Ann. Zool. Fennici 52: 257–279 ISSN 0003-455X (print), ISSN 1797-2450 (online)
Helsinki 18 December 2015

Using full models, stepwise regression and model selection in ecological data sets: Monte Carlo simulations

Gergely Hegyi[1,*] & Miklós Laczi[1]

*Behavioural Ecology Group, Department of Systematic Zoology and Ecology, Eötvös Loránd University, Pázmány Péter sétány 1/C, H-1117 Budapest, Hungary (*corresponding author's e-mail: gehegyi@caesar.elte.hu)*

Received 27 Aug. 2014, final version received 1 Sep. 2015, accepted 18 Sep. 2015

Hegyi, G. & Laczi, M. 2015: Using full models, stepwise regression and model selection in ecological data sets: Monte Carlo simulations. — *Ann. Zool. Fennici* 52: 257–279.

Most ecological studies with multiple independent variables use null hypothesis testing with full or stepwise models, or AICc-based model selection, but these approaches have not yet been compared using simulated data with known effect sizes. We compared these using ecologically relevant sample sizes, effect sizes, predictor numbers, collinearity and different degrees of explorative setups. Sample size and collinearity governed parameter identification success and parameter estimation accuracy, while the effect of the statistical modeling approach was comparatively smaller. Stepwise regression increased false detection rate compared with full models in settings where this error rate was overall low, but generally reduced the high detection failure rate in small samples. When reintroducing removed predictors to the final model, stepwise regression often improved the accuracy of point estimates relative to full models. The performance of AICc model selection and model averaging depended on the exact method, and did not differ overall from null hypothesis testing approaches.

Introduction

Explorative ecological research is based on the collection of data and on the identification of patterns in these data. The observed patterns can then be tested in confirmatory analyses, and explorative analysis therefore helps in generating new hypotheses (Johnson 2002, Guthery 2007). Unfortunately, the sample sizes are usually rather low while possible predictors are numerous. This widens confidence intervals around effect size estimates (Hocking 1976) and hampers the identification of influential predictors (Forstmeier & Schielzeth 2011), thereby reducing generality of the conclusions (Ginzburg & Jensen 2004). We are, therefore, inclined to adopt statistical practices that may yield a less complex (i.e., more parsimonious) model from which conclusions are subsequently drawn. Such practices can be classified as model simplification or model selection (Johnson & Omland 2004).

Model simplification uses metrics of predictor or model performance (most often null hypothesis significance testing, NHST) to adjust the model to the available data. In the case of NHST, model simplification involves multiple tests of the probability of obtaining the data given multiple, increasingly simple null hypotheses. Model selection, on the other hand, evaluates multiple models or statistical hypotheses in

a single step, using a model-specific metric to estimate the plausibility of each model given the data set. In both model simplification and model selection, the goal can be twofold. First, identifying the important predictors and discarding the unimportant predictors, which can be vital, e.g. for population management decisions. Second, reliably estimating the effects of a given variable, which can be important, e.g. for fine-tuning management actions.

In ecology, there are three typical methods used to solve statistical problems involving multiple independent variables. The first is fitting all predictors simultaneously and drawing conclusions from this saturated or full model. The second and most widely adopted approach is model simplification by stepwise selection (Miller 1992). This involves a sequential removal or reintroduction of terms, until neither the inclusion nor the exclusion of any term can be justified based on a given threshold criterion (most often p values). The third approach is information-theoretic (IT) model selection (Johnson & Omland 2004). IT model selection means the calculation of information criteria (most frequently the Akaike Information Criterion, AIC or its derivatives; Burnham & Anderson 2002) for each model in a pre-determined model set. Information criteria allow us to trade off model fit and parsimony in a search for the most suitable model given the data set (Ward 2008). Recently, IT methods were strongly advocated and NHST-based model simplification severely criticized (e.g. Johnson 1999, Anderson *et al*. 2000, Whittingham *et al*. 2006), but the use of IT methods in ecology is limited and the majority of studies continue to use NHST (Garamszegi 2011, Hegyi & Garamszegi 2011). There is also a mixture of model simplification and AIC, the "AIC-stepwise" method, but this is widely rejected in the IT literature for philosophical reasons and we do not discuss it further (*see* Burnham *et al*. 2011).

The caveats of stepwise regression are assumed to be well known and abundantly demonstrated (Burnham & Anderson 2002, Whittingham *et al*. 2006, Mundry & Nunn 2009, Forstmeier & Schielzeth 2011), although some of these problems can be mitigated while others can also be detected in IT methods (Hegyi & Garamszegi 2011). As a model simplification method, stepwise selection simplifies the model based on parameter estimates coming from the current data set. This is called "data dredging" and it may simultaneously reduce the generality of the conclusions and lead to parameter estimation bias (particularly with weak predictors), overly optimistic goodness of fit estimates and downward biased uncertainty estimates and p values (Pope & Webster 1972, Flack & Chang 1987, Copas & Long 1991). Stepwise regression, particularly with many predictors, may also increase the probability of detecting spurious ("uninformative") predictors and at the same time may fail to detect informative predictors with small effect sizes (Derksen & Keselman 1992, Murtaugh 1998). Furthermore, stepwise regression can be unstable in that the composition of the final model may depend on small changes in the data and on the details of the removal/introduction approach (James & McCulloch 1990, Derksen & Keselman 1992). Finally, stepwise regression focuses on a single final model and thereby ignores uncertainty in the choice of the final model, i.e. the fact that other models may fit the data similarly well (Draper 1995, Whittingham *et al*. 2006). It seems that parameter estimation problems with stepwise regression are attenuated with increasing sample size, while parameter identification problems remain even at large sample sizes (Austin 2008).

The performance of IT model selection received less scrutiny, which may stem from the fact that the detailed elaboration of the approach is a relatively recent development (Buckland *et al*. 1997, Burnham & Anderson 2002). Monte Carlo methods that generate data from known underlying effects were often used to compare AIC or its small-sample adjustment AICc (Hurvich & Tsai 1989) with other information criteria such as BIC or KIC (e.g. Mills & Prasad 1992, Kuha 2004, Kim & Cavanaugh 2005, Seghouane 2006). These studies almost uniformly noted the tendency of AIC(c) to select relatively complex 'best' models, in particular when samples sizes are low. We are aware of only two simulation studies that compared the parameter selection properties of AIC(c) and stepwise regression. One of them (Burnham & Anderson 2002: 121–124) used simulated data with a large number of

small to medium effects, i.e. a complex underlying process ("tapering effect sizes"). This study found that both stepwise selection and AICc underestimated the complexity of the underlying process, but this effect was more pronounced in AICc. Another study (Raffalovich *et al*. 2008) used a data set with a mixture of predictors with zero and nonzero effects and was thus based on the assumption that some of the predictors might be truly unrelated to the response (Mundry 2011 argues that this is likely to be a rather common situation in studies of ecology and behavior). Furthermore, some predictors were highly inter-correlated. Here, stepwise regression performed relatively well while AIC and AICc performed very poorly, mostly due to their strong tendency to select spurious predictors.

Fitting the full model, as the simplest approach, is often not explicitly included in simulation studies (*see* Austin 2008 and references therein). The few studies we are aware of that did so (Lovell 1983, Mundry & Nunn 2009, Forstmeier & Schielzeth 2011) used data in which all effects were set to zero at the population level. They showed an increased probability of finding significant predictors (Lovell 1983, Forstmeier & Schielzeth 2011) or significant models (Mundry & Nunn 2009) in stepwise regression than in the full model approach. One major benefit proposed for stepwise regression compared with full models is the higher probability to detect predictors with non-zero effects (e.g. Hocking 1976, Aiken & West 1991). These were absent from the above simulations, which calls for further work on this topic.

Surprisingly, to the best of our knowledge, there is no study that has tested the relative suitability of the three methods: full models, stepwise regression and the AIC(c)-IT approach together using simulated variables with known population-level effect sizes. Here, we attempt to fill this gap by using simulated data that include both informative predictors and uninformative predictors (we use the term uninformative predictors for those independent variables that are uncorrelated with the response at the population level and informative predictors for those that are correlated with the response). We use various combinations of sample size, number of predictors, collinearity among predictors, and the ratio of informative predictors among all predictors to cover some scenarios typically encountered in ecological and evolutionary studies. We test three different IT approaches to parameter identification, and model averaging as a method for parameter estimation using multiple models (Buckland *et al*. 1997, Richards *et al*. 2011). We repeat the same analyses with the Bayesian information criterion (BIC, Schwarz 1978). This criterion has been found to give better results in terms of parameter identification than AIC(c) (Ward 2008).

In the stepwise process, parameters will drop out from the model if they are below a certain threshold of effect size or statistical significance (Whittingham *et al*. 2006). Among significant effects that are retained, there is an increased chance of finding overestimated effect sizes due to sampling variation, since these will tend to become significant (Copas & Long 1991, Forstmeier & Schielzeth 2011). This bias will be more serious for predictors with weak true underlying effect sizes at the population level (Sauerbrei 1999), which applies to most ecological predictors. On the other hand, effect sizes for parameters that are not included in the reduced model are regarded as zero. This follows logically from the threshold-based philosophy of stepwise regression (Anderson *et al*. 2000, Martínez-Abrain 2007), but it certainly introduces massive parameter estimation bias in the presence of sampling error (Whittingham *et al*. 2006, Lukacs *et al*. 2010). It has recently been suggested that the well known extreme parameter estimation bias of stepwise regression can be mitigated by a simple adaptation of the process for parameter estimation: the reintroduction of removed terms one by one to the final model (Hegyi & Garamszegi 2011; hereafter called the SRPE method, standing for "stepwise-reintroduction for parameter estimation"). This gives estimates of effect sizes for all predictors, while still capitalizing on the reduced uncertainty intervals and potentially reduced bias of smaller models. Our present study employs the SRPE method for parameter estimation in stepwise regression and it therefore represents the first testing of the method.

