*Article*

# Multimodel approaches are not the best way to understand multifactorial systems

**Benjamin M. Bolker** [1,†,‡]

[1] Departments of Mathematics & Statistics and Biology, McMaster University, Hamilton, Ontario, Canada L8S4K1; bolker@mcmaster.ca

* Correspondence: 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 artificial 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, 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, giving overly narrow intervals. If researchers want 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 which predictors to include.

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

**Citation:** Bolker, B. M. Multimodel approaches and multifactorial systems. *Entropy* **2024**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

Modern scientific research often aims to quantify the ~~importance of multiple~~ 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 (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; Betini, Avgar, and Fryxell 2017).[1] 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 (Hastie, Tibshirani, and Friedman 2009). 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 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 give 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, 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 make sure

[1] While there is much interesting debate over the best methods for gathering evidence to distinguish among two or more particular, *intrinsically* discrete hypotheses (Taper and Ponciano 2015), that is not the focus of this paper.

that it describes the data reasonably well (e.g. by ~~computing $R^2$ values or estimating the degree of overdispersion~~examining model diagnostics and by making sure that the level of unexplained variation is not unacceptably large). (3) Construct possible submodels of the full model by setting subsets of parameters to zero. (4) Compute 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); possibly draw conclusions about the importance of different processes by summing the information-theoretic weights (Burnham and Anderson 2002). 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 Moritz, Batllori, and Bolker (2023)). 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"[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. Most critiques of MMA address technical concerns such as the influence of unobserved heterogeneity (Brewer, Butler, and Cooksley 2016), or criticize the misuse of information-theoretic methods by researchers (Mundry 2011; Cade 2015), but do not ask why we are comparing discrete models in the first place.

In contrast to averaging across discrete hypotheses, or treating a choice of discreting 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 Bernhardt and Jung (1979)). 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 (Jones and Tukey 2000; Dushoff, Kain, and Bolker 2019). ~~We perform these tests by statistically comparing a full model to a reduced model that pretends the effect is exactly zero.~~

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

[2] http://www.perlmonks.org/?node=XY+Problem

data in a multifactorial world. If we had enough information (even Big Data doesn't always provide the information we need: Meng (2018)), we could fit ~~just~~ 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 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 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 (Harrell 2001; Whittingham et al. 2006; Romano and Wolf 2005; Mundry and Nunn 2009). Information-theoretic 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. Some forms of model averaging provide ~~simple~~ *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 (Cade 2015). 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 (Freckleton (2011~~; but cf. Walker 2017~~)); when we are careless about the meaning of main effects in the presence of interactions; when we average model parameters rather than model predictions (Cade 2015); or when we use summed model weights to assess the relative importance of predictors (Galipaud et al. (2014), Cade (2015; but cf. Zhang, Zou, and Carroll 2015)).

Freckleton (2011) 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 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 epistemic 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 (Graham 2003; Morrissey and Ruxton 2018; Feng et al. 2019; Vanhove 2021). A full model might (correctly) tell us we can't 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 (Burnham and Anderson 1998, 2002; Johnson and Omland 2004). 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 effects in multifactorial systems. ~~These researchers encounter two problems~~ 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 ~~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 (or many)of these models and then model-average their parameters. We might be doing this in an effort to get good predictions, or to 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~~ model space. When we fit many different models as part of our analytical process (based on selection or averaging), the models are only means to an end~~, and we shouldn't fool ourselves into thinking that we are using the~~ ; 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 (1890). For example, Chamberlin ~~(1897, reprinted as Raup and Chamberlin (1995))~~ 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 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 (2013) 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 isn't), and how this matters to our inference (Bernardo and Smith 1994; Barker and Link 2015). 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) (Ponciano and Taper 2018). However, the fact that excluding some processes allows the fitted model to better match observation that does not mean that we should believe these processes are not affecting on our system — just that, with available data, dropping terms will give us better predictions than keeping the full model. If we are primarily interested in prediction, or in comparing qualitatively different, possibly non-nested hypotheses (Luttbeg, Langen, and Adams 2004), information-theoret methods do 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 (Dahlgren 2010) 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) (Lemoine 2019). 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.[3]

[3] Although they ~~often~~ may require a computationally expensive cross-validation step in order to choose the degree of penalization.

~~Here I am not concerned whether 'truth' is included in our model set (it isn't), and how this matters to our inference (Bernardo and Smith 1994; Barker and Link 2015). I am claiming the opposite, that our full model is usually as close to 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, unless prediction is our only goal~~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 (Rasmussen and Williams 2005; Wood 2017).

