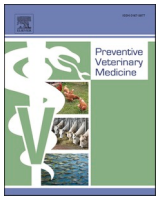
Preventive Veterinary Medicine 206 (2022) 105714

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed



Stability selection for mixed effect models with large numbers of predictor variables: A simulation study

Robert Hyde, Luke O'Grady, Martin Green *

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

ARTICLE INFO

Keywords:
Covariate selection
Mixed models
Selection stability
High dimensional clustered data
Permutation methods
False discovery rate

ABSTRACT

Covariate selection when the number of available variables is large relative to the number of observations is problematic in epidemiology and remains the focus of continued research. Whilst a variety of statistical methods have been developed to attempt to overcome this issue, at present very few methods are available for wide data that include a clustered outcome. The purpose of this research was to make an empirical evaluation of a new method for covariate selection in wide data settings when the dependent variable is clustered. We used 3300 simulated datasets with a variety of defined structures and known sets of true predictor variables to conduct an empirical evaluation of a mixed model stability selection procedure. Comparison was made with an alternative method based on regularisation using the least absolute shrinkage and selection operator (Lasso) penalty. Model performance was assessed using several metrics including the true positive rate (proportion of true covariates selected in a final model) and false discovery rate (proportion of variables selected in a final model that were non-true (false) variables). For stability selection, the false discovery rate was consistently low, generally remaining ≤ 0.02 indicating that on average fewer than 1 in 50 of the variables selected in a final model were false variables. This was in contrast to the Lasso-based method in which the false discovery rate was between 0.59 and 0.72, indicating that generally more than 60% of variables selected in a final model were false variables. In contrast however, the Lasso method attained higher true positive rates than stability selection, although both methods achieved good results. For the Lasso method, true positive rates remained ≥ 0.93 whereas for stability selection the true positive rate was 0.73–0.97. Our results suggest both methods may be of value for covariate selection with high dimensional data with a clustered outcome. When high specificity is needed for identification of true covariates, stability selection appeared to offer the better solution, although with a slight loss of sensitivity. Conversely when high sensitivity is needed, the Lasso approach may be useful, even if accompanied by a substantial loss of specificity. Overall, the results indicated the loss of sensitivity when employing stability selection is relatively small compared to the loss of specificity when using the Lasso and therefore stability selection may provide the better option for the analyst when evaluating data of this type.

1. Introduction

Identification of important or significant predictor variables when the number of variables (p) is relatively large compared to the number of observations (n), is problematic in epidemiology and remains the focus of continued research. In these circumstances, it is acknowledged that conventional procedures such as stepwise selection perform poorly, often resulting in over fit models (Hastie et al., 2015; Liu et al., 2015; Sirimongkolkasem and Drikvandi, 2019; Wasserman and Roeder, 2009). Whilst a number of variable selection and data reduction methods have been described (Chowdhury and Turin, 2020), many approaches are not applicable for wide data that contain a large number of genuine potential predictor variables. For example, pre-selection using a one-by-one univariable analysis of individual variables has been suggested (Hosmer et al., 2013) but with high dimensional, wide data, this is known to result in selection bias and inclusion of false positive covariates (Ambroise and McLachlan, 2002; Kuhn and Johnson, 2013). We provide an illustration of this effect in Section 2.1. An alternative pre-selection strategy commonly proposed is to use prior knowledge of possible causal relationships based on current beliefs or literature (Chowdhury and Turin, 2020; Royston et al., 2009), possibly including use of formal causal diagrams (Greenland et al., 1999). In this case,

* Corresponding author.
E-mail address: martin.green@nottingham.ac.uk (M. Green).

https://doi.org/10.1016/j.prevetmed.2022.105714
Received 28 March 2022; Received in revised form 8 July 2022; Accepted 10 July 2022
Available online 12 July 2022

Table 1
An overview of 3300 simulated datasets generated to conduct random effects stability selection analyses. U – uniform distribution.

Dataset Name	Number simulated sets	Size (rows / columns)	Number of higher level clusters of outcome variable	Number of subjects per cluster (* and cluster 1 per cluster 2)	True covariates at subject (a) cluster1 (b) cluster2 (c▲) level (a – b - c▲)	Distribution to draw error for cluster 1 (and cluster 2■)	Distribution to draw error at subject level
D1	300	500/500	1	20	6–2	U(0.5, 6)	U(0.5, 6)
D2	300	500/500	1	5	6–2	U(0.5, 6)	U(0.5, 6)
D3	300	500/500	1	20	4–4	U(0.5, 6)	U(0.5, 6)
D4	300	500/500	1	5	4–4	U(0.5, 6)	U(0.5, 6)
D5	300	500/ 5×10^3	1	5	4–4	U(0.5, 6)	U(0.5, 6)
D6	300	500/10^4	1	5	4–4	U(0.5, 6)	U(0.5, 6)
D7	300	500/500	1	20	10–10	U(0.5, 6)	U(0.5, 6)
D8	300	500/500	1	5	10–10	U(0.5, 6)	U(0.5, 6)
D9	300	500/500	2	20, 5 *	3 – 3–2▲	U(0.5, 4) U(0.5, 2)■	U(0.5, 6)
D10	300	500/500	2	5, 10 *	3 – 3–2▲	U(0.5, 4) U(0.5, 2)■	U(0.5, 6)
D11	300	500/10^6	1	5	4–4	U(0.5, 6)	U(0.5, 6)

however, care is required to avoid the problem of confirmation bias (Braithwaite et al., 2021) meaning only expected relationships are considered. Furthermore, this approach is not manageable when a large number of genuine potential predictors are available for analysis.