Mainstream advocates of IT methods may find our simulation exercise irrelevant. There is

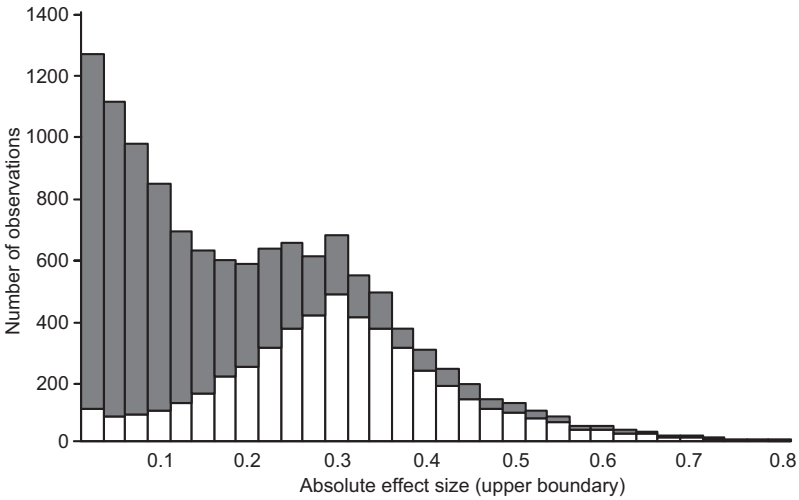

Fig. 1. Frequency distribution of absolute (i.e. unsigned) estimated effect sizes from our simulations (n = 12 000) when pooling all settings and applying full models. White and grey bars denote informative and uninformative predictors, respectively. The graph was truncated at 0.8 for better visibility because only 9 of the estimates fell above this value.

a philosophical difference between NHST and IT methods (model simplification versus model selection, *see* above). Due to this, some IT model selection criteria such as AIC(c) were designed for a situation where there are no true zero effects but rather a "tapering" frequency distribution of effect sizes. In this concept, "identification" of predictors is meaningless and the focus is on correct parameter estimation. In the case of AIC(c), this means the minimization of the Kullback-Leibler distance (Burnham & Anderson 2002). From another viewpoint, a trade-off may be assumed between parameter identification and estimation performance, and AIC(c) is on the estimation while BIC on the identification side (Yang 2005). We must keep this in mind when interpreting our results. However, comparing the approaches in the way we do here still makes sense and must be done because this is how AIC(c) is generally used in ecology: as a replacement for classical parameter identification approaches (reviewed in Hegyi & Garamszegi 2011).

Material and methods

Simulated data sets

1600 data sets were generated, all consisting of a dependent variable, informative predictors and uninformative predictors. Informative and uninformative predictors were generated so that their population-level bivariate correlations with the dependent variable were 0.3 and 0.0, respectively. Effect sizes in ecology are generally small (Møller & Jennions 2002), so we chose the lower threshold of medium effect size (Cohen 1988) for our informative predictors. Each data set is a random sample from a population with the respective known correlation value. As in any empirical data set, sampling variation causes sample correlations to deviate from the underlying population-level correlation. Indeed, when applying fully parameterized models and pooling informative predictors and uninformative predictors, the overall distribution of sample-level parameter estimates (n = 12 000) was continuous (though noticeably bimodal) (Fig. 1). Predictor variables and response values were sampled from normal distributions with unit variance throughout. The data sets belonged to 16 categories defined by the following four attributes:

1. Sample size was 30 or 200 which corresponded to relatively small and large ecological samples, respectively. There are sometimes much greater ecological samples than 200, but these typically involve repeated data points from the same entities. In a sample of 229 published effect sizes from various fields of ecology, evolution and physiology, only two had an effective sample size greater than 200 (A. P. Møller, pers. comm.). The sample size of 200 is also much above the interquartile range of any behavioral study

type reviewed by Taborsky (2010, *see* his Fig. 1; also note that behavioral samples are typically smaller than ecological samples). On the other hand, there are great numbers of ecological and behavioral samples smaller than 30 data, but using models with multiple predictors would be grossly inappropriate in these small samples (Green 1991, Stevens 2002). Recent simulations with a similar goal used similar or narrower sample size ranges (n = 170 to 230 in Mundry & Nunn 2009, n = 50 and n = 200 in Forstmeier & Schielzeth 2011).

2. Correlation among all predictors was uniformly 0 or 0.3, which implies no or moderate collinearity. There is to our knowledge no published information on average predictor correlations in ecology and evolution. We set these to a similar value as predictor–response correlations, as we believe that the two should be similar in magnitude if different predictors are not redundant but represent different causal or functional pathways. This point clearly deserves future investigation.
3. The number of predictors was 5 or 10, both of which are realistic values for an ecological study (Forstmeier & Schielzeth 2011).
4. The proportion of informative predictors among the complete set of predictors (hereafter informative predictor ratio, IPR) was either 0.2 or 0.6. These IPR values may be viewed as referring to situations with low *vs*. high amounts of prior information on the study system (Johnson & Omland 2004).

The uniform interrelation of predictors is a special case that was chosen as an example because of its interpretational simplicity. Several previous simulation studies used similar settings (e.g. Burnham & Anderson 2002, Freckleton 2011) and we refer to this analysis as the "uniform interrelation analysis". To confirm the generality of our findings under other within-model correlation structures, we also conducted another simulation exercise in which a separate array of 1600 data sets had 12 predictors, a fixed IPR of 0.5, and 2 × 3 different predictor interrelation types within a single data set ([informative, uninformative] × [uncorrelated, correlated with an informative predictor, correlated with an uninformative predictor]). This analysis ("varying interrelation analysis") was conducted with two sample sizes, and yielded similar results as the uniform interrelation analysis under the relevant settings (i.e. many predictors, large IPR). Therefore, we report only the uniform interrelation analysis in the main body of our paper, while the varying interrelation analysis and its interpre-

Table 1. The types of data sets used in our simulations.

Interrelation	Sample size	Correlation among predictors	Number of predictors	Informative predictor ratio
Uniform	30	Uniformly 0.0	5	0.2
Uniform	30	Uniformly 0.0	5	0.6
Uniform	30	Uniformly 0.0	10	0.2
Uniform	30	Uniformly 0.0	10	0.6
Uniform	30	Uniformly 0.3	5	0.2
Uniform	30	Uniformly 0.3	5	0.6
Uniform	30	Uniformly 0.3	10	0.2
Uniform	30	Uniformly 0.3	10	0.6
Uniform	200	Uniformly 0.0	5	0.2
Uniform	200	Uniformly 0.0	5	0.6
Uniform	200	Uniformly 0.0	10	0.2
Uniform	200	Uniformly 0.0	10	0.6
Uniform	200	Uniformly 0.3	5	0.2
Uniform	200	Uniformly 0.3	5	0.6
Uniform	200	Uniformly 0.3	10	0.2
Uniform	200	Uniformly 0.3	10	0.6
Varying	36	Six different types	12	0.5
Varying	240	Six different types	12	0.5

tation can be found in Appendix 1. All types of data sets used in our simulations are summarized in Table 1.

Analysis of simulated data sets

In our present simulations, interaction terms were not included because these generate extra complexity and interpretational problems (Aiken & West 1991, Engqvist 2005, Grueber *et al*. 2011). Interactions and non-linear terms will have to be examined in future simulations. All data sets were subject to three types of analyses, focusing on the main effects of variables. In the full model approach, hypothesis testing and parameter estimation were done in the saturated model containing all independent variables. In the stepwise approach, terms were automatically and sequentially removed and then reintroduced, starting with the full model and using removal and reintroduction thresholds of $p = 0.05$. Finally, for the model selection approach we calculated values of the bias-corrected AIC (AICc) and BIC for all combinations of the independent variables. This "all subsets" IT analysis is the most widely used IT approach in ecological studies (Hegyi & Garamszegi 2011, *see* also Whittingham *et al*. 2006, Lukacs *et al*. 2010, Symonds & Moussalli 2011). It signifies a situation where there is little or no prior information and hence characterizes an explorative analysis (*see* discussion in Eberhardt 2003, Hegyi & Garamszegi 2011). The results of these modeling approaches were then analyzed according to two different statistical paradigms, corresponding to the two main aims of statistical modeling: parameter identification and parameter estimation.

The parameter identification approach seeks to distinguish important and unimportant predictors. Following this paradigm in the three analysis types, we identified as important predictors (1) significant predictors ($p < 0.05$) in the full model, and (2) significant predictors (in this case, all predictors) in the final model of the stepwise process. For AICc or BIC, we employed three different methods as there is no general agreement in the literature. We chose (3) predictors in the model with the smallest AICc or BIC value (the "AICc-best" or "BIC-best" model), and predictors of AICc or BIC weights (W_i) greater than (4) 0.7 or (5) 0.9. From these five types of results, we calculated different identification error measures for each data set and result type. False detection probability (or Type I error rate, hereafter FPOS from "false positive") was defined as the number of uninformative predictors classified as important, divided by the total number of uninformative predictors. Detection failure probability (or Type II error rate, hereafter FNEG from "false negative") was defined as the number of informative predictors identified as unimportant, divided by the total number of informative predictors. Total or identification error probability (hereafter FTOT from "false total") was a composite error measure which gave the total number of misidentified predictors divided by the total number of predictors. FTOT is a combination of two different measures (FPOS and FNEG), and its value depends on the values of both constituents which are themselves negatively correlated with each other. Therefore, as the exact meaning of a given value of FTOT is unclear, FTOT could be regarded as a derived quantity from a statistical viewpoint, but it is relevant for an ecologist whose goal is to correctly assign all predictors of interest as either important or less important. In empirical studies, the identity of informative and uninformative predictors is always unknown, and the researcher may be equally interested in retaining informative and discarding uninformative predictors. FTOT measures the combined success of the two under the given settings. We finally had 3×5 different identification error measures for each data set, due to the five methods employed in parameter identification.

The parameter estimation approach aims at quantifying the effect sizes for all terms and is thus not exclusively focused on important predictors. Effect size estimates are critical for quantitatively evaluating our findings and reporting small (and thus non-significant) effect sizes can help mitigate publication bias (Nakagawa & Cuthill 2007). For the full model approach, effect size information came from the saturated model. In the stepwise process, we used the SRPE method that estimates effect sizes for the removed terms by reintroducing them into the final model one by one (Hegyi & Garamszegi 2011). Finally, for the AICc or BIC model selection approach, we

applied model averaging across the whole candidate model set (Buckland *et al*. 1997, Burnham & Anderson 2002). This procedure calculates average effect sizes across the models while weighting the contribution of a given effect by the plausibility (AICc or BIC weight) of the given model. We employed the unconditional estimator of model averaging (section 5.3 in Burnham & Anderson 2002) which uses information from all models and substitutes the effect size with zero for models without the respective parameter. After acquiring the effect size estimates for all individual parameters in each data set, we calculated mean effect sizes for all informative predictors and all uninformative predictors in each data set, which yielded two values (average effect sizes of informative and of uninformative predictors) for each combination of data set and modeling approach. We compared these averaged effect size values in the following.

Full model fitting and automatic stepwise analyses were conducted using the GLZ procedure and the SCL programming environment of Statistica 5.5 (StatSoft, Inc). Reintroductions for effect size estimation in the stepwise process, and all analyses involving AICc or BIC, were calculated in the REG procedure of SAS 9.1 (SAS Institute, Inc). AICc was *post-hoc* calculated in SAS from AIC output values given by the REG procedure. The different programs were using the same data sets, and their results for a given data set and procedure were the same. The reason for using multiple programs in the same simulation is purely practical (differences in data handling, programming and output options) and subjective. The findings were evaluated qualitatively, focusing on the magnitudes of identification errors and parameter estimation bias. Significance tests of the differences would be largely meaningless because quantitatively identical errors or estimates for any two methods are highly unlikely, and any small difference would become significant above a given sample size.

