Understanding multifactorial systems with limited data: implications of model selection, averaging and shrinkage for prediction accuracy and inference

Florian Hartig[a], Benjamin Bolker[b], Carsten F. Dormann[a]

[a] *University of Freiburg, Department of Biometry and Environmental System Analysis, Tennenbacherstrasse 4, 79106 Freiburg, Germany*
[b] *McMaster University, Departments of Mathematics & Statistics and Biology, 1280 Main St W, Hamilton ON Canada L8S 4K1*

Abstract

BB: abstract is long (≈ 474 words); is this OK? Do we have a target journal? MEE?

Understanding multifactorial systems (those driven by multiple ecological processes and their interactions) is challenging even in the best-case scenario of unlimited data. Ecologists nearly always face the worst-case scenario of noisy data, limited sample size, and a large number of potential (often correlated) explanatory variables. Although it is inferentially valid BB: unclear?, using all available variables in a regression model causes *overfitting*, leading to large predictive errors as well as large uncertainties in regression coefficients.

Standard remedies for this problem include model selection (MS: usually based on AIC), conditional or unconditional model averaging (MA), and penalized regression methods such as lasso and ridge regression. MS, popular in ecology, is well known to cause bias and overconfident (too-narrow) confidence intervals, especially when predictors are collinear. MA, also called multi-model inference, is better, but its inferential properties are still poorly understood: MA has been widely reported to mitigate but not solve the inferential problems of MS. Penalized regression is popular in the statistical literature, especially with extremely large predictor sets, but rare in ecology; effective inference for these methods is an open research problem.

BB: is there a way to foreground the no free lunch problem - data-driven model adjustment/shrinkage messes up inference - ?

Here, we use theoretical arguments and extensive simulations with linear, Poisson and logistic regression models to show that: (1) all three methods (MS, MA, penalized regression) can reduce predictive error, with MA and penalized regression typically superior to model selection. (2) All three methods lead to biased parameter estimates — least for regularization methods, and strongly dependent on model structure and predictor collinearity for MS and MA. (3) Unconditional MA shrinks parameters towards zero, similar to penalized regression. Penalized regression is clearly preferable with collinear predictors, while MA has the advantage that resulting parameters lead to predictions corresponding to averaging predictions for linear models. (4) The algorithms we assessed for constructing confidence intervals and deriving p-values are unreliable for all methods; the only known strategy for inference that is both simple and reliable is to reduce the set of predictor variables *a priori*, without using information about model goodness of fit.

All methods reduce predictive error, but model averaging and regularization methods work better than model selection. For inference, unconditional MA is preferable over conditional MA and MS, but usually worse than both the full model or penalized regression, both of which are theoretically better justified. BB: is inference based on lasso really better justified? without strong assumptions of sparsity etc.?

We end with a philosophical argument that complements our practical exploration: information-theoretic approaches such as MA improperly focus researchers' attention on the comparison among (or combination of) discrete, straw-man hypotheses, rather than on quantifying the strength and importance of the multifactorial processes that we know are *simultaneously* determining outcomes in ecological systems.

BB: Reduced emphasis on p-values: I know that people care about these, but hopefully we can discuss them in the main text. Even the *attempt* to quantify type I/II error distorts our understanding of statistical procedures, because it means that we have to simulate scenarios with effects of magnitude exactly zero. As much as possible I'd like to focus on coverage rather than type-I/II

BB: Reduced emphasis on bias: measuring bias implies that we know the correct scale on which to estimate parameters (e.g. unbiased estimates for β are biased for $\log(\beta)$, except asymptotically …). Should certainly cover it (as part of the set of bias/variance/RMSE/coverage/type-I/II), but I think it's less important/OK to de-emphasize in the abstract.

Keywords: Model selection, model averaging, conditional averaging, unconditional averaging, lasso regression, shrinkage coefficient averaging

1. Introduction

Even while supposedly entering an era of "big data", researchers in ecology and evolution are often confronted with observations that include far more potential predictors than one would prefer given the sample size at hand [rules of thumb suggest at least 10 events n per predictor variable k. See, e.g., **?**]. Manifold reasons exist for the commonness of this situation. Adequate control or randomization of potentially confounding variables may be impossible in observational studies, leading to the necessity to record many potentially confounding variables. Combining datasets may not only increase n but also k. And unfortunately, also a lack of adequately experimental planning frequently leads to an unfavorable ratio of n/k. Analyzing such data directly in a regression model (full model) may result in large type II error rates (failure to demonstrate existing effects), as well as in large variance and uncertainty of parameter estimates and predictions, which is clearly undesirable.

It is tempting but naive to expect that a "magic" method exists that would solve this problem without having to pay a price. If such a method existed, we could draw reliable inferences for an immensely complex systems (large k) with very few samples n. However, experience in statistics shows that one can often find methods that reduce the error in one inferential aspect, at the price of increasing the error in another aspect.

This idea is most explicit in regularization methods such as lasso or ridge regression. These methods introduce a deliberate bias towards zero (shrinkage) to reduce the variance of parameter estimates and predictions [**?** **?** , e.g.,].

