Ecology, 91(12), 2010, pp. 3493–3496

Model complexity and information in the data: Could it be a house built on sand?

Subhash R. Lele[1]

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta T6G 2G1 Canada

Heisey et al. (2010), in an interesting paper, try to address a very difficult problem of analyzing spatially referenced, age specific prevalence data. The general goal of the analysis is to understand how force of infection changes as a function of age, time, and space. To further complicate matters, all the data considered in the paper are censored observations. Binary data are notoriously difficult to analyze, especially when latent processes are involved and prevalence is very low. Frankly, I was surprised by the complexity of the models they consider and the limited amount of information available to fit these models. I would like to congratulate them for trying to address such a difficult problem and in the process bringing to the attention of the ecologists some important statistical models in survival analysis.

How does one generally deal with the conflicting issues of lack of information and desire to conduct inference about complex underlying processes? The standard approach is to compensate for lack of information by adding assumptions. This is done routinely in most statistical analyses by assuming a parametric model. For example, one can conduct inference in ANOVA without assuming any specific relationship between the treatment means if replicate observations are available at each treatment level. If such replicate data are not available, instead of giving up, we assume that there is a linear (or, some parametric) relationship between the covariates and the response, the regression approach. This is a smoothing assumption. Similarly, in one of the fundamental papers on statistical inference in the presence of nuisance parameters, Kiefer and Wolfowitz (1956) showed that simply assuming that the nuisance parameters arise from a distribution is enough of a smoothing assumption to estimate not only the parameters of interest but also the distribution function from which nuisance parameters are assumed to have arisen. Heisey et al. (2010) try to get away with the limited information available in the prevalence data, where all observations are censored, by imposing constraints on the log-hazard, a smoothing assumption of another sort. This is the easy part. The real questions are: (1) Given the limited amount of information in the data, what assumptions do we need until some inference is feasible? and (2) Are these inferences primarily driven by the data or by the assumptions? Technically, the answer to the first question is straightforward: add assumptions until the parameters in the model, at least the ones that are of scientific interest, are estimable given the data. The second question is qualitative. It is partially addressed by studying the sensitivity of the inferences on the parameters of interest (assuming they are identifiable) to the violations of the assumptions. I will discuss these issues in the remainder of the commentary. I assume readers are familiar with the basic descriptions in Heisey et al. (2010).

Perhaps the easiest way out of the limited information in the prevalence data is to assume a specific parametric model for the log-hazard function. This does not guarantee that the parameters will be identifiable but it has the best chance. Heisey et al. (2010) do not take this easy way out. They aspire to assume less about the form of the log-hazard function. As they point out, the "nonparametric" MLE of the log-hazard is very choppy and unstable. It is generally not consistent, at least not at the usual $\sqrt{n}$ rate. One way out of this is to assume that the log-hazard function is continuous or differentiable of a certain order; thus imposing some form of smoothness conditions. Another way out is to assume that the log-hazard values are arising from a random process, the "random effects" approach. These approaches are neither Bayesian nor non-Bayesian. These are simply different models. A non-Bayesian approach stops at this specification. Given these model assumptions, the likelihood approach computes the marginal distribution of the data as a function of the underlying parameters. The maximum-likelihood estimator finds the value of the parameter that maximizes this likelihood function. This task can be computationally challenging but certainly not impossible, even when the random effects approach is used (Lele et al. 2007).

The Bayesian approach goes one step further. It assumes known distributions, "the priors," on the parameters. With this additional assumption, the

Manuscript received 15 January 2010; revised 20 January 2010; accepted 25 January 2010. Corresponding Editor: M. Lavine. For reprints of this Forum, see footnote 1, p. 3487.

[1] E-mail: slele@ualberta.ca

marginal distribution of the data contains no unknown parameters. The inferences are based on the conditional distribution of the parameters given the data, the posterior distribution. These inferences, by construction, are affected by the choice of the prior distribution. One may try to choose priors that are as weakly informative as possible. However, the precise definition of "weakly informative" is unavailable. It appears from the literature that there are as many non-informative priors as there are Bayesians. For a discussion of the issues of informativeness of the priors, see, e.g., Press (2003), Lele and Dennis (2009), or Wasserman (2006). These priors, contrary to what is claimed in much of the ecological literature, are not always innocuous. With this background, now I will get into the details of Heisey et al. (2010) paper. Unfortunately not having the original data and adequate time, I am forced only to raise questions without attempting to answer them.