Results

Parameter identification

Mean identification errors for all combinations of our system settings are visualized in Appendix 2. Here we concentrate on major effects and interactions. The number of predictors had very small effects on parameter identification errors so it is neither plotted nor discussed here. When pooling the three methods (full model, stepwise and AICc-IT), sample size, predictor interrelation and IPR exerted a strong interactive effect on FPOS (Fig. 2). False detections always became more frequent when predictors were correlated, but there was an extreme at large samples and high IPR where mean FPOS was approximately 80% (Fig. 2A bottom right). FNEG, on the other hand, was low at large sample sizes regardless of other settings (Fig. 2B right-hand side), and it was overall highest with small samples, correlated predictors and high IPR (Fig. 2B bottom left). As a combined error measure, FTOT showed effects that were dominated by the dominant predictor type in the given setting (informative at high IPR and uninformative at low IPR). FTOT was low with large samples and uncorrelated predictors, while in other settings it was always higher at high than at low IPR.

The effects of model type (NHST and AICc-IT methods) on identification errors were largely independent of other system settings, with the exception that FNEG was uniformly low in large samples (*see* Appendix 2). Therefore, the model type effect is plotted separately (Fig. 3). With respect to FPOS (Fig. 3A and D), $W_i > 0.7$ gave consistently similar error rates to stepwise regression. Full model and stepwise also differed relatively little in FPOS, with a small overall advantage of full models in this error type. AICc-best always gave the highest FPOS while $W_i > 0.9$ the lowest. Where FNEG was present (small samples), it gave a roughly reversed ranking of model types compared with FPOS (Fig. 3B), i.e. in increasing order: (1) AICc-best, (2) stepwise and $W_i > 0.7$ (similar), (3) full model, (4) $W_i > 0.9$. The difference between full model and stepwise was more consistent for FNEG than for FPOS (*see* Appendix 2). Finally, FTOT as a combined error measure showed model type effects that were generally very small relative to the overall magnitude of the error (Fig. 3C and F). Full models, stepwise regression and $W_i > 0.7$ were always very similar in FTOT. At

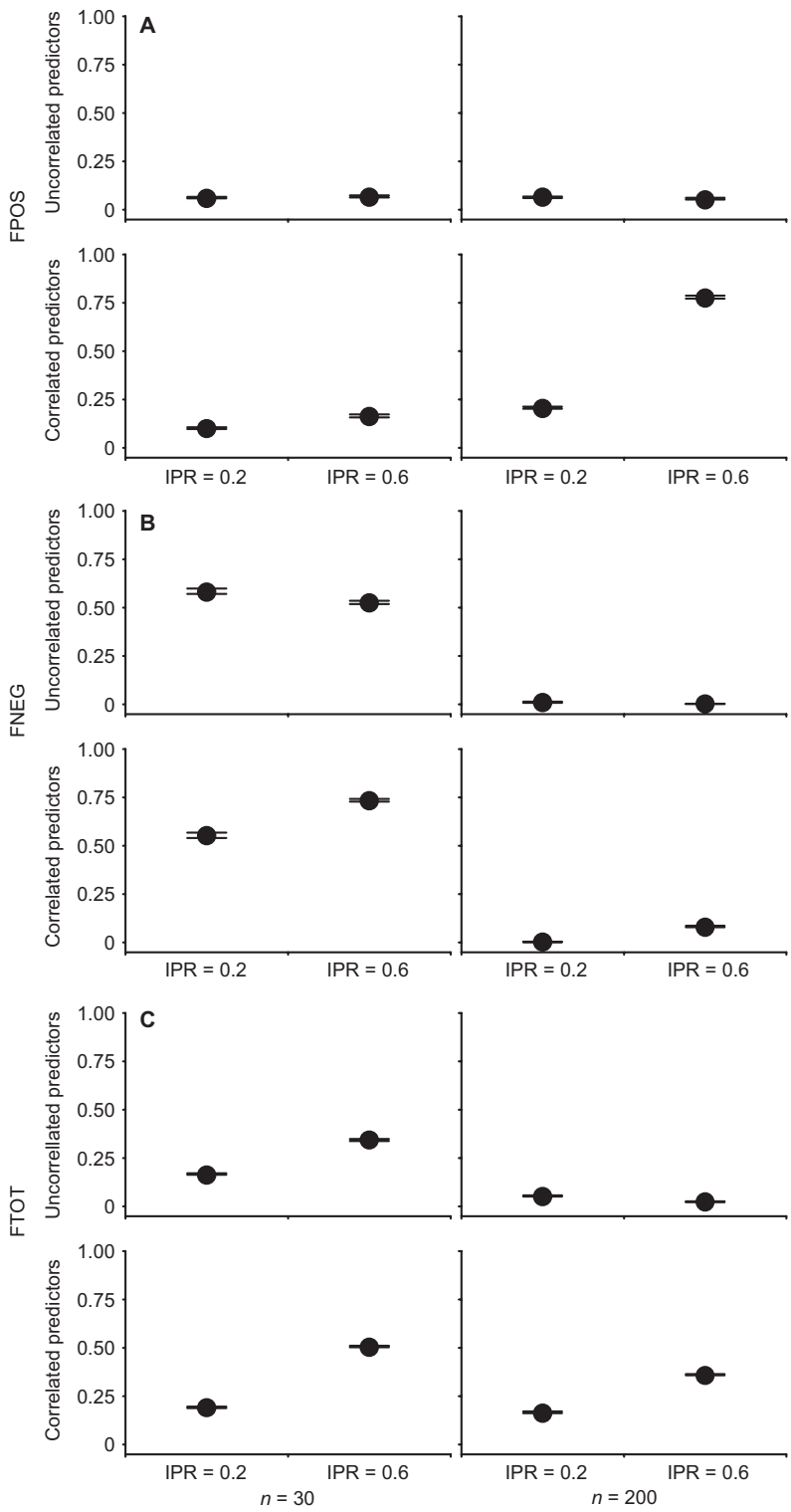

Fig. 2. Rates of (**A**) false detection, (**B**) detection failure, and (**C**) identification error in relation to simulation settings (means ± SEs). Results with AICc are shown; they are similar when using BIC. FPOS = false detection probability, FNEG = detection failure probability, FTOT = total identification error probability, IPR = informative predictor ratio.

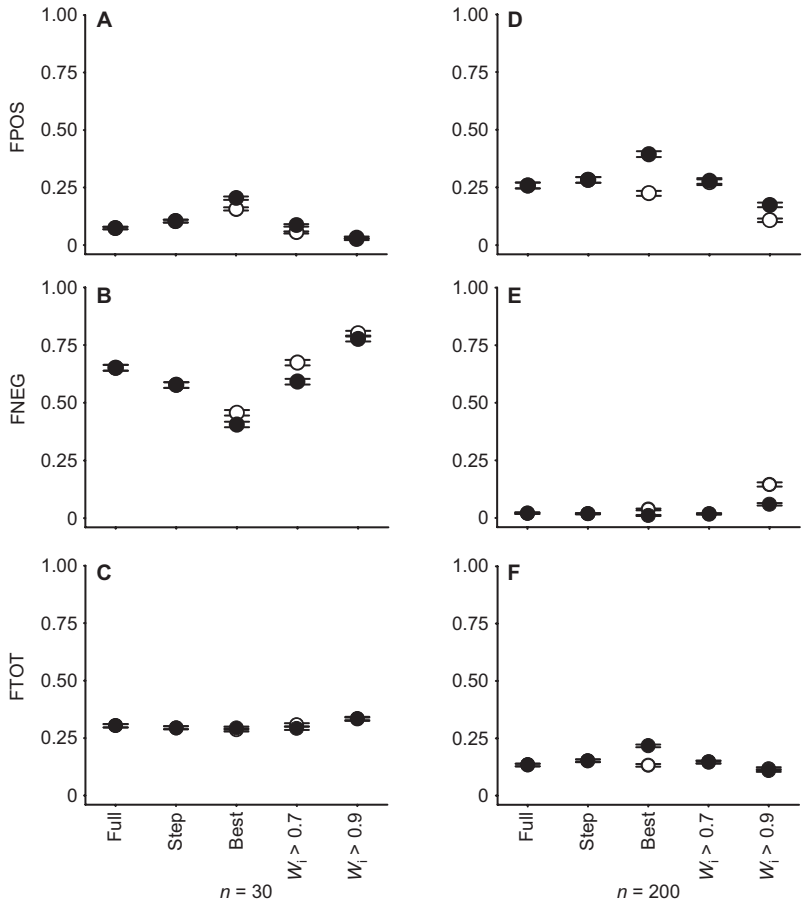

Fig. 3. Rates of (**A**, **D**) false detection, (**B**, **E**) detection failure, and (**C**, **F**) identification error in relation to model type at two sample sizes (shown are means ± SEs). There was little difference in the pattern with correlated and uncorrelated predictors so we plotted the pooled results here. Results for information theoretic methods refer to AICc (●) and BIC (○). Full = full model, Step = stepwise, Best = AICc or BIC best model.

large sample sizes, AICc best was the worst and $W_i > 0.9$ the best, while at small sample sizes, there was virtually no model type effect. When using BIC instead of AICc (open circles in Fig. 3; *see* also Appendix 3), FPOS generally decreased while FNEG increased. The only setting where FTOT was visibly affected was the analysis of the "best" model at large sample sizes, where BIC best performed better than AICc best.

Effect size estimation

As in the case of parameter identification, full details of the results are visualized in Appendix 2. For both informative and uninformative predictors, the accuracy of effect size estimation was dominated by predictor correlation and IPR, while statistical methodology played a minor role. There was hardly any estimation bias with uncorrelated predictors, irrespective of other settings (these results are not plotted here, *see* Appendix 2). With correlated predictors (Fig. 4), the direction of bias reversed from the low to high IPR for informative predictors (Fig. 4A), while it remained the same (negative) for uninformative predictors, but with a stronger bias at high IPR (Fig. 4B).

