entropy

MDPI

Article

Multimodel Approaches Are Not the Best Way to Understand Multifactorial Systems

Benjamin M. Bolker

Departments of Mathematics & Statistics and Biology, McMaster University, Hamilton, ON L8S4K1, Canada; bolker@mcmaster.ca

Abstract: Information-theoretic (IT) and multi-model averaging (MMA) statistical approaches are widely used but suboptimal tools for pursuing a multifactorial approach (also known as the method of multiple working hypotheses) in ecology. (1) Conceptually, IT encourages ecologists to perform tests on sets of artificially simplified models. (2) MMA improves on IT model selection by implementing a simple form of shrinkage estimation (a way to make accurate predictions from a model with many parameters relative to the amount of data, by "shrinking" parameter estimates toward zero). However, other shrinkage estimators such as penalized regression or Bayesian hierarchical models with regularizing priors are more computationally efficient and better supported theoretically. (3) In general, the procedures for extracting confidence intervals from MMA are overconfident, providing overly narrow intervals. If researchers want to use limited data sets to accurately estimate the strength of multiple competing ecological processes along with reliable confidence intervals, the current best approach is to use full (maximal) statistical models (possibly with Bayesian priors) after making principled, a priori decisions about model complexity.

Keywords: null-hypothesis significance testing; multi-model averaging; shrinkage estimators; Akaike information criterion; statistical inference

Citation: Bolker, B.M. Multimodel Approaches Are Not the Best Way to Understand Multifactorial Systems. *Entropy* **2024**, *26*, 506. https://doi.org/10.3390/e26060506

Academic Editors: Brian Dennis, Mark L. Taper and Jose Miguel Ponciano

Received: 19 April 2024
Revised: 25 May 2024
Accepted: 31 May 2024
Published: 11 June 2024

Modern scientific research often aims to quantify the effects of multiple simultaneously operating processes in natural or human systems. Some examples from my own work in ecology and evolution consider the effects of herbivory and fertilization on standing biomass [1]; the effects of bark, wood density, and fire on tree mortality [2]; or the effects of taxonomic and genomic position on evolutionary rates [3]. This multifactorial approach [4] complements, rather than replacing, the traditional hypothesis-testing or strong-inferential framework [5–7]. (While there is much interesting debate over the best methods for gathering evidence to distinguish among two or more particular, intrinsically discrete hypotheses [8], that is not the focus of this paper.) Such attempts to quantify the magnitude or importance of different processes also differ from predictive modeling, which dominates the fields of machine learning and artificial intelligence [9]. The prediction and quantification of process strength are closely related—if we can accurately predict outcomes over a range of conditions, then we can also predict the effects of changes in those conditions, and hence infer the strengths of processes, if the changes we are trying to predict are adequately reflected in our training data. However, predictive modelers are usually primarily concerned with predictions within the natural range of conditions, which may not provide us enough information to reliably make inferences about processes. The paper focuses on statistical modeling for estimation and inference, rather than prediction.

A standard approach to analyzing multifactorial systems, particularly common in ecology, is 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 make sure that it describes the data reasonably well (e.g., by examining model diagnostics and by ensuring that the level of unexplained variation is not unacceptably large). (3) Construct possible submodels of the full model by setting the subsets of parameters to zero. (4) Compute the information-theoretic measures of quality, such as the Akaike or Bayesian/Schwarz information criteria,

for every submodel. (5) Use multi-model averaging (MMA) to estimate model-averaged parameters and confidence intervals (CIs), and possibly draw conclusions about the importance of different processes by summing the information-theoretic weights [10]. I argue that this approach, even if used sensibly as advised by proponents of the approach (e.g., with reasonable numbers of candidate submodels), is a poor way to approach estimation and inference for multifactorial problems.

For example, suppose we want to understand the effects of ecosystem-level net primary productivity and fire intensity on species diversity (a simplified version of the analysis done in [11]). The model-comparison or model-averaging approach would construct five models: a null model with no effects of either productivity or fire, two single-factor models, an additive model, and a full model allowing for interactions between productivity and fire. We would then fit all of these models and model-average their parameters, and derive model-averaged confidence intervals.

