

Multimodel approaches are not the best way to understand multifactorial systems

Ben Bolker

10:58 05 January 2018

Abstract

Information-theoretic (IT) and multi-model averaging (MMA) statistical approaches are suboptimal tools for pursuing a multifactorial approach (the “method of multiple working hypotheses”) in ecology. (1) Conceptually, IT encourages ecologists to perform tests on sets of straw-man models. (2) MMA improves on IT model comparison; it represents a simple form of *shrinkage estimation* (a way to make accurate predictions from a model with many parameters, by “shrinking” parameter estimates toward zero). However, newer approaches to shrinkage estimation are faster and have more rigorous statistical underpinnings. (3) A major disadvantage of shrinkage methods is that confidence intervals are extremely difficult to derive. If researchers want accurate estimates of the strength of multiple competing ecological processes along with reliable confidence intervals, their best hope is to use full (maximal) statistical models after making principled, *a priori* decisions about which predictors to include.

Many modern ecological and evolutionary studies try to quantify the strength and importance of multiple processes in ecological systems: for example, the effects of herbivory and fertilization on standing biomass (Gruner et al. 2008); the effects of bark, wood density, and fire on tree mortality (Brando et al. 2012); or the effects of taxonomic and genomic position on evolutionary rates (Ghenu et al. 2016). This *multifactorial* approach (McGill 2016) complements, rather than replacing, the traditional hypothesis-testing or strong-inferential framework (Platt 1964; Fox 2016).¹

¹While there is much interesting debate over the best methods for gathering evidence to distinguish among two or more particular, *intrinsically* discrete hypotheses (Taper and

A standard statistical approach to analyzing multifactorial systems, particularly common in wildlife and conservation ecology, goes as follows: (1) Construct a full model that encompasses as many of the processes (and their interactions) as is feasible. (2) Fit the full model and check that it describes the data reasonably well (e.g. by computing R^2 values or estimating overdispersion). (3) Construct all, or most, of the possible submodels of the full model by setting subsets of parameters to zero. (4) If one model clearly dominates the ensemble of models, draw conclusions based on it. Otherwise, either (a) use multi-model averaging to estimate model-averaged parameters and confidence intervals or (b) draw conclusions about the importance of different processes either informally, by seeing which parameters are contained in the best (lowest-AIC) models, or formally, by summing the AIC weights. [CITATION?] I claim that this approach, even if used sensibly (e.g. with reasonably small numbers of predictors and submodels)

Our goal is to tease apart the contributions of many processes, *all* of which we believe are affecting the populations and communities we study to some degree. If our scientific question is (something like) “How important is this factor, in an absolute sense or relative to other factors?”, not “Which of these factors are actually doing *anything at all* in my system?”, why are we working so hard to fit many models of which only one (the full model) incorporates all of the factors? If we are not really interested (at the moment) in particular discrete hypotheses about our system, why does so much of our data-analytic effort go into various ways to test between, or combine and reconcile, multiple discrete models? This is an “XY problem”²: rather than thinking about the best way to solve our real problem X (understanding multifactorial systems), we have gotten bogged down in the details of how to make a particular tool, Y (multimodel approaches) provide the answers we need.

One legitimate reason to fit multiple models is as one step in a null-hypothesis significance testing (NHST) procedure. While much maligned, NHSTs are a useful part of data analysis — *not* to decide whether we really think a null hypothesis is false (they almost always are), but to see if we can reliably determine the *direction of effects* of ecological processes — that is, not whether a parameter is zero, but whether we can tell unequivocally that it is negative or positive³. We can perform these tests by statistically comparing

Ponciano 2015), that is not my focus here.

²<http://www.perlmonks.org/?node=XY+Problem>

³Thanks to J. Dushoff for this perspective on NHST.

a full model to a reduced model that pretends the effect is exactly zero.