The price for getting better predictions and estimates is less explicit in the currently most widely used remedies for dealing with small data in ecology and evolution: model selection and model averaging [e.g., **?** **?**].

The idea of model selection (MS) is to identify the single-best among a set of candidate models according to some criterion (model selection score) that typically considers model fit as well as model complexity. Common scores are AIC, BIC, cross-validation or the Bayes factor [**?** , see, e.g,]. For variable selection in a regression context, a typical approach is to define the largest model one is willing to consider (full model), and then either simplify step-wise, or consider all possible submodels. Either way, MS tends to reduce variance in estimates and predictions, and typically also the predictive error. MS also tends tends to improve reported p-values for parameter estimates, but this is usually a statistical artifact that can be understood when considering that in a MS procedure, multiple models are tested and typically those are chosen for which parameter estimates are more significant, even if this is not the selection criterion. The result is that, without correction, p-values of the best model will tend to be too low and type I error too high [**?** **?**]. Moreover, as we show later, effect sizes may be biased by MS. This is particularly the case if predictors show strong collinearity, because the effect of removed predictors will be absorbed by the remaining predictors. Nevertheless, MS is widely applied in ecology and evolution, arguably also because of a lack of simple alternatives.

Model averaging (MA) can be motivated from various viewpoints. One popular motivation is that considering only the single best model neglects model selection uncertainty, especially when model selection scores are very similar between models. As a remedy, model averaging methods produce a weighted average of all considered models. This is a natural idea from a Bayesian perspective, where the Bayes factor is available as a theoretically consistent weighting factor [**?**]. From the frequentist perspective, however, it is less clear why and how such a weighting should be chosen. A common argument is quite generally that it would not be parsimonious to only use the best model [**?**]. A further argument is that mixing model predictions should result in better estimates, assuming that errors of the different models vary around the truth. Connected to the less straightforward justification for frequentist model averaging is the question of the appropriate weighting factor. The BIC would be most close to the Bayes Factor, which, as said above, has good theoretical properties from the Bayesian perspective. However, arguably due to the popularity of the AIC in frequentist statistics, it has become common practice to weight models by their AIC or AICc in ecology and evolution studies.

A further question for MA is: what is to be weighted? The original idea was to keep all models as they are, and only weight their predictions. Typically, however, researchers are also interested in effect sizes and significance of the models. In recent years, it has become more and more common to use MA also for generating averaged parameters in nested linear and nonlinear models such as GLMs. There are a number of issues with the approach of averaging model parameters in this way. First of all, averaging model parameters leads the same predictions as averaging predictions only if the model is linear, and if unconditional weighting (explanation later) is used [**?**]. In practice, however, parameter averaging is applied to GLMs or other nonlinear models, as well as with conditional weighting, often with no clear theoretical justification. Another issue is the interpretation of the averaged parameters. The motivation of MA was to reduce the variance in the predictions. The success of this method seems have led to belief that it must work for reducing variance and error of parameter estimates, often with (in our opinion wrong) reference to **?**]. As a result, there are frequent statements in the literature that averaged parameters are ecologically more meaningful, better interpretable, and generally to be preferred over the parameters of the full model or the best (sensu model selection) model.

Given that there is a price for each of this method, it's worth asking: is it worth paying? From anecdotal experience, we understand that some reviewers insist on model selection or model averaging, even if the study has sufficient power, apparently under the pretext that a model selection is always "more objective". We will show in the course of this paper that this is not always so - for some question, even when very little data is available, the methods described above struggle to

Email addresses: florian.hartig@biom.uni-freiburg.de (Florian Hartig), bolker@mcmaster.ca (Benjamin Bolker)

outperform the full model. To examine and demonstrate these issues, we

1. **Review and explain** the current theory and practice of model selection and model averaging, and show that there are at the moment three non-identical model averaging methods, namely averaging of predictions, as well as conditional and unconditional averaging.
2. **Simulate and compare** the expected bias, estimation error and predictive error for a) the full model b) the AIC/BIC best model c) the three variants of MA d) and lasso regression for 1) linear, 2) logistic and 3) poisson regression when applied to a large number of predictors, low sample size, and problems such as collinearity, nonlinearities and interactions.
3. **Give recommendation**, based on theory and simulations, for the usability and interpretability of results obtained by MA, MS, or regularization methods.

We stress that our results to not necessarily apply to MA and MS in for situations where where the power is appropriate in the first place (large N), and the goal is to select the true or the best model structure among a set of candidates.

2. Review of model selection and averaging methods

2.1. Model selection

Model selection (MS) aims to identify the "most appropriate" among a number of candidate models. The fact that a large number of non-identical MS methods exist suggest that the definition of "appropriate" is ambiguous to some extent. The commonly used methods agree on one aspect at least: model fit cannot be a sole criterion for model selection, because larger models are more flexible and therefore always tend to fit better, even when error in predictions and parameter estimates increases through the complexity (overfitting).