The goal of a multifactorial analysis is to tease apart the contributions of many processes, all of which we believe are affecting our study system to some degree. If our scientific questions are (something like) "How important is this factor, in an absolute sense or relative to other factors?" (or equivalently, "How much does a change in this factor change the system in absolute or relative terms?"), rather than "Which of these factors are having any effect at all on 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 do not have particular, a priori discrete hypotheses about our system (such as "process A influences the outcome but process B has no effect at all"), why does so much of our data-analytic effort go into various ways to test between, or combine and reconcile, multiple discrete models? In software development, this is called an "XY problem" (http://www.perlmonks.org/?node=XY+Problem, accessed on 30 May 2024): rather than thinking about the best way to solve our real problem X (understanding multifactorial systems), we have become bogged down in the details of how to make a particular tool, Y (multimodel approaches), provide the answers we need. Most critiques of MMA address technical concerns such as the influence of unobserved heterogeneity [12] or criticize the misuse of information-theoretic methods by researchers [13,14], but do not ask why we are comparing discrete models in the first place.

In contrast with averaging across discrete hypotheses or treating a choice of discrete hypotheses as an end goal, fitting and comparing multiple models as a step in a null-hypothesis significance testing (NHST) procedure is defensible. In the biodiversity analysis described above, we might fit the full model and then assess the significance of individual terms by comparing the fit of the full model to models with those terms dropped (taking particular care with the interpretation of dropping a lower-level effect in models with interactions, e.g., see [15]). 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 distinguish signal from noise. Another interpretation is that NHSTs can test whether we can reliably determine the direction of effects—that is, not whether the effect of a predictor on some process is zero, but whether we can tell unequivocally that it has a particular sign—positive or negative [16,17].

However, researchers using multimodel approaches are not fitting one-step-reduced models to test hypotheses; rather, they are fitting a wide range of submodels, typically in the hope that model choice or multimodel averaging will help them deal with insufficient data in a multifactorial world. If we had enough information (even Big Data does not always provide the information we need [18], we could fit only the full model, drawing our conclusions from the estimates and CIs with all of the factors considered simultaneously. But we nearly always have too many predictors, and not enough data; we do not want to overfit, (which will inflate our CIs and p-values to the point where we cannot tell anything for sure), but at the same time we are afraid of neglecting potentially important effects.

Stepwise regression, the original strategy for separating signals from noise, is now widely deprecated because it interferes with correct statistical inference [19–22]. Information-theoretic tools mitigate the instability of stepwise approaches, allow for the simultaneous

comparison of many non-nested models, and avoid the stigma of NHST. A further step forward, multi-model averaging [10], accounts for model uncertainty and avoids focusing on a single best model. Some forms of model averaging provide shrinkage estimators; averaging the strength of effects between models where they are included and models where they are absent adjusts the estimated effects toward zero [14]. More recently, model averaging is experiencing a backlash, as studies point out that multimodel averaging may run into trouble when variables are collinear and/or have differential levels of measurement error [23], when we are careless about the meaning of main effects in the presence of interactions, when we average model parameters rather than model predictions [14], or when we use summed model weights to assess the relative importance of predictors ([14,24]; but cf. [25]).

Freckleton [23] makes the point that model averaging will tend to shrink the estimates of multicollinear predictors toward each other, so that estimates of weak effects will be biased upward and estimates of strong effects will be biased downward. This is an unsurprising (in hindsight) consequence of shrinkage estimation. With other analytical methods such as lasso regression, or selection of a single best model by AIC, the weaker of two correlated predictors, or more precisely the one that appears weaker based on the available data, could be eliminated entirely, leading all of its effects to be attributed to the stronger predictor. Researchers often make a case for dropping correlated terms in this way because collinearity of predictors inflates parameter uncertainty and complicates interpretation. However, others have repeatedly pointed out that collinearity is a problem of intrinsic uncertainty—we are simply missing the data that would tell us which combination of collinear factors really drives the system. The confidence intervals of parameters from a full model estimated by regression or maximum likelihood will correctly identify this uncertainty; modeling procedures that automatically drop collinear predictors (by model selection or sparsity-inducing penalization) not only fail to resolve the issue, but can lead to inaccurate predictions based on new data [26–29]. A full model might (correctly) tell us we cannot confidently assess whether either productivity or fire decrease or increase species diversity, because their estimated effects are strongly correlated. However, by comparing the fit of the full model to one that dropped both productivity and fire, we could conclude that their joint effect is highly significant.