The effects of model type (NHST and AICc-IT methods) were inconsistent and generally very small relative to the above effects. These results again refer to correlated predictors as bias was negligible with uncorrelated predictors. For informative predictors, a robust model type effect appeared at low IPR, where bias increased from SRPE to AICc model averaging to full models (Fig. 4A top half). For uninformative predictors, SRPE had the largest bias of the three methods with large samples and low IPR (Fig. 4B top right), while it had the lowest bias

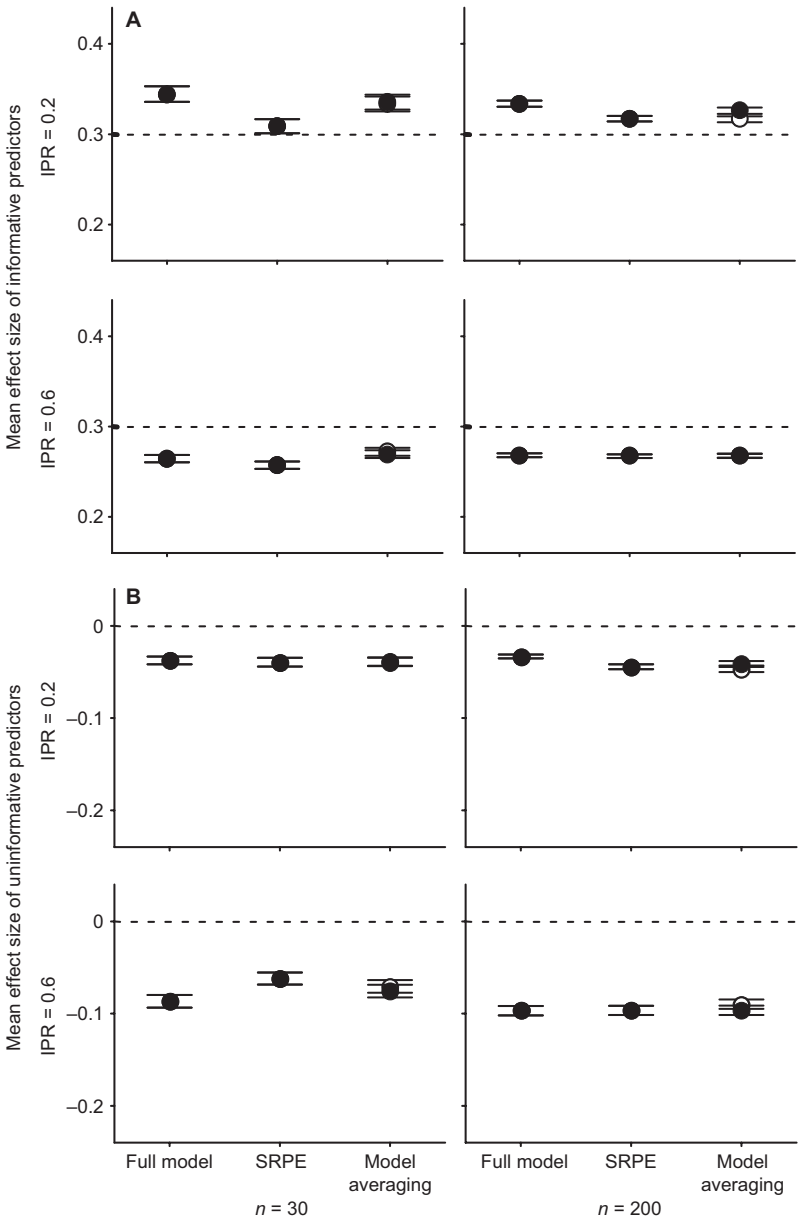

Fig. 4. Mean estimated effect sizes of (**A**) informative and (**B**) uninformative predictors in relation to model type and simulation settings (shown are means ± SEs) in settings with correlated predictors. There is very little bias for uncorrelated predictors and those results are presented in Appendix 2. Results for model averaging refer to AICc (●) and BIC (○). Dotted lines indicate the effect size specified in the simulation (i.e. the true population level effect size of the given predictor type). IPR = informative predictor ratio, SRPE = stepwise-reintroduction for parameter estimation.

of the three with small samples and high IPR (Fig. 4B bottom left). The AICc model averaging method was intermediate in both cases. There was little model type effect in the remaining settings, although the magnitude of the bias was sometimes very marked. Using BIC instead of AICc changed the results very slightly (*see* also Appendix 3). In nearly all cases with a visible difference, BIC reduced parameter estimation error compared to AICc, but model averaging still showed a similar performance as the NHST methods, particularly SRPE.

Discussion

Identification error in relation to sample size and collinearity

Our focus here was to compare three statistical

model types under a number of realistic settings, but the effect of these settings on parameter identification and estimation error generally overshadowed the effect of model type itself. As expected, false detection and detection failure were roughly inversely proportional (*see* also Derksen & Keselman 1992, Raffalovich *et al*. 2008). False detection dominated in large samples and detection failure in small samples. Ecological studies may sometimes have even more extreme sample sizes than our "large" and "small" samples, so our simulations suggest that ecologists may quite often run into problems with identification errors. Correlation between predictors strongly amplified these issues. The adverse effects of multicollinearity have repeatedly been underlined in the literature (Graham 2003, Freckleton 2011), but predictors are generally considered collinear only from an r value of 0.5 upwards. Here, we showed that predictor correlation as low as 0.3 can severely impair parameter identification success.

The increase of the dominant detection error rate with collinearity depended on IPR. The identity of informative predictors is by definition unknown to the researcher, so it seems crucial to remove predictor interrelation as much as possible from the data before analyzing them. This is best done in the design, e.g. by applying experimental treatments in a balanced and crossed manner. For cases in which experimental control is impractical, at least two methods have been recommended: residual regression and principal component regression (Graham 2003). Removing collinearity can be difficult with small samples in which moderate correlations between predictors may often become non-significant and therefore "undetected" (Jennions & Møller 2003, Nakagawa 2004), and if the underlying correlation structure is weak and the true values of correlations are unclear due to their broad error margins (García-Berthou 2002, Budaev 2010). Further studies are urgently needed to identify and test strategies to reduce bias in the analysis of such data sets. Finally, we note that the statistical removal of collinearity will lead to correct statistical results, but it will not clarify the relative importance and the causal relations of multiple, correlated explanatory variables. Again, the experimental manipulation of correlated factors is the most effective solution to this problem.

Identification error and model type

When discussing the model type comparisons here, we focus on overall patterns. Looking at the detailed results (*see* Appendix 2), the model type specific relative magnitudes of identification error and estimation bias can often be predicted from (1) the ratio of informative and uninformative predictors in the given setting and model type, (2) predictor interrelation, and (3) the effect size estimates of the given predictor type in the varying interrelation analysis.

Where both FPOS and FNEG were present (small samples), the two yielded a roughly inverse ordering of model types, again reflecting the trade-off between the two error types (*see* previous section). The first point of particular interest was the relative performance of full models and stepwise regression. Recent methodological papers criticized the use of stepwise regression techniques for parameter identification, and recommended alternative methods such as full models (Mundry & Nunn 2009, Forstmeier & Schielzeth 2011) and information theory (Whittingham *et al*. 2006). Among IT methods, the historically most-widely used one in ecology is the AICc-best approach which performed rather poorly in our simulations (and it was among the worst in the study of Raffalovich *et al*. 2008). It exhibited considerably lower detection failure in small samples than full or stepwise models, but much higher false detection rates at both small and large sample sizes, thereby often having higher overall identification error rate (FTOT) than the two NHST methods. However, these findings must be viewed with caution owing to our low confidence in the AICc-best model. Even with the extremely permissive threshold difference of dAICc < 2 from the best model (Symonds & Moussalli 2011), the AICc-best model was "truly" the best in generally less than 10% of the cases (results not shown here). We also simulated multimodel inference using parameter weights, which give a quantitative measure of support for a certain parameter in the candidate model set and thereby take into account model uncertainty (Burnham & Anderson 2002). Using the parameter weight threshold of $W_i = 0.7$ reduced FPOS and increased FNEG compared with AICc-best, producing very similar results to the NHST

methods. The threshold of W_i = 0.9 seemed suboptimal, yielding very low FPOS but considerable FNEG. In settings where the given error type reached notable levels, these differences between the AICc-based approaches were relatively immune to data set characteristics. Therefore, using parameter weights with a moderately high threshold value could provide a viable alternative to NHST methods, with the added benefits of enabling direct parameter and model comparisons (Symonds & Moussalli 2011).

Note that model simplification using AIC(c) may to some extent be doomed to fail because this information criterion was born under the philosophy of no true zero effects (Burnham & Anderson 2002). In other words, AIC(c) is less well designed for parameter identification than for parameter estimation (Burnham & Anderson 2002). It is therefore an important question whether the results would change qualitatively when using other information criteria with different characteristics. A straightforward choice would be BIC which may be slightly better designed for parameter identification and less well for predicting future observations than AIC(c), although the two differ only in their penalty term for model complexity (Yang 2005, Ward 2008). We therefore recalculated all analyses using BIC instead of AICc. The results exhibited a surprising similarity to those with AICc. As expected, BIC often shows lower FPOS and higher FNEG than AICc in parameter identification, but the total identification error rate only differs between AICc best and BIC best and at large sample sizes. However, despite its sample size specific advantage over AICc best, BIC best still does not represent the ideal solution for parameter identification, and the overall difference (across sample sizes) between the BIC based and the NHST methods is small.

Full models are used in a similar number of studies as stepwise methods, although we note that it is often not clear from a paper if and how model selection or simplification was conducted. Interestingly, one of the central reasons why stepwise regression has originally been recommended was the reduction of parameter detection uncertainty caused by the presence of uninformative predictors in the full model, i.e. reduced FNEG or Type II error rate (*see* e.g. Hocking 1976, Aiken & West 1991). However, reviews of the topic (Anderson *et al*. 2001, Whittingham *et al*. 2006) as well as simulation studies (Lovell 1983, Mundry & Nunn 2009, Forstmeier & Schielzeth 2011) emphasized the other side of the coin, which is increased FPOS (i.e. Type I error rate) in stepwise regression compared with full models. In our study, increased FPOS of stepwise regression was most marked in situations with low overall FPOS. There was very little FPOS difference between stepwise and full models in the case with highest overall FPOS (*see* Appendix 2). On the other hand, stepwise regression consistently reduced FNEG compared with full models in small samples, where FNEG was high, which is in line with our expectations (Hocking 1976, Miller 1992). When looking at overall identification error rate (FTOT), differences between stepwise and full models were generally small. If FPOS and FNEG are equally important to us (which will depend on the goal of the given study; Burnham & Anderson 2002), we cannot dismiss stepwise regression based on its performance in parameter identification. Indeed, reduction of a high FNEG when using stepwise regression can be of practical importance in small samples, which are commonplace in some fields of ecology and evolution (*see* also Moran 2003, Nakagawa 2004, Martínez-Abrain 2007). It might be proposed that higher among-predictor correlations than used here would make stepwise selection more prone to misidentifying predictors. However, a previous study using high predictor correlations (one-third of them exceeding 0.5) yielded very similar conclusions to ours, with better performance of stepwise regression than AIC(c) (Raffalovich *et al*. 2008). On the other hand, we believe that between-predictor correlations higher than 0.5 are principally not due to biological relationships but wrong parameterization such as the use of mathematically or mechanistically interdependent variables, and they can and should be avoided (*see* also Freckleton 2011). The true value of predictor correlations and its effects must be studied further. As a guideline for future simulation work, there are alternative model simplification methods such as ridge regression or the lasso that are not routinely used in our field, and some of these might reduce overall identification error compared with

stepwise or full models (Murtaugh 2009, Dahlgren 2010).

Effect size estimation bias

The most important message from the results of the effect-size based approach is that bias in the estimates is mainly due to correlation among predictors (Freckleton 2011). With uncorrelated predictors, there was hardly any bias irrespective of sample size, predictor number, IPR or statistical approach (*see* Appendix 2). With correlated predictors, even at the low correlation level of 0.3 we used, there was sometimes serious estimation bias which varied in magnitude and also in direction in response to system settings, particularly IPR.

