Selection Stability in High Dimensional Statistical Modelling: Defining a Threshold for Robust Model Inference

Martin Green (martin.green@nottingham.ac.uk)
University of Nottingham

Eliana Lima
European Food Safety Authority

Robert Hyde
University of Nottingham

Research Article

Keywords: Covariate selection, selection stability, stability threshold, high dimensional data, statistical triangulation.

Posted Date: August 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-738092/v1

Title

Selection stability in high dimensional statistical modelling: Defining a threshold for robust model inference

Author names and affiliations

Martin Green[a*], Eliana Lima[a,b], Robert Hyde[a]

[a] School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington Campus, Leicestershire, United Kingdom

[b] Present address: European Food Safety Authority, Via Carlo Magno 1A, 43126, Parma, Italy

***Corresponding author**

Martin Green

E-mail: martin.green@nottingham.ac.uk

School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington Campus, Leicestershire, LE12 5RD, United Kingdom

Tel +44 115 951 6116

Declarations of interest: none

Abstract

Epidemiological research commonly involves identification of causal factors from within high dimensional (wide) data, where predictor variables outnumber observations. In this situation, however, conventional stepwise selection procedures perform poorly. Selection stability is one method to aid robust variable selection, by refitting a model to repeated resamples of the data and calculating the proportion of times each covariate is selected. A key problem when applying selection stability is to determine a threshold of stability above which a covariate is deemed 'important'.

In this research we describe and illustrate a two-step process to implement a stability threshold for covariate selection. Firstly, covariate stability distributions were established with a permuted model (randomly reordering the outcome to sever the relationship with predictors) using a cumulative distribution function. Subsequently, covariate stability was estimated using the true model outcome and covariates with a stability above a threshold defined from the permuted model, were selected in a final model. The proposed method performed well across 22 varied, simulated datasets with known outcomes; selection error rates were consistently lower than conventional implementation of equivalent models. This method of covariate selection appears to offer substantial advantages over current methods, to accurately identify the correct covariates from within a large, complex parameter space.

Keywords

Covariate selection; selection stability; stability threshold; high dimensional data; statistical triangulation.

1. Introduction

Epidemiological research increasingly involves identification of potentially causal factors from within relatively wide datasets, that is, when the number of variables (p) is relatively large compared to the number of observations (n). Identification of causal variables for inference from within a high dimensional, wide dataspace is problematic because conventional stepwise selection procedures perform poorly, resulting in over fit models [1–4]. Statistical methods have been developed to improve variable selection with such data, including modifications to AIC/BIC [5], and a range of regularisation methods based on functions that penalise model coefficients to balance over and under fitting, the so called variance-bias trade off [6–8].

In addition to regularisation, it is acknowledged that robustness in model selection is improved through use of covariate selection stability [9–11]. Covariate stability is estimated by refitting a model to repeated resamples of the data (for example through Bootstrapping) and calculating the proportion of times each covariate is selected across all resamples. Whilst it is known that the most stable variables (those selected in most subsampled models) are least likely to be false positives [9], the optimal threshold for stability, above which a covariate should be deemed 'important' or 'significant', has not been determined. A stability threshold originally proposed for use with lasso regression [9] has been shown to be too conservative, resulting in true causal variables being missed [12,13]. An empirical method of stability selection, proposed for genetic data using elastic net regression [13], appeared to improve upon the method proposed by Meinshausen and Bühlmann [9], but the issue of missing many true causal variables (false negative results) remained. To date, since a clear, generalisable method to identify a stability threshold is unavailable, arbitrary thresholds have been employed for stability analyses in veterinary epidemiology [14,15] and the need for research to

establish a suitable cut-off for the covariate selection in high dimensional data is clear and has been recently re-emphasised in human epidemiology [16].

In this paper we build on principles described by Meinshausen and Bühlmann [9]and Kim et al [13], to develop a general approach to identify a threshold for covariate selection stability and hence to aid robust model selection for high dimensional, correlated data. To evaluate and illustrate the method, we use 22 simulated datasets with known outcomes and increasingly complex correlation structures, alongside five methods of covariate selection to illustrate generalisability. For each statistical method we compare the performance of covariate selection based on a conventional approach to that using the application of selection stability with the newly defined threshold.

2. **Materials and Methods**

Five statistical methods were used to analyse 22 simulated datasets that were constructed with known underlying relationships and correlation structures. For each method, a conventional approach to covariate selection was conducted followed by implementation of selection stability by bootstrapping [14]. The modelling methods chosen are acknowledged approaches for inferential modelling and were stepwise selection based on Akaike information criterion (sAIC) [17], stepwise selection based on a modified Bayesian Information Criterion (mBIC) [5], elastic net regression (enet) [8], minimax convex penalty regression (MCP) [18] and a combination method that synthesised results from all four methods [19]. These modelling techniques are described in detail in Section 2.1 and the new approach to stability selection, including the determination of a threshold for inference, is described in Section 2.2. All datasets used for analysis are described in Section 2.3.

2.1. Statistical models

The following statistical approaches were used to evaluate and compare covariate selection.

2.1.1. Stepwise selection using AIC (sAIC)

A conventional linear regression model was implemented using the 'stepwise' function in the bigstep package [20] in R [21]. The regression equation took the form;

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + e \quad (1)$$

where y was the response variable specified in the simulated data, β_0 an intercept term, x_j represented the j[th] of p covariates with an estimated coefficient β_j and e the residual model error. To speed computation, variable selection was conducted by first removing explanatory variables with a relatively poor correlation with the outcome (Pearson correlation, P >0.05) followed by an iterative stepwise procedure (forward and backward) with selection of the final set of variables through minimisation of the Akaike information criterion (AIC). The AIC loss function is defined as; $2k - 2\ln(\hat{L})$, where k is the number of parameters in the model and $\hat{L}$ the likelihood function.

2.1.2. Stepwise selection using a modified Bayesian Information Criterion (mBIC)

A modified Bayesian Information Criterion was used for model selection, also implemented in the bigstep package [20] with the 'stepwise' function being used for covariate selection. The modified BIC increases the penalty on the number of parameters selected beyond the conventional BIC, producing a sparser model. The mBIC loss function was implemented as;

$$mBIC = logL(Y|M_{i,}\theta_i) - \frac{1}{2} k_i logn - k_i \log(\frac{1-p}{p}) \quad (2)$$

where $logL(Y|M_{i,}\theta_i)$ represented the log likelihood given model M_i and parameter values θ_i, k_i was the number of predictors in the selected model, n the sample size, and p

the probability that a randomly chosen predictor influenced Y. As the number of available predictors increases relative to the number of samples (n), p decreases and $k_i \log(\frac{1-p}{p})$ becomes of increasing importance as the penalty term [5].