In ecology, information criteria were introduced by applied ecologists who were primarily interested in making the best possible predictions to inform conservation and management; they were less concerned with inference or quantifying the strength of underlying processes [10,30,31]. Rather than using information criteria as tools to identify the best predictive model, or to obtain the best overall (model-averaged) predictions, most current users of information-theoretic methods use them either to quantify variable importance, or, by multimodel averaging, to have their cake and eat it too—to avoid either over- or underfitting while quantifying the effects in multifactorial systems. There are two problems with this approach—one conceptual and one practical.

The conceptual problem with model averaging reflects the original sin of unnecessarily discretizing a continuous model space. When we fit many different models as part of our analytical process (based on selection or averaging), the models are only a means to an end; despite the claims of some information-theoretic modelers, we are not really using the submodels in support of the method of multiple working hypotheses as described by Chamberlin [32]. 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 [preglacial 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 mechanisms fits the data best (in whatever sense), but instead that we acknowledge that the world is multifactorial. In a similar vein, Gelman and Shalizi [33] advocate "continuous model expansion", creating models that include all components of interest (with appropriate

Bayesian priors to constrain the overall complexity of the model) rather than selecting or averaging across discrete sets of models that incorporate subsets of the processes.

Here, I am not concerned whether 'truth' is included in our model set (it is not), and how this matters to our inference [34,35]. I am claiming the opposite, that our full model—while certainly not the true model—is usually the closest thing we have to a true model. This claim seems to contradict the information-theoretic result that the best approximating model (i.e., the minimum-AIC model) is expected to be closest to the true (generating) model in a predictive sense (i.e., it has the smallest expected Kullback–Leibler distance) [36]. However, the fact that excluding some processes allows the fitted model to better match the observation does not mean that we should believe these processes are not affecting on our system—just that, with the available data, dropping terms will provide us better predictions than keeping the full model. If we are primarily interested in prediction, or in comparing qualitatively different, possibly non-nested hypotheses [37], information-theoretic methods match our goals well.

The technical problem with model averaging is its computational inefficiency. 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 [38] are well known in some scientific fields; in a Bayesian setting, informative priors centered at zero have the same effect of regularizing—pushing weak effects toward zero and controlling model complexity (more or less synonymous with the shrinkage of estimates described above) [39]. Developed for optimal (predictive) fitting in models with many parameters, penalized 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. (However, they may require a computationally expensive cross-validation step in order to choose the degree of penalization.) Furthermore, penalized approaches underlie modern nonparametric methods such as additive models and Gaussian processes that allow models to expand indefinitely to match the available data [40,41].

Penalized models have their own challenges. A big advantage of information-theoretic methods is that, like wrapper methods for feature selection in machine learning [42], we can use model averaging, as long as we can fit component models and extract the log-likelihood and number of parameters—we never need to build new software. 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`), quantile regression [43], software for some more exotic models (e.g., zero-inflated models, models for censored data) may not be readily available. Fitting these models requires the user to choose the strength of penalization. This process is conveniently automated in tools like `glmnet`, but correctly assessing the out-of-sample accuracy (and hence the correct level of penalization) is tricky for data that are correlated in space or time [44,45]. Penalization (or regularization) can also be achieved by imposing Bayesian priors on subsets of parameters [46], but this converts the choice of strength of penalization to a similarly challenging choice of appropriate priors.

Finally, frequentist inference (computing p-values and CIs) for parameters in penalized models—one of the basic outputs we want from a statistical analysis of a multifactorial system—is a current research problem; statisticians have proposed a variety of methods [47–50], but they typically make extremely strong asymptotic assumptions and are far from being standard options in software. Scientists should encourage their friends in statistics and computer science to build tools that make penalized approaches easier to use.