To enable safe exploration of data that contain a large number of covariates, a variety of statistical methods have been developed including modifications to AIC/BIC (Bogdan et al., 2008) and regularisation methods based on functions that penalise model coefficients to balance over and under fitting (Fan and Peng, 2004; Tibshirani, 1996; Zou and Hastie, 2005). More recent research has focussed on further reducing the detection of false positive covariates and has included the use of the 'knockoff filter' (Barber and Candès, 2015) and 'stability selection' (Lima et al., 2021a; Meinshausen and Bühlmann, 2010). The 'knockoff filter' method involves the creation of an equivalent number of additional predictor variables to that in the original set, which are designed to imitate the correlation structure of the original covariates, but with no relationship to the outcome; they are effectively negative controls for each variable. Each knockoff variable is compared with the equivalent real variable to evaluate statistically whether it differs from its knockoff in terms of association with the outcome (Barber and Candés, 2019). Stability selection uses the principle of multiple resampling. The concept is that covariates most frequently selected in a model repeatedly constructed under resampling of the original data are most likely to be truly associated with the outcome (Meinshausen and Bühlmann, 2010). As with the knockoff approach, comparison can be made with a 'control' scenario, in this case through permutation of the outcome variable, to determine a stability threshold to minimise selection of false positive covariates (Green et al., 2021).

At present, an important omission from the research base is that very few methods are available for, or applicable to, wide data that include a clustered dependent variable. Whilst limited research has been conducted for mixed models that incorporates regularisation (Eliot et al., 2011; Li et al., 2018; Schelldorfer et al., 2014), such methods tend to be unwieldy to employ and are rarely used in practice. Clustered dependent variables are common in epidemiologic research (e.g., repeated measurements within subject or multiple subjects within a group), therefore, a straightforward method to manage robust variable selection with wide data and a clustered outcome would be of value.

The purpose of this research was to make an empirical evaluation of a stability selection method for covariate selection in wide data settings when the dependent variable was clustered. To achieve this, we used multiple simulations and a variety of wide data structures to investigate the performance of a mixed model stability selection procedure and compare this to the performance of an alternative current method, incorporating regularisation.

2. Materials and methods

Before considering stability selection, in Section 2.1 we present an illustrative example of conventional covariate selection that comprises a univariable pre-selection procedure followed by a final backward stepwise selection. We apply this method to increasingly wide, randomly generated data, purposefully with no signal. Subsequently, in Section 2.2 we describe a new method for the analysis of similar data through use of stability selection.

2.1. Illustration of selection bias: an example of covariate selection using random effects models when no signal is present

To illustrate the impact of covariate selection bias associated with use of a univariable pre-filtration step, a simulation study was conducted using 1200 hypothetical datasets. Data generation and analysis were conducted using the R platform (R Core Team, 2020). All explanatory variables within the data consisted of randomly generated vectors (columns) drawn from random standard normal distributions such that each had a mean = 0 and standard deviation = 1. The number of rows (n = 500, 1000) and columns (p = 30, 50, 100, 300, 500, 1000) varied by dataset; the latter constituted the independent variables. A further vector that constituted the dependent variable was added to each set of data. Therefore, each simulated dataset contained a dependent variable which was a single column of length n (equivalent to the number of rows in that dataset) and p independent variables (where p sequentially equalled 30, 50, 100, 500, 1000 or 5000). The dependent variable had a clustered structure (and was therefore appropriate for a random effects analysis) and was generated as the sum of a draw from a standard normal distribution (mean = 0, standard deviation = σ_a) and a draw from a standard normal distribution at cluster level (mean = 0, standard deviation = σ_b). The values of σ_a and σ_a were drawn from a uniform distribution U(0.5, 6) to allow the degree of higher level clustering to vary between simulations. The dependent variable was clustered in groups of 5; that is there were 5 lower level units (rows of data) within each higher level cluster. The structure of the simulated datasets was similar to that of dataset D2 in later analysis (see below and Table 1) except that the dependent variable had no relationship with the independent variables and was simply a randomly generated, clustered vector. For example, for the size of dataset with 500 rows of data (n = 500), 300 independent variables (p = 300) and a clustered, dependent variable (of length 500), data were simulated at random 50 times and the 50 separate datasets subjected to an identical analytic procedure to allow comparison of results.