2.1.3. Elastic net regression (enet)

Elastic net is a form of regularised regression that incorporates a mixture of lasso (L1) and ridge (L2) penalties [8] and was implemented as;

$$SSE_{enet} = \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 + \lambda_E \left[\sum_{j=1}^{P} \frac{1}{2} (1-\alpha)\ \beta_j^{\,2} + \alpha\ |\beta_j|\right] \quad (3)$$

where SSE_{enet} represented the elastic net loss function to be minimised, i denoted each observation and n the sample size, y_i and $\hat{y}_i$ were the simulated true and model predicted outcomes respectively for the ith observation, *j* denoted a predictor variable with p the number of predictor variables in total, and |β| represented absolute values of the regression coefficients. The hyperparameters that provide the penalty (λ_E) and the relative proportion of penalisation on either the sum of the square of the coefficients or the unsquared coefficients (α) were optimised using 10 x 10-fold cross validation to minimise mean absolute error (MAE). Elastic net models were built using the glmnet package [22] using the caret package platform [23] in R [21].

2.1.4. Minimax convex penalty (MCP)

Minimax convex penalty (MCP) [18] is a form of regularised regression in which the size of the penalty function varies with the size of variable coefficient, β. It was implemented as;

$$SSE_{mcp} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 + \sum_{j=1}^{p} P\ \left(\beta_j | \lambda, \gamma\right) \quad (4)$$

where SSE_{mcp} represented the loss function to be minimised, i, y_i, $\hat{y}_i$, j, p and n were as defined in equation (3) and $P(\beta_j|\lambda,\gamma)$ represented a penalty function as follows;

$$P(\beta|\lambda,\gamma) = \begin{cases} \lambda|\beta| - \frac{\beta^2}{2\gamma}, & \text{if } |\beta| \leq \gamma\lambda \\ 0.5\gamma\lambda^2, & \text{if } |\beta| > \gamma\lambda \end{cases} \quad (5)$$

where γ and λ were hyperparameters optimised using 10 x 10-fold cross validation to minimise the MAE. MCP models were estimated using the R package ncvreg [24].

2.2. Estimation of covariate selection stability and coefficient distributions

Conventional bootstrapping [25] was used to estimate covariate stability for all analytic approaches, according to methods previously described [26]. In brief, selection stability [9,10,27] was evaluated for each model as the percentage of times each covariate was selected across 200 bootstrap samples. For elastic net and MCP, model hyperparameters used in the bootstrap sampling were those identified using 10 x10-fold cross validation in an initial model using the full dataset. The distributions of variable coefficients were calculated from all non-zero values of the coefficient in the bootstrap samples and a bootstrap P value (BPV) was calculated for each covariate [14]. The BPV was calculated from all non-zero coefficient values of the bootstrap sample (i.e. when the variable was selected in the model) and defined as the smallest proportion of (non-zero) coefficient values to one side of zero. For example, if a covariate was selected in the model in 100 bootstrap samples and 95 of these were either greater or less than zero, then the Bootstrap P value would be; (100-95)/100 = 0.05. Variable selection was visualised by plotting covariate selection stability against Bootstrap P value.

2.3. Multiple method synthesis

Covariate selection stabilities were combined across all modelling methods using an approach described previously [19]. In brief, the bootstrap matrices from each individual modelling method were aligned by covariate to create a single matrix containing an equal number of bootstrap samples from each method. This combined matrix was used to calculate overall covariate selection stabilities and Bootstrap P values as described above. Therefore this method provided an estimate for selection stability and Bootstrap P values for all covariates synthesised across all four methods of covariate selection.

2.4. Stability threshold calculation

A stability threshold above which each covariate was deemed to be important and selected in a final model ('selection stability threshold (T^S)'), was calculated as follows. The concept was to initially determine a distribution of covariate stability for each dataset on the basis that there was *no casual relationship* between the outcome and explanatory variables and to compare this to the stability of covariates in the actual data. The distribution of covariate stability when no causal relationship existed we name here, the 'baseline stability distribution'. When the stability of a covariate in the actual data had a specified probability of exceeding values in the baseline stability distribution, it was deemed to be important and selected in the final model.

For each model and dataset, the baseline stability distribution was determined by randomly permuting the outcome variable to sever any relationship between outcome and explanatory variables. Using the permuted outcome, a stability analysis was conducted as described in Section 2.2 and the baseline stability of all covariates estimated. An empirical cumulative distribution function (ECDF) of the baseline stability distribution was formulated and

stability thresholds were tested at different specified probability values of the ECDF. The specified probability values were calculated as;

$$1- x/p \quad (6)$$

where x was either equal to 0, 1, 3, 5, or 10 and p the number of covariates available in a dataset. For example, when x = 1 and 500 potential covariates were available in the dataset, the probability value of the ECDF at which the stability threshold would be set was;

$$1 – (1/500) = 0.998$$

Therefore, x/p represented the probability that the stability of any covariate would exceed the defined stability threshold.

To clarify the method, we list the steps used to calculate T^S;

i. For a specified dataset and model, the outcome variable was randomly permuted to remove the relationship between the outcome and explanatory variables.
ii. Covariate selection stability was conducted (as described in Section 2.2) using a set of 20 bootstrap samples.
iii. The distribution of covariate stability values for all ‘p’ covariates was defined using an empirical cumulative distribution function (‘ecdf’ function, base R) and the (1-x/p) probability value of the distribution calculated, to provide the stability threshold T^S. The values tested for ‘x’ were 0, 1, 3, 5 and 10.
iv. Steps i. to iii. were repeated ten times and the mean and standard error for T^S calculated across the ten repeats.
v. Covariate selection stability was conducted on the full dataset (i.e. without permuting the outcome variable) using 200 bootstrap samples and the mean value of T^S , as calculated in steps i – iv, used as the threshold above which variables were considered to be ‘true’ and therefore selected in the final model.

Therefore, the stability threshold was set at a value at which there was a low probability (x/p) of a covariate reaching this level of stability if it was not truly associated with the outcome. Importantly, since location of the threshold is likely to be dependent upon the method of modelling and specific dataset used, a new stability threshold was calculated for each dataset and analysis undertaken.