Statisticians derived confidence intervals for ridge regression long ago [51]—surprisingly, they are identical to the confidence intervals one would have achieved from the full model without penalization. Wang and Zhou [52] similarly proved that model-averaging CIs derived as suggested by Hjort and Claeskens [53] are asymptotically (i.e., for arbitrarily large data sets) equivalent to the CIs from the full model. Analytical and simulation

studies [54–59] have shown that a variety of alternative methods for constructing CIs are overoptimistic, i.e., that they generate too-narrow confidence intervals with coverage lower than the nominal level. Simulations from several of the studies above show that MMA confidence intervals constructed according to the best known procedures typically include the true parameter values only about 80% or 90% of the time. In particular, Kabaila et al. [58] say that constructing CIs that take advantage of shrinkage but still achieve correct coverage will be very difficult to achieve using model averaged confidence intervals. (The only examples I have been able to find of MMA confidence intervals with close to nominal coverage are from Chapter 5 of [10].) In short, it seems difficult to find model-averaged confidence intervals that compete successfully with the standard confidence interval based on the full model.

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 known way to gain statistical power for testing hypotheses, or narrowing our uncertainty about our predictions, is to limit the scope of our models a priori [19], to add information from pre-specified Bayesian priors (or equivalent regularization procedures), or to collect more data. Burnham and Anderson [60] defined a "savvy" prior that reproduces the results of AIC-based multimodel averaging in a Bayesian framework, but it is a weak conceptual foundation for understanding multifactorial systems. Because it is a prior on discrete models, rather than on the magnitude of continuous parameters that describe the strength of different processes, it induces a spike-and-slab type prior on parameters that assigns a positive probability to the unrealistic case of a parameter being exactly zero; furthermore, the prior will depend on the particular set of models being considered.

Multimodel averaging is probably most popular in ecology (in May 2024, Google Scholar returned $\approx$ 65,000 hits for "multimodel averaging" alone and 31,000 for "multimodel averaging ecology"). However, multifactorial systems—and the problems of approaching inference through comparing and combining discrete models that consider artificially limited subsets of the processes we know are operating — occur throughout the sciences of complexity, those involving biological and human processes. In psychology, economics, sociology, epidemiology, ecology, and evolution, every process that we can imagine has some influence on the outcomes that we observe. Pretending that some of these processes are completely absent can be a useful means to an inferential or computational end, but it is rarely what we actually believe about the system (although see [13] for a counterargument). We should not let this useful pretense become our primary statistical focus.

If we have sensible scientific questions and good experimental designs, muddling through with existing techniques will often provide reasonable results [61]. But researchers should at least be aware that the roundabout statistical methods they currently use to understand multifactorial systems were designed for prediction, or the comparison of discrete hypotheses, rather than for quantifying the relative strength of simultaneously operating processes. When prediction is the primary goal, penalized methods can work better (faster and with better-understood statistical properties) than multimodel averaging. When estimating the magnitude of effects or judging variable importance, penalized or Bayesian 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 to data to fit and interpreting the full model.

Funding: This research was funded by NSERC Discovery grants 2016-05488 and 2023-05400.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: There is no data in the article.

Acknowledgments: Thanks to Jonathan Dushoff for conversations on these topics over many years. Dana Karelus, Daniel Turek, and Jeff Walker provided useful input: Noam Ross encouraged me to

finally submit the paper; Tara Bolker gave advice on straw men; three anonymous reviewers gave useful feedback. This work was supported by multiple NSERC Discovery grants.

Conflicts of Interest: The author declares no conflicts of interest. The funder had no role in the writing of the manuscript or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

CI	confidence interval
MMA	multi-model averaging
NHST	null-hypothesis significance testing

