A guide to Bayesian model checking for ecologists

Paul B. Conn[1,5], Mevin B. Hooten[2,3,4], Devin S. Johnson[1], and Peter L. Boveng[1]

[1]*National Marine Mammal Laboratory, NOAA, National Marine Fisheries Service, Alaska Fisheries Science Center, 7600 Sand Point Way NE, Seattle, WA 98115 USA*

[2]*U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, Fort Collins, CO 80523 USA*

[3]*Department of Fish, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, CO 80523 USA*

[4]*Department of Statistics, Colorado State University, Fort Collins, CO 80523 USA*

Abstract. Checking that models adequately represent data an essential component of applied statistical inference. Ecologists increasingly use hierarchical, Bayesian statistical models in their research. The appeal of this modeling paradigm is undeniable, as researchers can build and fit models that embody complex ecological processes while simultaneously controlling for potential biases arising from sampling artifacts. However, ecologists tend to be less focused on checking model assumptions and assessing potential lack-of-fit when applying Bayesian methods than when they apply frequentist methods such as maximum likelihood. There are also multiple ways of assessing goodness-of-fit for Bayesian models, each of which has strengths and weaknesses. For instance, in ecological applications, the "Bayesian p-value" is probably the most widely used approach for assessing lack of fit. Such p-values are relatively easy to compute, but they are well known

[5]Email: paul.conn@noaa.gov

to be conservative, producing p-values biased towards 0.5. Alternatively, lesser known approaches to model checking, such as prior predictive checks, probability integral transforms, and pivot discrepancy measures may produce more accurate characterizations of goodness-of-fit but are not as well known to ecologists. In addition, a suite of visual and targeted diagnostics can be used to examine violations of different model assumptions and lack-of-fit at different levels of the modeling hierarchy, and to check for residual temporal or spatial autocorrelation. In this review, we synthesize existing literature in order to guide ecologists to the many available options for Bayesian model checking. We illustrate methods and procedures with several ecological case studies, including i) explaining variation in spatio-temporal counts of bearded seals in the eastern Bering Sea, (ii) modeling the distribution of a herbaceous plant in in the Ozark Highlands of Missouri (USA), and (iii) using resource selection functions to model habitat preferences of XXX. We argue that model checking is an essential component of scientific discovery and learning that should accompany Bayesian analyses whenever they are used to analyze ecological datasets.

Bayesian p-value, Bayesian qq-plot, count data, goodness-of-fit diagnostic check, hierarchical model, model checking, occupancy, resource selection, pivot discrepancy, predictive distribution, probability interval transform, resource selection, spatio-temporal model

Introduction

Ecologists increasingly use Bayesian methods to analyze complex hierarchical models for natural systems (Hooten and Hobbs 2015). Adoption of a Bayesian perspective requires that one specify prior distributions for model parameters, a process some have criticized for

introducing unneeded subjectivity into the scientific process (Lele and Dennis 2009). However, there is a clear upside of making this Bayesian bargain: one can entertain models that were previously intractable using common modes of frequentist statistical inference (e.g., maximum likelihood). Ecologists are using Bayesian modes of inference to fit richer classes of models to their datasets, allowing them to model features such as temporal or spatial autocorrelation, individual level random effects, hidden states, and to separate the effects of process and measurement error (Link et al. 2002, Clark and Bjørnstad 2004, Cressie et al. 2009).

Ultimately, the reliability of inferences from a fitted model (Bayesian or otherwise) are dependent on how well the model approximates reality. There are multiple ways of assessing a model's performance in representing the system being studied. A first step is often to examine diagnostics that compare observed data to model output to pinpoint if and where any systematic differences occur. This process, which we term *model checking*, is an integral part of statistical inference, as it helps diagnose assumption violations and illuminate places where a model might be amended to more faithfully represent gathered data. Following this step, one might proceed to compare the performance of alternative models embodying different hypotheses using any number of model comparison or out-of-sample predictive performance metrics (see Hooten and Hobbs 2015, for a review) in order to gauge the support for alternative hypotheses or optimize predictive ability. Note that scientific inference can still proceed if models do not fit the data well, but conclusions need to be tempered; one approach in such situations is to compute a variance inflation factor to adjust precision levels downward (e.g. Burnham and Anderson 2002).

