Published in final edited form as:

Epidemiology. 2011 May ; 22(3): 365–367. doi:10.1097/EDE.0b013e318212b31a.

Turning the Bayesian Crank

Richard F. MacLehose^{a,b}, J. Michael Oakes^b, and Bradley P. Carlin^a

^aDivision of Biostatistics, University of Minnesota, Minneapolis, MN

^bDivision Epidemiology and Community Health, University of Minnesota, Minneapolis, MN

Bayesian statistics is the conceptually natural process of updating a prior distribution with observed data to produce a posterior distribution. In simple terms, we start with a belief, examine some new evidence (ie, data), and update our belief by combining what we used to believe with the new evidence. Bayesian statistical inference actually predates formal frequentist inference by more than 150 years. Through the second half of the 18th century, Laplace made use of Bayesian statistics to aid his scientific inference. By the early 20th century, Bayesian inference had become so routine that Fisher felt compelled to distinguish his new methods for frequentist inference from it.^{1,2} The appeal of Fisher's maximum likelihood techniques, coupled with the difficulty in computing Bayesian posterior distributions for any but the simplest, low-dimensional models, led frequentist inference to supplant Bayesian inference as the statistical technique of choice for most of the 20th century. Fisher's vehement objection toward Bayesian inference, once referring to it as "fallacious rubbish,"³ no doubt, hastened its decline as well (although it's interesting to note that Fisher's greatest statistical failure, fiducialism, was essentially an attempt to "enjoy the Bayesian omelet without breaking any Bayesian eggs"⁴).

It was into the world of burgeoning frequentist statistics that modern epidemiology was born. Not surprisingly, then, the brief history of modern epidemiology has been overwhelmingly frequentist. Generations of epidemiologists have been trained exclusively in frequentist statistics, many without ever realizing it. More recently, however, prominent epidemiologists and statisticians have argued for the inclusion of Bayesian techniques in the epidemiologist's toolkit.^{5–8} The appeal to Bayesian methods has often rested on its subjective nature; a prior distribution is nothing more than a mathematical representation of an investigator's best guess about the magnitude of the parameter of interest as well as the uncertainty surrounding that guess. If the prior distribution is specified based on previous research, then a simple Bayesian analysis can be thought of as approximately equivalent to a meta-analysis, a procedure with which most epidemiologists are quite comfortable. Incorporating substantive knowledge in prior distributions can improve estimation and offers the benefit of producing interval estimates (credible intervals in the Bayesian jargon) that have the interpretation most scientists want to attach to frequentist confidence intervals.⁹

Correspondence: Richard F. MacLehose, Division of Epidemiology and Community Health, University of Minnesota, 1300 S 2nd St, Ste 300, Minneapolis, MN 55454. mac10029@umn.edu.

RICHARD MACLEHOSE is an assistant professor of epidemiology and biostatistics at the University of Minnesota with research interests in applied Bayesian methods. J. MICHAEL OAKES is an associate professor of epidemiology at the University of Minnesota engaged in social epidemiology and methods research. BRADLEY CARLIN is Mayo professor and head of biostatistics at the University of Minnesota. He is actively engaged in research in Bayesian inference and spatial statistics.

Moreover, the availability of a full posterior distribution also enables direct probability statements about key model quantities, such as the probabilities that each census tract will be assigned to the highest risk quintile, attractively mapped by Marí-Dell’Olmo and colleagues¹⁰ in their Figure 3.

Increasingly often, however, epidemiologists and statisticians are drawn to Bayesian analysis because it represents the easiest and sometimes the only way to obtain results. Much of Bayesian inference depends on modern computing power, in particular, on Markov chain Monte Carlo algorithms that draw repeated samples from the posterior distribution of interest. Once a prior distribution and a likelihood (often a regression model) are specified, all that is required is to generate samples from a posterior distribution by implementing the algorithm. These algorithms often have an automatic nature (indeed, the WinBUGS [MRC Biostatistics Unit, Cambridge, UK] and SAS [SAS Institute, Cary NC] software packages both offer generic Markov chain Monte Carlo algorithms). This step in Bayesian analysis has often been labeled as “turning the Bayesian crank.” Interestingly, no matter how much more difficult or high dimensional the prior or likelihood gets, turning this crank often does not get that much harder—hence the allure of Bayesian analysis in complex models. In contrast, frequentist inference tends to rely on asymptotic approximations that become increasingly unreliable, forcing the user to make unattractive and perhaps unjustified compromises in the modeling.