References

1. Gruner, D.S.; Smith, J.E.; Seabloom, E.W.; Sandin, S.A.; Ngai, J.T.; Hillebrand, H.; Harpole, W.S.; Elser, J.J.; Cleland, E.E.; Bracken, M.E.; et al. A Cross-System Synthesis of Consumer and Nutrient Resource Control on Producer Biomass. *Ecol. Lett.* **2008**, *11*, 740–755. [CrossRef] [PubMed]
2. Brando, P.M.; Nepstad, D.C.; Balch, J.K.; Bolker, B.; Christman, M.C.; Coe, M.; Putz, F.E. Fire-Induced Tree Mortality in a Neotropical Forest: The Roles of Bark Traits, Tree Size, Wood Density and Fire Behavior. *Glob. Chang. Biol.* **2012**, *18*, 630–641. [CrossRef]
3. Ghenu, A.-H.; Bolker, B.M.; Melnick, D.J.; Evans, B.J. Multicopy Gene Family Evolution on Primate Y Chromosomes. *BMC Genom.* **2016**, *17*, 157. [CrossRef]
4. McGill, B. Why Ecology Is Hard (and Fun)—Multicausality. *Dynamic Ecology*; 2016. Available online: https://dynamicecology.wordpress.com/2016/03/02/why-ecology-is-hard-and-fun-multicausality/ (accessed on 30 May 2024).
5. Platt, J.R. Strong Inference. *Science* **1964**, *146*, 347–353. [CrossRef] [PubMed]
6. Fox, J. Why Don't More Ecologists Use Strong Inference? *Dynamic Ecology*. (Blog Post). 2016. Available online: https://dynamicecology.wordpress.com/2016/06/01/obstacles-to-strong-inference-in-ecology/ (accessed on 30 May 2024).
7. Betini, G.S.; Avgar, T.; Fryxell, J.M. Why Are We Not Evaluating Multiple Competing Hypotheses in Ecology and Evolution? *R. Soc. Open Sci.* **2017**, *4*, 16056. [CrossRef] [PubMed]
8. Taper, ML.; Ponciano, J.M. Evidential Statistics as a Statistical Modern Synthesis to Support 21st Century Science. *Popul. Ecol.* **2015**, *58*, 9–29. [CrossRef]
9. Hastie, T.; Tibshirani, R.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, USA, 2009.
10. Burnham, K.P.; Anderson, D.R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*; Springer: Berlin/Heidelberg, Germany, 2002.
11. Moritz, M.A.; Batllori, E.; Bolker, B.M. The Role of Fire in Terrestrial Vertebrate Richness Patterns. *Ecol. Lett.* **2023**, *26*, 563–574. [CrossRef]
12. Brewer, M.J.; Butler, A.; Cooksley, S.L. The Relative Performance of AIC, AICC and BIC in the Presence of Unobserved Heterogeneity. *Methods Ecol. Evol.* **2016**, *7*, 679–692. [CrossRef]
13. Mundry, R. Issues in Information Theory-Based Statistical Inference—A Commentary from a Frequentist's Perspective. *Behav. Ecol. Sociobiol.* **2011**, *65*, 57–68. [CrossRef]
14. Cade, B.S. Model Averaging and Muddled Multimodel Inference. *Ecology* **2015**, *96*, 2370–2382. [CrossRef]
15. Bernhardt, I.; Jung, B.S. The Interpretation of Least Squares Regression with Interaction or Polynomial Terms. *Rev. Econ. Stat.* **1979**, *61*, 481–483. [CrossRef]
16. Jones, L.V.; Tukey, J.W. A Sensible Formulation of the Significance Test. *Psychol. Methods* **2000**, *5*, 411–414. [CrossRef]
17. Dushoff, J.; Kain, M.P.; Bolker, B.M. I Can See Clearly Now: Reinterpreting Statistical Significance. *Methods Ecol. Evol.* **2019**, *10*, 756–759. [CrossRef]
18. Meng, X. Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election. *Ann. Appl. Stat.* **2018**, *12*, 685–726. [CrossRef]
19. Harrell, F. *Regression Modeling Strategies*; Springer: Berlin/Heidelberg, Germany, 2001
20. Romano, J.P.; Wolf, M. Stepwise Multiple Testing as Formalized Data Snooping. *Econometrica* **2005**, 73, 1237–1282. [CrossRef]
21. Whittingham, M.J.; Stephens, P.A.; Bradbury, R.B.; Freckleton, R.P. Why Do We Still Use Stepwise Modelling in Ecology and Behaviour? *J. Anim. Ecol.* **2006**, *75*, 1182–1189. [CrossRef] [PubMed]
22. Mundry, R.; Nunn, C.L. Stepwise Model Fitting and Statistical Inference: Turning Noise into Signal Pollution. *Am. Nat.* **2009**, *173*, 119–123. [CrossRef] [PubMed]
23. Freckleton, R.P. Dealing with Collinearity in Behavioural and Ecological Data: Model Averaging and the Problems of Measurement Error. *Behav. Ecol. Sociobiol.* **2011**, *65*, 91–101. [CrossRef]