Non-Bayesian statistical software often include a suite of goodness-of-fit diagnostics that allow practitioners to assess how well different models fit their data. For instance,

when fitting generalized linear (McCullagh and Nelder 1989) or additive (Wood 2006) models in the R programming environment (R Development Core Team 2013), one can easily access diagnostics such as quantile-quantile, residual, and leverage plots. These diagnostics allow one to assess the reasonability of the assumed probability model, to examine whether there is evidence of heteroskedasticity, and to pinpoint outliers. Likewise, in capture-recapture analysis there are established procedures for assessing overall fit as well as departures from specific model assumptions which are codified in user-friendly software such as U-CARE (Choquet et al. 2009). Results of such goodness-of-fit tests are routinely reported when publishing analyses in the ecological literature.

Somehow, the implicit requirement that one conduct model checking exercises does seems not to apply when reporting results of Bayesian analyses in the ecological literature. For instance, a search of recent volumes of Ecology indicated that only 25% of articles employing Bayesian analysis on real datasets reported any model checking or goodness-of-fit testing (Fig. 1). We can think of several reasons why this might be the case. First, it likely has to do with momentum; the lack of precedent in ecological literature may lead some authors looking for templates on how to publish Bayesian analyses to conclude that model checking is unnecessary. Second, when researchers seek to publish new statistical methods, applications may be presented more as proof-of-concept exhibits than as definitive analyses that can stand up to scrutiny on their own. In such studies (and textbooks; see e.g., Royle and Dorazio 2008), topics like goodness-of-fit and model checking are often reserved for future research, presumably in journals with smaller impact factors. We (the authors) are certainly guilty of presenting our research in this fashion. Third, all of the articles we examined did a commendable job in reporting convergence diagnostics to support their contention that Markov chains from MCMC runs had reached their

stationary distribution. Perhaps there is a mistaken belief among authors and reviewers that convergence to a stationary distribution, combined with a lack of prior sensitivity, implies that a model fits the data? Finally, it may just be that those publishing Bayesian analyses in ecological literature ". . . like artists, have the bad habit of falling in love with their models" (to borrow a quote attributed to G.E.P. Box and referenced by Link and Barker (2010) with regard to model checking). We are certainly guilty of this fault as well. Indeed this monograph can be viewed as a partial atonement for unrequited love.

If we accept the premise that Bayesian models in ecology should be routinely checked for compatibility with data, a logical next question is how best to conduct such checks. Unfortunately, there is no single best answer. Most texts in ecology (e.g. King et al. 2009, Link and Barker 2010, Kéry and Schaub 2012) focus on posterior predictive checks, as pioneered by Guttman (1967), Rubin (1981, 1984) and Gelman et al. (1996) (among others). These procedures are also the main focus of the popular Bayesian analysis text by Gelman et al. (2004) and are based on the intuitive notion that data simulated from the posterior distribution should be similar to the data one is analyzing. However, "Bayesian p-values" generated from these tests tend to be conservative (biased towards 0.5) since the data are in effect used twice (once to fit the model and once to test the model; Bayarri and Berger 2000, Robins et al. 2000). By contrast, other approaches less familiar to ecologists (e.g. such as prior predictive checks, probability integral transforms, and pivot discrepancy measures) may produce more accurate characterizations of goodness-of-fit but may require extra data (for out-of-sample prediction), or may be more difficult to implement.