When looking at the effect of model type, the most surprising pattern is the relative performance of full model and stepwise regression. If we are to follow the philosophy of effect size based inference in stepwise regression (Nakagawa 2004, Nakagawa & Cuthill 2007), we must obtain meaningful (i.e. non-zero) effect size estimates also for the removed terms, for example by reintroducing them one by one into the final model (Hegyi & Garamszegi 2011). Here we tested this SRPE method for the first time. Surprisingly, the method did not increase the effect-size bias compared with the full model in most settings with an observable model type effect (*see* the top and the bottom left of Fig. 4). It therefore seems that the increased effect-size bias in stepwise regression compared with full models is largely due to the failure to present non-zero estimates for the removed terms (Whittingham *et al*. 2006, Lukacs *et al*. 2010). However, in our simulations the effect-size bias was sometimes even lower in stepwise regression (SRPE) than in full models, and this requires further explanation. The reduced bias likely reflects the reduction of parameter number, and therefore a reduced distortion due to collinearity (Hocking 1976, Graham 2003, Raffalovich *et al*. 2008) in stepwise compared with full models. Our results indicate that the reintroduction of removed parameters into the final model largely eliminates and may even reverse the effect-size bias caused by stepwise regression when compared with full models (*see* also detailed results in Appendix 2).

The estimation accuracy of AICc model averaging was always intermediate between the full model and stepwise methods. Possible ways of further improving the performance of model averaging are to establish confidence model sets (Whittingham *et al*. 2006) and to further refine these by removing the unsupported, complex counterparts of nested submodels (Richards 2008). Establishing such methods in ecological statistics will require agreement on threshold values for confidence sets, and publicly available statistical software for implementing Richards' algorithm in large candidate sets. Future studies should also examine the performance of information criteria other than AIC(c) (box 1 in Grueber *et al*. 2011). When we recalculated our model averaging with BIC, even the setting-specific differences from AICc were very small. Interestingly, BIC seemed to relatively consistently reduce the estimation bias compared with AICc, although the differences were extremely small. The direction of the difference therefore does not agree with the "trade-off" principle of Yang (2005) as BIC seemed to outperform AICc in both identification and estimation. Similarly to what we found for parameter identification, BIC model averaging did not stand out as a better method overall for parameter estimation than the NHST methods. We can therefore conclude that the surprisingly small overall difference in identification and estimation error we detected between full model, stepwise and IT approaches does not change when replacing AICc with BIC.

Conclusions and future directions

Our results highlight that the choice of the statistical method may often be less important for obtaining reliable results than aspects of the quality of the data set such as sample size and predictor interrelation (Garamszegi *et al*. 2009). This finding must be borne in mind during an active debate over competing ways of statistical modeling (Stephens *et al*. 2007, Guthery 2008), when factors that are perhaps more influential than model type, such as the collinearity of predictors, are seldom mentioned (Zuur *et al*. 2010,

Freckleton 2011). Stepwise regression showed similar overall parameter identification accuracy to full models and the exact outcome depended on data set attributes. The SRPE approach involving parameter reintroduction (Hegyi & Garamszegi 2011) seemed to mitigate the well-known extreme parameter estimation bias of classical stepwise regression and even surpassed the estimation accuracy of full models in some settings with correlated predictors. The performance of the AICc- and BIC-based approach was extremely sensitive to the exact method. The parameter identification success of the parameter weight method with a moderately high threshold value was similar to the NHST methods, and the same applied to the parameter estimation accuracy of AICc and BIC model averaging.

An important extension of our present work will be the testing of effect size bias in the presence of interaction terms. Full models with interactions provide an interpretable output for main effects only if the predictors are uncorrelated, centered, bivariate normal and perfectly balanced (Aiken & West 1991). These conditions are hardly ever met in ecology or evolution, especially in the case of categorical predictors (although *see* recommendations for centering in Schielzeth 2010). Future studies will particularly have to show whether effect sizes from the SRPE approach are more or less biased than those from full models in the presence of non-significant interaction terms (Hegyi & Garamszegi 2011).

Acknowledgments

We are indebted to Holger Schielzeth for preparing the data sets and for detailed comments on the manuscript in multiple stages of this work. We thank the organizers and participants of the statistical symposia at ISBE 2008 Cornell and ISBE 2010 Perth for inspiration. We are grateful to Benjamin M. Bolker for helpful comments on a previous version of the manuscript, and Bob O'Hara and an anonymous reviewer for comments on the current version. The study was supported by OTKA grants (PD72117 and K101611) and a Bolyai fellowship to G.H.

References

Aiken, L. S. & West, S. G. 1991: *Multiple regression: testing and interpreting interactions*. — Sage, Newbury Park, London.

Anderson, D. R., Burnham, K. P. & Thompson, W. L. 2000: Null hypothesis testing: problems, prevalence, and an alternative. — *Journal of Wildlife Management* 64: 912–923.

Anderson, D. R., Burnham, K. P., Gould, W. R. & Cherry, S. 2001: Concerns about finding effects that are actually spurious. — *Wildlife Society Bulletin* 29: 311–316.

Austin, P. C. 2008: The large-sample performance of backwards variable elimination. — *Journal of Applied Statistics* 35: 1355–1370.

Buckland, S. T., Burnham, K. P. & Augustin, N. H. 1997: Model selection: an integral part of inference. — *Biometrics* 53: 603–618.

Budaev, S. V. 2010: Using principal components and factor analysis in animal behaviour research: some caveats and guidelines. — *Ethology* 116: 472–480.

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

Burnham, K. P., Anderson, D. R. & Huyvaert, K. P. 2011: AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. — *Behavioral Ecology and Sociobiology* 65: 23–35.

Cohen, J. 1988: *Statistical power analysis for the behavioural sciences*, 2nd ed. — Lawrence Erlbaum Associates, Hillsdale, NJ.

Copas, J. B. & Long, T. 1991: Estimating the residual variance in orthogonal regression with variable selection. — *Statistician* 40: 51–59.

Dahlgren, J. P. 2010: Alternative regression methods are not considered in Murtaugh (2009) or by ecologists in general. — *Ecology Letters* 13: E7–E9.

Derksen, S. & Keselman, H. J. 1992: Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. — *British Journal of Mathematical and Statistical Psychology* 45: 265–282.

Draper, D. 1995: Assessment and propagation of model uncertainty (with discussion). — *Journal of the Royal Statistical Society Series B* 57: 45–97.

Eberhardt, L. L. 2003: What should we do about hypothesis testing? — *Journal of Wildlife Management* 67: 241–247.

Engqvist, L. 2005: The mistreatment of covariate interaction terms in linear model analyses of behavioural and evolutionary ecology studies. — *Animal Behaviour* 70: 967–971.

Flack, V. F. & Chang, P. C. 1987: Frequency of selecting noise variables in subset regression analysis: a simulation study. — *American Statistician* 14: 84–86.

Forstmeier, W. & Schielzeth, H. 2011: Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. — *Behavioral Ecology and Sociobiology* 65: 47–55.

Freckleton, R. P. 2011: Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. — *Behavioral Ecology and Sociobiology* 65: 91–101.

Garamszegi, L. Z. 2011: Information-theoretic approaches

to statistical analysis in behavioral ecology: an introduction. — *Behavioral Ecology and Sociobiology* 65: 1–11.

Garamszegi, L. Z., Calhim, S., Dochtermann, N., Hegyi, G., Hurd, P. L., Jørgensen, C., Kutsukake, N., Lajeunesse, M. J., Pollard, K. A., Schielzeth, H., Symonds, M. R. E. & Nakagawa, S. 2009: Changing philosophies and tools for statistical inferences in behavioural ecology. — *Behavioral Ecology* 20: 1363–1375.

García-Berthou, E. 2002: On the misuse of residuals in ecology: testing regression residuals vs. the analysis of covariance. — *Journal of Animal Ecology* 70: 708–711.

Ginzburg, L. R. & Jensen, C. X. J. 2004: Rules of thumb for judging ecological theories. — *Trends in Ecology and Evolution* 19: 121–126.

Graham, M. H. 2003: Confronting multicollinearity in ecological multiple regression. — *Ecology* 84: 2809–2815.

Green, S. B. 1991: How many subjects does it take to do a regression analysis? — *Multivariate Behavioral Research* 26: 499–510.

Grueber, C. E., Nakagawa, S., Laws, R. J. & Jamieson, I. G. 2011: Multimodel inference in ecology and evolution: challenges and solutions. — *Journal of Evolutionary Biology* 24: 699–711.

Guthery, F. S. 2007: Deductive and inductive methods of accumulating reliable knowledge in wildlife science. — *Journal of Wildlife Management* 71: 222–225.

Guthery, F. S. 2008: Statistical ritual versus knowledge accrual in wildlife science. — *Journal of Wildlife Management* 72: 1872–1875.

Hegyi, G. & Garamszegi, L. Z. 2011: Using information theory as a substitute for stepwise regression in ecology and behavior. — *Behavioral Ecology and Sociobiology* 65: 69–76.

Hocking, R. R. 1976: The analysis and selection of variables in linear regression. — *Biometrics* 32: 1–49.

Hurvich, C. M. & Tsai, C.-L. 1989: Regression and time series model selection in small samples. — *Biometrika* 76: 297–307.

James, F. C. & McCulloch, C. E. 1990: Multivariate analysis in ecology and systematics: panacea or Pandora's box? — *Annual Review of Ecology and Systematics* 21: 129–166.

Jennions, M. D. & Møller, A. P. 2003: A survey of the statistical power of research in behavioral ecology and animal behavior. — *Behavioral Ecology* 14: 438–445.

Johnson, D. H. 1999: The insignificance of statistical significance testing. — *Journal of Wildlife Management* 63: 763–772.

Johnson, D. H. 2002: The importance of replication in wildlife research. — *Journal of Wildlife Management* 66: 919–932.

Johnson, J. B. & Omland, K. S. 2004: Model selection in ecology and evolution. — *Trends in Ecology and Evolution* 19: 101–108.

Kim, H. J. & Cavanaugh, J. E. 2005: Model selection criteria based on Kullback information measures for nonlinear regression. — *Journal of Statistical Planning and Inference* 134: 332–349.

Kuha, J. 2004: AIC and BIC — comparisons of assumptions and performance. — *Sociological Methods and Research* 33: 188–229.

Lovell, M. C. 1983: Data mining. — *Review of Economics and Statistics* 65: 1–12.

Lukacs, P. M., Burnham, K. P. & Anderson, D. R. 2010: Model selection bias and Freedman's paradox. — *Annals of the Institute of Statistical Mathematics* 62: 117–125.

Martínez-Abrain, A. 2007: Are there any differences? A non-sensical question in ecology. — *Acta Oecologica* 32: 203–206.

Miller, A. J. 1992: *Subset selection in regression*. — Chapman and Hall, Boca Raton, FL.

Mills, J. A. & Prasad, K. 1992: A comparison of model selection criteria. — *Economical Reviews* 11: 201–233.

Møller, A. P. & Jennions, M. D. 2002: How much variance can be explained by ecologists and evolutionary biologists? — *Oecologia* 132: 492–500.