The paper in this issue of EPIDEMIOLOGY by Marí-Dell’Olmo and colleagues¹⁰ is an intriguing example of this vein of Bayesian analysis. The authors propose a factor-analysis model to estimate an index of deprivation in 3 cities in Spain. Using factor models to estimate a latent-deprivation index is common in social epidemiologic research. Marí-Dell’Olmo and colleagues augment this approach by specifying conditional autoregressive priors to allow the possibility of spatial correlation of the latent-deprivation index (as well as spatial correlation of the residual variance) across census tracts. Rather than incorporate substantive information into their analysis, the authors use noninformative priors and turn the Bayesian crank to readily obtain Markov chain Monte Carlo estimates in a complex spatial data setting that would be very challenging for a frequentist. The authors do not offer technical details (or even much intuition) concerning how the conditional autoregressive operates; those wishing such background may wish to consult the book by Banerjee et al (Sections 3.3 and 5.4).¹¹

Factor analysis is commonly used in social epidemiology as a two-step process. First, a model is fit to summarize the observed data with a small number of latent factors. In this article, the observed indicators of deprivation are assumed to be related to a single latent factor, the index of deprivation. Second, for a given dataset, the predicted value of the latent factor is computed for each observation in the dataset. That imputed latent factor can then be used in a separate regression model, optimally in a different set of data, to predict or explain some outcome. For example, Laraia et al¹² used imputed latent factors representing neighborhood incivility, territoriality, and social spaces to predict physical activity during pregnancy in North Carolina. Among the difficulties with such an approach is that the second regression model assumes the factor is known with certainty, even though there is clearly uncertainty related to it. It follows that all interval estimates in the second regression

model will be overly precise. The approach of Marí-Dell’Olmo and colleagues allows the estimation of the entire distribution of the factor. It would be relatively straightforward to join this spatial factor model to an outcome model and incorporate the uncertainty in the latent factor in the whole estimation procedure. Such a model would be considered a structural equations model and, although these models are common in the statistics and psychology literature,^{13,14} they are relatively uncommon in epidemiology (see the paper by McKeown-Eyssen¹⁵ for an exception).

Of course, the ability of Bayesian methods to allow estimation of parameters in increasingly complex models is not without its own set of problems. Some of these problems are technical issues often inherent in complicated models. The authors of the current paper, for example, mention a “flip-flop problem” that is a result of a fundamentally unidentified model; the data do not contain enough information to uniquely estimate each parameter in the model. Identifying assumptions are commonly made for factor models in both Bayesian and frequentist inference, allowing inference to proceed without undue concern regarding model convergence. The authors’ solution for identifiability using equations (10) and (11) is reminiscent of the approach (originally proposed by Besag et al¹⁶) of imposing awkward constraints numerically as part of the Markov chain Monte Carlo algorithm, rather than the analytically safe but often tedious traditional approach. Other complicated models (such as bias models) may also suffer from nonidentifiability, but identifying assumptions may be impossible (or very unwise) to make. In this setting, Bayesian inference through Markov chain Monte Carlo algorithm holds a host of problems, as described elsewhere.^{17,18}

Although technical hurdles such as nonidentified models or poorly functioning MCMC algorithms are not uncommon in complicated Bayesian models, the greater concern comes with interpretation. Simply being able to estimate a model does not imply that we should do so. Rather, as applied researchers, we need to be comfortable that the assumptions of the model are not substantively unreasonable, and that models are in fact estimating parameters of importance. In the present case, it is not obvious that the estimated deprivation score is substantively helpful. It is unclear why manual labor should be considered a measure of deprivation. The inclusion of manual labor, as well as the exclusion of factors such as indoor plumbing and the availability of healthy foods, results in a potentially misleading measure. Furthermore, the model yields city-specific deprivation scores that make between-city comparisons untenable, as what constitutes deprivation in one city is not the same as what constitutes deprivation in another city. Finally, applied researchers will be interested in estimating the effect of the latent deprivation index on some health outcome. Regardless of whether the researcher incorporates the uncertainty surrounding the latent factor in a structural equation model or simply imputes one latent deprivation score for each observation in the dataset, it is instructive to think of this regression model in causal terms. That is, if we consider designing an intervention to change the amount of deprivation by some amount, what happens to our outcome of interest? The difficulty with a factor model in this example is the difficulty in imagining how we could intervene on “deprivation” directly. Instead, we would need to intervene on the manifest factors that determine the deprivation score. From a public health (or perhaps public policy) standpoint, estimating the effect of each of these factors on the health outcome of interest may be more closely related to the question at hand.