24. Galipaud, M.; Gillingham, M.A.F.; David, M.; Dechaume-Moncharmo, F.X. Ecologists Overestimate the Importance of Predictor Variables in Model Averaging: A Plea for Cautious Interpretations. *Methods Ecol. Evol.* **2014**, *5*, 983–991. [CrossRef]
25. Zhang, X.; Zou, G.; Carroll, R.J. Model Averaging Based on Kullback-Leibler Distance. *Stat. Sin.* **2015**, *25*, 1583–1598. [CrossRef] [PubMed]
26. Graham, M.H. Confronting Multicollinearity in Ecological Multiple Regression. *Ecology* **2003**, *84*, 2809–2815. [CrossRef]
27. Morrissey, M.B.; Ruxton, G.D. Multiple Regression Is Not Multiple Regressions: The Meaning of Multiple Regression and the Non-Problem of Collinearity. *Philos. Theory Pract. Biol.* **2018**, *10*, 3. [CrossRef]
28. Feng, X.; Park, S.D.; Liang, Y.; Pandey, R.; Papeş, M. Collinearity in Ecological Niche Modeling: Confusions and Challenges. *Ecol. Evol.* **2019**, *9*, 10365–10376. [CrossRef] [PubMed]
29. Vanhove, J. Collinearity Isn't a Disease That Needs Curing. *Meta-Psychology* **2021**, *5*, 1–11. [CrossRef]
30. Burnham, K.P.; Anderson, D.R. *Model Selection and Inference: A Practical Information-Theoretic Approach*; Springer: New York, NY, USA, 1998.
31. Johnson, J.B.; Omland, K.S. Model Selection in Ecology and Evolution. *Trends Ecol. Evol.* **2004**, *19*, 101–108. [CrossRef] [PubMed]
32. Chamberlin, T.C. The Method of Multiple Working Hypotheses. *Science* **1890**, *15*, 92–96. [CrossRef] [PubMed]
33. Gelman, A.; Shalizi, C.R. Philosophy and the Practice of Bayesian Statistics. *Br. J. Math. Stat. Psychol.* **2013**, *66*, 8–38. [CrossRef]
34. Bernardo, J.M.; Smith, A.F.M. *Bayesian Theory*, 1st ed.; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 1994. [CrossRef]
35. Barker, R.J.; Link, W.A. Truth, Models, Model Sets, AIC, and Multimodel Inference: A Bayesian Perspective. *J. Wildl. Manag.* **2015**, *79*, 730–738. [CrossRef]
36. Ponciano, J.; Mark L. Taper. Multi-Model Inference Through Projections in Model Space. *arXiv* **2018**, arXiv:1805.08765.
37. Luttbeg, B.; Langen, T.A. Comparing alternative models to empirical data: Cognitive models of western scrub-jay foraging behavior. *Am. Nat.* **2004**, *163*, 263–276. [CrossRef]
38. Dahlgren, J.P. Alternative Regression Methods Are Not Considered in Murtaugh (2009) or by Ecologists in General. *Ecol. Lett.* **2010**, *13*, E7–E9. [CrossRef] [PubMed]
39. Lemoine, N.P. Moving Beyond Noninformative Priors: Why and How to Choose Weakly Informative Priors in Bayesian Analyses. *Oikos* **2019**, *128*, 912–928. [CrossRef]
40. Rasmussen, C.E.; Williams, C.K. *Gaussian Processes for Machine Learning*; The MIT Press: Cambridge, MA, USA, 2015.
41. Wood, S.N. *Generalized Additive Models: An Introduction with R*; CRC Texts in Statistical Science; Chapman & Hall: London, UK, 2017.
42. Chandrashekar, G.; Sahin, F. A Survey on Feature Selection Methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
43. Koenker, R. Quantile Regression: 40 Years on. *Annu. Rev. Econ.* **2017**, *9*, 155–176. [CrossRef]
