25:421–58

Berk R, Western B, Weiss R. 1995b. Statistical inference for apparent populations—reply. *Sociol. Methodol.* 25:481–85

Billari FC, Graziani R, Melilli E. 2014. Stochastic population forecasting based on combinations of expert evaluations within the Bayesian paradigm. *Demography* 51:1933–54

Bollen K. 1989. *Structural Equations with Latent Variables*. New York: Wiley

Bollen K. 1995. Apparent and nonapparent significance tests. *Sociol. Methodol.* 25:459–68

Bollen KA, Ray S, Zavisca J, Harden JJ. 2012. A comparison of Bayes factor approximation methods including two new methods. *Sociol. Methods Res.* 41:294–324

Butts CT. 2011. Bayesian meta-analysis of social network data via conditional uniform graph quantiles. *Sociol. Methodol.* 41:257–98

Cheng S, Xi Y, Chen MH. 2008. A new mixture model for misclassification with applications for survey data. *Sociol. Methods Res.* 37:75–104

Eliason SR. 1993. *Maximum Likelihood Estimation: Logic and Practice*. Thousand Oaks, CA: SAGE

Fallesen P, Breen R. 2016. Temporary life changes and the timing of divorce. *Demography* 53:1377–98

Feng XN, Wu HT, Song XY. 2017. Bayesian adaptive lasso for ordinal regression with latent variables. *Sociol. Methods Res.* 46:926–53

Firebaugh G. 1995. Will Bayesian inference help? A skeptical view. *Sociol. Methodol.* 25:469–72

Freese J, Peterson D. 2017. Replication in social science. *Annu. Rev. Sociol.* 43:147–65

Gamerman D, Lopes H. 2006. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Boca Raton, FL: Chapman & Hall/CRC. 2nd ed.

Gelfand AE, Smith AFM. 1990. Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* 85:398–409

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2013. *Bayesian Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC. 3rd ed.

Gelman A, Hill J. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge Univ. Press

Gelman A, Rubin D. 1995. Avoiding model selection in Bayesian social research. *Sociol. Methodol.* 25:165–73

Gill J. 2002. *Bayesian Methods: A Social and Behavioral Sciences Approach*. Boca Raton, FL: Chapman & Hall/CRC. 1st ed.

Gill J. 2007. *Bayesian Methods: A Social and Behavioral Sciences Approach*, *Second Edition*. Boca Raton, FL: Chapman and Hall/CRC. 2nd ed.

Gill J. 2014. *Bayesian Methods: A Social and Behavioral Sciences Approach*. Boca Raton, FL: Chapman & Hall/CRC. 3rd ed.

Hauser R. 1995. Better rules for better decisions. *Sociol. Methodol.* 25:175–83

Hoeting JA, Madigan D, Raftery AE, Volinsky CT. 1999. Bayesian model averaging: a tutorial. *Stat. Sci.* 14(4):382–401

Ioannidis JP. 2005. Why most published research findings are false. *PLOS Med.* 2:696–701

Jackman S. 2009. *Bayesian Analysis for the Social Sciences*. New York: Wiley

Johnson DR, Young R. 2011. Toward best practices in analyzing datasets with missing data: comparisons and recommendations. *J. Marriage Fam.* 73:926–45

Johnson VE, Albert J. 1999. *Ordinal Data Modeling*. New York: Springer-Verlag

Jones K, Johnston R, Manley D, Owen D, Charlton C. 2015. Ethnic residential segregation: a multilevel, multigroup, multiscale approach exemplified by London in 2011. *Demography* 52:1995–2019

Kuiper RM, Buskens V, Raub W, Hoijtink H. 2013. Combining statistical evidence from several studies: a method using Bayesian updating and an example from research on trust problems in social and economic exchange. *Sociol. Methods Res.* 42:60–81

Lee B, Conley D. 2016. Does the gender of offspring affect parental political orientation? *Soc. Forces* 94:1103–27

Little RJA, Rubin DB. 2002. *Statistical Analysis with Missing Data*. New York: Wiley. 2nd ed.

Lynch SM, Taylor M. 2016. Trajectory models for aging research. In *Handbook of Aging and the Social Sciences*, ed. L George, K Ferraro, pp. 23–51. London: Elsevier. 8th ed.

Lynch SM. 2007. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. New York: Springer

Lynch SM, Brown JS. 2005. A new approach to estimating life tables with covariates and constructing interval estimates of life table quantities. *Sociol. Methodol.* 35:177–225

Lynch SM, Brown JS. 2010. Obtaining multistate life table distributions for highly refined subpopulations from cross-sectional data: a Bayesian extension of Sullivan's method. *Demography* 47:1053–77

Lynch SM, Western B. 2004. Bayesian posterior predictive checks for complex models. *Sociol. Methods Res.* 32

Madigan D, Raftery A. 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Am. Stat. Assoc.* 89:1535–46

Matsueda R, Kreager D, Huizinga D. 2006. Deterring delinquents: a rational choice model of theft and violence. *Am. Sociol. Rev.* 71:95–122

McGrayne S. 2011. *The Theory That Would Not Die*. New Haven, CT: Yale Univ. Press