Hence, to avoid overfitting, MS methods have to penalize model complexity. The most widely applied method in ecology is the Akaike-Information Criterion (AIC), defined by **?**] as

$$AIC = -2(LL - k) \tag{1}$$

Here, LL is the log-Likelihood (which can be thought of as a measure of fit), reduced by a penalty proportional to k, the number of parameters. The model that has the lowest AIC is considered the AIC-best model.

Akaike derived the AIC in **?**] as an approximation of the Kullback–Leibler (KL) distance, an information-theoretic measure of the distance between the model prediction and the hypothetical "true model". Another way to approximate the KL distance is leave-one-out cross-validation (CV), which basically tries to estimate the predictive error of the model on new data. Asymptotically, AIC and CV therefore lead to identical results [**?**]. They may differ, however for finite sample sizes. Still, one may say that AIC has a similar goal than cross-validation. There are a number of noteworthy alternatives to the AIC for model selection, most importantly the Bayesian Information Criterion (BIC), an approximation of the Bayes Factor [**?**], and likelihood-ratio tests. These methods have a different motivations and definitions that AIC, and hence they will also select different models. We concentrate here on the AIC because it is currently without doubt the most widely score for model selection and model averaging.

A problem with applying AIC or MS methods on small datasets is that the penalty term will result in models that include a small number of predictors, even if much more predictors do have an influence on the response. Basically, the model selection will choose the predictors that show the strongest effect first, and would increase the number of predictors as the size of the available dataset increases. This raises two concerns: 1) frequently, the predictors identified by MS are often interpreted as the "ecologically meaningful" variables. This, however, seems questionable when the probability of a variable to be identified depends highly on a) the sample size, and b) the number of other predictors. 2) Especially when the number of explanatory variables is large compared to the sample size, models that are structurally completely different may have very similar AIC or BIC scores.2) MS basically chooses the largest effects first. If all variables have the same effect, it will chose that first for which estimates deviate randomly towards larger effect. The results are effects that are biased upwards.

2.2. Averaging of model predictions

Model averaging (MA) has often been promoted as a potential remedy for the problems of MS that we just discussed. The idea of MA is to create a weighted average $\bar{y}$ of the predictions of all models considered by calculating

$$\bar{y} = \sum_i w_i f_i(\alpha_i * X) \tag{2}$$

Here, f_i are the models with parameters $alpha_i$ and predictors X_i, and w_i is the weight given to each model. In Bayesian inference, the logical weight would be the Bayes factor [**?**]. In a frequentist mindset, it is a priori less clear how weights should be chosen. AIC, CV, BIC or other weights could be considered. We are not aware of any formal argument that explains why a weighting by one of those should be optimal, and if so, what quality that weighting would optimize. **?**] give an ad-hoc argument for a weighting by AIC, and due to the popularity of this source, this weighting is widely applied. We think that an AIC weighting is more or less sensible weighting for smaller data sets, as there are good arguments for the fact that it reduces some of the problems of MS discussed above when using the model for predictions. It remains to be show if AIC is the optimal weighting. Moreover, we question if AIC is ideal for model averaging on large data sets, due to the known tendency of AIC to select overcomplicated models, but this is not the setting we consider in this study.

2.3. Averaging of model parameters

The "traditional" MA, as described above, averages only model predictions. It does not provide information about significance and effect sizes, which is obviously a disadvantage compared to MS.

To address this problem, two approaches have been applied. The first is to calculate how often a parameter is included in the selected models, weighted by the model selection score. We conceit that this has some diagnostic value, but otherwise support the opinion of [paper] that there is very little support for interpreting these values systematically in terms of significance etc. In the best scenario, the weights calculated with these methods should have the same problems as described above for MS.

The second approach is what we call parameter averaging (PA). We think the practice of PA started by noting that, for linear models, the weights on the models f_i can be pushed through to the predictors.

$$\sum_i w_i \cdot f_i(\alpha_i * X_i) = \sum_i f_i(w_i \cdot \alpha_i * X) \tag{3}$$

if all parameters α_i that are not included in the submodel f_i are set to zero. Thus, we can calculate averaged parameters similar to averaged predictions as

$$\bar{\alpha} = \sum_i w_i \cdot \alpha_i \tag{4}$$

Obviously, all comments made regarding the problem of appropriate weights apply for this as well. If parameter estimates α_i for predictors that are not included in the submodel f_i are set to zero, one speaks about 'unconditionally parameter averaging" (UPA), because estimates are always included in the average, unconditional on whether their corresponding predictor was selected or not. An alternative way of averaging the parameters is "conditional parameter averaging" (CPA), where the estimates for predictors that were not selected are simply omitted in the average. One can directly make two conclusions from this definition

1. Estimates for UPA are always smaller or equal than those for CPA, because the only difference is that CPA drops the zeros from the average.
2. For linear models, predictions of MA are identical to UPA, while predicting with the parameters obtained by CPA has no immediate correspondence to weighting of predictions. For nonlinear models, predictions obtained by parameters from UPA, and CPA are in general different from MA. To see the latter, note that if f in eq.**??** is nonlinear, the equality is not valid any more.

