Let me begin by welcoming the paper by Cressie et al. (2009) as an insightful overview of hierarchical statistical modeling that will be valuable from both Bayesian and classical perspectives. Statistical conclusions hinge on the appropriateness of the mathematical models used to represent hypotheses, and Cressie et al. explain many merits of hierarchical models. My comments will highlight the shared challenges and common ground of Bayesian and classical analysis of hierarchical models; give a likelihood theory complement to Cressie et al.’s explanations of their merits; summarize methods for maximum likelihood and “empirical Bayes” estimation and incorporation of uncertainty; and explore some of the claims of Bayesian advantages and classical limitations.

Both Bayesian and classical analyses can and must address the inherent challenges of inference and prediction from noisy data, including comparing models, selecting a best model or combination of models, deciding if a model is acceptable at all, avoiding overfitting and statistical “fishing” or “data dredging,” and incorporating uncertainty. In practice, an appropriate modeling framework trumps many other issues, including choice of Bayesian or classical analysis.

The difference between Bayesian and classical philosophies is that Bayesian analysis uses the mathematics of probability distributions for model parameters, \( P \) as defined by Cressie et al., while classical analysis does not. It is generally agreed that Bayesian analysis must define “probability of \( P \)” as “degree of belief for \( P \)” (or “subjective probability of \( P \)” or other synonymous terms), whereas classical “probability” refers to a frequency of outcomes over a long run (O’Hagan 1994). Bayesian analysis makes “probability” statements about hypotheses, given data, that are formally weighted degree-of-belief comparisons, while classical analysis makes statements about frequencies with which different hypotheses would have produced the data and, in some approaches, hypothetical unobserved data. Classical analysis encompasses much more than Neyman-Pearson and/or Fisher hypothesis testing (which have been critiqued for both their actual logical limitations when correctly interpreted and their potential for misinterpretation; see Mayo and Spanos [2006] for extensions of NP logic). For example, model comparison using Akaike’s Information Criterion (AIC; Burnham and Anderson 2002) takes a classical approach to parameters. While Bayesian and classical approaches are philosophically different, one can interpret results from them in tandem (Efron 2005).

It is worth emphasizing that choosing a hierarchical model is separate from choosing a Bayesian or classical approach for parameters. In a hierarchical model, one has parameters (\( P \), either Bayesian or classical) that define the distribution of unknown ecological states (\( E \)) that define the distribution of data (\( D \)). I will call \( E \) the latent variables and/or random effects, which are among the many names for random variables whose values are not directly known but in theory define relationships among data values, to emphasize that hierarchical models are mixed models (McCulloch and Searle...
2001); as Cressie et al. point out, their example uses a
generalized linear mixed model. Having at least one $E$
level make a model hierarchical. Treating the param-
ters, $P$, with degree-of-belief probability makes an
analysis Bayesian. Cressie et al. illustrate a general route
to a Bayesian analysis. A general route to a classical
analysis would typically include maximum likelihood
estimation of $P$ under various models, with comparisons
and estimates of uncertainty made by likelihood ratios
(Royall 1997), likelihood ratio hypothesis tests and
confidence regions, AIC, parametric or nonparametric
bootstrapping, or other approaches.

Why are hierarchical models such a good idea for
either Bayesian or classical analysis? Cressie et al.
explain the sensibility of conditionally nested hierarch-
ies of probability models in terms of ecological and
sampling processes. A complementary view is that the
likelihood of the parameters, $P = (P_D, P_E)$, given $D$, is
the integral of Cressie et al.'s Eq. 1 with respect to $E$:  

$$L(P_D, P_E) = [D|P_D, P_E] = \int [D|E, P_D][E|P_E]dE.$$  (1)

This states that the probability of $D$ given $P$ is the sum
over all possible $E$ values of the probability of (i) those $E$
values given $P_E$ and (ii) $D$ given $P_D$ and those $E$ values.
This likelihood is the "[data |parameters]" mentioned by
Cressie et al. (2009). The asymptotic (as the amount of
data increases) behavior of this likelihood guarantees
convergence to the best $P$, a Gaussian shape for $L$, and
other properties (McCulloch and Searle 2001). By "best"
$P$, one can mean either the "true" $P$ or the $P$ that
minimizes the theoretical Kullback-Leibler discrepancy
(Burnham and Anderson 2002). Likelihood asymptotics
give a theoretical bedrock for both Bayesian and
classical analysis. Thus, a more formal appeal of
hierarchical models is that they often define appropriate
likelihoods for $P$, which drive the soundness of results
from either Bayesian or classical analysis and often
make the two approaches yield similar conclusions.