Discretization approach.—As described by Heisey et al. (2010), working with the likelihood in terms of the underlying continuous time, continuous space process involves evaluating high dimensional integrals. To avoid the integration, the authors use the mean value theorem to approximate the integrals. They consider seven age classes, eight time points, and approximately 200 spatial locations. The full model has about 215 or so parameters. The first question that arises is how does the unit of discretization affect inferences? This is known as modifiable areal unit problem (MAUP) in geographical literature. The authors have neither pointed out the problem nor addressed it to any extent. It would have been interesting to try different age classes, temporal and spatial units and see how different the inferences are.

Nonspatial models.—The authors start with the simpler nonspatial model. This model, as far as I can see, involves only 14 parameters. To me it was somewhat surprising that the maximum-likelihood approach was not even attempted in this simple situation. It would have been instructive to compare the maximum-likelihood estimates with the ones obtained using the flat priors. Instead, the authors assume that these parameters arise out of a random process, the random effects approach. Is it really reasonable or necessary to assume a random effects model for such a small number of parameters? In the traditional mixed effects models (Searle et al. 1992), the random effects approach is usually deployed when the ratio of number of parameters to number of observations does not converge to zero as the number of observations increases. I found the authors' reluctance to use the likelihood approach in this simple situation somewhat puzzling. It may be that I have computed the number of parameters incorrectly.

Prior specification.—To conduct inference, authors further impose completely known prior distributions on the parameters of the random effects distribution. These priors, although euphemistically called non-informative, are not always innocuous. A flat prior on one scale is guaranteed to be non-flat on any other scale. Thus, they are not parameterization invariant. This is widely known in statistics and has been a major reason for developing other "non-informative" priors such as the Jeffrey's priors or reference priors (Press 2003). Unfortunately, computation of these priors is usually difficult and involves knowing the likelihood function which itself is nearly impossible to compute for hierarchical models. How important is the parameterization invariance? In an unpublished manuscript, we report ecological consequences of using flat priors. One example we consider is in the context of estimation of the probability of occupancy in the presence of detection error. In this situation, a uniform prior on the scale of probability between 0 and 1 gave us somewhat reasonable answers. On the other hand, for the same data, when we put uniform prior on the odds, the answers were extremely biased. Even for such a simple situation, under the uniform prior on the odds scale, it took nearly 100 000 observations before we could obtain reasonable answers. On the other hand, the likelihood estimators were excellent with samples sizes of 100 or so. We found common situations in population dynamics that also show similar problems. There is no criterion that suggests the "correct" scale on which such uniform or flat priors should be used. I wonder how different authors' results would be if they put flat priors on different transformations of the parameters. I would like to note that likelihood inference is guaranteed to be parameterization invariant.

Parameter identifiability.—Given a model, one can suggest any number of methods of estimation. Such estimation methods are nothing more than computational algorithms unless and until the estimators are shown to have reasonable properties. One important, nay essential, criterion is that the estimators are consistent under the true model. Parameters have to be identifiable for them to be consistently estimable. If the parameters are non-identifiable, the likelihood function is constant over a subset of the parameter space. Thus, no amount of data can discriminate between the parameters in such a subset. The spatial random effects model considered in the paper consists of non-identifiable parameters, namely $SD(\beta_s)$ and $SD(\beta_h)$. It is obvious that if these parameters are non-identifiable, the parameter $\psi = [SD(\beta_s)]/[SD(\beta_s) + SD(\beta_h)]$ cannot be identifiable either. What exactly does it mean to make scientific statements based on this non-identifiable parameter? It is very disturbing to learn that Bayesians do not care for identifiability. This could be true only if they are willing to base their inference on belief alone. As a scientist, I cannot agree with the sentiment. In my opinion, scientifically valid inference can only be based on identifiable features of the model.