2.5. Simulated datasets

The purpose of simulating data was to construct a set of realistic datasets, of varying size and complexity, in which true underlying causal relationships were known. Therefore, model performance could be evaluated objectively by comparing the accuracy with which each method (including the stability threshold method) correctly selected the true causal variables. For each dataset, 8 covariates were used to calculate an outcome variable; we refer to these as the 'true' covariates. Additional noise variables were generated for each dataset, randomly and independent of the outcome, we refer to these as 'false' variables. In each dataset, true and false variables were set to be correlated with each other to a different extent such that each statistical method could be evaluated with different but realistic complexity added to the data (Table 1).

In all datasets, an outcome variable, "y_out", was calculated from the true explanatory variables as follows;

$$y\_out = Intercept + \sum_{k=1}^{r} 2.5\, x_k + \sum_{j=1}^{q} 2.5\, x_j + e \quad (7)$$

where $Intercept$ =1, x_k represented the kth of r true covariates simulated with no specified correlation with other covariates, x_j represented the jth of q true covariates that were specified to have correlations with a set of ten false explanatory covariates, and e was a random

variable that represented all other true but unknown effects that causally influenced the outcome, y_out. e was drawn from a Normal distribution with mean = 0 and standard deviation which varied between dataset to allow the amount of variability explained by the true covariates to differ between datasets (Table 1).

All explanatory covariates were drawn from distributions with mean = 0, SD = 1; they represented standardised variables. Covariates that were simulated to be correlated (x_j) were drawn from a multivariate normal distribution (using the mvrnorm function in the MASS package [28] in R). In this case, each variable drawn from a distribution with mean = 0, SD = 1 and with covariance matrix specified such that variables were generated with a correlation between 0.3 – 0.7, depending on the dataset (Table 1).

The characteristics of all datasets are summarised in Table 1. For each dataset, conventional linear regression was used to model solely the true covariates, to illustrate the underlying relationship between the outcome and true variables and estimate the total amount of variation explained. True variables were simulated in all datasets such that their partial coefficients in these regression models were individually significant ($P<0.05$); that is their signal was sufficiently strong to be deemed 'significant' when modelled alone. The purpose of datasets containing more highly correlated covariates was to evaluate the extent to which true variables could be retrieved when a substantial quantity of random and correlated noise was added to the data and to explore whether selection stability enhanced retrieval.

For four of the 22 generated datasets a binary outcome was simulated instead of a continuous outcome so that the analytic methods could be evaluated using an alternative outcome distribution. A similar approach was used as in equation (6) except that an inverse logit function followed by a draw from a binomial distribution was used to alter the outcome to be 1 or 0. As for the datasets with normally distributed outcomes, correlations were included between true and false explanatory covariates to provide an increasing complexity within the

data. The variation explained by the true variables was estimated using an adapted R^2 [29]; further details of the binomial outcome datasets are provided in Table 1.

{Table 1 approx here}

2.6. Evaluation and comparison of model performance

Model performance was assessed based on the correct selection of true variables and correct omission of the false variables, and comparisons were made between methods for all datasets. Graphical analysis was used to visualise model outcomes; scatterplots were used to illustrate results of the stability analysis by plotting the stability (%) against BPV for covariates selected in the model, as previously reported [19] but with the additional annotation of the selection stability threshold.

3. **Results**

3.1. Error rates for full (non-bootstrapped) models

A summary of the performance of all full (non-stability threshold) modelling methods, for each dataset is provided in Table 2. Across all datasets, model selection using sAIC, elastic net regression and MCP included a higher number of incorrect (false positive and false negative) variables than selection using mBIC. The incorrect covariates selected using sAIC, enet and MCP were mostly false positives indicating these models generally over fit the data. Whilst the final models produced using mBIC were more sparse than for other methods (fewer variables selected in the final models), the incorrectly classified covariates were mostly false negatives indicating a tendency towards under fitting with this method. In general, for most methods, more covariates were incorrectly classified as the complexity of correlation structure within the data increased (Table 2).

293

294 {Table 2 approx here please}

295

296 3.2. Error rates for models incorporating covariate selection stability

297 Results of the covariate selection stability analyses using stability thresholds calculated with
298 x = 0, are provided in Table 3. Results for all other thresholds, x = 1, 3, 5 and 10, are
299 provided in Supplementary Materials (Tables S1-S6). The stability thresholds that produced
300 fewest incorrectly classified covariates (false positives and false negatives) across all models
301 and all datasets were those calculated using x = 0 or x = 1 but the threshold calculated with x
302 = 0 selected fewest false positive covariates and was therefore judged the optimal threshold to
303 use. As expected, with increasing values of x, the number of false positive variables selected
304 tended to increase and the number of false negatives decreased. The one exception to x=0
305 being the optimal threshold was the highly complex dataset IC2, in which the stability
306 threshold at x=3 produced very slightly better results than x=0 or x=1 (Table S6).
307 Importantly, for all modelling methods, use of bootstrap selection stability at a threshold
308 using x=0 resulted in substantially fewer covariates being incorrectly selected (Table 3)
309 compared to the conventional implementation of each technique (Table 2). For sAIC, enet
310 and MCP the improvement in accuracy of covariate selection was marked; the number of
311 false positive covariates was greatly reduced with no increase in false negatives. For mBIC
312 the improvement in performance was less marked but still evident. The combination method,
313 a synthesis of methods incorporating results from all four model types, produced the best
314 overall performance compared to other methods, with the lowest number of false positive
315 variables selected whilst maintaining a low number of false negative covariates.

316

317 {Table 3 approx here}

Across all model types, it was noticeable that as complexity of correlations within the simulated data increased, accuracy of covariate selection slightly deteriorated. Importantly, in the datasets that were simulated with no or relatively few underlying correlations (Datasets A-C), with a selection stability threshold set at x = 0 or x = 1, all model methods achieved near-perfect selection accuracy when stability selection was used (Table 3). This was in contrast to the performance of the models without selection stability (Table 2) in which only mBIC models performed well on non-correlated data. As the size and number of correlations within the data increased, performance of all methods slightly deteriorated, however, model performance remained much better with the use of covariate selection stability than for conventional implementation of each model type.

3.3. Diagrammatic illustration of model performance

Covariate selection stability thresholds for all values of x are illustrated graphically for datasets MS1, MS2 and LS4, in Figures 1 – 3 respectively. The graphs provide an illustration of the positioning of the stability thresholds (T^S) relative to the known true and false covariates, for different values of x. Use of 10 replicates of 20 sets of bootstrap samples was sufficient to produce relatively small 95% confidence intervals for T^S although the number of replicates can be increased to further reduce the confidence interval as required.