Using these randomly generated datasets with no pre-defined relationships between predictor and outcome variables, a commonly used

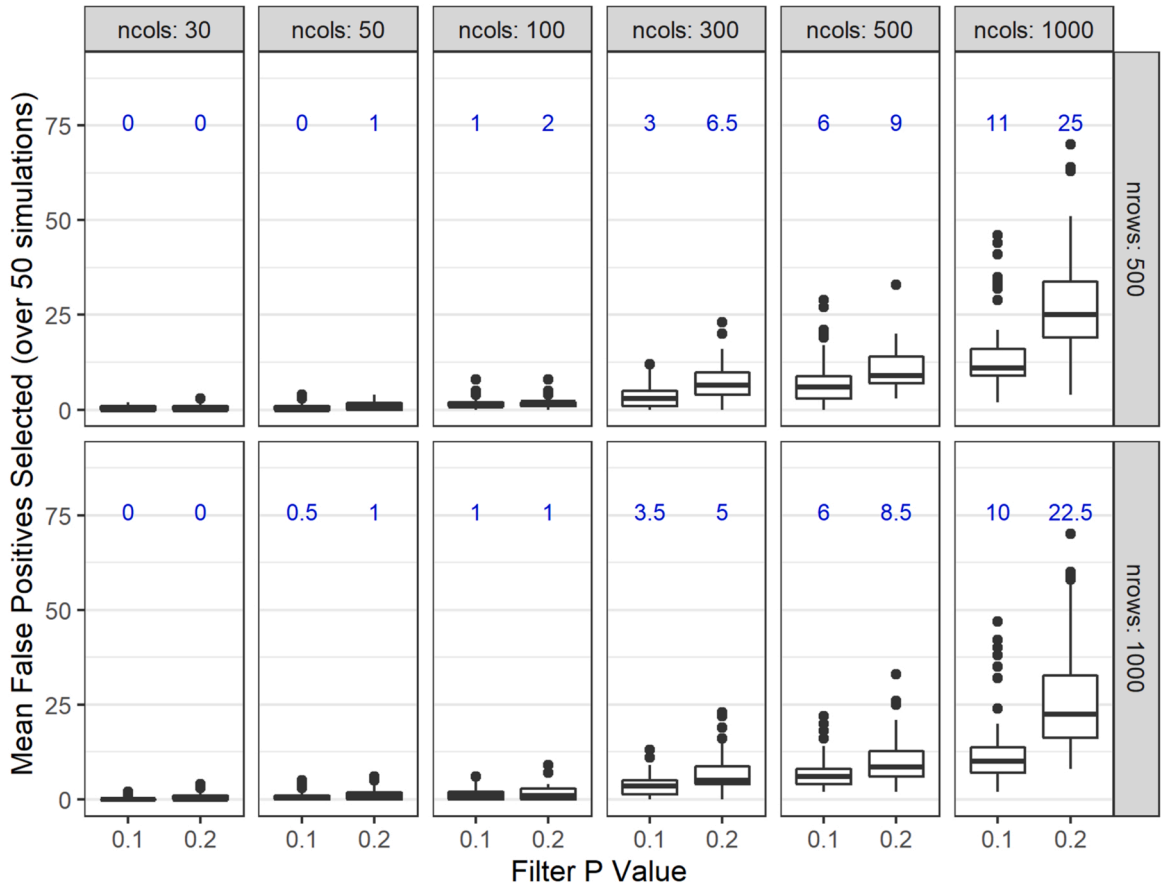


Fig. 1. An illustration of the distribution of false positive covariates selected from 1200 simulated datasets with different numbers of rows (nrows) and columns (ncols), in which a clustered dependent and standard normal independent variables were generated at random. i.e., no relationship between dependent and independent variables existed. The modelling procedure comprised an initial univariable filter based on a Pearson correlation P value (with a threshold for variable inclusion defined on the x-axis as 'Filter P Value') followed by conventional backward stepwise mixed effect linear regression based on $P < 0.05$. The number above each boxplot represents the median value, over 50 simulations, of the number of false positive variables selected.

modelling procedure was implemented as follows. An initial data reduction step was applied using a univariable Pearson correlation. The correlation between the dependent and each independent covariate was calculated, and independent covariates retained for further modelling when the correlation P value was less than a defined threshold (set as either 0.1 or 0.2; this is standard practice when a univariable pre-filtration method is used). Remaining variables were carried forward to a random effects model constructed using the lme4 package (Bates et al., 2015), with final variables selected using conventional backward stepwise regression based on $P < 0.05$. Backward stepwise regression was implemented using the lmerTest package (Kuznetsova et al., 2017) in R. For each of the simulated dataset structures described above, the procedure was repeated 50 times (i.e., with 50 different sets of randomly generated data) and the number of independent covariates selected in the final model recorded. Results were displayed graphically to illustrate the distribution of the number of false discovery variables selected in each dataset. Code to reproduce this analysis is provided in Supplementary Materials.

Results are presented in Fig. 1 and demonstrate that after a pre-selection procedure to reduce the number of variables, when no underlying association between dependent and independent variables were present in the data, a substantial number of false discovery variables were selected. As expected, as the number of columns of randomly generated independent variables increased, the number of false discovery covariates selected for each dataset also increased. Whilst it is well known that an increase in independent variables results in multiple testing that requires some statistical adjustments to control the false discovery rate (Curran-Everett, 2017; Simas et al., 2014), a univariable pre-filtration step does not alleviate the issue. Importantly from these examples it can be seen that once the number of columns reaches ≥ 300, the number of false discovery covariates selected is substantial and this effect occurs whether the dependent variable is clustered or not (analysis not shown). Yet in epidemiological research, it is common practice to use univariable pre-filtration followed by conventional stepwise selection with $P < 0.05$ to determine final covariates.

2.2. *Stability selection for mixed effect models*

In this section we describe a procedure for implementing stability selection for data with a clustered outcome based on a linear mixed effect, random intercept model. In Section 2.2.1 we describe the creation of simulated datasets, with a clustered outcome 'y_out' dependent on either eight or twenty known 'true' predictor variables, and which incorporated a large set of additional random 'noise' variables. In Section 2.2.2 we describe the stability selection procedure and in Section 2.2.3 define the approach to evaluation of model performance. In Section 2.2.4 we describe an alternative current method of variable selection, based on regularisation, to compare results with stability selection.