In this monograph, we seek to collate and digest relevant statistical literature with the goal of providing ecologists with a practical guide to Bayesian model checking. We start by defining a consistent notation that we use throughout the paper. Next, we work to compile

a bestiary of Bayesian model checking procedures, providing positives and negatives associated with each approach. After describing several ways in which model checking results can sometimes be misleading (as with hierarchically centered models), we illustrate Bayesian model checking using three case studies. These include a species distribution model (SDM) developed from bearded seal counts (*Erignathus barbatus*) in the Chukchi Sea, an SDM developed from presence-absence data of a herbaceous plant (*Genus species*) in Missouri, and analysis of animal telemetry data. We conclude with several recommendations on how model checking results should be presented in the ecological literature.

Background and notation

Model checking procedures

Posterior predictive checks (and the Bayesian p-value)

algorithm description

example of p-values not being uniformly distributed

More power to detect differences if discrepancy measure does not depend on unknown parameters (**?**, Chapter 8). Plug in (i.e. $\hat{\theta}$

Prior predictive checks

Box (1980) argued that the hypothetico-deductive process of scientific learning can be embodied through successive rounds of model formulation and testing. According to his view, models are built to represent current theory and an investigator's knowledge of the

system under study; data are then collected to evaluate how well the existing theory (i.e., model) matches up with reality. If necessary, the model under consideration can be amended, and the process repeats itself. These Model checking is an integral part of the scientific enterprise in the modern era, including the study of ecology. For instance,

From a Bayesian standpoint, such successive rounds of *estimation* and *criticism* are embodied through posterior inference and model checking, respectively (Box 1980).

If one views a model, complete with all its set of assumptions and prior beliefs

relationship to approximate Bayesian computation?

Direct chi-square testing

Johnson (2004)

Probability integral transforms

Pivoting discrepancy

Yuan and Johnson (2012)

Just build a bigger model! Tradeoffs between fit and prediction

Avoiding potential traps with model checking

Mean structure vs. dispersion - not always obvious where misspecification occurs.

Hierarchical centering

Examples

Modeling the distribution of a herbaceous plant

Spatio-temporal bearded seal counts

Resource selection of XXX

Discussion

Focus on prior sensitivity, convergence diagnostics and sometimes model comparison (e.g. DIC or cross validation) - not as much focus on GOF.

GOF on most general model, then model selection/comparison/averaging (**?**).

path structure - directional/separation tests

Acknowledgments

Funding for Bering Sea aerial surveys was provided by the U.S. National Oceanic and Atmospheric Administration and by the U.S. Bureau of Ocean Energy Management (Interagency Agreement M12PG00017). The views and conclusions in this article represent the views of the authors and the U.S. Geological Survey but do not necessarily represent findings or policy of the U.S. National Oceanic and Atmospheric Administration. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Literature Cited

Bayarri, M., and J. O. Berger. 2000. P values for composite null models. Journal of the American Statistical Association **95**:1127–1142.

Beaumont, M. A., W. Zhang, and D. J. Balding. 2002. Approximate Bayesian computation in population genetics. Genetics **162**:2025–2035.

Box, G. E. 1980. Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society. Series A (General) pages 383–430.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach, 2nd Edition. Springer-Verlag, New York.

Choquet, R., J.-D. Lebreton, O. Gimenez, A.-M. Reboulet, and R. Pradel. 2009. U-CARE: Utilities for performing goodness of fit tests and manipulating CApture–REcapture data. Ecography **32**:1071–1074.

Clark, J. S., and O. N. Bjørnstad. 2004. Population time series: process variability, observation errors, missing values, lags, and hidden states. Ecology **85**:3140–3150.

Cressie, N., C. Calder, J. Clark, J. Ver Hoef, and C. Wikle. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecological Applications **19**:553–570.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian Data Analysis, 2nd Edition. Chapman and Hall, Boca Raton.

Gelman, A., X.-L. Meng, and H. Stern. 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statistica sinica **6**:733–760.

Guttman, I. 1967. The use of the concept of a future observation in goodness-of-fit problems. Journal of the Royal Statistical Society. Series B (Methodological) pages 83–100.