Parameter estimability.—The method of data cloning (Lele et al. 2010) can be used to study the estimability of the parameters. It can be shown that as the number of

clones increases, the posterior variance converges to zero if and only if the parameters are estimable. On the other hand, if the posterior distribution converges to a nondegenerate distribution, it implies non-estimability of the parameters. This nondegenerate distribution can be, and usually is, different than the prior distribution; there can be "Bayesian learning" without identifiability. Consider a simple example. Let $Y_i \mid \mu_i \sim \mathcal{N}(\mu_i, \sigma^2)$ and $\mu_i \sim \mathcal{N}(\mu, \tau^2)$. Then it is obvious that $Y_i \sim \mathcal{N}(\mu, \sigma^2 + \tau^2)$. The parameters σ^2 and τ^2 are individually non-identifiable. Suppose we put priors $\sigma^2 \sim$ Unif(0, 100) and $\tau^2 \sim$ Unif(0, 100). Suppose the truth is such that $\sigma^2 + \tau^2 = 10$. Then the marginal posterior distributions for σ^2 and τ^2 necessarily get concentrated on the interval (0, 10) as the sample size increases. Their joint distribution will be concentrated along the diagonal of the square defined by the coordinates (0, 0), (0, 10), (10, 10), and (10, 0). This distribution is different than the prior distribution. Thus, there is "Bayesian learning" but clearly existence of Bayesian learning does not imply that the parameters are identifiable or even that legitimate inferences can be drawn about the parameters for which Bayesian learning happens. If a part of the model is non-identifiable, it can make estimators of other parameters inconsistent. They converge to a single, but wrong point. An example is given in Lele et al. (2010). The only way we can be sure that we are learning something sensible is by showing that the posterior distribution becomes degenerate at the true value as the sample size increases, namely the consistency of the estimator. The consistency of the estimators of the parameter of interest, especially in the presence of non-identifiable components of the model, needs to be established. Until such a result is proved, the spatial model discussed in this paper and inferences thereof should be considered suspect. I would have liked to see at least some simulation results indicating that the inferences are likely to be legitimate.

Maximum-likelihood analysis.—I would have liked to try data cloning to obtain maximum-likelihood estimates of the parameters for Heisey et al. (2010). However, original data are not available publicly. Hence I decided to use a somewhat different model and a data set provided in Heisey's training materials for survival analysis using WinBUGS (*available online*).[2] I considered an example in Chapter 3, *Bugs3_1*. The data are about nest survival. They are interval censored and hence far more informative than the fully censored data in the Heisey et al. (2010) paper. This model is similar to the models used in Heisey et al. (2010) and so I hoped to learn something about the behavior of the models and estimators in their paper. I did data cloning based analysis of these data using "rjags" (Plummer 2009), instead of WinBUGS. The original model, identical to the nonspatial EX model in the Heisey et al. paper, gave me substantial trouble with convergence. To get it to converge, I had to use a specific parameterization and also give fairly informative priors. I decided to use a slightly different prior, the AR(1) prior where $f_{t+1} = cf_{t+1} + \varepsilon_{t+1}$, $|c| \leq 1$, and $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ are independent random variables. This model was very stable in terms of convergence. The results are presented in Table 1. Here K denotes the number of clones used in data cloning. Priors were gamma0 $\sim \mathcal{N}(-3, 1)$, $c \sim$ Uniform(−1, 1), and $\sigma \sim$ Uniform(0, 4). Multivariate Gelman-Rubin statistics were 1.28 ($K = 1$), 1.01 ($K = 9$), and 1.08 ($K = 16$) indicating convergence of MCMC. Notice that as we increase the number of clones, the posterior variances do not converge to zero. This indicates non-estimability of the parameters (Lele et al. 2010). From these results, it appears that the parameters in this model are non-estimable. Furthermore, to see the effect of parameterization on the estimates, I put flat priors on different parameterizations. The results in Table 2 show that the estimates are quite different depending on which parameterization is used. When I tried to use $\sigma \sim$ Lognormal(0, 10), MCMC failed to converge. Surprisingly, initially for a shorter burn-in, it looked like it may converge (Rhat $\sim$ 1.9) but further burn-in made it worse (Rhat $\sim$ 6.8). Such behavior usually suggests problems with identifiability. Both these aspects make me wonder how much one should rely on the estimates and inferences described in Heisey et al. (2010). It is, of course, possible that I have misinterpreted the example in Heisey's notes and my results on non-estimability are incorrect. I would prefer that to be the case. At the same time, I would like to see some evidence that the parameters in the models used in Heisey et al. (2010) are, in fact, identifiable given the meager amount of information available in the observations. Bayesian learning is only a necessary but definitely not a sufficient condition for identifiability.