Penalized models have their own challenges. A big advantage of information-theoretic methods is that, like wrapper methods for feature selection in machine learning (Chandrashekar and Sahin 2014), 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 (Koenker 2017), software for some more exotic models (e.g. zero-inflated models, ~~quantile regressions,~~ models for censored data) may not be readily available. Fitting these models requires the user to choose the ~~degree~~ strength of penalization. This process is conveniently automated in tools like `glmnet`, but correctly assessing out-of-sample accuracy (and hence the correct level of penalization) is tricky for data that are correlated in space or time (Wenger and Olden 2012; Roberts et al. 2016). Penalization (or regularization) can also be done by imposing Bayesian priors on subsets of parameters (Chung et al. 2013), 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 (Pötscher and Schneider 2010; Javanmard and Montanari 2014; Lockhart et al. 2014; Taylor and Tibshirani 2018), 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 (Obenchain 1977) — surprisingly, they are identical to the confidence intervals one would have gotten from the full model without penalization! Wang and Zhou (2013) similarly proved that model-averaging CIs derived as suggested by Hjort and Claeskens (2003) are asymptotically (i.e. for arbitrarily large data sets) equivalent to the CIs from the full model. Analytical and simulation studies (D. Turek and Fletcher 2012; Fletcher and Turek 2012; D. B. Turek 2013; D. Turek 2015; Kabaila, Welsh, and Abeysekera 2016; Dormann et al. 2018) 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, Welsh, and Abeysekera (2016) 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 Burnham and Anderson (2002).) 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* (Harrell 2001), to add information from pre-specified Bayesian priors (or equivalent regularization procedures), or to collect more data. Burnham and Anderson (2004) 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 ~~marginal~~ 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 (~~Google Scholar returns~~ in May 2024, Google Scholar returned ≈ ~~60~~65,000 hits for "multimodel averaging" alone and ~~30~~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 ~~(almost) never~~ rarely what we actually believe about the system (although see Mundry (2011) 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 give reasonable results (Murtaugh 2009). 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.

**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. I used Microsoft Copilot to remind me of the term "epistemic uncertainty".

**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. 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. https://doi.org/10.1002/jwmg.890.
2. Bernardo, José M., and Adrian F. M. Smith. 1994. *Bayesian Theory*. 1st ed. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470316870.
3. Bernhardt, Irwin, and Bong S. Jung. 1979. "The Interpretation of Least Squares Regression with Interaction or Polynomial Terms." *The Review of Economics and Statistics* 61 (3): 481–83. https://doi.org/10.2307/1926085.
4. Betini, Gustavo S., Tal Avgar, and John M. Fryxell. 2017. "Why Are We Not Evaluating Multiple Competing Hypotheses in Ecology and Evolution?" *Royal Society Open Science* 4 (1): 16056. https://doi.org/10.1098/rsos.160756.
5. 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. https://doi.org/10.1111/j.1365-2486.2011.02533.x.
6. Brewer, Mark J., Adam Butler, and Susan L. Cooksley. 2016. "The Relative Performance of AIC, AICC and BIC in the Presence of Unobserved Heterogeneity." *Methods in Ecology and Evolution* 7 (6): 679–92. https://doi.org/10.1111/2041-210X.12541.
7. Burnham, Kenneth P., and David R. Anderson. 1998. *Model Selection and Inference: A Practical Information-Theoretic Approach*. New York: Springer.
8. ———. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer.
9. ———. 2004. "Multimodel Inference: Understanding AIC and BIC in Model Selection." *Sociological Methods & Research* 33 (2): 261–304. https://doi.org/10.1177/0049124104268644.
10. Cade, Brian S. 2015. "Model Averaging and Muddled Multimodel Inference." *Ecology*. https://doi.org/10.1890/14-1639.1.
11. Chamberlin, T. C. 1890. "The Method of Multiple Working Hypotheses." *Science* XV (366): 92–96. https://doi.org/10.1126/science.ns-15.366.92.
12. Chandrashekar, Girish, and Ferat Sahin. 2014. "A Survey on Feature Selection Methods." *Computers & Electrical Engineering* 40 (1): 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.
13. Chung, Yeojin, Sophia Rabe-Hesketh, Vincent Dorie, Andrew Gelman, and Jingchen Liu. 2013. "A Nondegenerate Penalized Likelihood Estimator for Variance Parameters in Multilevel Models." *Psychometrika* 78 (4): 685–709. https://doi.org/10.1007/s11336-013-9328-2.
14. Dahlgren, Johan P. 2010. "Alternative Regression Methods Are Not Considered in Murtaugh (2009) or by Ecologists in General." *Ecology Letters* 13 (5): E7–9. https://doi.org/10.1111/j.1461-0248.2010.01460.x.
15. Dormann, Carsten F., Justin M. Calabrese, Gurutzeta Guillera-Arroita, Eleni Matechou, Volker Bahn, Kamil Bartoń, Colin M. Beale, et al. 2018. "Model Averaging in Ecology: A Review of Bayesian, Information-Theoretic and Tactical Approaches for Predictive Inference." *Ecological Monographs*. https://doi.org/10.1002/ecm.1309.
16. Dushoff, Jonathan, Morgan P. Kain, and Benjamin M. Bolker. 2019. "I Can See Clearly Now: Reinterpreting Statistical Significance." *Methods in Ecology and Evolution* 10 (6): 756–59. https://doi.org/10.1111/2041-210X.13159.
17. Feng, Xiao, Daniel S. Park, Ye Liang, Ranjit Pandey, and Monica Papeş. 2019. "Collinearity in Ecological Niche Modeling: Confusions and Challenges." *Ecology and Evolution* 9 (18): 10365–76. https://doi.org/10.1002/ece3.5555.
18. Fletcher, David, and Daniel Turek. 2012. "Model-Averaged Profile Likelihood Intervals." *Journal of Agricultural, Biological, and Environmental Statistics* 17 (1): 38–51.
19. 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/.
20. 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.
21. 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. https://doi.org/10.1111/2041-210X.12251.
22. Gelman, Andrew, and Cosma Rohilla Shalizi. 2013. "Philosophy and the Practice of Bayesian Statistics." *British Journal of Mathematical and Statistical Psychology* 66 (1): 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x.