However, ecologists pursuing multimodel approaches are not just fitting one-step-reduced models to test hypotheses; they are fitting *all* of the reduced models. Their usual motivation is a hope that multimodel averaging will help them deal with insufficient data in a multifactorial world. If we had enough information (even “big data” doesn’t always provide as much information as we need), we could fit just the full model, drawing our conclusions from the estimates and confidence intervals (CIs) with all of the factors considered simultaneously. But we always have too many predictors, and not enough data; we don’t want to overfit (which will inflate our CIs and p-values to the point where we can’t tell anything for sure), but at the same time we are scared of neglecting potentially important effects.

Stepwise regression, the original strategy for separating signals from noise, is now widely deprecated (Harrell 2001; Whittingham et al. 2006)⁴. Since the mid-1990s, ecologists have adopted information-theoretic approaches (Burnham and Anderson 1998). These tools mitigate the instability of stepwise approaches, allow simultaneous comparison of many, non-nested models, and avoid the stigma of NHST. A further step forward, multi-model averaging (Burnham and Anderson 2002) accounts for model uncertainty and avoids focusing on a single best model. More recently, however, model averaging is experiencing a backlash, as statistically savvy ecologists point out that multimodel averaging runs into trouble whenever variables are collinear (Freckleton 2011); when we are careless about the meaning of main effects in the presence of interactions; when we average model parameters on the way to predicting nonlinear functions of the parameters (Cade 2015); and when we use summed model weights to assess the relative importance of predictors (Galipaud et al. 2014).

Rather than using information criteria as tools to identify the best predictive model, or to obtain the best overall (model-averaged) predictions, most users of information criteria are using them either to (dubiously) quantify variable importance (Galipaud et al. 2014), or, by multimodel averaging, to have their cake and eat it too — to avoid either over- or underfitting while quantifying effects in multifactorial systems. They encounter two problems, one conceptual and one practical.

⁴Although it may sometimes be adequate for selecting a single best model for prediction (Murtaugh 2009).

Conceptually, many of the difficulties of model averaging come from the original sin of unnecessarily dichotomizing a continuous world. Suppose we want to understand the effects of temperature and precipitation on biodiversity. The model-comparison or model-averaging approach would construct five models: a null model with no effects of either temperature or precipitation, two single-factor models, an additive model, and a full model allowing for interactions between temperature and precipitation. We would then fit all of these models and then model-average their parameters. We might be doing this in an effort to get good predictions, or to test our confidence that we know the signs of particular effects (measured in the context of whatever processes are included in the reduced and the full models), but they are only means to an end, and we certainly shouldn't fool ourselves into thinking that we are using the method of multiple working hypotheses. For example, Chamberlin argued that in teaching about the origin of the Great Lakes we should urge students "to conceive of three or more great agencies [pre-glacial erosion, glacial erosion, crust deformation] working successively or simultaneously, and to estimate how much was accomplished by each of these agencies." Chamberlin was *not* suggesting that we test which individual mechanism or combination of mechanism fits the data best (in whatever sense), but instead that we acknowledge that the world is multifactorial.

Practically, model averaging is slow. Individual models can take minutes or hours to fit, and we may have to fit dozens or scores of sub-models in the multi-model averaging process. There are efficient tools available for fitting "right-sized" models that avoid many of the technical problems of model averaging. Penalized methods such as ridge and lasso regression (Dahlgren 2010) are well known outside of ecology; in a Bayesian setting, informative priors centered at zero have the same effect of *regularizing* — pushing weak effects toward zero and controlling model complexity. Developed specifically for optimal (predictive) fitting in models with many parameters, these models have well-understood statistical properties; they avoid the pitfalls of model-averaging correlated or nonlinear parameters; and, by avoiding the need to fit many sub-models in the model-averaging processes, they are much faster.⁵

Here I am not tackling the issue of whether 'truth' is included in our model set (it isn't), and how this matters to our inference (Barker and Link 2015). I am claiming the opposite, that our full model is usually about as close to

⁵Although they often require a computationally expensive cross-validation step in order to choose the degree of penalization.

truth as we can get; we don't really believe any of the less complex models. If we are trying to get the best predictions, or to compare the strength of various processes in a multifactorial context, there may be better ways to do it. In situations where we really want to compare qualitatively different, non-nested hypotheses (Luttbeg, Langen, and Adams 2004), AIC or BIC or any appropriate model-comparison tool is fine; however, if the models are *really* qualitatively different, perhaps we shouldn't be trying to merge them by averaging.