Marí-Dell’Olmo and colleagues¹⁰ incorporate spatial dependence in their model by assuming that the latent deprivation score in one census tract depends on the deprivation scores in adjacent census tracts. This is an extremely popular and common assumption in spatial statistics; it implies that the deprivation scores in neighboring census tracts will borrow information and be shrunk toward one another. If the deprivation score in one census tract is high, the census tracts next to it will tend to be high as well. This type of shrinkage estimation tends to stabilize and improve estimation if the conditional autoregressive assumption is reasonable. The plausibility of this assumption will vary with substantive considerations. It seems quite reasonable when it comes to natural landscapes such as vegetation growth. But when considering human developments such as socioeconomic deprivation, sharp discontinuities will often be expected (implied by the phrase “the other side of the tracks”), and the conditional autoregressive assumption will likely bias estimates on either side of the physical division. Interestingly, all 3 deprivation maps in Figure 1 show regions where census tracts with high deprivation are adjacent to census tracts with low deprivation. Of more interest, perhaps, is Figure 2 which compares the spatial factor model with a crude principal component analysis without a spatial component. The 2 methods perform comparably for census tracts with low deprivation scores; however, for census tracts with high deprivation scores, the spatial factor method produces lower scores than the principal component analysis. Some of this may be due to the (appropriate) smoothing of unstable high deprivation scores. On the other hand, this could as easily result from shrinking deprivation scores across discontinuities—a result that could mislead policymakers. Methods that allow for large discontinuities using mixture priors^{19–21} may have some appeal in future spatial analyses of deprivation.

The potential benefits of Bayesian analysis are many: incorporation of substantive information, directly interpretable interval estimates, and improved model performance, to name a few. The ability of Markov chain Monte Carlo algorithms to tackle the most challenging model forms (such as the one in the paper by Marí-Dell’Olmo and colleagues¹⁰) may well end up being one of the most important aspects of Bayesian analysis. We caution, however, that although the models may be more complicated than those typically used in epidemiology, our approach to them should be the same: scrutinize and report the assumptions and make sure the model is estimating what we want.

Acknowledgments

Supported by National Institutes of Health (grant 1U01-HD061940) (to R.M.) and National Cancer Institute (grant 2-R01-CA095955-05A2) (to B.C.).

References

1. Fisher RA. On the mathematical foundations of theoretical statistics. *Philos Trans R Soc A*. 1922; 222:309–368.
2. Aldrich J. RA Fisher on Bayes and Bayes’ theorem. *Bayesian Anal*. 2008; 3:161–170.
3. Fisher RA. The nature of probability. *Centen Rev*. 1958; 2:261–274.
4. Morris CN. Parametric empirical Bayes inference: Theory and applications. *J Am Stat Assoc*. 1983; 78:47–65.
5. Cornfield J. Recent methodological contributions to clinical trials. *Am J Epidemiol*. 1976; 104:408–421. [PubMed: 788503]

6. Greenland S. Probability logic and probabilistic induction. *Epidemiology*. 1998; 9:322–332. [PubMed: 9583426]
7. Greenland S. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *Int J Epidemiol*. 2006; 35:765–775. [PubMed: 16446352]
8. Dunson DB. Practical advantages of Bayesian analysis of epidemiologic data. *Am JEpidemiol*. 2001; 153:1222–1226. [PubMed: 11415958]
9. Carlin, BP., Louis, TA. *Bayesian Methods for Data Analysis*. 3rd. Boca Raton: Chapman and Hall/CRC Press; 2009.
10. Mari-Dell’Olmo M, Martínez-Beneito MA, Borrell C, et al. Bayesian factor analysis to calculate a deprivation index and its uncertainty. *Epidemiology*. 2011; 22:356–364. [PubMed: 21423017]
11. Banerjee, S., Carlin, BP., Gelfand, AE. *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: Chapman and Hall/CRC Press; 2004.
12. Laraia B, Messer L, Evernson K, Kaufman JS. Neighborhood factors associated with physical activity and adequacy of weight gain during pregnancy. *J Urban Health*. 2007; 84:793–806. [PubMed: 17710552]
13. Bollen, KA. *Structural Equations With Latent Variables*. New York: Wiley-Interscience; 1989.
14. Lee, SY. *Structural Equation Modeling: A Bayesian Approach*. West Sussex, UK: John Wiley & Sons; 2007.
15. McKeown-Eyssen G. Methodologic issues for the study of obesity. *Epidemiology*. 2006; 17:134–135. [PubMed: 16477251]
16. Besag J, Green P, Higdon D, Mengersen K. Bayesian computation and stochastic systems (with discussion). *Stat Sci*. 1995; 10:3–66.
17. Gustafson, P. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Boca Raton: Chapman and Hall/CRC Press; 2004.
18. Greenland S. Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Stat Sci*. 2009; 24:195–210.
19. Lawson A, Clark A. Spatial mixture relative risk models applied to disease mapping. *Stat Med*. 2002; 21:359–370. [PubMed: 11813223]
20. Militino A, Ugarte M, Dean C. The use of mixture models for identifying high risks in disease mapping. *Stat Med*. 2001; 20:2035–2049. [PubMed: 11427958]
21. Lu H, Reilly CS, Banerjee S, Carlin BP. Bayesian areal wombling via adjacency modeling. *Environ Ecol Stat*. 2007; 14:433–452.