Moran, M. D. 2003: Arguments for rejecting the sequential Bonferroni in ecological studies. — *Oikos* 100: 403–405

Mundry, R. 2011: Issues in information theory based statistical inference: a commentary from a frequentist's perspective. — *Behavioral Ecology and Sociobiology* 65: 57–68.

Mundry, R. & Nunn, C. L. 2009: Stepwise model fitting and statistical inference: turning noise into signal pollution. — *American Naturalist* 173: 119–123.

Murtaugh, P. A. 1998: Methods of variable selection in regression modeling. — *Communications in Statistical Simulation and Computing* 27: 711–734.

Murtaugh, P. A. 2009: Performance of several variable selection methods applied to real ecological data. — *Ecology Letters* 12: 1061–1068.

Nakagawa, S. 2004: A farewell to Bonferroni: the problems of statistical power and publication bias. — *Behavioral Ecology* 15: 1044–1045.

Nakagawa, S. & Cuthill, I. C. 2007: Effect size, confidence interval and statistical significance: a practical guide for biologists. — *Biological Reviews* 82: 591–605.

Pope, P. T. & Webster, J. T. 1972: Use of an F-statistic in stepwise regression procedures. — *Technometrics* 14: 327–340.

Raffalovich, L. E., Deane, G. D., Armstrong, D. & Tsao, H. 2008: Model selection procedures in social research: Monte Carlo simulation results. — *Journal of Applied Statistics* 35: 1093–1114.

Richards, S. A. 2008: Dealing with overdispersed count data in applied ecology. — *Journal of Applied Ecology* 45: 218–227.

Richards, S. A., Whittingham, M. J. & Stephens, P. A. 2011: Model selection and model averaging in behavioural ecology: the utility of the IT-AIC framework. — *Behavioral Ecology and Sociobiology* 65: 77–89.

Sauerbrei, W. 1999: The use of resampling methods to simplify regression models in medical statistics. — *Journal of the Royal Statistical Society Series C* 48: 313–329.

Schielzeth, H. 2010: Simple means to improve the interpretability of regression coefficients. — *Methods in Ecology and Evolution* 1: 103–113.

Schwarz, G. 1978: Estimating the dimension of a model. — *Annals of Statistics* 6: 461–464.

Seghouane, A. K. 2006: A note on overfitting properties of KIC and KIC_c. — *Signal Processing* 86: 3055–3060.

Stephens, P. A., Buskirk, S. W. & Martínez del Rio, C. 2007: Inference in ecology and evolution. — *Trends in Ecol-*

ogy and Evolution 22: 192–197.
Stevens, J. 2002: *Applied multivariate statistics for the social sciences* (4th ed.). — Lawrence Erlbaum Associates, Mahwah, NJ.
Symonds, M. R. E. & Moussalli, A. 2011: A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's Information Criterion. — *Behavioral Ecology and Sociobiology* 65: 13–21.
Taborsky, M. 2010: Sample size in the study of behaviour. — *Ethology* 116: 185–202.
Ward, E. J. 2008: A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. — *Ecological Modelling* 211: 1–10.
Whittingham, M. J., Stephens, P. A., Bradbury, R. B. & Freckleton, R. P. 2006: Why do we still use stepwise modelling in ecology and behaviour? — *Journal of Animal Ecology* 75: 1182–1189.
Yang, Y. 2005: Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. — *Biometrika* 92: 937–950.
Zuur, A. F., Ieno, E. N. & Elphick, C. S. 2010: A protocol for data exploration to avoid common statistical problems. — *Methods in Ecology and Evolution* 1: 3–14.

Appendix 1: Varying interrelation analysis

Material and methods

In this analysis, we modeled the presence of different types of predictor correlations in a single data set and its effect on the performance of the three different statistical model types. We simulated 1600 data sets with 12 predictors in each of them. To maintain the ratios of predictors to observations similar to those in the uniform interrelation analysis presented in the main body of the paper, we set $n = 36$ as a small sample size (800 data sets) and $n = 240$ as a large sample size (800 data sets). All predictor correlations or effects were set to either $r = 0.3$ (informative predictors) or $r = 0.0$ (uninformative predictors). All data sets included six informative and six uninformative predictors, representing six different combinations of cases (with two predictors per category): (i) informative predictors uncorrelated with other predictors, (ii) informative predictors correlated with an informative predictor, (iii) informative predictors correlated with an uninformative predictor, (iv) uninformative predictors uncorrelated with other predictors, (v) uninformative predictors correlated with an informative predictor, (vi) uninformative predictors correlated with an uninformative predictor. All data sets were processed for parameter identification error (false detection probability FPOS and detection failure probability FNEG) and effect sizes in the same way as in the main analysis.

Results

False detection probability (FPOS)

False detections were mainly determined by an interactive effect of sample size and correlation type. The "correlated with informative" group showed higher FPOS than the groups "uncorrelated" and "correlated with uninformative" with the latter two being similar to each other. The elevated FPOS of "correlated with informative" was much more pronounced at the larger sample size. The effect of model type was relatively small, with generally higher FPOS for AICc best. The two NHST-based methods were similar to each other and to $W_i > 0.7$, with occasionally higher FPOS for stepwise than for full models. The $W_i > 0.9$ setting showed very low FPOS (Fig. A1_1A).

Detection failure probability (FNEG)

With large samples, there were almost no detection failures irrespective of setting. With small samples, the "correlated with informative" group again dominated the other two correlation types. Regarding model type, AICc best always gave the lowest and full models and especially $W_i > 0.9$ the highest FNEG, with stepwise regression showing lower FNEG than full models (Fig. A1_1B).

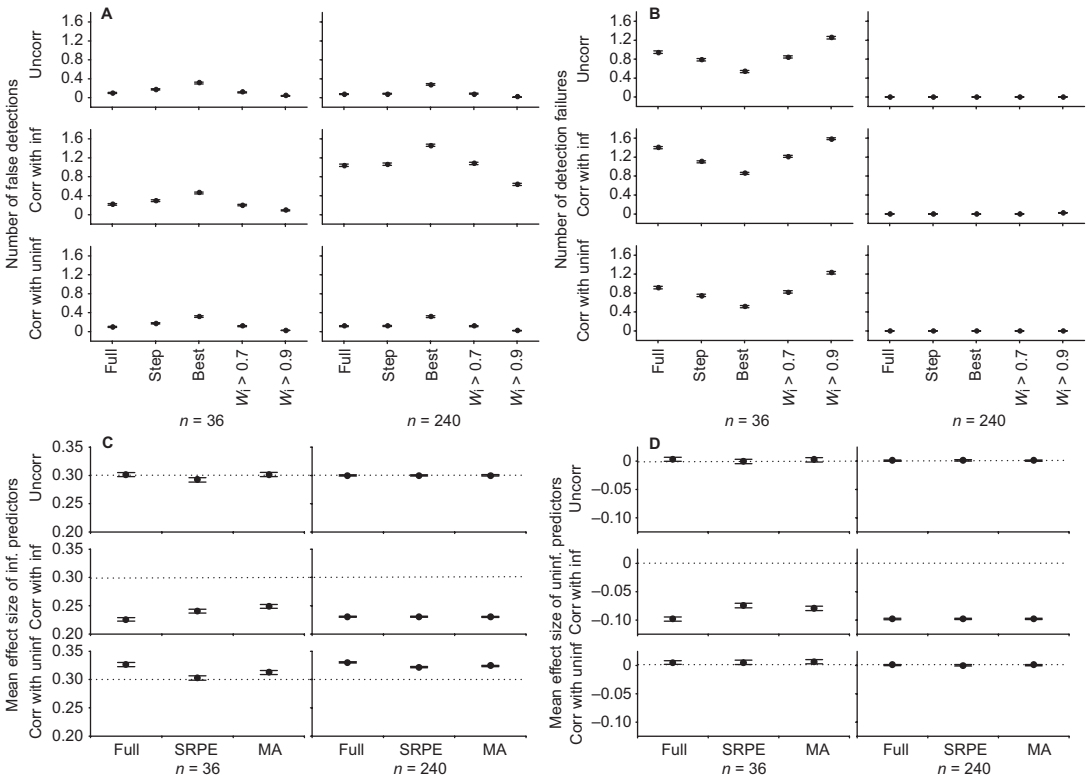

Fig. A1_1. Parameter identification error rates and the effect sizes of informative and uninformative predictors in the varying interrelation analysis, in relation to model type and simulation settings (means ± SEs). Dotted lines indicate the effect size specified in the simulation (i.e. the true population level effect size of the given parameter type). (**A**) false detection, (**B**) detection failure, (**C**) effect size of informative predictors, and (**D**) effect size of uninformative predictors. Corr = correlated, uncorr = uncorrelated, inf = informative, uninf = uninformative, Full = full model, Step = stepwise regression, Best = AICc best, SRPE = stepwise-reintroduction for parameter estimation, MA = AICc model averaging.

Effect size of informative predictors

Both the overall parameter estimation bias and the model type effect varied strongly with correlation type and sample size (Fig. A1_1C). Informative predictors correlated with informative predictors showed a large downward bias, while those correlated with uninformative predictors exhibited a slight upward bias. Overall bias was negligible with uncorrelated predictors. Where a model type effect was visible (in three of four settings of the two categories of correlated predictors), full models always gave the largest bias. SRPE only slightly differed from AICc model averaging and the direction of the difference also depended on the setting.

Effect size of uninformative predictors

Among uninformative predictors, parameter estimation bias was very small or absent in uncorrelated predictors and those correlated with other uninformative predictors (Fig. A1_1D). Uninformative predictors correlated with an informative predictor gave drastically downward biased estimates, i.e. negative effect sizes. Model type influenced bias only at the small sample size, where full models showed the largest and SRPE the smallest bias.

Appendix 2: Detailed results.