Penalized models have their own problems. Although powerful computational tools exist for fitting penalized versions of linear and generalized linear models (e.g. the `glmnet` package for R) and mixed models (`glmmLasso`), software for some of the more exotic models used by ecologists (e.g. zero-inflated models) may not be readily available. Fitting these models requires the user to choose the degree of penalization; although this process is conveniently automated in tools like `glmnet`, it may be tricky for data that are correlated in space or time (Wenger and Olden 2012). Finally, computing CIs for parameters in penalized models — one of the most basic outputs we need from a statistical analysis of a multifactorial system — is a current research problem; statisticians have proposed methods for deriving CIs (Pötscher and Schneider 2008; Lee et al. 2013; Javanmard and Montanari 2014; Lockhart et al. 2014), but they are far from being standard options in software. Ecologists should encourage their quantitatively savvy friends to build tools that make penalized approaches easier to use.

Statisticians derived confidence intervals for ridge regression long ago (Obenchain 1977) — but, paradoxically, they are identical to the confidence intervals one would have gotten from the full model without penalization! More recently, work by Turek improving and evaluating different methods for constructing multi-model averaged confidence intervals (D. Turek and Fletcher 2012; Fletcher and Turek 2012; D. B. Turek 2013; D. Turek 2015) shows that multi-model averaged confidence intervals are nearly always too narrow; in simulations, 95% confidence intervals constructed according to even the best multi-model averaging algorithm may typically include the true parameter values only about 80% of the time. Free lunches do not exist in statistics, any more than anywhere else. We can use penalized approaches to improve prediction accuracy without having to sacrifice any input variables (by trading bias for variance), but the *only* way to gain statistical power for testing hypotheses, or narrowing or uncertainty about our predictions, is to limit the scope of our models *a priori* (Harrell 2001) — or to collect more data.

If we have good experimental designs and sensible scientific questions, muddling through with existing techniques will give us reasonable results, most of the time (Murtaugh 2009). But ecologists should be aware that the roundabout statistical methods they currently rely on to understand multifactorial systems are *not* designed for those purposes; they were developed for prediction rather than inference. When prediction is the primary goal, penalized methods can work better (faster and with better-known properties) than multimodel averaging. When estimating the magnitude of effects or judging variable importance, penalized methods may be appropriate — or we may have to go back to the difficult choice of focusing on a restricted number of variables for which we have enough data to fit and interpret the full model.

Acknowledgments

Dana Karelus, Daniel Turek, Jeff Walker, ...

Additional refs

- (Fieberg and Johnson 2015; Walker 2017; Nilsen, Asche, and Tveterås 2005; Lukacs, Burnham, and Anderson 2010)
- shrinkage estimators: van Houwelingen 2001 *Statistica Neerlandica*, Hooten and Hobbs 2015 *Eco Monographs* (Bayesian model selection)
- multi-model averaging: Claeskens and Carroll 2007 *Biometrika*, Wang and Zhou 2013 *Communications in Statistics*; Claeskens and Hjort (2008)
- note critique of Galipaud et al.

Interestingly, the model averaged variance is smaller than the stepwise selected variance, yet produces a confidence interval with better coverage because of the reduced bias. Therefore, model averaging was better here than stepwise selection in terms of both bias and coverage in the case of weakly correlated variables.