McKinlay J, Lin T, Freund K, Moskowitz M. 2002. The unexpected influence of physician attributes on clinical decisions: results of an experiment. *J. Health Soc. Behav.* 43:92–106

Morgan JH, Rogers KB, Hu M. 2016. Distinguishing normative processes from noise: a comparison of four approaches to modeling impressions of social events. *Soc. Psychol. Q.* 79:311–32

Oravecz Z, Faust K, Batchelder WH. 2014. An extended cultural consensus theory model to account for cognitive processes in decision making in social surveys. *Sociol. Methodol.* 44:185–228

Papineau D. 2018. Thomas Bayes and the crisis in science. *TLS/Footnotes to Plato Blog*, June 28. **https://www.the-tls.co.uk/articles/public/thomas-bayes-science-crisis/**

Raftery AE. 1995a. Bayesian model selection in social research. *Sociol. Methodol.* 25:111–63

Raftery AE. 1995b. Rejoinder: Model selection is unavoidable in social research. *Sociol. Methodol.* 25:185–95

Raftery AE, Chunn JL, Gerland P, Sevcikova H. 2013. Bayesian probabilistic projections of life expectancy for all countries. *Demography* 50:777–801

Raudenbush SW. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: SAGE. 2nd ed.

Rendall MS, Handcock MS, Jonsson SH. 2009. Bayesian estimation of Hispanic fertility hazards from survey and population data. *Demography* 46:65–83

Richardson K, Jatrana S, Tobias M, Blakely T. 2013. Migration and Pacific mortality: estimating migration effects on Pacific mortality rates using Bayesian models. *Demography* 50:2053–73

Rubin DB. 1976. Inference and missing data. *Biometrika* 63:581–92

Rubin DB. 1977. Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J. Am. Stat. Assoc.* 72:538–43

Rubin DB. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Stat.* 12:1151–72

Rubin DB. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley

Rubin DB. 1995. Bayes, Newman, and calibration. *Sociol. Methodol.* 25:473–79

Savitz NV, Raudenbush SW. 2009. Exploiting spatial dependence to improve measurement of neighborhood social processes. *Sociol. Methodol.* 39:151–83

Schroeder T, Hoey J, Rogers KB. 2016. Modeling dynamic identities and uncertainty in social interactions: Bayesian affect control theory. *Am. Sociol. Rev.* 81:828–55

Stamey JD, Beavers DP, Sherr ME. 2017. Bayesian analysis and design for joint modeling of two binary responses with misclassification. *Sociol. Methods Res.* 46:772–92

StataCorp. 2017. *Stata 15 Base Reference Manual*. College Station, TX: Stata

Sutton JR. 2013. The transformation of prison regimes in late capitalist societies. *Am. J. Sociol.* 119:715–46

Tighe E, Livert D, Barnett M, Saxe L. 2010. Cross-survey analysis to estimate low-incidence religious groups. *Sociol. Methods Res.* 39:56–82

Van de Schoot R, Winter SD, Ryan O, Zondervan-Zwijnenburg M, Depaoli S. 2017. A systematic review of Bayesian articles in psychology: the last 25 years. *Psychol. Methods* 22:217–39

Vanlandingham M, Suprasert S, Grandjean N, Sittitrai W. 1995. Two views of risky sexual practices among northern Thai males: the health belief model and the theory of reasoned action. *J. Health Soc. Behav.* 36:195–212

Von Hippel PT. 2013. The bias and efficiency of incomplete-data estimators in small univariate normal samples. *Sociol. Methods Res.* 42:531–58

Warren JR, Grodsky E, Lee JC. 2008. State high school exit examinations and postsecondary labor market outcomes. *Sociol. Educ.* 81:77–107

Weakliem DL. 1999. A critique of the Bayesian Information Criterion for model selection. *Sociol. Methods Res.* 27:359–97

Weakliem DL. 2004. Introduction to the special issue on model selection. *Sociol. Methods Res.* 33:167–87

Western B. 1994. Unionization and labor-market institutions in advanced capitalism, 1950–1985. *Am. J. Sociol.* 99:1314–41

Western B. 1999. Guest editor's introduction to the special issue on Bayesian methods in the social sciences. *Sociol. Methods Res.* 28:3–6

Western B. 2001. Bayesian thinking about macrosociology. *Am. J. Sociol.* 107:353–78

Whalen T, Boeri M. 2014. Measuring discontinuity in binary longitudinal data applications to drug use trajectories. *Sociol. Methods Res.* 43:248–79

Winship C. 1999. Editor's introduction to the special issue on the Bayesian Information Criterion. *Sociol. Methods Res.* 27:355–58

Wisniowski A, Smith PWF, Bijak J, Raymer J, Forster JJ. 2015. Bayesian population forecasting: extending the Lee-Carter method. *Demography* 52:1035–59

Xu H. 2014. Comparing spatial and multilevel regression models for binary outcomes in neighborhood studies. *Sociol. Methodol.* 44:229–72

Zhou X. 2015. Shrinkage estimation of log-odds ratios for comparing mobility tables. *Sociol. Methodol.* 45:320–56

Ziliak ST, McCloskey DN. 2008. *The Cult of Statistical Significance*. Ann Arbor: Univ. Mich. Press