44. Wenger, S.J.; Olden, J.D. Assessing Transferability of Ecological Models: An Underappreciated Aspect of Statistical Validation. *Methods Ecol. Evol.* **2012**, *3*, 260–267. [CrossRef]
45. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. *Ecography* **2016**, *40*, 913–929. [CrossRef]
46. Chung, Y.; Rabe-Hesketh, S.; Dorie, V.; Gelman, A.; Liu, J. A Nondegenerate Penalized Likelihood Estimator for Variance Parameters in Multilevel Models. *Psychometrika* **2013**, *78*, 685–709. [CrossRef] [PubMed]
47. Pötscher, B.M.; Schneider, U. Confidence Sets Based on Penalized Maximum Likelihood Estimators in Gaussian Regression. *Electron. J. Stat.* **2010**, *4*, 334–360. [CrossRef]
48. Javanmard, A.; Montanari, A. Confidence Intervals and Hypothesis Testing for High-Dimensional Regression. *J. Mach. Learn. Res.* **2014**, *15*, 2869–2909. Available online: http://dl.acm.org/citation.cfm?id=2697057 (accessed on 30 May 2024).
49. Lockhart, R.; Taylor, J.; Tibshirani, R.J.; Tibshirani, R. A Significance Test for the Lasso. *Ann. Stat.* **2014**, *42*, 413. Available online: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4285373/ (accessed on 30 May 2024). [CrossRef] [PubMed]
50. Taylor, J.; Tibshirani, R. Post-Selection Inference for L1-penalized Likelihood Models. *Can. J. Stat.* **2018**, *46*, 41–61. [CrossRef] [PubMed]
51. Obenchain, R. Classical *F*-Tests and Confidence Regions for Ridge Regression. *Technometrics* **1977**, *19*, 429–439. [CrossRef]
52. Wang, H.; Zhou, S.Z.F. Interval Estimation by Frequentist Model Averaging. *Commun. Stat.-Theory Methods* **2013**, *42*, 4342–4356. [CrossRef]
53. Hjort, N.L.; Claeskens, G. Frequentist Model Average Estimators. *J. Am. Stat. Assoc.* **2003**, *98*, 879–899. [CrossRef]
54. Turek, D.; Fletcher, D. Model-Averaged Wald Confidence Intervals. *Comput. Stat. Data Anal.* **2012**, *56*, 2809–2815. [CrossRef]
55. Fletcher, D.; Turek, D. Model-Averaged Profile Likelihood Intervals. *J. Agric. Biol. Environ. Stat.* **2012**, *17*, 38–51. [CrossRef]
56. Turek, D.B. Frequentist Model-Averaged Confidence Intervals. Ph.D. Thesis, University of Otago, Dunedin, New Zealand, 2013. https://www.otago.ourarchive.ac.nz/bitstream/handle/10523/3923/TurekDanielB2013PhD.pdf (accessed on 30 May 2024).
57. Turek, D. Comparison of the Frequentist MATA Confidence Interval with Bayesian Model-Averaged Confidence Intervals. *J. Probab. Stat.* **2015**, *2015*, 420483 . [CrossRef]
58. Kabaila, P.; Welsh, A.H.; Abeysekera, W. Model-Averaged Confidence Intervals. *Scand. J. Stat.* **2016**, *43*, 35–48. [CrossRef]
59. Dormann, C.F.; Calabrese, J.M.; Guillera-Arroita, G.; Matechou, E.; Bahn, V.; Bartoń, K.; Beale, C.M.; Ciuti, S.; Elith, J.; Gerstner, K.; et al. Model Averaging in Ecology: A Review of Bayesian, Information-Theoretic and Tactical Approaches for Predictive Inference. *Ecol. Monogr.* **2018**, *88*, 485–504. [CrossRef]

60. Burnham, K.P.; Anderson, D.R. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociol. Methods Res.* **2004**, 33, 261–304. [CrossRef]
61. Murtaugh, P.A. Performance of Several Variable-Selection Methods Applied to Real Ecological Data. *Ecol. Lett.* **2009**, *12*, 1061–1068. [CrossRef]