From Kabaila, Welsh, and Abeysekera (2016): > An ideal confidence interval should have minimal coverage equal to its nominal coverage and, to show a benefit of model selection, have scaled expected length that (a) is substantially less than 1 under M_1 , (b) has a maximum value that [is] not much larger than 1 and (c) is close to 1 if the data happens to strongly contradict

the model M_1 . The results of the present paper and the fact that the Hjort & Claeskens' (2003) interval turns out to be essentially the same as the standard confidence interval based on fitting a full model suggest that this will be very difficult to achieve using model averaged confidence intervals. In other words, it seems difficult to find model-averaged confidence intervals that compete successfully with the standard confidence interval based on the full model.

References

- Barker, Richard J., and William A. Link. 2015. "Truth, Models, Model Sets, AIC, and Multimodel Inference: A Bayesian Perspective." *The Journal of Wildlife Management* 79 (5): 730–38. doi:10.1002/jwmg.890.
- Brando, P.M., D.C. Nepstad, J.K. Balch, B. Bolker, M.C. Christman, M. Coe, and F.E. Putz. 2012. "Fire-Induced Tree Mortality in a Neotropical Forest: The Roles of Bark Traits, Tree Size, Wood Density and Fire Behavior." *Global Change Biology* 18 (2): 630–41. doi:10.1111/j.1365-2486.2011.02533.x.
- Burnham, Kenneth P., and David R. Anderson. 1998. *Model Selection and Inference: A Practical Information-Theoretic Approach*. New York: Springer.
- . 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer.
- Cade, Brian S. 2015. "Model Averaging and Muddled Multimodel Inference." *Ecology*, March. doi:10.1890/14-1639.1.
- Claeskens, Gerda, and Nils Lid Hjort. 2008. *Model Selection and Model Averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. https://www.amazon.com/dp/0521852250/_encoding=UTF8?coliid=I3GVZ4E3JY7YLN&colid=3TNO30DO5701R.
- Dahlgren, Johan P. 2010. "Alternative Regression Methods Are Not Considered in Murtaugh (2009) or by Ecologists in General." *Ecology Letters* 13 (5): E7–E9. doi:10.1111/j.1461-0248.2010.01460.x.
- Fieberg, John, and Douglas H. Johnson. 2015. "MMI: Multimodel Inference or Models with Management Implications?: Multimodel Inference and Models for Management." *The Journal of Wildlife Management* 79 (5): 708–18.

doi:10.1002/jwmg.894.

Fletcher, David, and Daniel Turek. 2012. "Model-Averaged Profile Likelihood Intervals." *Journal of Agricultural, Biological, and Environmental Statistics* 17 (1). Springer: 38–51.

Fox, Jeremy. 2016. "Why Don't More Ecologists Use Strong Inference?" *Dynamic Ecology*. <https://dynamicecology.wordpress.com/2016/06/01/obstacles-to-strong-inference-in-ecology/>.

Freckleton, Robert P. 2011. "Dealing with Collinearity in Behavioural and Ecological Data: Model Averaging and the Problems of Measurement Error." *Behavioral Ecology and Sociobiology* 65 (1): 91–101.

Galipaud, Matthias, Mark A. F. Gillingham, Morgan David, and François-Xavier Dechaume-Moncharmont. 2014. "Ecologists Overestimate the Importance of Predictor Variables in Model Averaging: A Plea for Cautious Interpretations." *Methods in Ecology and Evolution* 5 (10): 983–91. doi:10.1111/2041-210X.12251.

Ghenu, Ana-Hermina, Benjamin M. Bolker, Don J. Melnick, and Ben J. Evans. 2016. "Multicopy Gene Family Evolution on Primate Y Chromosomes." *BMC Genomics* 17: 157. doi:10.1186/s12864-015-2187-8.

Gruner, D. S., J. E. Smith, E. W. Seabloom, S. A. Sandin, J. T. Ngai, H. Hillebrand, W. S. Harpole, et al. 2008. "A Cross-System Synthesis of Consumer and Nutrient Resource Control on Producer Biomass." *Ecology Letters* 11 (7): 740–55.

Harrell, Frank. 2001. *Regression Modeling Strategies*. Springer.