UPA and CPA is commonly applied in the ecological literature, but we are not aware of any systematic studies of its properties, or any theoretical justification. We think that there is reason for concern for a number of reasons.

The first concern is about predictions: as said, for linear models, predictions with parameters estimates from CPA are identical to MA, but different from UPA. For nonlinear models, all three options lead to different results. Although B&A note that parameter averaging is only valid for linear models, it seems that in the literate it has tacitly been extended to generalized linear models, which, despite having the word "linear" in their name, are linear on the scale of the linear predictors (link scale), but on the scale of the model predictions, because typical link functions that are used such as the exponential and the logit link are nonlinear. As a result, we have three methods that make different predictions, and it is therefore important to understand whether one of these methods is clearly best, or whether they perform differently in different conditions.

The second concern applies to the interpretation of the averaged parameters. Claims have been made in the literature that

1. The averaged coefficients are "better" model than fitted models (REF).
2. Averaged coefficients are ecologically better interpretable (REF).
3. Substituting terms not in a model by NA *or* 0 is both fine (since there are no guidelines, either approach is fine, while obviously only one can be best for a given data set and question) (REF).
4. The weighted proportion of models in which a term occurs is a sound measure of its importance (i.e. importance is interpreted as statistical significance) (REF).

We have not been able to locate any formal proof for such claims. It seems that the general line of thought is that an average over multiple options is always more certain and parsimonious than a single estimate. Yet, note the following points:

1) The model built with the "averaged" parameters combines all predictors that have been selected in one of the models, and is therefore often considerably larger than the largest model that was AIC selected.

2) If we assume that the larger number of parameters included in PA is correct, we would know an alternative unbiased estimator for their values - it is obtained simply by fitting this model to the data. Given the definition of parameter averaging above, it's easy to see that for linear effects only

$$|E(\bar{\alpha}^{uncond})| < |true| < |E(\bar{\alpha}^{cond})| \tag{5}$$

meaning that one would expect that UPA leads to a shrinkage (a bias towards zero), and CPA leads to a bias towards larger values. Moreover, we would expect the bias of CPA to be highest and the bias of UPA to be lowest under strong collinearity, and the reverse for low collinearity in the predictors.

2.4. Lasso and ridge regression

Such a bias, if small, may be acceptable if it reduces the variance of the estimators, or makes the model computable in the first place (regularization). In fact, two other estimators that we already mentioned have similar properties: MS leads to a bias towards larger estimates, similar to CPA, and lasso regression leads to a bias towards smaller estimates, similar to UPA. Lasso and MS are theoretically much better understood however. More specifically, it is unclear of PA leads to any improvement over the existing method. Moreover, the argument above was made only for linear models. It is less trivial to see

how UPA and CPA would behave if predictors affect the response in a nonlinear way (e.g. quadratic effects), or if there are interactions between predictors.

CIs and p-values not well understood [**?**]

3. Methods simulation study

The situation we consider is regression, many predictors, few data. In this context, the full model refers, broadly speaking, to the model that includes all predictors as well as possible interactions or nonlinear effects.

In fact, two other estimators that we already mentioned have similar properties: MS leads to a bias towards larger estimates, similar to CPA, and lasso regression leads to a bias towards smaller estimates, similar to UPA. Lasso and MS are theoretically much better understood however. More specifically, it is unclear of PA leads to any improvement over the existing method. Moreover, the argument above was made only for linear models. It is less trivial to see how UPA and CPA would behave if predictors affect the response in a nonlinear way (e.g. quadratic effects), or if there are interactions between predictors.

We therefore chose to compare all the option that were discussed so far as possibilities to estimate effect sizes and make predictions when the number of predictors is large given the available sample size.

As methods, we consider: using the full model, MS, MA, UPA, CPA and lasso regression. MS, MA, UPA and CPA were calculated for AICc and BIC, but we concentrate the presentation of the results on AICc because this is more widely applied.

We use this methods on linear, logistic and Poisson regression models with 10 parameters and small datasets, comparing 4 cases (equal true parameters, unequal true parameters, quadratic effects, interactions) under no and high collinearity between the predictors. Details and code are provided in the appendix. Our expectation, summarized in Table.3 are following the discussion of the preceding section.

From each of the combination above, we obtained parameter estimates from 100 generated datasets with known properties. We then calculated bias of the parameter estimates, and the error in terms of RMSE (linear), AUC (logistic) or deviance (Poisson) when predicting on new data. Our expectations, motivated by the discussion in the preceding section, are summarized in Table.3.

4. Results

4.1. Linear model, linear effects

Parameter estimates for the linear model with equal effect sizes (Fig 4) largely follow the pattern expected in Table. 3.

If predictors are independent, the full model and CPA are unbiased. UPA and lasso lead to shrinkage towards 0. MS leads to a bias towards larger estimates. There may be a slight tendency towards smaller variance in parameter estimates for MS, while there is no discernible reduction of variance through parameter averaging.