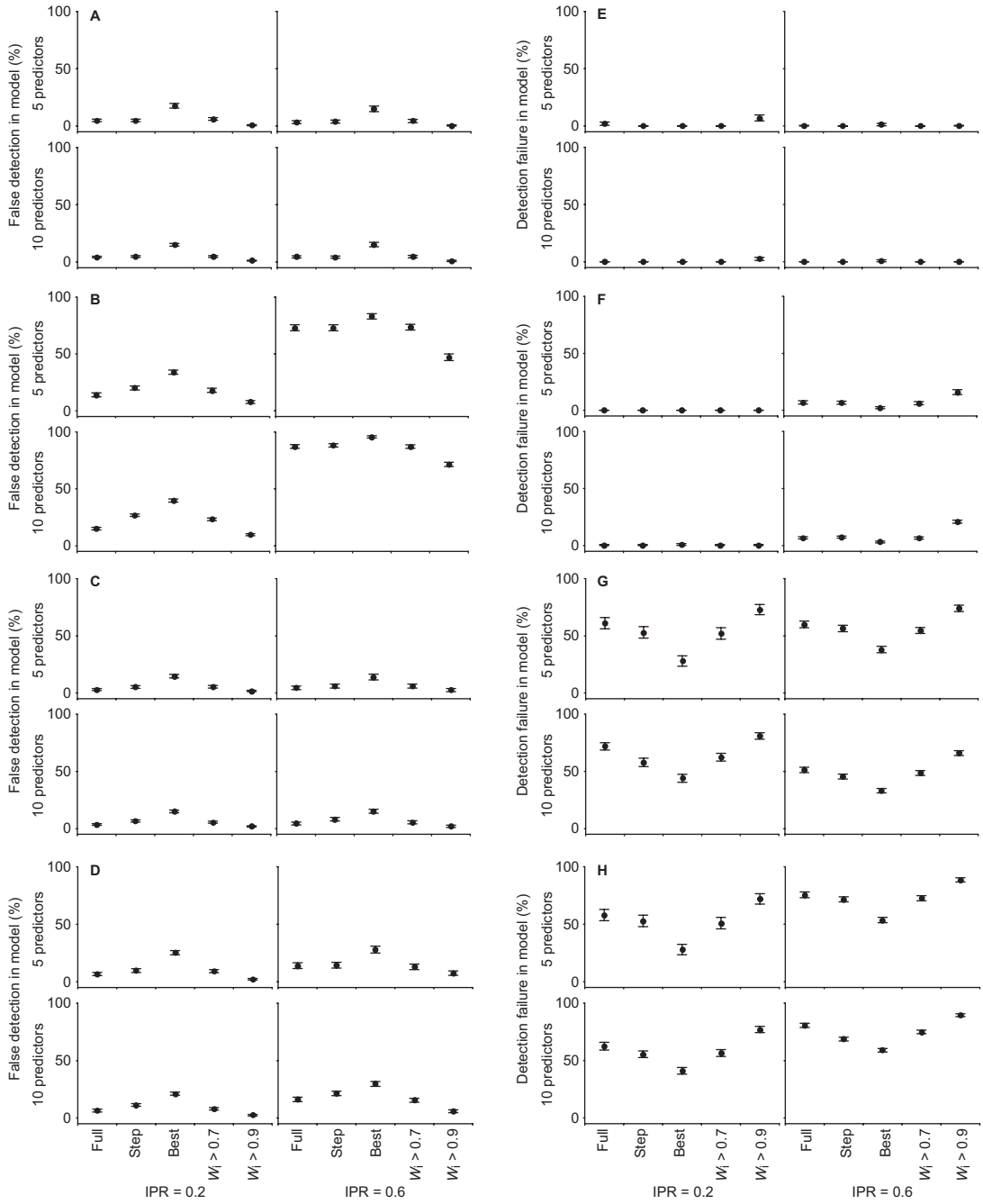

Fig. A2_1. Rates of (**A**–**D**) false detection and (**E**–**H**) detection failure in relation to the model type and simulation settings (means ± SEs). **A** and **E**: large sample and uncorrelated predictors; **B** and **F**: large sample and correlated predictors; **C** and **G**: small sample and uncorrelated predictors; **D** and **H**: small sample and correlated predictors. IPR = informative predictor ratio, Full = full model, Step = stepwise regression, Best = AICc best model.

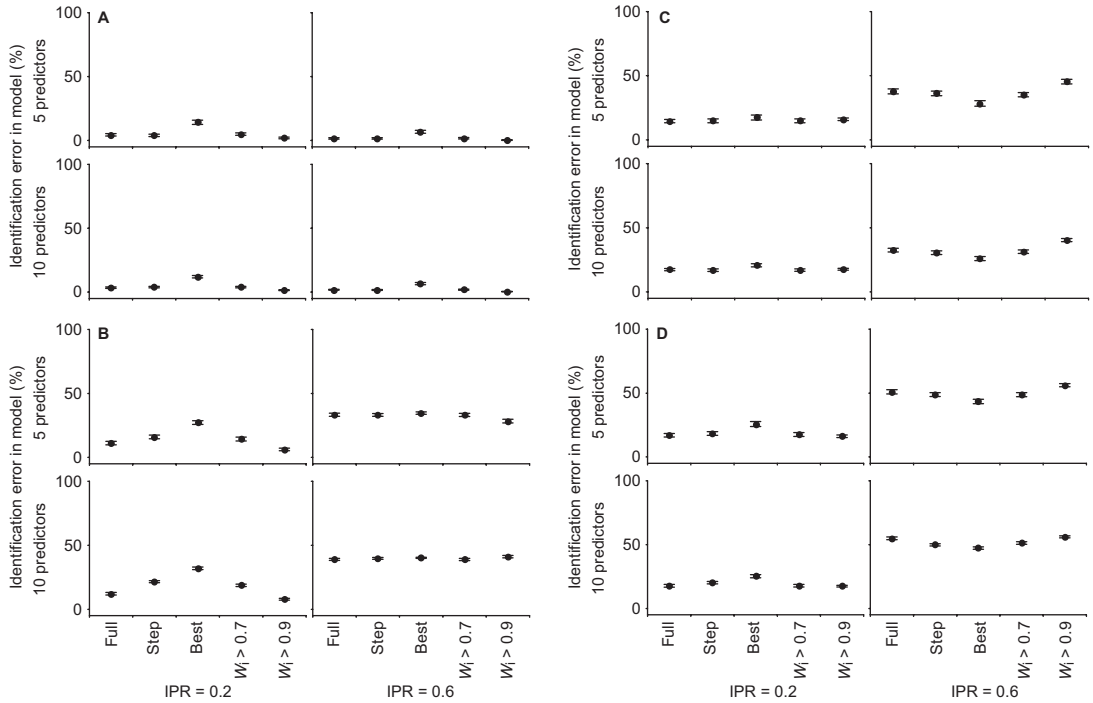

Fig. A2_2. Total parameter identification error rate (FTOT) in relation to model type and simulation settings (means ± SEs). (**A**) Large sample and uncorrelated predictors, (**B**) large sample and correlated predictors, (**C**) small sample and uncorrelated predictors, and (**D**) small sample and correlated predictors. IPR = informative predictor ratio, Full = full model, Step = stepwise regression, Best = AICc best model.

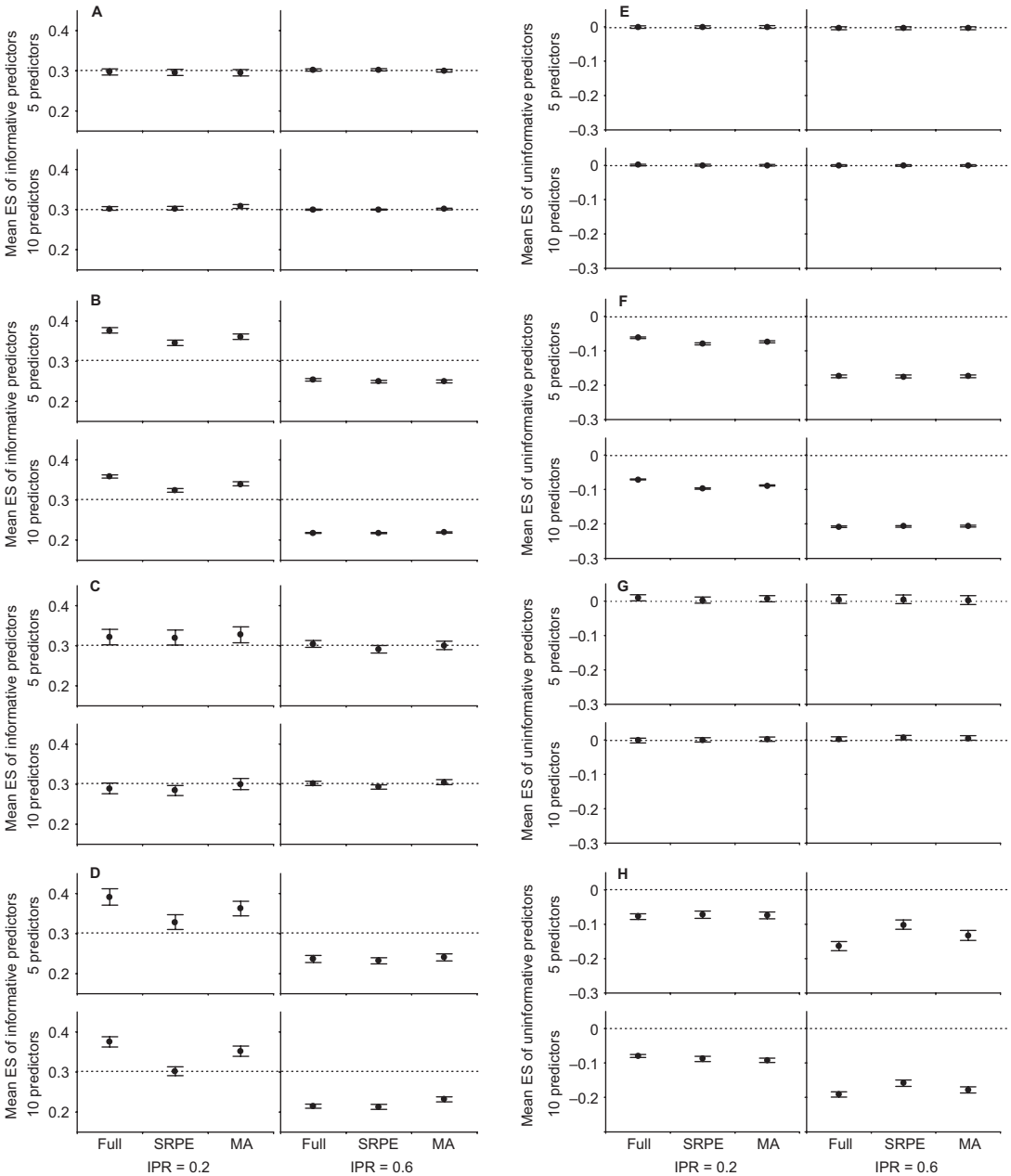

Fig. A2_3. Estimated effect sizes of (**A**–**D**) informative predictors and (**E**–**H**) uninformative predictors in relation to model type and simulation settings (means ± SEs). Dotted lines indicate the effect size specified in the simulation (i.e. the true effect size of the given predictor type). **A** and **E**: large sample and uncorrelated predictors, **B** and **F**: large sample and correlated predictors, **C** and **G**: small sample and uncorrelated predictors, **D** and **H**: small sample and correlated predictors. ES = effect size, IPR = informative predictor ratio, Full = full model, SRPE = stepwise-re-introduction for parameter estimation, MA = AICc model averaging.