{Figure 1 here, to be reproduced in colour please}

Figure 1. Graphical representation of model results from dataset MS1. Each scatterplot illustrates results of the stability analysis for one of five statistical methods; stepwise selection using Akaike information criterion (AIC), stepwise selection using a modified Bayesian

Information Criterion (mBIC), elastic net regression (enet), Minimax Convex Penalty regression (MCP) and the four methods combined (combi). The y-axis defines the negative bootstrap P value for each covariate (-BPV) and the x-axis represents covariate stability (%). The dashed lines on each graph (from right to left) represent the stability threshold calculated at x=0, 1, 3, 5 and 10 respectively. The shaded grey area depicts the 95% confidence interval for the stability threshold calculated at x=0.

{Figure 2 here please, to be reproduced in colour please}

Figure 2. Graphical representation of model results from dataset MS2. Each scatterplot illustrates results of the stability analysis for one of five statistical methods; stepwise selection using Akaike information criterion (AIC), stepwise selection using a modified Bayesian Information Criterion (mBIC), elastic net regression (enet), Minimax Convex Penalty regression (MCP) and the four methods combined (combi). The y-axis defines the negative bootstrap P value for each covariate (-BPV) and the x-axis represents covariate stability (%). The dashed lines on each graph (from right to left) represent the stability threshold calculated at x=0, 1, 3, 5 and 10 respectively. The shaded grey area depicts the 95% confidence interval for the stability threshold calculated at x=0.

{Figure 3 here, to be reproduced in colour please}

Figure 3. Graphical representation of model results from dataset LS4. Each scatterplot illustrates results of the stability analysis for one of five statistical methods; stepwise selection using Akaike information criterion (AIC), stepwise selection using a modified Bayesian Information Criterion (mBIC), elastic net regression (enet), Minimax Convex Penalty regression (MCP) and the four methods combined (combi). The y-axis defines the negative bootstrap P value for each covariate

(-BPV) and the x-axis represents covariate stability (%). The dashed lines on each graph (from right to left) represent the stability threshold calculated at x=0, 1, 3, 5 and 10 respectively. The shaded grey area depicts the 95% confidence interval for the stability threshold calculated at x=0.

4. **Discussion**

Our results using simulated datasets with known relationships highlight the value of selection stability as a means to identify the correct explanatory variables when the number of covariates (*p*) is relatively large with respect to the number of observations (*n*). Whilst the concept of selection stability is not new [9,27], its use within epidemiological fields has been relatively limited. It is generally considered that selection stability is a necessary and beneficial approach to enhance inferential modelling with high dimensional data [11,30] and that without selection stability results of different selection methods vary dramatically leading to confusion in inference [26,31].

An important consideration, however, when employing selection stability is to identify a stability threshold above which covariates are deemed 'important' or 'significant' and selected in a final model; this was the primary purpose of the current study. Whilst it is known that a higher threshold for selection stability will result in the selection of fewer false positive covariates [9], a generalisable approach to identify an optimal threshold suitable for multiple model types has not been reported. A threshold proposed for one model type, lasso regression [9] has been shown to be too conservative with too few true variables being selected [13,32]. A further problem with the threshold proposed by Meinshausen and Bühlmann [9] is that it cannot be defined at a value <50% which, as we have shown in this research, is certainly necessary with some types of models and data (Figures 1-3).

In this study we defined and evaluated a straightforward approach to identify a covariate selection stability threshold that can be readily used with different datasets and model types. The intuitive approach of using an expected or 'baseline' stability distribution from a model with no underlying relationships and using the same data to compare this to a distribution from a model with causal relationships, worked well across all five model types and all 22 complex datasets. Even for the highly correlated datasets, error rates were relatively low and, importantly, were markedly superior to the equivalent conventional (non-stability selection) models. Therefore, for safety of inference, when using relatively wide, high dimensional datasets, covariate selection stability with the defined threshold will minimise identification of false positive and false negative variables compared to conventional modelling techniques. Furthermore, although highly correlated data are generally recognised to distort model parameters [33], our results indicate that selection stability with an appropriate choice of threshold provides a method to mitigate this issue.

The concept of permuting variables to enhance statistical inference is not new. It has been used in machine learning to facilitate estimation of variable importance [34,35] and as a replacement for cross validation to identify covariates in elastic net regression using non-sparse genetic data [13]. In the setting we describe, permuting the outcome allowed identification of an optimal covariate stability threshold to identify true variables for a variety of modelling methods and types of data. Whilst using covariate selection stability was more computationally expensive than conventional approaches, this was not prohibitive. For example, for a set of 10 repeats of 20 bootstrap samples (to define the baseline stability threshold) followed by 200 bootstrap samples (to select variables above the baseline threshold) using a dataset with 500 observations and 1000 covariates, the total time taken for computation of the mBIC method was 3 minutes and 16 seconds using a 15 core i9 processor with 32GB RAM. The equivalent time for the elastic net model was 28 minutes and 4

418 seconds. In terms of the substantial improvements in accuracy of covariate selection, the
419 additional time required would appear to be very worthwhile.
420 Also worthy of note in our results was the excellent performance of the combination
421 modelling method; this is in agreement with previous research [19]. For all datasets this method
422 provided the most accurate variable selection and this is likely to be due to the principle of
423 triangulation. Triangulation is based on the concept that reliability of results is enhanced
424 through integration of different approaches, particularly when each approach has a different,
425 unrelated source of bias [36]. For this reason, it has been suggested that triangulation should be
426 more widely adopted to ensure robustness and reproducibility of scientific results [37] although
427 it has rarely been employed in animal health research.
428
429 4.1. Study limitations
430 There were several limitations of this study worthy of consideration. Although substantial
431 complexity was incorporated into the simulated datasets, no clustering of the outcome
432 variable was considered. Therefore, extension of this stability threshold method to mixed
433 effect models, although likely to be valid, requires additional investigation. Similarly, non-
434 linearities within predictor variables were not considered in the simulated data and this also
435 warrants additional exploration in terms of selection stability thresholds. Four methods of
436 modelling were evaluated, and a combination of all four together, which suggests our
437 proposed method of selection stability may be generalisable, however, further research using
438 additional model types would be worthwhile. In addition, although one group of simulated
439 dataset incorporated discrete (binary) outcome variables (Datasets BIN1-4), further research
440 would be valuable to investigate the value of these methods with outcome distributions of
441 varying types. Finally, it should be noted that underlying true solutions in our simulated data
442 were relatively sparse; eight covariates were chosen to have true causal influences. This