23. 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. https://doi.org/10.1186/s12864-015-2187-8.
24. Graham, Michael H. 2003. "Confronting Multicollinearity in Ecological Multiple Regression." *Ecology* 84 (11): 2809–15. https://doi.org/10.1890/02-3114.
25. 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.
26. Harrell, Frank. 2001. *Regression Modeling Strategies*. Springer.
27. Hastie, Trevor, Robert Tibshirani, and J. H Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.
28. Hjort, Nils Lid, and Gerda Claeskens. 2003. "Frequentist Model Average Estimators." *Journal of the American Statistical Association* 98 (464): 879–99. https://doi.org/10.1198/016214503000000828.
29. Javanmard, Adel, and Andrea Montanari. 2014. "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression." *The Journal of Machine Learning Research* 15 (1): 2869–909. http://dl.acm.org/citation.cfm?id=2697057.
30. Johnson, Jerald B., and Kristian S. Omland. 2004. "Model Selection in Ecology and Evolution." *Trends in Ecology & Evolution* 19 (2): 101–8. https://doi.org/10.1016/j.tree.2003.10.013.
31. Jones, Lyle V., and John W. Tukey. 2000. "A Sensible Formulation of the Significance Test." *Psychological Methods* 5 (4): 411–14. https://doi.org/10.1037//1082-989X.5.4.411.
32. Kabaila, Paul, A. H. Welsh, and Waruni Abeysekera. 2016. "Model-Averaged Confidence Intervals." *Scandinavian Journal of Statistics* 43 (1): 35–48. https://doi.org/10.1111/sjos.12163.
33. Koenker, Roger. 2017. "Quantile Regression: 40 Years On." *Annual Review of Economics* 9 (Volume 9, 2017): 155–76. https://doi.org/10.1146/annurev-economics-063016-103651.
34. Lemoine, Nathan P. 2019. "Moving Beyond Noninformative Priors: Why and How to Choose Weakly Informative Priors in Bayesian Analyses." *Oikos* 128 (7): 912–28. https://doi.org/10.1111/oik.05985.
35. 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/.
36. 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. https://doi.org/10.1086/381319.
37. 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/.
38. Meng, Xiao-Li. 2018. "Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election." *Annals of Applied Statistics* 12 (2): 685–726. https://doi.org/10.1214/18-AOAS1161SF.
39. Moritz, Max A., Enric Batllori, and Benjamin M. Bolker. 2023. "The Role of Fire in Terrestrial Vertebrate Richness Patterns." *Ecology Letters* 26 (4): 563–74. https://doi.org/10.1111/ele.14177.
40. Morrissey, Michael B. ; Ruxton, and Graeme D. Ruxton. 2018. "Multiple Regression Is Not Multiple Regressions: The Meaning of Multiple Regression and the Non-Problem of Collinearity." *Philosophy, Theory, and Practice in Biology* 10. https://doi.org/http://dx.doi.org/10.3998/ptpbio.16039257.0010.003.
41. Mundry, Roger. 2011. "Issues in Information Theory-Based Statistical Inference—a Commentary from a Frequentist's Perspective." *Behavioral Ecology and Sociobiology* 65 (1): 57–68.
42. Mundry, Roger, and Charles L. Nunn. 2009. "Stepwise Model Fitting and Statistical Inference: Turning Noise into Signal Pollution." *The American Naturalist* 173 (1): 119–23. https://doi.org/10.1086/593303.
43. Murtaugh, Paul A. 2009. "Performance of Several Variable-Selection Methods Applied to Real Ecological Data." *Ecology Letters* 12 (10): 1061–68. https://doi.org/10.1111/j.1461-0248.2009.01361.x.
44. Obenchain, R. 1977. "Classical *F*-Tests and Confidence Regions for Ridge Regression." *Technometrics* 19 (4): 429–39.
45. Platt, John R. 1964. "Strong Inference." *Science*, New Series, 146 (3642): 347–53. https://doi.org/10.2307/1714268.
46. Ponciano, Jose-Miguel, and Mark L. Taper. 2018. "Multi-Model Inference Through Projections in Model Space." arXiv. https://arxiv.org/abs/1805.08765.