Javanmard, Adel, and Andrea Montanari. 2014. "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression." *The Journal of Machine Learning Research* 15 (1): 2869–2909. <http://dl.acm.org/citation.cfm?id=2697057>.

Kabaila, Paul, A. H. Welsh, and Waruni Abeysekera. 2016. "Model-Averaged Confidence Intervals." *Scandinavian Journal of Statistics* 43 (1): 35–48. doi:10.1111/sjos.12163.

Lee, Jason D., Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. 2013. "Exact Post-Selection Inference with the Lasso," November. <http://xxx.tau.ac.il/abs/1311.6238v4>.

Lockhart, Richard, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani.

2014. "A Significance Test for the Lasso." *Annals of Statistics* 42 (2): 413. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4285373/>.

Lukacs, Paul M., Kenneth P. Burnham, and David R. Anderson. 2010. "Model Selection Bias and Freedman's Paradox." *Annals of the Institute of Statistical Mathematics* 62 (1): 117–25. doi:10.1007/s10463-009-0234-4.

Luttbeg, Barney, Tom A. Langen, and Associate Editor: Eldridge S. Adams. 2004. "Comparing Alternative Models to Empirical Data: Cognitive Models of Western Scrub-Jay Foraging Behavior." *The American Naturalist* 163 (2): 263–76. doi:10.1086/381319.

McGill, Brian. 2016. "Why Ecology Is Hard (and Fun) – Multicausality." *Dynamic Ecology*. <https://dynamicecology.wordpress.com/2016/03/02/why-ecology-is-hard-and-fun-multicausality/>.

Murtaugh, Paul A. 2009. "Performance of Several Variable-Selection Methods Applied to Real Ecological Data." *Ecology Letters* 12 (10): 1061–8. doi:10.1111/j.1461-0248.2009.01361.x.

Nilsen, Odd Bjarte, Frank Asche, and Ragnar Tveterås. 2005. "Confidence Intervals for the Shrinkage Estimator." Working paper. SNF. <https://brage.bibsys.no/xmlui/handle/11250/165518>.

Obenchain, R. 1977. "Classical F -Tests and Confidence Regions for Ridge Regression." *Technometrics* 19 (4): 429–39.

Platt, John R. 1964. "Strong Inference." *Science*, New Series, 146 (3642): 347–53. doi:10.2307/1714268.

Pötscher, Benedikt M., and Ulrike Schneider. 2008. "Confidence Sets Based on Penalized Maximum Likelihood Estimators." MPRA Paper. <http://mpra.ub.uni-muenchen.de/16013/>.

Taper, Mark L., and José Miguel Ponciano. 2015. "Evidential Statistics as a Statistical Modern Synthesis to Support 21st Century Science." *Population Ecology* 58 (1): 9–29. doi:10.1007/s10144-015-0533-y.

Turek, Daniel. 2015. "Comparison of the Frequentist MATA Confidence Interval with Bayesian Model-Averaged Confidence Intervals." *Journal of Probability and Statistics* 2015. Hindawi Publishing Corporation.

Turek, Daniel Bernard. 2013. "Frequentist Model-Averaged Confidence Intervals." PhD thesis, University of Otago.

Turek, Daniel, and David Fletcher. 2012. "Model-Averaged Wald Confidence

Intervals." *Computational Statistics & Data Analysis* 56 (9). Elsevier: 2809–15.

Walker, Jeffrey A. 2017. "A Defense of Model Averaging." *BioRxiv*, May, 133785. doi:10.1101/133785.

Wenger, Seth J., and Julian D. Olden. 2012. "Assessing Transferability of Ecological Models: An Underappreciated Aspect of Statistical Validation." *Methods in Ecology and Evolution* 3 (2): 260–67. doi:10.1111/j.2041-210X.2011.00170.x.

Whittingham, Mark J., Philip A. Stephens, Richard B. Bradbury, and Robert P. Freckleton. 2006. "Why Do We Still Use Stepwise Modelling in Ecology and Behaviour?" *Journal of Animal Ecology* 75 (5): 1182–9.