TABLE 1. Non-estimability of the parameters for the EX model with interval censored data.

Parameter and statistics	$K = 1$	$K = 9$	$K = 16$
gamma0			
Posterior mean	−4.71	−6.69	−7.20
Posterior SD	0.26	0.44	0.40
c			
Posterior mean	0.079	0.102	0.085
Posterior SD	0.52	0.22	0.22
σ			
Posterior mean	0.27	3.02	3.36
Posterior SD	0.22	0.47	0.38

Notes: K denotes the number of clones used in data cloning The parameter gamma0 relates to the survival probability in an interval, c corresponds to the auto-correlation between the random effects that vary from interval to interval and parameter, and σ corresponds to the variation between random effects.

[2] ⟨http://www.nwhc.usgs.gov/staff/dennis_heisey.jsp⟩

TABLE 2. Bayesian estimators depend on the parameterization.

Parameter and statistics	$\theta \sim$ Uniform(0, 4)	$\log(\theta) \sim \mathcal{N}(0, 10)$
gamma0		
Mean	−4.834	−4.69
SD	0.2682	0.2567
c		
Mean	0.078	0.1927
SD	0.3696	0.4863
σ		
Mean	0.646	0.2716
SD	0.1353	0.1837

Notes: Let $1/\sigma^2 = \theta$ be the precision parameter. Putting flat priors on θ or $\log(\theta)$ give different estimates. These are also different from the estimates reported in Table 1, $K = 1$ case, where the prior was on the σ scale.

Ecologists know a great deal about the processes. While constructing mathematical models, they have a strong and admirable desire to include all the nuances. Unfortunately the data are not always informative enough to conduct inferences on all the complexities of the model. As a consequence, either the model parameters become non-identifiable or non-estimable. If estimation is possible, estimates tend to be extremely uncertain with large standard errors, thus precluding their use in effective decision making. I would urge ecologists to establish identifiability of the parameters in their models before conducting any scientific inferences. Data cloning automatically informs the user if the parameters are estimable and if they are, it gives the estimates and their standard errors. The Bayesian inference procedure, used by Heisey et al. (2010), has a potential to provide answers that are misleading without any clear warning. As scientists, we need to learn to balance the desire to incorporate all the complexity in nature against the available information in the data. A mismatch can only lead to a house built on sand.

ACKNOWLEDGMENTS

This work was partially supported by an NSERC grant.

LITERATURE CITED

Heisey, D. M., E. E. Osnas, P. C. Cross, D. O. Joly, J. A. Langenberg, and M. W. Miller. 2010. Linking process to pattern: estimating spatiotemporal dynamics of a wildlife epidemic from cross-sectional data. Ecological Monographs 80:221–240.

Kiefer, J., and J. Wolfowitz. 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Annals of Mathematical Statistics 27:886–906.

Lele, S. R., and B. Dennis. 2009. Bayesian methods for hierarchical models: Are ecologists making a Faustian bargain? Ecological Applications 19:581–584.

Lele, S. R., B. Dennis, and F. Lutscher. 2007. Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. Ecology Letters 10:551–563.

Lele, S. R., K. Nadeem, and B. Schmuland. 2010. Estimability and likelihood inference for generalized linear mixed models using data cloning. Journal of the American Statistical Association, *in press*.

Plummer, M. 2009. rjags: Bayesian graphical models using MCMC. R package version 1.0.3-12. ⟨http://CRAN.R-project.org/package=rjags⟩

Press, S. J. 2003. Subjective and objective Bayesian statistics. Second edition. John Wiley, New York, New York, USA.

Searle, S., G. Casella, and C. E. McCulloch. 1992. Variance components. John Wiley and Sons, New York, New York, USA.

Wasserman, L. 2006. Frequentist Bayes is objective. Bayesian Analysis 1:451–456.