If we move to collinear predictors, the full model stays unbiased. MS and CPA lead to considerable bias towards larger values. Lasso and UPA show shrinkage towards zero, but the bias is small. Also, both lasso and upa lead to a substantial reduction of variance compared to the full model and the other options.

If we look at the predictive error (RMSE) of the former cases (Fig.4.1), there is again a clear difference between the independent and the collinear case: for independent predictors, the full model does at least as well as all other alternatives, and differences are overall relatively small. For collinear predictors, CPA leads to highly unreliable predictions, while UPA (identical to MA) and lasso do best.

The case of all predictors having equal effects may or may not be realistic. An alternative scenario is that half of the predictor have an effect, and the other half does not. The results are displayed in Fig. 4.1. The main difference to the previous results is that both UPA and lasso are very effective in correctly identifying predictors that have no effect, regardless of collinearity, while MS is very ineffective for the same task. Presumably because of the effective shrinkage, UPA (identical to MA) and lasso are now doing as good or better than other methods for both independent and collinear predictors

Linear model, quadratic effects and interactions. The addition of quadratic effects and interactions did not differ substantially from the behavior of linear effects described in the previous section. We therefore show only the results for fitting models with interactions (Fig.4.1). The general observation is that UPA and lasso are very successful in reducing the variance on estimates of the interaction, specially for the case of collinearity, but are biased towards zero. In contrast, the full model and CPA show substantial variance, particularly when dealing with collinear predictors. MS does reduce the variance, but remains to show a large bias towards larger values.

Regarding prediction, UPA (identical to MA) and lasso did again overall best. CPA remained to show large errors when predicting on collinear data. Given the large bias in the parameter estimates, MS did surprisingly well regarding predictions.

Generalized linear models. In general, the patterns found for the linear models remains valid for GLMs. However, they are overlaid by three known problems: first of all, maximum likelihood estimates for GLMs, in particular logistic regression, are biased for small sample sizes [**? ?**]. This means that the full model is not unbiased any more, and that for the remaining methods, we see a mix between this inherent MLE bias, and the bias induced by MS, MA or lasso. Secondly, particularly for small sample sizes, parameter estimates for GLMs may diverge. Thirdly, unlike the linear model, GLMs are nonlinear through their link function, which we mentioned as a theoretical problem above. The nonlinearity of the link functions is such that biases from parameter averaging may change with both the true intercept and the effect size that, which, together with all other combinations, creates a large array of different possible settings in which the methods could be compared.

Method	Parameter		Predictions	
	Bias	Error	Bias	Error
Full	0	large	0	large
Model selection	larger values the more collinear	?		
Averaging Pred.	-	-	0?	smaller
Conditional Av.	towards larger values, the more collinearity	medium		
Unconditional Av.	towards 0 the less collinear	medium		
Lasso	towards 0	medium		

Table 1: Expectations for bias and variance of parameter estimates and predictions for the different options.