443 aligns with the ‘bet on sparsity principle’ [4], which dictates that for causal variables to be
444 identifiable (in any model), a relatively small number of predictors have to be responsible for
445 most of the effect on the outcome of interest; this is why such data structures were chosen in
446 this study.
447
448 4.2. Conclusions
449 In this research we have proposed a new method to conduct covariate selection for use with
450 relatively wide data. The method is based on a new approach to covariate selection stability
451 that incorporates a stability threshold to define which covariates should be included in a final
452 model. Results indicate the approach offers substantial reductions in covariate selection error
453 rates compared with conventional model selection methods.
454
455 **Acknowledgements**
456 The author Eliana Lima is employed with the European Food Safety Authority (EFSA) in the
457 Unit ALPHA that provides scientific and administrative support to EFSA's Scientific
458 Activities in the area of animal health and welfare. However, the present article is published
459 under the sole responsibility of the author Eliana Lima and may not be considered as an
460 EFSA scientific output. The positions and opinions presented in this article are those of the
461 author alone and are not intended to represent any views/any official position or scientific
462 work of EFSA. To know about the views or scientific outputs of EFSA, please consult its
463 website under http://www.efsa.europa.eu.

References

1. Wasserman, L. & Roeder, K. High Dimensional Variable Selection. *Ann. Stat.* **1**, 2178–2201 (2009).
2. Sirimongkolkasem, T. & Drikvandi, R. On Regularisation Methods for Analysis of High Dimensional Data. *Ann. Data Sci.* **6**, 737–763 (2019).
3. Liu, J. Y., Zhong, W. & Li, R. Z. A selective overview of feature screening for ultrahigh-dimensional data. *Sci. China Math.* **58**, 2033–2054 (2015).
4. Statistical Learning with Sparsity: the Lasso and Generalizations. Available at: https://web.stanford.edu/~hastie/StatLearnSparsity/. (Accessed: 2nd March 2020)
5. Bogdan, M., Ghosh, J. K. & Zak-Szatkowska, M. Selecting explanatory variables with the modified version of the bayesian information criterion. in *Quality and Reliability Engineering International* **24**, 627–641 (2008).
6. Fan, J. & Peng, H. Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Stat.* **32**, 928–961 (2004).
7. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1996).
8. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **67**, 301–320 (2005).
9. Meinshausen, N. & Bühlmann, P. Stability selection. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **72**, 417–473 (2010).
10. Baldassarre, L., Pontil, M. & Mourão-miranda, J. Sparsity Is Better with Stability : Combining Accuracy and Stability for Model Selection in Brain Decoding. *Front. Neurosci.* **11**, 62 (2017).
11. Heinze, G., Wallisch, C. & Dunkler, D. Variable selection – A review and recommendations for the practicing statistician. *Biometrical Journal* **60**, 431–449

(2018).

12. Alexander, D. H. & Lange, K. Stability selection for genome-wide association. *Genet. Epidemiol.* **35**, 722–728 (2011).

13. Kim, K., Koo, J. & Sun, H. An empirical threshold of selection probability for analysis of high-dimensional correlated data. *J. Stat. Comput. Simul.* **90**, 1606–1617 (2020).

14. Lima, E. *et al.* Use of bootstrapped, regularised regression to identify factors associated with lamb-derived revenue on commercial sheep farms. *Prev. Vet. Med.* **174**, 104851 (2020).

15. Hyde, R. M., Green, M. J., Hudson, C. & Down, P. M. Factors associated with daily weight gain in preweaned calves on dairy farms. *Prev. Vet. Med.* **190**, 105320 (2021).

16. Spooner, A. *et al.* A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Sci. Rep.* **10**, 20410 (2020).

17. R Tibshirani, D. W. G. J. T. H. *An Introduction to Statistical Learning with Applications in R (older version). Springer Texts in Statistics* (New York : Springer, [2013] ©2013, 2013).

18. Zhang, C. H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **38**, 894–942 (2010).

19. Lima, E., Hyde, R. & Green, M. Model selection for inferential models with high dimensional data: synthesis and graphical representation of multiple techniques. *Sci. Rep.* **11**, 412 (2021).

20. Piotr Szulc. bigstep: Stepwise Selection for Large Data Sets. R package. (2019).

21. R Core Team. R: A language and environment for statistical computing (version 1.1.463). (2018).

22. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).

514 23. Kuhn, M. *et al.* caret: Classification and Regression Training. (2019).
515 24. Breheny, P. & Huang, J. Coordinate descent algorithms for nonconvex penalized
516 regression, with applications to biological feature selection. *Ann. Appl. Stat.* **5**, 232–
517 253 (2011).
518 25. Breiman, L. Bagging Predictors. *Mach. Learn.* **24**, 123–140 (1996).
519 26. Lima, E., Davies, P., Kaler, J., Lovatt, F. & Green, M. Variable selection for
520 inferential models with relatively high-dimensional data: Between method
521 heterogeneity and covariate stability as adjuncts to robust selection. *Sci. Rep.* **10**, 1–11
522 (2020).
523 27. Sauerbrei, W. The Use of Resampling Methods to Simplify Regression Models in
524 Medical Statistics. *J. R. Stat. Soc. Ser. C (Applied Stat.* **48**, 313–329 (1999).
525 28. Venables, W. N. & Ripley, B. D. *Modern applied statistics with S. Statistics and*
526 *computing* **45**, (Springer-Verlag New York, 2002).
527 29. Nagelkerke, N. J. D. A note on a general definition of the coefficient of determination.
528 *Biometrika* **78**, 691–692 (1991).
529 30. Wang, F., Mukherjee, S., Richardson, S. & Hill, S. M. High-dimensional regression in
530 practice: an empirical study of finite-sample prediction, variable selection and ranking.
531 *Stat. Comput.* **30**, 697–719 (2020).
532 31. Pfeiffer, R. M., Redd, A. & Carroll, R. J. On the impact of model selection on
533 predictor identification and parameter inference. *Comput. Stat.* **32**, 667–690 (2017).
534 32. Hofner, B., Boccuto, L. & Göker, M. Controlling false discoveries in high-dimensional
535 situations: Boosting with stability selection. *BMC Bioinformatics* **16**, 144 (2015).
536 33. Dormann, C. F. *et al.* Collinearity: a review of methods to deal with it and a simulation
537 study evaluating their performance. *Ecography (Cop.).* **36**, 27–46 (2013).
538 34. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

35. Kursa, M. B., Jankowski, A. & Rudnicki, W. R. Boruta - A system for feature selection. *Fundam. Informaticae* **101**, 271–285 (2010).

36. Lawlor, D. A., Tilling, K. & Smith, G. D. Triangulation in aetiological epidemiology. *Int. J. Epidemiol.* **45**, 1866–1886 (2016).