Hooten, M., and N. Hobbs. 2015. A guide to Bayesian model selection for ecologists. Ecological Monographs **85**:3–28.

Johnson, V. E. 2004. A Bayesian χ^2 test for goodness-of-fit. Annals of Statistics pages 2361–2384.

Kéry, M., and M. Schaub. 2012. Bayesian population analysis using WinBUGS: a hierarchical perspective. Academic Press.

King, R., B. Morgan, O. Gimenez, and S. Brooks. 2009. Bayesian analysis for population ecology. CRC Press, Boca Raton, Florida.

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

Link, W., and R. Barker. 2010. Bayesian Inference with Ecological Applications. Academic Press, London, U.K.

Link, W., E. Cam, J. Nichols, and E. Cooch. 2002. Of BUGS and birds: Markov chain Monte Carlo for hierarchical modeling in wildlife research. Journal of Wildlife Management **66**:277–291.

Lunn, D., C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter. 2013. The BUGS book: A practical introduction to Bayesian analysis. Chapman & Hall/CRC, Boca Raton, Florida.

McCullagh, P., and J. A. Nelder. 1989. Generalized Linear Models. Chapman and Hall, New York.

R Development Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`.

Robins, J. M., A. van der Vaart, and V. Ventura. 2000. Asymptotic distribution of P values in composite null models. Journal of the American Statistical Association **95**:1143–1156.

Royle, J., and R. Dorazio. 2008. Hierarchical Modeling and Inference in Ecology. Academic Press, London, U.K.

Rubin, D. B. 1981. Estimation in parallel randomized experiments. Journal of Educational and Behavioral Statistics **6**:377–401.

Rubin, D. B., et al. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. The Annals of Statistics **12**:1151–1172.

Wood, S. N. 2006. Generalized additive models. Chapman & Hall/CRC, Boca Raton, Florida.

Yuan, Y., and V. E. Johnson. 2012. Goodness-of-fit diagnostics for Bayesian hierarchical models. Biometrics **68**:156–164.

218 TABLES

TABLE 1. Example discrepancy functions for predictive checks.

Name	Definition	
A. Omnibus discrepancy functions		
χ^2	$T(\mathbf{y}, \boldsymbol{\theta}) = \sum_i (y_i - E(y_i\|\boldsymbol{\theta}))^2 / \text{var}(y_i\|\boldsymbol{\theta})$	Often used for count data for count d
Deviance (D)	$T(\mathbf{y}, \boldsymbol{\theta}) = -2 \log[\mathbf{y}\|\boldsymbol{\theta}]$	
Likelihood ratio statistic	$T(\mathbf{y}, \boldsymbol{\theta}) = 2 \sum_i y_i \log(y_i / E(y_i\|\boldsymbol{\theta}))$	
B. Targeted discrepancy functions		
Proportion of zeros	$T(\mathbf{y}) = \sum_i I(y_i = 0)$	
Skewness checks	$T(\mathbf{y}) = y_{p\%}$	Using the $p\%$ quantile

FIGURE CAPTIONS

FIGURE 1. Type of model checking procedures used in $n = 31$ articles published in the journal Ecology during 2014 and 2015. Articles were found via a Web of Science for articles including the topic "Bayesian" (search conducted 10/1/2015). Six articles were determined to be non-applicable (N/A) because they either (1) were simulation studies, or (2) used approximate Bayesian computation, which is conceptually different than traditional Bayesian inference (see e.g. Beaumont et al. 2002). Of the remaining 25, 20 did not report any model checking procedures. Five articles reported specific model checking procedures, which included a combination of Bayesian cross validation (*Cross.val*,), frequentist software (*Non-Bayes*, posterior predictive p-values (*Pp.pval*), and posterior predictive graphical checks (*Pp.gc*).