2.2.1. *Simulated datasets*

Simulated datasets were constructed with a specified outcome variable dependent on a set of fixed effects ($n = 8$ or 20) and in which error in the outcome arose from combinations of unknown effects at both subject and cluster level. Steps in dataset generation were as follows:

Table 2
Performance metrics for stability selection models across 3300 simulated datasets that contained a clustered outcome variable. TPR - true positive rate, FDR - false discovery rate, FDER - false discovery noise error rate, NIR - no information rate, fFDR - filtered false discovery rate, iqr – interquartile range.

Dataset Name	Number simulated sets	TPR Mean (median, iqr)	FDR Mean (median, iqr)	FDER Mean (median, iqr)	NIR Mean (median, iqr)	fFDR Mean (median, iqr)
D1	300	0.96 (1.00, 1.00–1.00)	0.02 (0, 0–0)	0.0004 (0, 0–0)	0.001 (0, 0–0.002)	0.35 (0.33, 0.20 –0.47)
D2	300	0.97 (1.00, 1.00–1.00)	0.02 (0, 0–0)	0.0004 (0, 0–0)	0.001 (0, 0–0.002)	0.32 (0.33, 0.20 –0.47)
D3	300	0.92 (1.00, 0.88–1.00)	0.03 (0, 0–0)	0.0006 (0, 0–0)	0.001 (0, 0–0.002)	0.35 (0.33, 0.20 –0.47)
D4	300	0.95 (1.00, 0.87–1.00)	0.02 (0, 0–0)	0.0004 (0, 0–0)	0.001 (0, 0–0.002)	0.33 (0.33, 0.20 –0.47)
D5	300	0.80 (1.00, 0.75–1.00)	0.05 (0, 0–0)	0.000003 (0, 0–0)	2×10^{-5} (0, 0–0)	0.47 (0.50, 0.33 – 0.62)
D6	300	0.73 (1.00, 0.63–1.00)	0 (0, 0–0)	0 (0, 0–0)	3×10^{-6} (0, 0–0)	0.49 (0.53, 0.37 – 0.64)
D7	300	0.78 (0.78, 0.65–0.90)	0.02 (0, 0–0)	0.0005 (0, 0–0)	0.002 (0, 0–0.002)	0.26 (0.27, 0.11–0.38)
D8	300	0.95 (1.00, 0.95–1.00)	0.02 (0, 0–0)	0.0004 (0, 0–0)	0.002 (0.002, 0–0.004)	0.31 (0.33, 0.20–0.42)
D9	300	0.84 (0.88, 0.75–1.00)	0.02 (0, 0–0)	0.0005 (0, 0–0)	0.001 (0, 0–0.002)	0.31 (0.33, 0.20–0.43)
D10	300	0.91 (1.00, 0.87–1.00)	0.03 (0, 0–0)	0.0005 (0, 0–0)	0.001 (0, 0–0.002)	0.34 (0.33, 0.27–0.47)
D11	300	0.63 (0.70, 0.45–0.85)	0 (0, 0–0)	0 (0, 0–0)	0 (0, 0–0)	0.34 (0.35, 0.19–0.47)

i. Size of dataset specified as comprising 500 rows and either 500, 5000, 10,000 or 100,000 predictor variables
ii. A hierarchical clustered structure of y was specified as either:
   a. 1 level of clustering with either 5 subjects (rows) per cluster or 20 subjects per cluster
   b. 2 levels of clustering with either 5 subjects per level 1 cluster and 10 level 1 clusters per level 2 cluster, or 20 subjects per level 1 cluster and 5 level 1 cluster per level 2 cluster.
iii. A sets of 'true' predictor variables (fixed effects) were defined at both subject and cluster level. The number of true variables at each level varied by dataset (see Table 1).
iv. For simulated datasets in which the dependent variable had one higher level cluster, the outcome variable, "y_out", was specified as follows;

$$y_{out_{ij}} = Intercept + \sum_{k=1}^{r} 2x_{ij} + \sum_{l=1}^{q} 2x_j + cl_j + e_{ij} \quad (1)$$

$$cl_j \sim N(0, \quad \sigma_{cl})$$

$$e_{ij} \sim N(0, \quad \sigma_e)$$

Where $y_{out_{ij}}$ represented an outcome consisting of values for subjects i within clusters j, $Intercept = 1$, x_{ij} represented the k^{th} of r true covariates at subject level drawn from a standard normal distribution, x_j represented the l^{th} of q true covariates at cluster level drawn from a standard normal distribution, cl_j was a random variable at cluster level drawn from a Normal distribution with mean = 0 and standard deviation drawn randomly from a uniform distribution U(0.5, 6) and e_{ij} a random variable that represented all other unknown effects on $y_{out_{ij}}$ drawn from a Normal distribution with mean = 0 and standard deviation also drawn randomly from a uniform distribution U(0.5, 6) (Table 1). No covariance was assumed between cl_j and e_{ij}.