37. Munafò, M. R. & Davey Smith, G. Robust research needs many lines of evidence. *Nature* **553**, 399–401 (2018).

Author contributions statement

EL, RH and MG contributed to development of statistical techniques and statistical analysis.

MG, RH and EL all contributed to the writing of the final manuscript.

Table 1. Description of the simulated datasets used to evaluate five modelling techniques incorporating covariate selection stability.

Dataset name	p	True covariates	n	Description of correlation structure	R^2
HS: CONTINUOUS NORMALLY DISTRIBUTED OUTCOME; HIGH SIGNAL					
HS1	500	8	500	all true and false variables uncorrelated	0.70
HS2	500	8	500	4 true covariates correlated individually with 10 false covariates at 0.3	0.71
HS3	500	8	500	4 true covariates correlated individually with 10 false covariates at 0.5	0.69
HS4	500	8	500	4 true covariates correlated individually with 10 false covariates at 0.7	0.71
MS: CONTINUOUS NORMALLY DISTRIBUTED OUTCOME ; MEDIUM SIGNAL					
MS1	500	8	500	all true and false variables uncorrelated	0.51
MS2	500	8	500	4 true covariates correlated individually with 10 false covariates at 0.3	0.46
MS3	500	8	500	4 true covariates correlated individually with 10 false covariates at 0.5	0.45
MS4	500	8	500	4 true covariates correlated individually with 10 false covariates at 0.7	0.45
LS: CONTINUOUS NORMALLY DISTRIBUTED OUTCOME; LOW SIGNAL					
LS1	500	8	500	all true and false variables uncorrelated	0.29
LS2	500	8	500	4 true covariates correlated individually with 10 false covariates at 0.3	0.26
LS3	500	8	500	4 true covariates correlated individually with 10 false covariates at 0.5	0.28
LS4	500	8	500	4 true covariates correlated individually with 10 false covariates at 0.7	0.29
WID: CONTINUOUS NORMALLY DISTRIBUTED OUTCOME; WIDER DATASET					
WID1	1000	8	500	all true and false variables uncorrelated	0.51
WID2	1000	8	500	4 true covariates correlated individually with 10 false covariates at 0.3	0.46
WID3	1000	8	500	4 true covariates correlated individually with 10 false covariates at 0.5	0.45
WID4	1000	8	500	4 true covariates correlated individually with 10 false covariates at 0.7	0.45
BIN: BINARY OUTCOME MEDIUM SIGNAL					
BIN1	500	8	500	all true and false variables uncorrelated	0.48
BIN2	500	8	500	4 true covariates correlated individually with 10 false covariates at 0.3	0.48
BIN3	500	8	500	4 true covariates correlated individually with 10 false covariates at 0.5	0.47
BIN4	500	8	500	4 true covariates correlated individually with 10 false covariates at 0.7	0.47
IC: CONTINUOUS NORMALLY DISTRIBUTED OUTCOME; INCREASED COMPLEXITY					
IC1	500	8	500	4 true covariates correlated with each other and 20 false covariates at 0.9	0.56
IC2	500	8	500	4 true covariates correlated with each other and 92 false covariates at 0.7	0.42

Key; p – total number of covariates offered to the model, True covariates – number of covariates that were used to directly calculate the outcome and therefore deemed casual, n – number of observations (rows) of data, R^2 – amount of total variation in the outcome variable explained by the true covariates (for binary outcome data, an adapted R^2 was used (Nagelkerke, 1991).

Table 2. Results of full (non-bootstrapped) models conducted on 22 simulated datasets (Table 1) to illustrate the number of false positive and false negative covariates selected in final models for each method.

	Model type							
Dataset name	sAIC		mBIC		enet		MCP	
	FP	FN	FP	FN	FP	FN	FP	FN
HIGH SIGNAL								
HS1	11	1	0	1	22	0	1	1
HS2	15	1	0	1	31	0	2	1
HS3	15	1	0	1	31	0	3	1
HS4	17	1	0	1	35	0	1	1
MEDIUM SIGNAL								
MS1	14	1	0	1	9	0	7	1
MS2	13	1	0	1	11	0	6	1
MS3	25	1	0	1	27	0	15	1
MS4	25	1	0	1	27	0	4	1
LOW SIGNAL								
LS1	15	1	0	3	38	0	17	1
LS2	23	1	0	3	48	0	28	1
LS3	19	1	0	2	52	0	14	1
LS4	31	1	1	4	57	0	15	2
WIDER DATASET								
WID1	24	1	0	1	12	0	6	1
WID2	24	1	0	1	19	0	1	1
WID3	39	1	0	1	66	0	8	1
WID4	46	1	0	1	45	0	8	1
BINARY OUTCOME								
BIN1	17	1	0	1	20	0	5	1
BIN2	17	1	0	1	33	0	8	1
BIN3	28	1	0	1	94	0	9	1
BIN4	20	2	1	2	43	0	8	2
INCREASED COMPLEXITY								
IC1	15	1	0	2	30	0	24	2
IC2	25	2	0	3	51	0	14	2

Key; sAIC – covariate selection based on stepwise Akaike information criterion, mBIC - covariate selection based on a stepwise on a modified Bayesian Information Criterion, enet – covariate section based on elastic net regression, MCP – covariate selection based on minimax convex penalty regression, FP – number of false positive covariates selected, FN – number of false negative variables selected

Table 3. Results of models incorporating covariate selection stability (with selection threshold set at x = 0), conducted on 22 simulated datasets (Table 1), to illustrate the number of false positive and false negative covariates selected in final models for each method.

	Model type									
	sAIC		mBIC		enet		MCP		Combi	
Dataset name	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
HIGH SIGNAL										
HS1	0	0	0	0	0	0	0	0	0	0
HS2	0	0	0	0	0	0	0	0	0	0
HS3	0	0	0	0	0	0	0	0	0	0
HS4	0	0	1	0	0	0	0	0	0	0
MEDIUM SIGNAL										
MS1	0	0	0	0	0	0	0	0	0	0
MS2	0	0	0	0	0	0	0	0	0	0
MS3	0	0	0	0	0	0	0	0	0	0
MS4	1	0	0	0	0	0	0	0	0	0
LOW SIGNAL										
LS1	0	2	0	1	0	2	0	1	0	0
LS2	0	2	1	0	0	2	0	2	0	0
LS3	0	3	2	0	0	2	0	1	1	0
LS4	1	1	1	1	0	2	0	1	0	1
WIDER DATASET										
WID1	0	0	0	0	0	0	0	0	0	0
WID2	0	0	0	0	0	0	0	0	0	0
WID3	0	0	0	0	0	0	0	0	0	0
WID4	0	0	0	0	0	0	0	0	0	0
BINARY OUTCOME										
BIN1	0	0	2	0	0	0	0	0	0	0
BIN2	0	1	1	0	0	0	0	0	0	0
BIN3	1	0	1	0	0	0	0	0	1	0
BIN4	0	1	1	1	0	1	0	1	0	1
INCREASED COMPLEXITY										
IC1	0	4	1	0	0	2	0	3	0	3
IC2	0	3	0	2	0	3	0	4	0	3