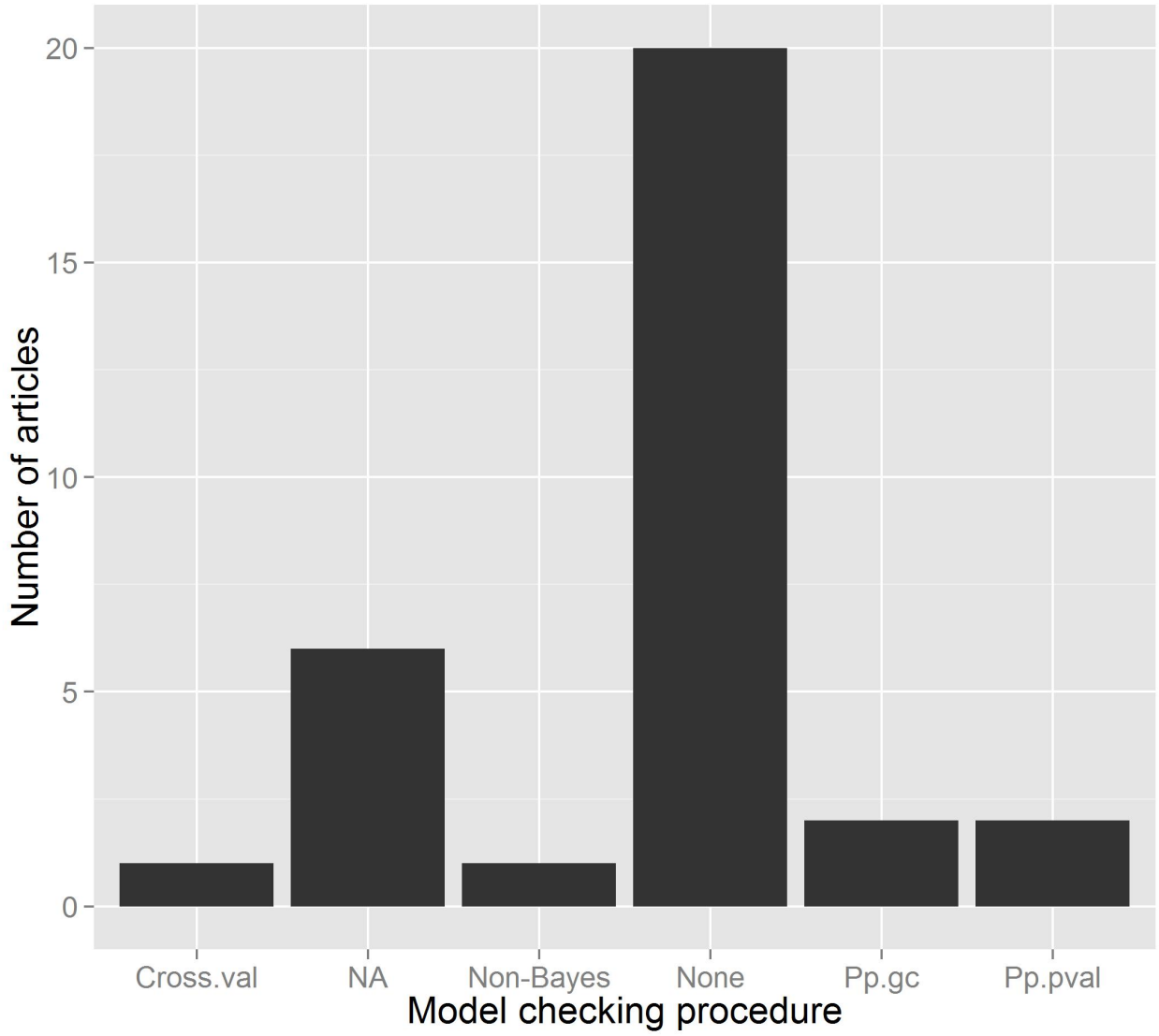
20
15
10
5
0
Number of articles
Cross.val
NA
Non-Bayes
None
Pp.gc
Pp.pval
Model checking procedure

Fig 1