The effect size of the true covariates was set at 2 because, alongside the random variation incorporated in the outcome as defined by the distributions of σ_{cl} and σ_e, this provided sufficient signal for all true covariates (x_{ij} and x_j) to be 'significant' in a conventional analysis ($P < 0.05$) when modelled alone (i.e., without additional noise variables).

For two simulated datasets in which the dependent variable had two higher level clusters, similar principles were used to specify the dependent variable with the addition of an extra random variable to define error of the higher-level cluster.

v. To complete each dataset, additional 'noise' variables were generated, each from a standard normal distribution. For each data set, additional noise variables were added such that the total number of variables available for selection were either 500, 5000, 10,000 or 100,000.

Therefore, each simulated dataset contained a clustered outcome variable dependent on a set of true covariates, with randomly assigned associated error at both subject and cluster level and a large number of additional noise variables generated at random. The characteristics of all datasets are summarised in Table 1. Code to reproduce example simulated datasets is available at https://github.com/cran/stabiliser (see Supplementary Materials).

2.2.2. Stability model construction

All analyses were conducted using R statistical software (R Core Team, 2020). A detailed, step by step description of the stability selection procedure, including calculation of the stability threshold, is

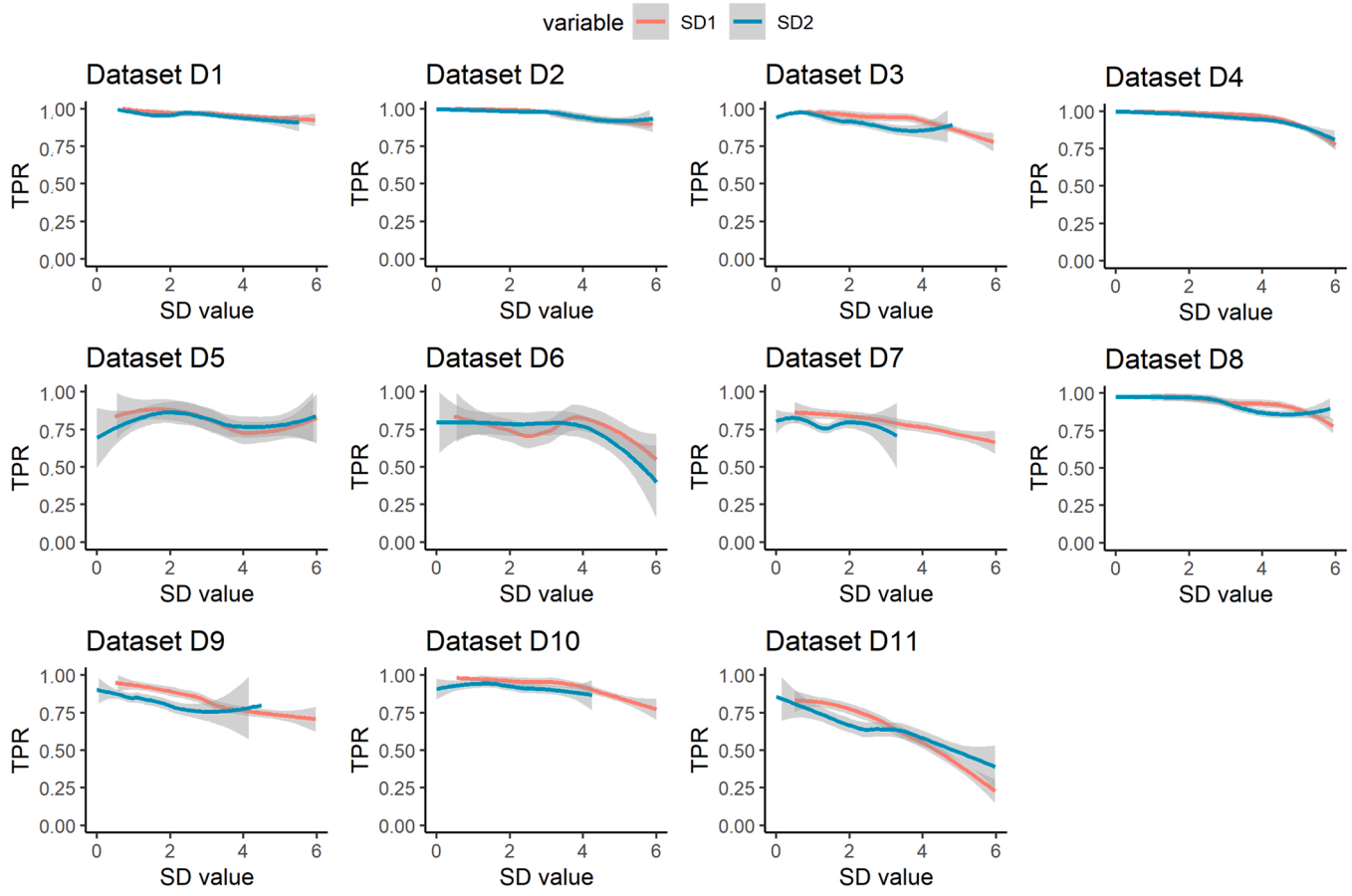


Fig. 2. Smoothed line plots (using locally estimated scatterplot smoothing: LOESS) depicting the mean and 95% confidence interval (shaded) proportion of true covariates selected in each of 300 models by stability selection (true positive rate (TPR)) with increasing model error at subject level (SD1) or cluster 1 level (SD2). The smoothed line and 95% confidence intervals were calculated by fitting local polynomial regression using the default loess function in the stats package of R (R Core Team, 2020). In general, the TPR tended to decrease as SD values increased.