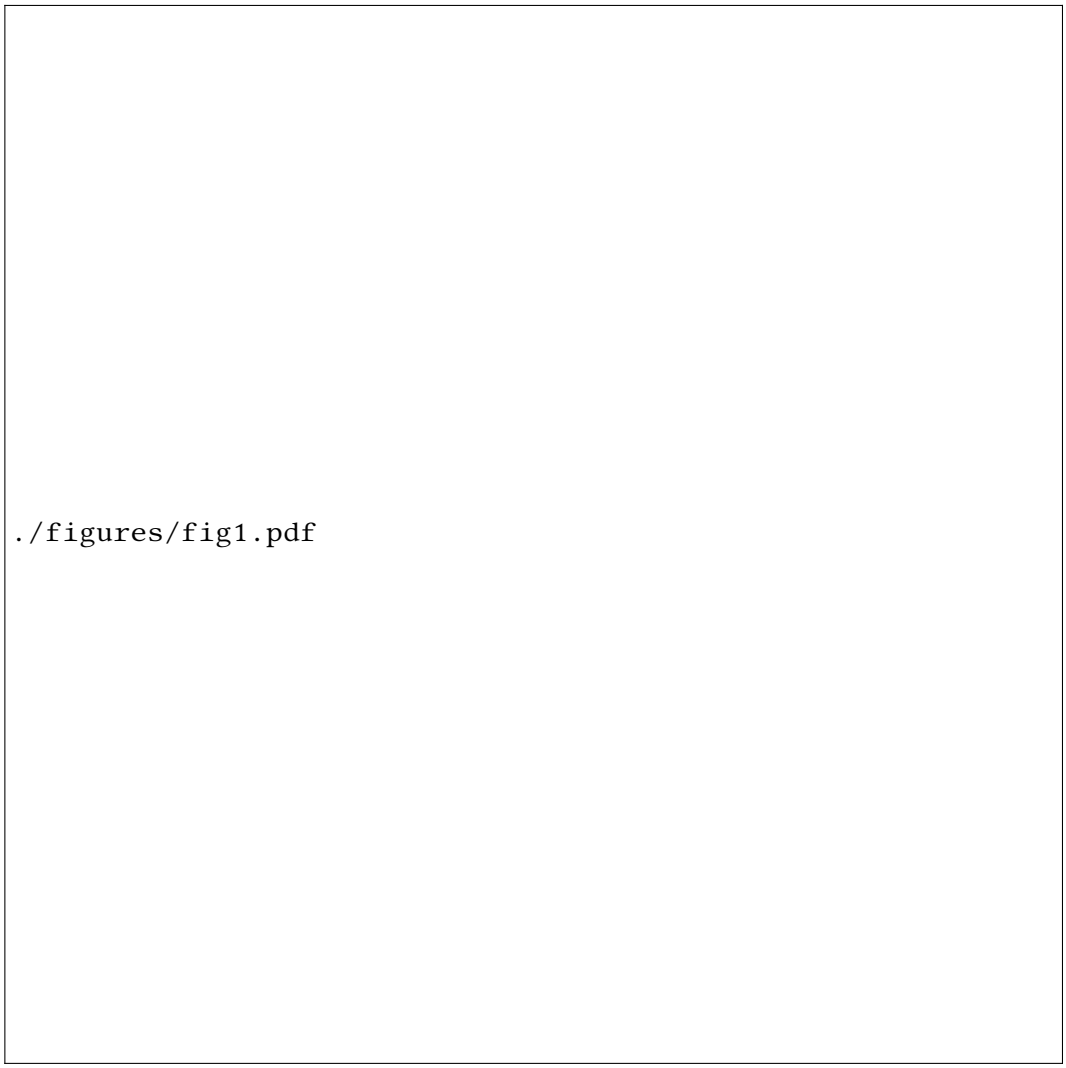

Figure 1: Bias of parameter estimates for linear regressions with independent (left) and collinear predictors.