What do I mean by saying that a good model trumps
many other considerations? Given a choice between
Bayesian analysis of a set of appropriate models and
classical analysis with a set of plainly inappropriate
models, or vice versa, I’ll usually take the one with the
appropriate models and then judge the pros and cons of
the analysis. Recently, such a choice has led to a
common pragmatic appeal of Bayesian analysis: for
many people it is currently the easiest framework for
analyzing a hierarchical model. With time, I suspect that
advances in software and algorithms will move us closer
to being able to match any model structure to any
analysis approach. Then one will be able to choose a set
of hierarchical models based on the ecology and conduct
either a Bayesian or classical analysis based on scientific
goals.

How would one obtain maximum likelihood results
for a hierarchical model such as Cressie et al.’s seal
example, so one can use likelihood ratios or AIC, for
example? For relatively simple models, generalized
linear mixed model software can accomplish this. For
more general models, one can use numerical integration
of (1) via either grid-based (Efron 1996) or Monte Carlo
integration approaches, reviewed by de Valpine (2004).
The latter include Monte Carlo expectation maximiza-
tion, Monte Carlo Newton-Raphson iterations, “direct”
Monte Carlo integration with importance sampling,
sequential Monte Carlo integration (“particle filters”),
iterated Monte Carlo likelihood ratio approximations,
and Monte Carlo kernel likelihood approximations (de
Valpine 2004). More recent approaches include iterated
particle filtering (Ionides et al. 2006) and data cloning
(Lele et al. 2007). Each approach has pros and cons, just
as Monte Carlo simulation of Bayesian posteriors can be
done with relatively efficient or inefficient flavors of
Markov chain Monte Carlo (MCMC) algorithms, with
the best choice depending on the specific problem.
Monte Carlo kernel likelihood (MCKL) uses the full
Bayesian posterior, an appealing feature for those who
want to compare Bayesian and maximum likelihood
results.

Most of these algorithms use calculations that omit a
part of the likelihood known as the “normalizing
constant,” which must be estimated as a separate step.
This problem is mathematically identical to estimating
the marginal likelihood used in Bayes Factors after
simulating an MCMC posterior sample (de Valpine
2008). In summary, maximum likelihood and normaliz-
ing constant algorithms are highly feasible for many
hierarchical models, but current software makes Baye-
sian analysis more easily available for some models.

To a newcomer, the terminology that “empirical
Bayes” is a “non-Bayesian” analysis may seem baffling.
Typically, the empirical Bayes parameters would be the
maximum likelihood estimates, sometimes called the
“plug-in” parameters (Cressie et al. 2009). They do not
require degree-of-belief probabilities and so are not
Bayesian. The confusing distinction is illustrated by
contrasting Efron (1986), who discussed empirical Bayes
as a frequentist method, and Little (2006), who called it
Bayesian in a “broad view” that encompasses “a large
class of practical frequentist methods with a Bayesian
interpretation.”

Another potential confusion is that some purported
contrasts of Bayesian and classical results are con-
found with contrasts of hierarchical and nonhierarch-
ical models, respectively. Empirical Bayes has its roots in
the estimation of $E$ (not $P$): if a hierarchical model is
appropriate, then using the conditional distribution of $E$
given $D$ for maximum likelihood $P$ (empirical Bayes)
or a posterior for $P$ (Bayes) can be better than using a
nonhierarchical maximum likelihood approach for $E$ for
each unit of $D$ (Morris 1983). Some contrasts use a
nonhierarchical model for the classical side and a
hierarchical model for the Bayesian (or empirical Bayes)
side. This issue is reflected in Cressie et al.’s (2009)
reference to a nonhierarchical analysis as the “usual
Bayesian analysis that treats parameters (Morris 1983, Rubin 1984, Robins et al. 2000). That is, the performance of Bayesian analysis that treats parameters (P) with subjective probability can be evaluated based on the frequencies of outcomes over a “true” distribution of hypothetical data, such as coverage of a “true” P or E value by a credible region, which is naturally related to P-value concepts (not to be confused here with model parameters P). The motivation from a Bayesian view is that “frequency calculations are useful for making Bayesian statements scientific, ... capable of being shown wrong by empirical test” (Rubin 1984). Cressie et al. (2009) appeal to this frequentist justification of “accurate” credible intervals, and such accuracy is driven by the likelihood. The common ground that all models should be scientifically rejectable using the same performance currencies seems valuable. In contrast, a “pure” Bayesian would eschew any model testing based on frequencies of hypothetical, unobserved data.