provided in Supplementary Materials Table S1. The R code to reproduce stability selection analyses is available at https://github.com/cran/-stabiliser (see Supplementary Materials). In brief, steps in the stability selection procedure were as follows:

i. A univariable Pearson correlation between each individual covariate and the outcome was calculated and the 100 covariates with the largest absolute Pearson correlation coefficients were caried forward for further analysis. Although this step was not essential, it reduced computation time without resulting in selection bias (this was tested by computing the no-information rate, as descried below).
ii. A backward stepwise, mixed effects model was built using the lme4 package (Bates et al., 2015). A conventional bootstrapping procedure (sampling of rows with replacement, to a sample size equal to that of the original data (Efron, 1979)) was used to estimate covariate stability as previously described (Lima et al., 2020). In brief, stability selection was defined as the percentage of times each covariate was selected in the mixed effect model across 100 bootstrap samples.
iii. Variables were selected in a final model when their stability exceeded a specified stability threshold (see Supplementary Materials Table S1 for full details). In brief, the stability threshold represents the maximum stability that would be expected *when no true covariates are present* in the data. To estimate covariate stability when no true covariates are present, the outcome variable is randomly permuted to sever any relationships between outcome and explanatory variables. Using the permuted outcome, a bootstrap resampling procedure is repeated as described above in ii) and the stability of all covariates estimated; this represents a null or baseline stability distribution for the data. The permuted bootstrap resampling is repeated ten times and the mean of the maximum covariate stability in each permuted run is calculated and set as the stability threshold for the full model analysis.

2.2.3. *Model performance*

Performance of the stability selection procedure across simulated datasets was judged using the following criteria:

a) True positive rate (TPR); the proportion of true covariates selected in the final model. This was effectively a measure of the power of analysis; the probability that a true covariate was identified.
b) False discovery rate (FDR); the number of noise variables selected as a proportion of the total number of variables selected in the final model.
c) False discovery noise error rate (FDER); the number of noise variables selected in the final model as a proportion of the total noise variables available for selection.
d) To evaluate whether selection bias was present in the stability selection procedure, the no information rate (NIR) was calculated for each dataset. To achieve this, for each simulated dataset and modelling procedure, *the dependent variable was permuted to remove underlying relationships to the true covariates.* Having rerun the stability selection procedure with a permuted outcome, the mean NIR was defined as the number of noise variables selected in the final