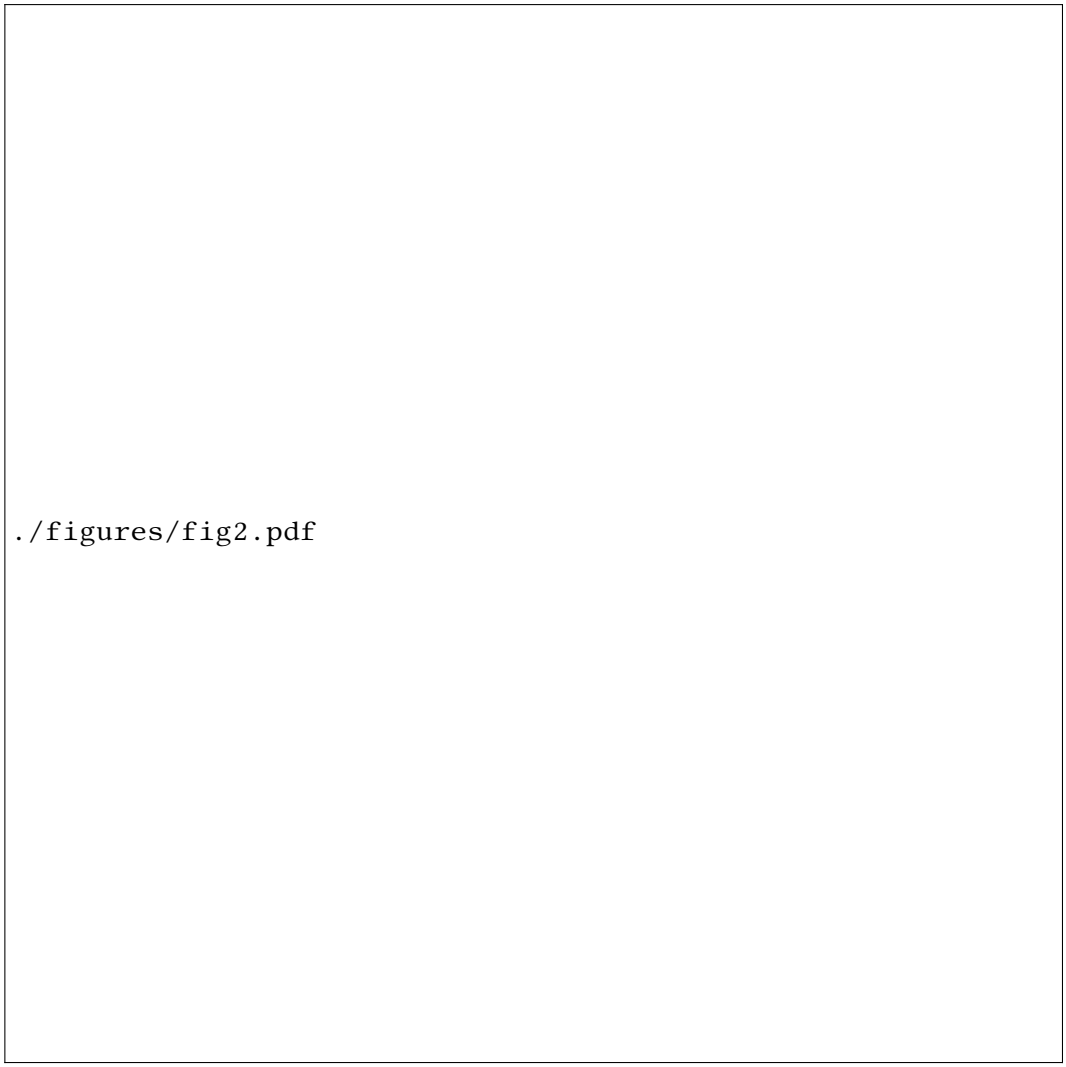

Figure 2: Predictive error of parameter estimates for linear regressions with independent (left) and collinear predictors.

Figure 3: Bias and predictive error of parameter estimates for linear regressions with independent (left) and collinear predictors.

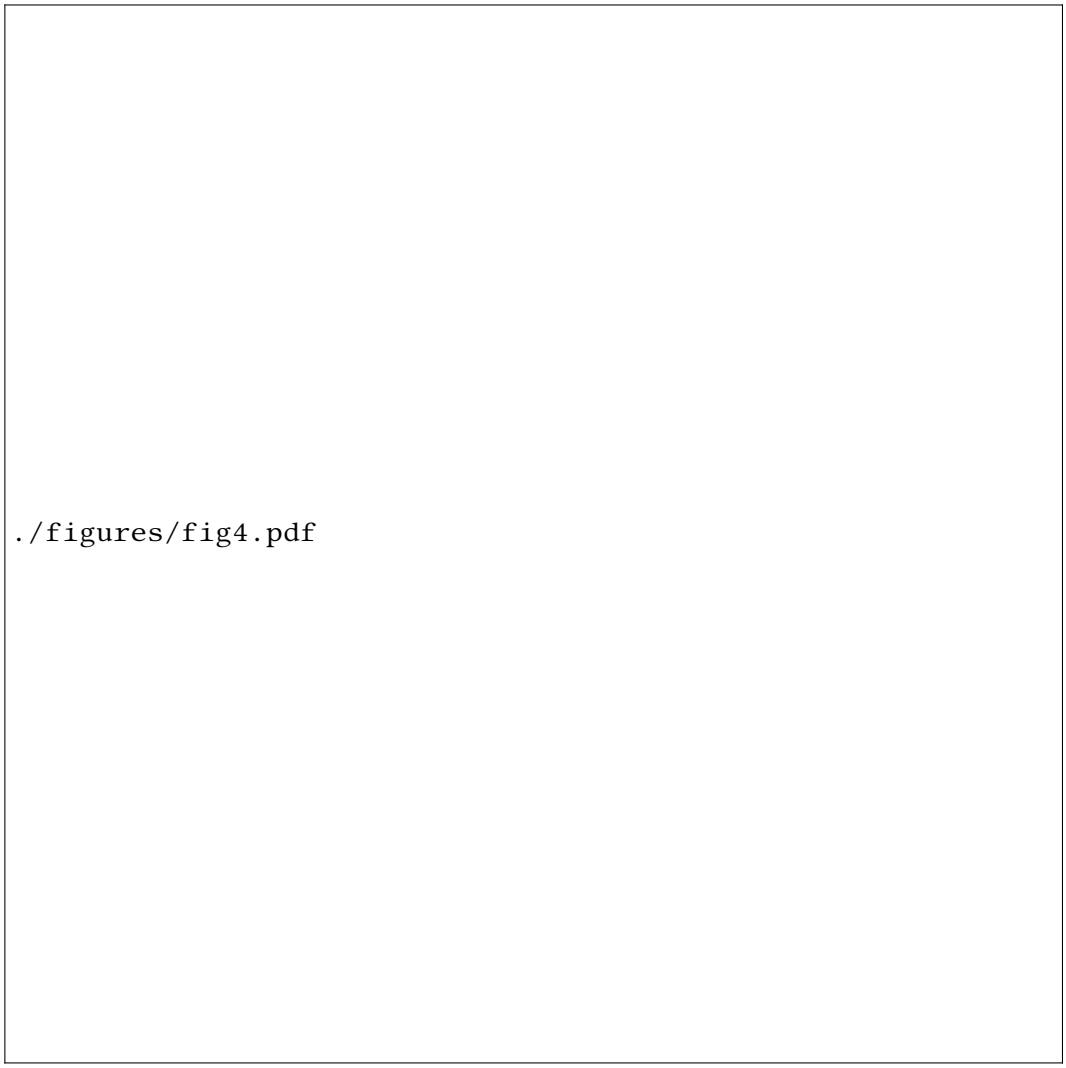

Figure 4: Bias and predictive error of parameter estimates for linear regressions with half of the parameters being interactions, calculated for independent (left) and collinear predictors.