How do Bayesian and classical analyses of hierarchical models each “incorporate uncertainty,” and how justified are the statements of Cressie et al. (2009) that the Bayesian approach “captures the variability in the parameters [P]” while plug-in estimates “do not account properly for the variability,” and that frequentist incorporation of uncertainty is limited in complexity? Frequentist thinking has long recognized that estimated parameters by definition give an optimistic picture of model fit, and one must account for this in inference and prediction. Cressie et al. mention quadratic likelihood approximations, bootstrapping and cross-validation. One could extend this list by including profile likelihoods, AIC, bootstrapping in empirical Bayes (Efron 1996), generalized cross-validation, generalized degrees of freedom to account for over-fitting due to model selection (Ye 1998), general covariance penalties (Efron 2004), and more. Nevertheless, in many situations a Bayesian posterior sample will be the easiest picture of uncertainty in P one can quickly generate, and indeed the only feasible picture for relatively complex models. Since neither approach is inherently superior at incorporating uncertainty, and since this topic touches the core of much statistical research, the pragmatic boundaries between approaches are likely to change with time.

Although a Bayesian picture of uncertainty is valuable and practical, it is not necessarily better for all purposes than even simply “plugging-in” the maximum likelihood estimate for P. For example, it turns out that using posterior predictive intervals to test overall model fit can be too conservative in the sense that they are guided by the same data used to evaluate the model, so the model is not rejected as often, in a frequency sense, as intended by the analyst (Bayarri 2003, Gelfand 2003). Robins et al. (2000) compared seven classical and “Bayesian” P value approaches for model assessment. Maximum likelihood (“plug-in”) P values were more accurate than posterior predictive P values. The best approaches were newer ones developed by Bayarri and Berger (2000), who also concluded that the P value from maximum likelihood estimates “seems superior” to the posterior predictive P value, “which would seem to contradict the common Bayesian intuition that it is better to account for parameter uncertainty by using a posterior than by” plugging in the maximum likelihood estimate.

The related topics of model selection and model validation or criticism are more important than ever with the Pandora’s box of computational model-fitting opened: we can fit a huge range of models to data—a good problem to have—but must avoid being misled by them. O’Hagan (2003) reviewed Bayesian “model criticism” starting from a “growing unease that the power of HSSS [hierarchical] modeling ... was tempting us to build models that we did not know how to criticize effectively.” To give ecologists context about using DIC for model selection (Cressie et al. 2009): Spiegelhalter et al. (2002) presented DIC “tentatively” on “somewhat heuristic” grounds that were motivated by classical theory. It is also motivated pragmatically by using only information available from MCMC output. AIC is defined for a hierarchical model in terms of the likelihood (1) and parameters, P, and BIC is an asymptotic estimate of a Bayesian marginal likelihood, but neither is available simply from MCMC output (de Valpine 2008). The discussants of Spiegelhalter et al. (2002) largely praised the DIC but raised many cautions and questions about its performance. These fields are sure to see both Bayesian and classical advances.

What should ecologists make of Cressie et al.’s claims that Bayesian methods incorporate information “in a coherent fashion” and give a “conceptually holistic approach to inference”? Based on other Bayesian usages of incorporating information “coherently” (Carlin and Louis 2000, Efron 2005), the first claim seems to refer largely to formulating a likelihood function that includes data related by random effects or from multiple sources (studies or sampling procedures), so it is really a feature of hierarchical models more than of Bayesian analysis. I interpret at least an aspect of the second appeal to refer to the handy practical situation that once you have a posterior sample of all P and E dimensions at your fingertips, much of the rest of your analysis involves summarizing it in various ways without worrying about breaking Bayesian rules. I think this is in part a valuable, practical view, and in part overoptimistic. I have already given the examples that calculating marginal likelihoods for model comparisons requires additional computations; that many conceptually meaningful P values for model criticism can be defined; that model selection will see further development; and that the easily generated posterior predictive intervals may represent over-fits to the data. As an example of a more advanced approach, Cressie et al. (2009) mention cross-validation, a form of...
which was recently used to criticize a Bayesian population model for conservation (Snover 2008).