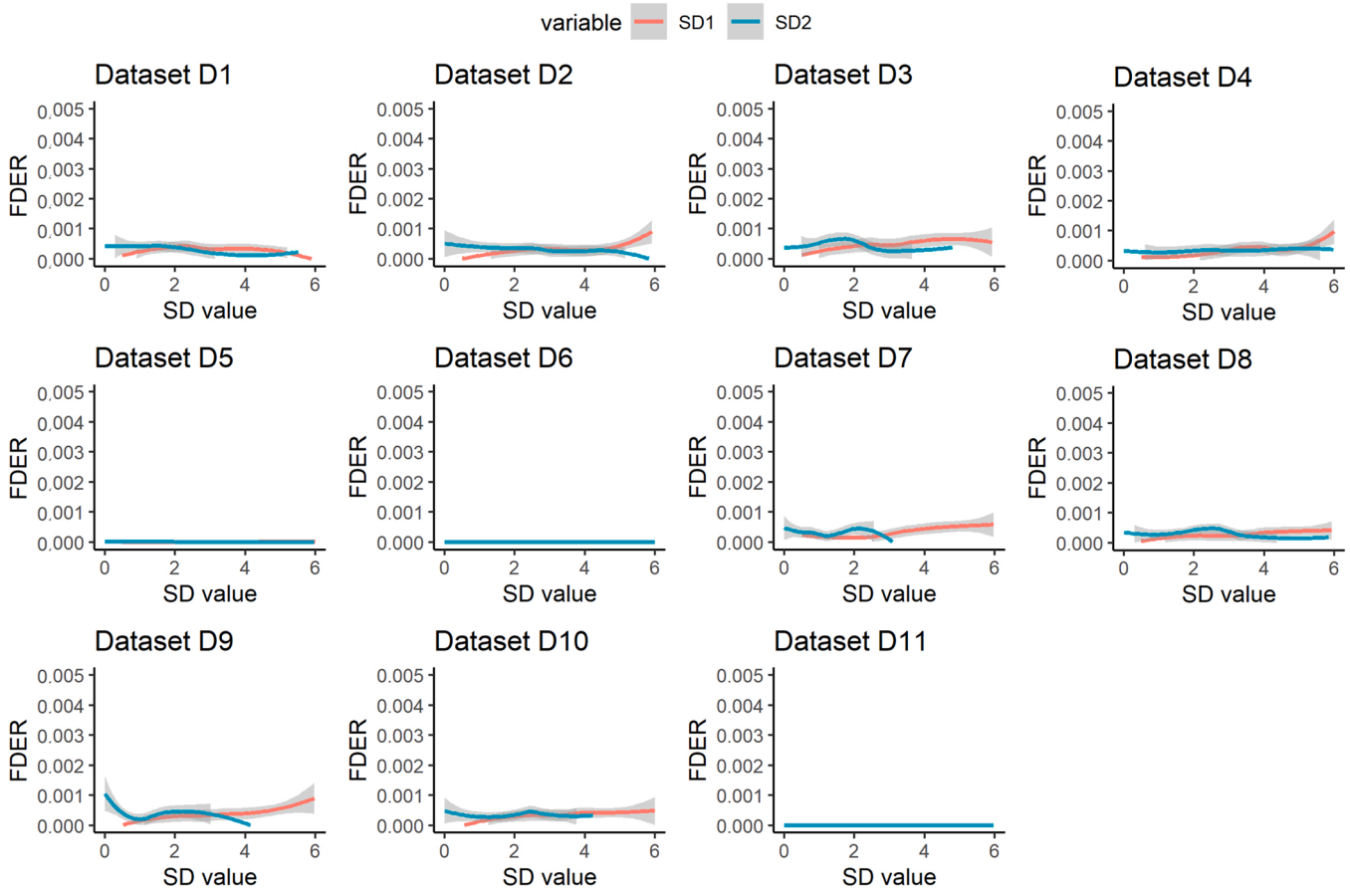


Fig. 3. Smoothed line plots (using locally estimated scatterplot smoothing: LOESS) depicting the mean and 95% confidence interval (shaded) proportion of false positive covariates (false discovery noise error rate; FDER) selected in each of 300 models by stability selection and increasing model error at subject level (SD1) or cluster 1 level (SD2). The smoothed line and 95% confidence intervals were calculated by fitting local polynomial regression using the default loess function in the stats package of R (R Core Team, 2020). It was notable that FDER remained below 1 in 1000 across all models and values of model error incorporated in the data.

model as a proportion of the total noise variables available. Therefore, the NIR represented the false discovery error rate *when no true covariates were present in the data*.

In addition, to provide a comparison with a common conventional method of variable selection, a 'pre-filtration' approach was used for the analysis of all simulated datasets. This comprised undertaking an initial univariable Pearson correlation between each covariate and the outcome to identify the 100 covariates with the highest absolute correlation coefficients and subsequently conducting a backward stepwise selection at $p < 0.05$ to determine covariates in a final model. The mean number of noise variables selected as a proportion of the total number of variables selected in each final model across 300 data simulations was calculated and termed the 'filtered false discovery rate' (fFDR).

2.2.4. Comparison with glmmlasso

To compare results obtained from stability selection with a published alternative method of covariate selection, a regularized linear mixed model with the least absolute shrinkage and selection operator (Lasso) penalty was implemented, as previously described (Schelldorfer et al., 2014). In brief, the method includes a parameter, λ, that governs the size of a penalty applied to the sum of the covariate coefficients, which has the effect of reducing coefficient size. Coefficients of some variables are set to zero meaning they are not selected in the model, hence the Lasso provides in-built covariate selection (Tibshirani, 1996). In this study the value of lambda was optimised using five-fold cross validation to identify the minimum mean absolute error (Kuhn and Johnson, 2013). The method was implemented in R using the glmmLasso package (Groll, 2017). For computational reasons (the algorithm is very slow to implement on very wide data), it was not possible to implement the algorithm on datasets with 100,000 covariates and only 100 datasets were used with 10,000 covariates. Comparisons between stability selection and glmmLasso were made using the metrics TPR, FDR and FDER.

3. Results

3.1. Stability selection; model performance

A summary of stability selection model performance for all 3300 simulated datasets is shown in Table 2. For stability selection, across all dataset classes, the false discovery rate and false discovery noise error rates were consistently low. The mean FDR remained ≤ 0.05 indicating that on average fewer than 1 in 20 of the variables selected in a final model were noise variables. This was in contrast to the mean filtered false discovery rate which ranged from 0.15 to 0.47 indicating that, as expected, a univariable filtration followed by conventional backward stepwise selection resulted in a substantial number of noise variables being selected in final models. With stability selection, in addition to the FDR and fFDR remaining low ($\leq$0.05), the mean true positive rate was 0.73–0.97 (median 0.78–1.00) with the exception of the dataset using 100,000 covariates in which the mean (median) TPR was 0.63 (0.70). Therefore, the power of analysis remained reasonably high; on average the large majority, although not all of the true covariates were identified. An evaluation of the distributions of TPR across simulations (Fig. 2), identified that in general, selection of true variables decreased

Table 3
Performance metrics for glmmLasso models across 2800 simulated datasets that contained a clustered outcome variable. TPR - true positive rate, FDR -false discovery rate, FDER - false discovery noise error rate, iqr – interquartile range.