Appendix 3: Using BIC instead of AICc

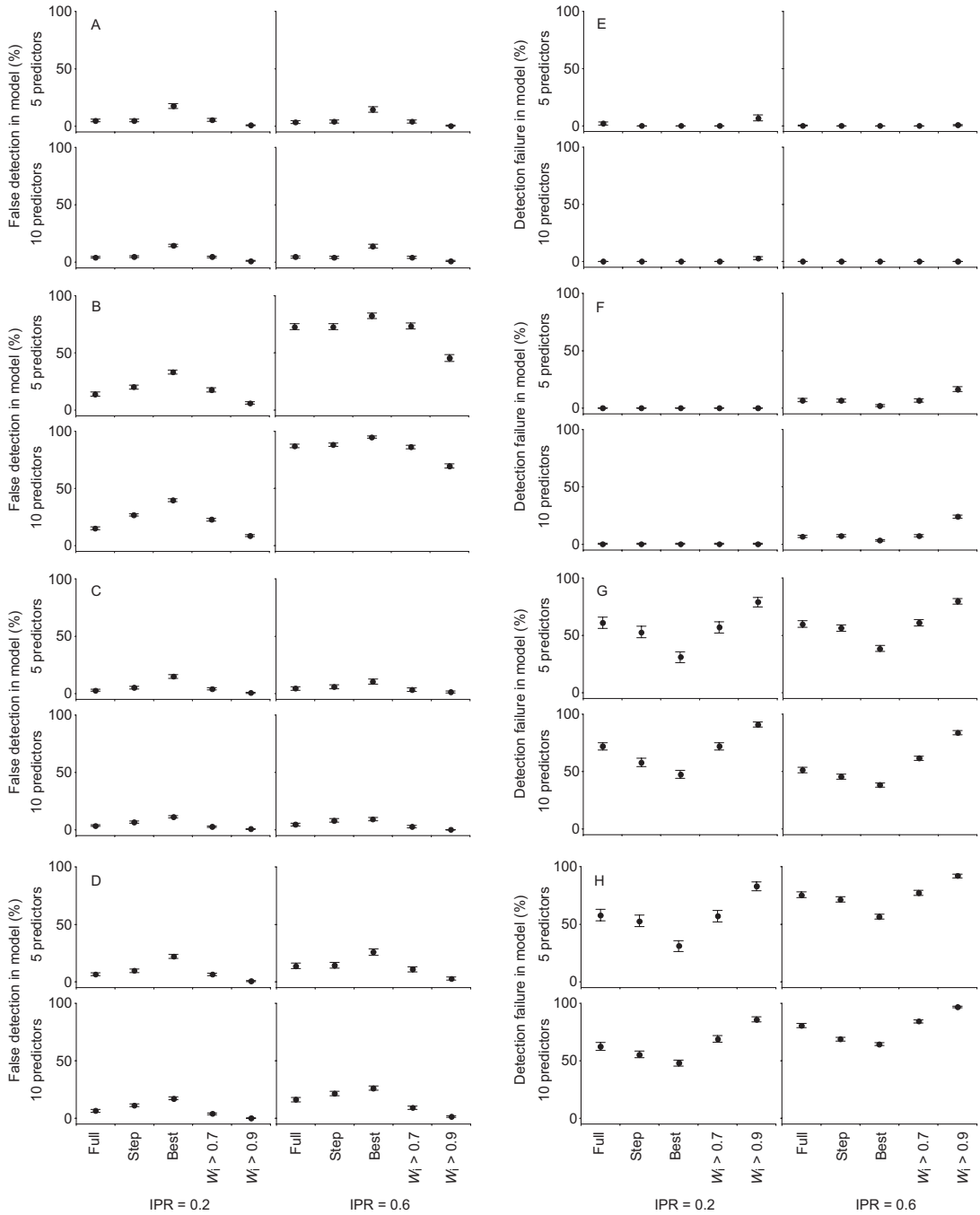

Fig. A3_1. Rates of (**A**–**D**) false detection and (**E**–**H**) detection failure in relation to model type and simulation settings (means ± SEs) when using BIC. **A** and **E**: large sample and uncorrelated predictors, **B** and **F**: large sample and correlated predictors, **C** and **G**: small sample and uncorrelated predictors, **D** and **H**: small sample and correlated predictors. IPR = informative predictor ratio, Full = full model, Step = stepwise regression, Best = BIC best model.

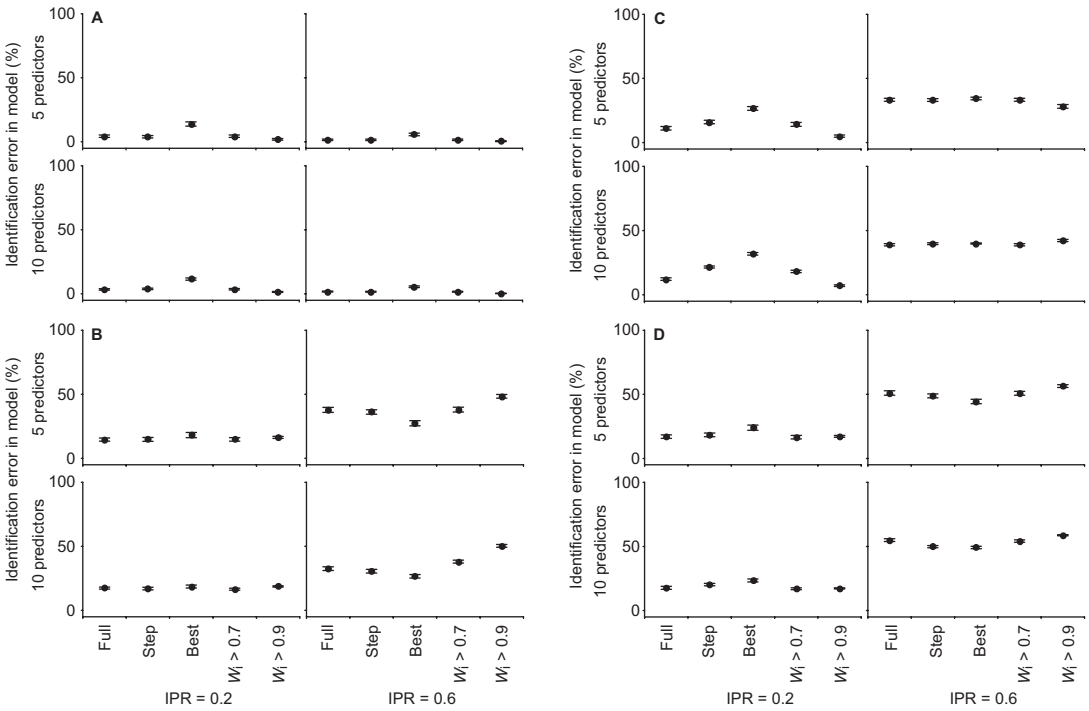

Fig. A3_2. Total parameter identification error rate (FTOT) in relation to model type and simulation settings (means ± SE) when using BIC. (**A**) large sample and uncorrelated predictors; (**B**) large sample and correlated predictors; (**C**) small sample and uncorrelated predictors; (**D**) small sample and correlated predictors. IPR = informative predictor ratio, Full = full model, Step = stepwise regression, Best = BIC best model.

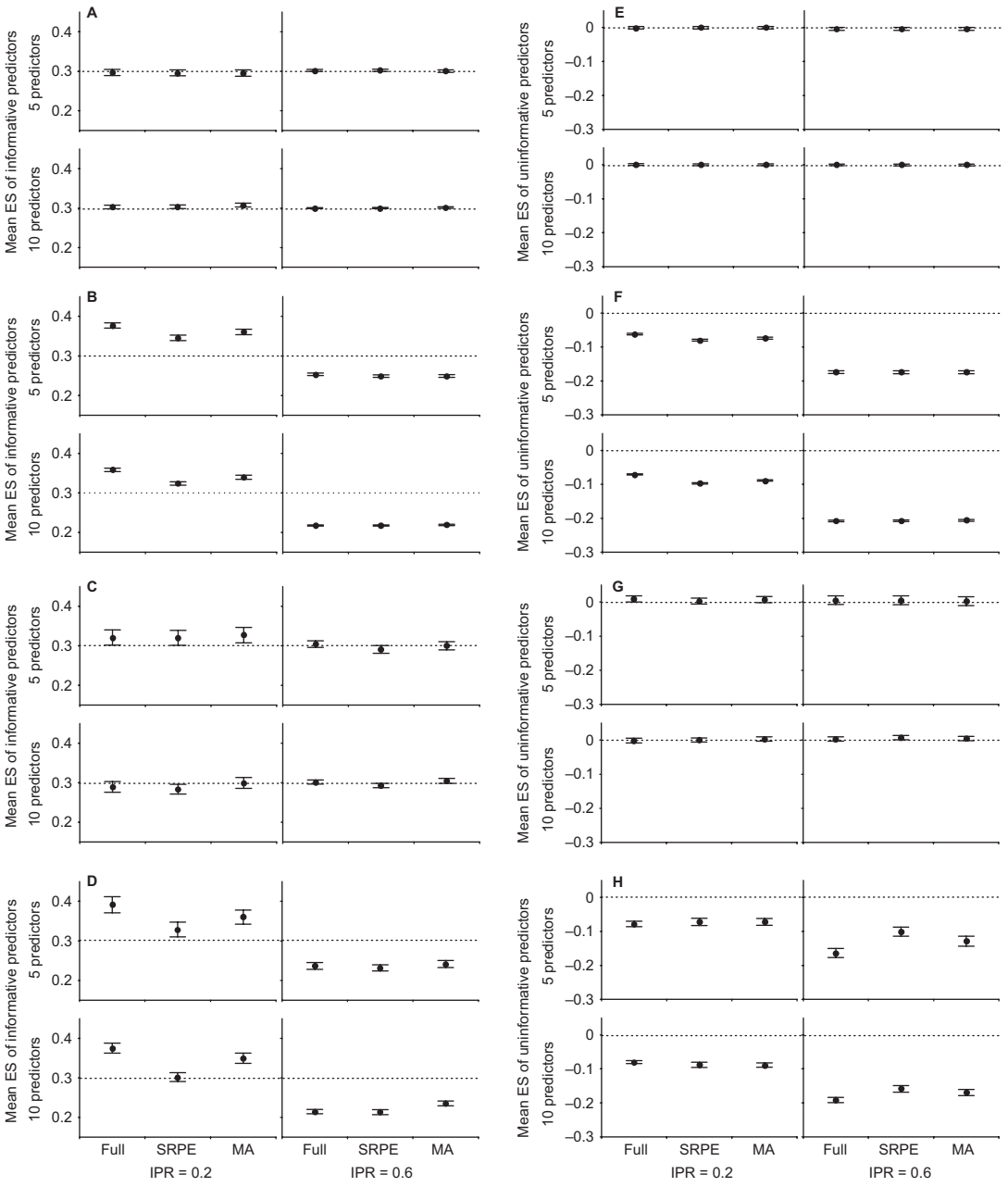

Fig. A3_3. Estimated effect sizes of (**A**–**D**) informative and (**E**–**H**) uninformative predictors in relation to model type and simulation settings (means ± SEs) when using BIC. Dotted lines indicate the effect size specified in the simulation (i.e. the true effect size of the given predictor type). **A** and **E**: large sample and uncorrelated predictors, **B** and **F**: large sample and correlated predictors, **C** and **G**: small sample and uncorrelated predictors, **D** and **H**: small sample and correlated predictors. ES = effect size, IPR = informative predictor ratio, Full = full model, SRPE = stepwise-reintroduction for parameter estimation, MA = BIC model averaging.