47. Pötscher, Benedikt M., and Ulrike Schneider. 2010. “Confidence Sets Based on Penalized Maximum Likelihood Estimators in Gaussian Regression.” *Electronic Journal of Statistics* 4 (January): 334–60. https://doi.org/10.1214/09-EJS523.

48. ~~Raup, David C., and T. C. Chamberlin. 1995. “The Methodof MultipleWorkingHypotheses. ”*The Journal of Geology* 103 (3): 349–54.~~

49. Rasmussen, Carl Edward, and Christopher K. I. Williams. 2005. *Gaussian Processes for Machine Learning*. Cambridge, Mass: The MIT Press.

50. Roberts, David R., Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, et al. 2016. “Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure.” *Ecography*, December, 913–29. https://doi.org/10.1111/ecog.02881.

51. Romano, Joseph P., and Michael Wolf. 2005. “Stepwise Multiple Testing as Formalized Data Snooping.” *Econometrica* 73 (4): 1237–82. https://doi.org/10.1111/j.1468-0262.2005.00615.x.

52. 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. https://doi.org/10.1007/s10144-015-0533-y.

53. Taylor, Jonathan, and Robert Tibshirani. 2018. “Post-Selection Inference for L1-penalized Likelihood Models.” *Canadian Journal of Statistics* 46 (1): 41–61. https://doi.org/10.1002/cjs.11313.

54. Turek, Daniel. 2015. “Comparison of the Frequentist MATA Confidence Interval with Bayesian Model-Averaged Confidence Intervals.” *Journal of Probability and Statistics* 2015. https://doi.org/ 10.1155/2015/420483.

55. Turek, Daniel Bernard. 2013. “Frequentist Model-Averaged Confidence Intervals.” PhD thesis, University of Otago. https://www.otago.ourarchive.ac.nz/bitstream/handle/10523/3923/TurekDanielB2013PhD.pdf.

56. Turek, Daniel, and David Fletcher. 2012. “Model-Averaged Wald Confidence Intervals.” *Computational Statistics & Data Analysis* 56 (9): 2809–15. https://doi.org/10.1016/j.csda.2012.03.002.

57. ~~Walker, Jeffrey A. 2017.~~

58. Vanhove, Jan. 2021. “~~A Defense of Model Averaging~~Collinearity Isn’t a Disease That Needs Curing.” ~~*bioRxiv*~~*Meta-Psychology* ~~, 133785.~~ 5 (April). https://doi.org/10.15626/MP.2021.2548.

59. Wang, Haiying, and Sherry Z. F. Zhou. 2013. “Interval Estimation by Frequentist Model Averaging.” *Communications in Statistics - Theory and Methods* 42 (23): 4342–56. https://doi.org/10.1080/03610926.2011.647218.

60. 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. https://doi.org/10.1111/j.2041-210X.2011.00170.x.

61. 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): 1182–89. https://doi.org/10.1111/j.1365-2656.2006.01141.x.

62. Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with R*. CRC Texts in Statistical Science. Chapman & Hall.

63. Zhang, Xinyu, Guohua Zou, and Raymond J. Carroll. 2015. “Model Averaging Based on Kullback-Leibler Distance.” *Statistica Sinica* 25: 1583–98. https://doi.org/10.5705/ss.2013.326.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.