The above variety of considerations illustrate that a thorough Bayesian analysis, beyond just a posterior, can lead to a fairly complicated set of results requiring careful judgment. This seems not very different from the situation in classical analysis that structured probability models define likelihoods, which provide a theoretical core connecting related analyses. An example of how the beguiling unity of treating both random effects and parameters as Bayesian “parameters” led to mistakes in justifying and using state-space fisheries models is explained by de Valpine and Hilborn (2005).

Cressie et al. (2009) give an example of the practical appeal of posterior probability summaries, such as 95% credible intervals, for an E value. Credible intervals are informative, but it is useful to note that such claims about their interpretation revolve around the contrast between degree-of-belief and frequency “probability.”

I will touch briefly on some other aspects of hierarchical model analysis set up by Cressie et al. While the challenge of generating sound MCMC results varies greatly between problems, I do not see computational convergence as a major concern for the overall approaches. For full Bayesian analysis, a practical way to assess the computational error for posterior summaries is the Moving Block Bootstrap (Mignani and Rosa 2001). The subjectivity of Bayesian priors seems to me a more serious issue, but not for the most commonly mentioned reasons. A more subtle problem than sensitivity to prior parameters is that there is no such thing as a universally flat prior, because flatness depends on the parameterization of the model. Flat priors for $\sigma^2$, $\sigma$, or $1/\sigma^2$ will all give different results, as will flat priors for $\lambda$ (population growth rate) or $\log(\lambda)$ in a population dynamics model. In many cases, the difference in results may be small, but nevertheless considering this issue should be a standard reason. A more subtle problem than sensitivity to prior parameters is that there is no such thing as a universally flat prior, because flatness depends on the parameterization of the model. Flat priors for $\sigma^2$, $\sigma$, or $1/\sigma^2$ will all give different results, as will flat priors for $\lambda$ (population growth rate) or $\log(\lambda)$ in a population dynamics model. In many cases, the difference in results may be small, but nevertheless considering this issue should be a standard step in applications. Finally, while I appreciate the contrast between “curve fitting” and “formal statistical modeling” (Cressie et al. 2009) as a gentle warm-up to the rationale for hierarchical models, the distinction seems limited. Even in a hierarchical model, one is estimating, or “fitting,” parameters of “curves” as reflected by the “smooth curve” language of Cressie et al.’s seal example; the advantage is doing so with a better model.

In summary, hierarchical models are an excellent framework for analyzing data, and Cressie et al. should go a long way towards helping ecologists adopt them. Learning to formulate and interpret hierarchical models involves learning to think clearly about stochastic relationships among variables in complex systems. Likelihood theory represents a major common ground for most Bayesian and classical analysis methods and is the reason they often give practically similar insights. Both classical and Bayesian approaches have pros and cons for pragmatism, performance, and interpretation. My comments are not intended to tally points in a debate, but rather to emphasize that inference and prediction from noisy data are very hard problems. Bayesian parameter distributions can give useful accountings of uncertainty in many contexts. However, some claims of Bayesian advantages based partly on appeals to intuition do not hold up to theoretical analysis, even if the intuitions have real merits.

While Bayesian approaches are practical for current use, many methods exist for maximum likelihood estimation and related analyses, including incorporation of parameter uncertainty, that would work well for many ecological hierarchical models. Gradually these approaches will allow choice of analysis philosophy to be based on scientific needs rather than having it tied by pragmatism to choice of a hierarchical model structure. Bayesian analysis will still be chosen for some important problems. It will be important for more ecologists to understand relationships between analysis methods to reach shared understanding of statistical results, and therefore of the evidence for hypotheses and support for predictions from data.

Acknowledgments

I thank Kevin Gross and Ray Hilborn for helpful comments on a draft. The perspectives and remaining flaws are my own.

Literature Cited