Key; sAIC – covariate selection based on stepwise Akaike information criterion, mBIC - covariate selection based on a stepwise on a modified Bayesian Information Criterion, enet – covariate section based on elastic net regression, MCP – covariate selection based on minimax convex penalty regression, Combi – covariate selection based on a synthesised combination of sAIC, mBIC, enet and MCP, FP – number of false positive covariates selected, FN – number of false negative variables selected

Figure 1.

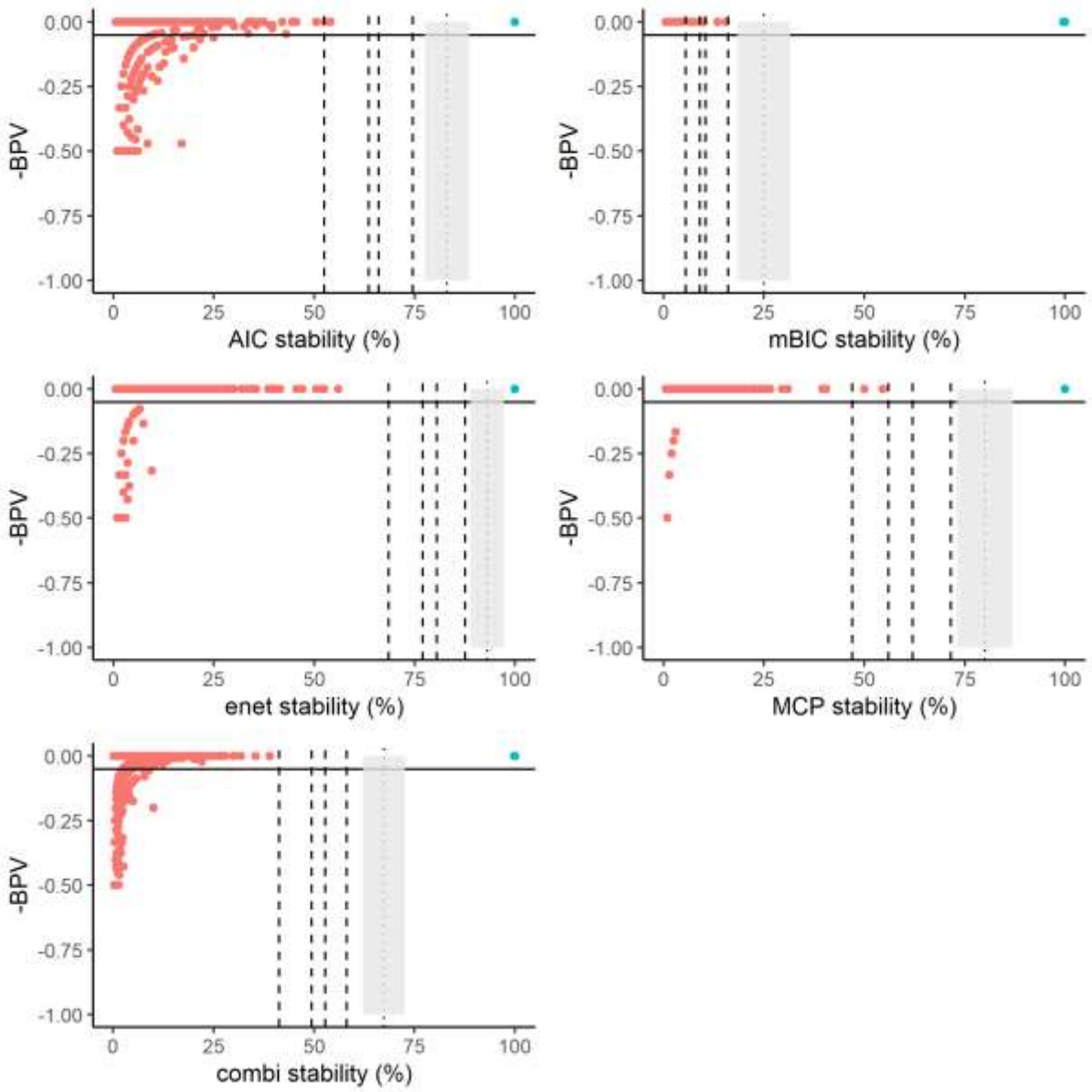
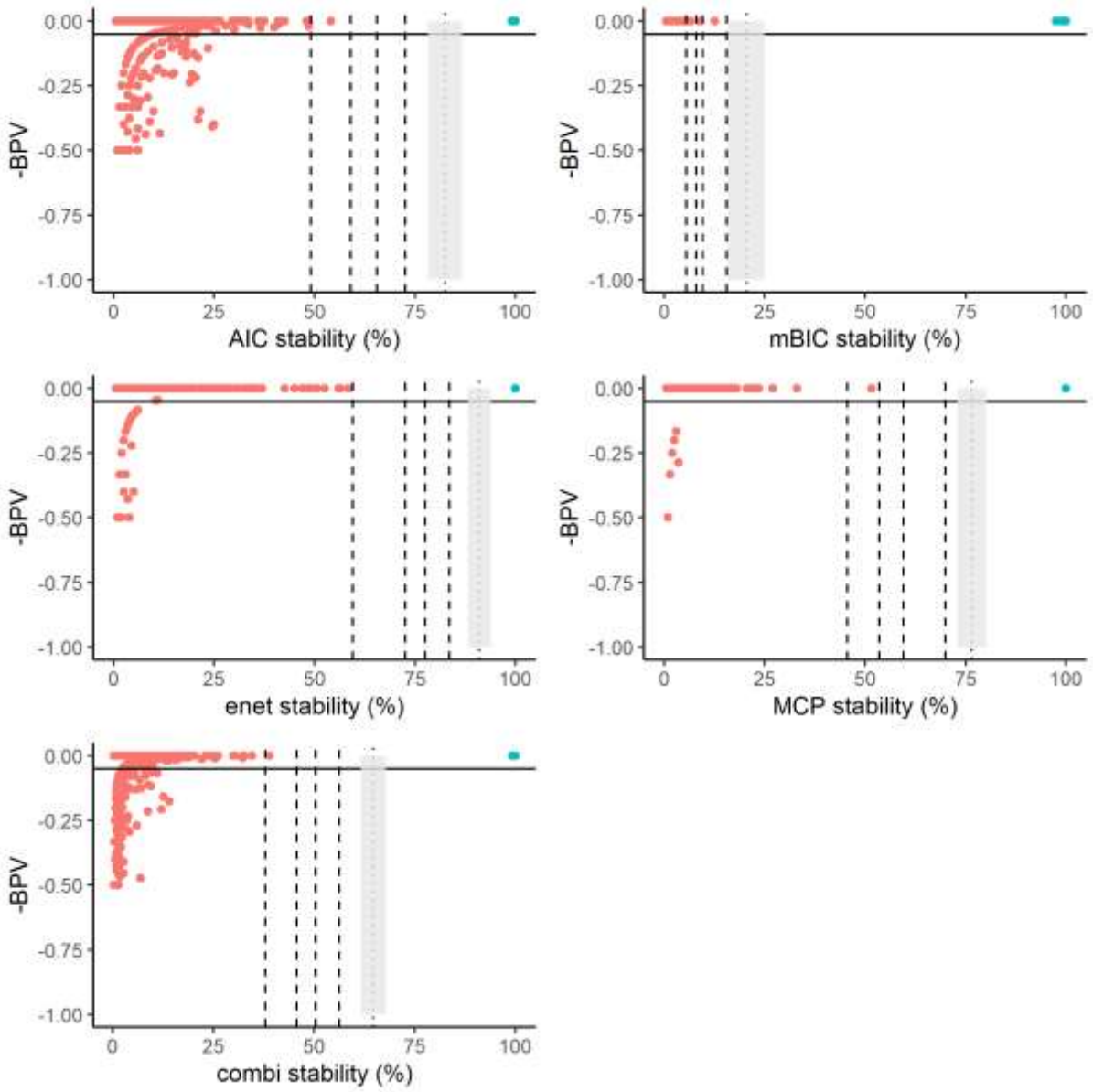
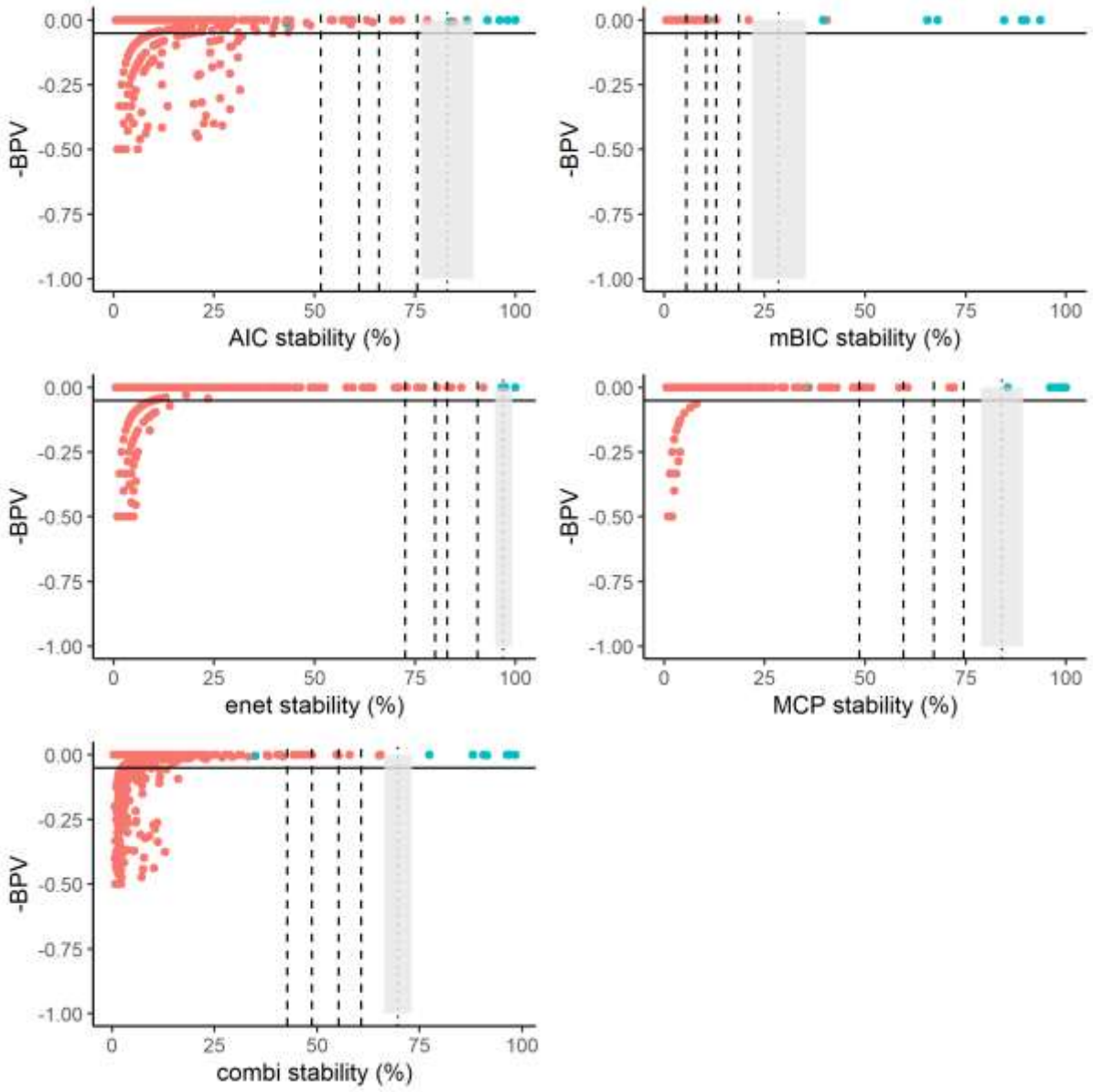

-BPV
0.00
-0.25
-0.50
-0.75
-1.00
0
25
50
75
100
AIC stability (%)
mBIC stability (%)
enet stability (%)
MCP stability (%)
combi stability (%)

Figure 2.


0.00
-0.25
-0.50
-0.75
-1.00
-BPV
0
25
50
75
100
AIC stability (%)
mBIC stability (%)
enet stability (%)
MCP stability (%)
combi stability (%)

615 Figure 3.
616



617
618
619

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterials.docx