As an example, we show the case for the logistic model with a sample size of 100 and low values of the intercept, meaning that we have few positive observations (Fig.4.1). The results are similar as for the linear case, except that CPA does comparatively better for both estimates and predictions than in the linear case, which seemed to be a general trend for the binomial models. The somewhat unfavorable predictions for the lasso were no general trend. They may have something to do with the way the shrinkage of the lasso is adopted through cross-validation.

AIC vs BIC. Although not the focus of this study, we conducted all calculations for BIC as well. The results were somewhat context-dependent, as one might expect from the fact that AIC and BIC describe a different trade-off between data size and model complexity. We would carefully conclude that BIC seems to do slightly better for the linear models that we examined. For the GLMs, results were inconclusive.

5. Discussion

We used theoretical arguments and simulations to compare the utility of model averaging of predictions (MA) and conditional and unconditional model averaging of parameters (CPA / UPA) for situations where the number of predictors is large for the size of the dataset to the theoretically better founded alternatives of using the full model, model selection (MS), or lasso regression .

For reducing the predictive error, we find no general advantage of parameter averaging over the traditional averaging of model predictions. In fact, conditional parameter averaging (CPA) often lead to extremely unfavorable predictions, particularly for linear models. Unconditional parameter averaging (UPA), which is identical to normal averaging of the predictions for linear models, generally fared better and lead to results comparable to lasso regression. Our recommendation is nevertheless that one should generally refrain from parameter averaging for the purpose of making predictions, and instead use either normal averaging of the predictions, or a shrinkage methods such as the lasso. Our results show that the full model is often also doing surprisingly well, as long as collinearity is moderate.

Regarding parameter estimates, we have to distinguish between bias and variance of the estimator. Bias refers to a systematic difference of the estimated effects to the true values. Variance refers to the typical difference of the estimates effects to the true value.

Broadly, all methods of model simplification or averaging discussed here introduce a bias. A particularly strong bias is created by MS. There were few if any situations in which MS seemed to be a suitable method for generating reliable estimates of effect sizes. For MA, PA and lasso, biases were smaller, but still substantial. This creates a larger concern for linear models, because here the full model is truly unbiased. For linear models, one should therefore consider carefully whether any other gains (predictions, or variance discussed in the next paragraph) make it worthwhile to deviate from the full model. Generalized linear models, on the other hand, are biased also for the full model, in the same order of magnitude than biases introduced by the methods consider here. In many cases, biases cancel or at least don't add up. In general, the loss of precision seems less clear as for the linear models.

The main reason for accepting some bias would be to reduce the variance of the estimators. For simple cases with independent predictors of similar effect sizes, none of the methods led to considerable improvements over the full model. However, when faced with a) collinear predictors b) predictors with no effect c) interactions and quadratic effects, or combinations thereof, UPA and lasso did lead to substantial reduction of variance in the parameter estimates, at the cost of a bias, however. When one expects this situations to be present, both of them seem viable alternatives. We would tend to prefer the lasso because it is computationally more efficient and theoretically better understood.

We did not systematically analyze differences between using AICc and BIC for model and parameter averaging. Our simulations show that differences exist. There seemed to be a small advantage of BIC overall. However, more systematic tests are needed to confirm this pattern. In this case, for example CV, could be considered.

6. Conclusions

When the amount of data is to small for the question that is asked, it is clear from the start that only second-best solutions can be provided. The basic problem is that

One should be under no illusion that all of the methods compared here are suboptimal compared to a properly balanced experiment with sufficient statistical power.

For making predictions only, our results suggest that model averaging and regularization methods generally result in better predictive accuracy than either the full model or a model constructed by model selection. We also show that predictions based on parameter averaging are not identical on to averaged predictions. Although

PA does not lead to an improvement over MA, and as it is less well founded, we would not recommend it for making predictions d) that lasso regression is often working as well as MA. Hence, for making predictions, traditional MA or lasso regression seem the best alternatives.

For obtaining overall good and ecologically interpretable parameter estimates, our results suggest that the full model may still be fine as long as collinearity is low and there are few interactions and non-effective parameters in the data. If the opposite is the case, UPA and lasso work fine. MS and CPA are less robust, and we do not recommend them in this situation.

We hold that model averaging of predictions has a shaky, and model averaging of parameters has in general no sound theoretical foundation. That, however, does not mean that they may not be useful tools of inference.

Overall, our conclusions are that MA works well for predictions. For obtaining reasonable parameter estimates, we saw little reason to deviate from the full model as long as the situations permits it. If we have to, lasso regression and UPA

Figure 5: Bias and predictive error of parameter estimates for logistic regressions with independent (left) and collinear predictors.

are alternatives with similar properties. The difference is that lasso regression is computationally faster, theoretically better understood, and has a clear correspondence also for Bayesian inference. For this reason, we would recommend to use shrinkage estimators such as the lasso or the ridge over the use of UPA.

6.1. BOX - GLOSSARY

candidate model = one of the > 1 models that is considered for MS or MA full model = the largest of all candidate models nested models = all candidate models are submodels of the full model