Dataset Name	Number simulated sets	TPR Mean (median, iqr)	FDR Mean (median, iqr)	FDER Mean (median, iqr)
D1	300	1.00 (1.00, 1.00–1.00)	0.66 (0.68, 0.56–0.75	0.04 (0.03 0.02–0.05
D2	300	0.99 (1.00, 1.00–1.00)	0.64 (0.65, 0.56–0.75	0.04 (0.03 0.02–0.05
D3	300	0.99 (1.00, 1.00–1.00)	0.68 (0.70, 0.62–0.77	0.04 (0.04 0.03–0.05
D4	300	0.93 (1.00, 1.00–1.00)	0.59 (0.67, 0.56–0.74	0.04 (0.03 0.02–0.05
D5	300	0.99	0.79 (0.71, 0.62–0.78	0.01 (0.04 0.004–0.02
D6	100	0.92 (1.00, 1.00–1.00)	0.30 (0.25, 0–0.58	0.001 (0.0003 0–0.001
D7	300	0.98 (1.00, 0.95–1.00)	0.72 (0.72 0.66–0.78	0.12 (0.11 0.08–0.135
D8	300	1.00 (1.00, 1.00–1.00)	0.71 (0.70 0.66–0.88	0.10 (0.11 0.08–0.13
D9	300	0.98 (1.00, 1.00–1.00)	0.68 (0.71, 0.62–0.78	0.05 (0.04 0.02–0.06
D10	300	0.98 (1.00, 1.00–1.00)	0.68 (0.86, 0.69–0.93)	0.05 (0.04 0.03–0.06
D11	NA	NA	NA	NA

as model error increased. Increased model error effectively dilutes the signal attributable to the true covariates and therefore this pattern is unsurprising. Importantly, despite this reduction in power with a higher degree of error, the FDR did not increase (Fig. 3) suggesting that even when true covariates cannot be identified because the associated error is too high, these are not replaced by noise covariates.

Importantly the mean no information rates for the stability selection procedure remained low at ≤ 0.001 across all simulations and data structures indicating that even when no relationships were present in the data, on average only 1 in 1000 covariates were incorrectly selected as false positives.

Performance metrics for the glmmLasso method are presented in Table 3. Across all datasets the mean TPR remained ≥ 0.93 indicating that for the vast majority of datasets, all true variables were selected. The false discovery rate and false discovery noise error rates, however, were consistently high, the FDR was between 0.59 and 0.72, and the FDER between 0.04 and 0.12. Therefore, generally more than 60% of variables selected in a final model were noise variables.

4. Discussion

The purpose of this study was to evaluate a mixed model stability selection procedure for covariate selection in data with a clustered outcome variable. The proposed method appeared to work well, with good overall performance metrics, and notably better than the conventional pre-filtration method (Table 2) that resulted in very high false discovery rates. Indeed, a key strength of stability selection appeared to be a strong control of false discovery rates resulting in a high degree of certainty across all data types and error structures tested, that the variables selected were indeed true positives (>95% certainty). This was in contrast to the glmmLasso method which had a tendency to over fit data resulting in the selection of substantial numbers of noise variables and false discovery rates in the region of 0.60–0.70 (as opposed to <0.05 for stability selection). Whilst stability selection provided superior control of false discovery rates than the Lasso method, this was associated with a slight reduction in TPR (ability to detect the true variables). For most classes of dataset, this difference was relatively small at < 0.1 (i.e., the Lasso method identified 1 in 10 additional true variables compared to stability selection) although as dataset width increased, this difference increased to around 0.3. However, as noted previously, the relatively high TPR achieved using glmmLasso was accompanied by a very poor FDR indicating the two methods appear to have differing strengths. The Lasso method tended to over fit, identify a high proportion of true positive variables but at the cost of selecting large numbers of noise variables, whilst stability selection tended to slightly underfit resulting in sparse models that contained nearly entirely only true variables, but at the cost of omitting a small proportion of the true variables. This suggests the two methods may both be useful in different situations. When high specificity is needed in terms of identification of true covariates (i.e., high certainty is needed that selected variables are truly associated with an outcome), stability selection appears to offer a good solution, although with a slight loss of sensitivity. Conversely when high sensitivity is needed (i.e., a high certainty is required that all true covariates associated with an outcome are identified, even with additional noise variables) the Lasso approach may be useful, even if accompanied by a substantial loss of specificity. Results from this simulation study suggest that the loss of sensitivity when employing stability selection is relatively small compared to the loss of specificity when using the Lasso and therefore, in general, stability selection may provide a better choice for the analyst when evaluating wide data with a clustered outcome.

We conducted additional analyses of the distributions of TPR and FDR for stability selection models (Figs. 1 and 2) and identified, as may be expected, that TPR tended to decrease as the magnitude of error associated with the outcome variable increased. This relationship was not surprising since additional error in the outcome variable will dilute the direct relationship between the true covariates and outcome meaning some may be missed from selection. Whilst such a loss of power is not ideal, importantly it was not accompanied by an increase in FDR (Fig. 2) which provides confidence in the variables selected using this approach; when covariates are selected, there was a high level of certainty that they were true positives.

The methods implemented in this research, stability selection and glmmLasso, have some similarity to statistical correction for multiple testing; all of the methods aim to limit identification of false positive covariates. Although a detailed comparison is beyond the scope of this paper, modern methods for minimising false discovery rates in multiple testing have been evaluated recently (Korthauer et al., 2019). In summary, the more recent methods described for multiple comparison correction, such as independent hypothesis weighting (Ignatiadis et al., 2016) and Adaptive p-value Thresholding (Lei and Fithian, 2018), showed only mild advantages over the traditional approaches such as the Bonferroni correction (Dunn, 1961) or the Benjamini-Hochberg Procedure (Benjamini and Hochberg, 1995). In fact all methods showed a tendency for conservatism (reduction in detection of true positive covariates (Korthauer et al., 2019)). A further problem arises with these approaches regarding the validity in a complex model-based setting and this is one reason for continued research in this field (Barber and Candés, 2019; Barber and Candès, 2015). It is therefore clear that currently these methods do not provide a full solution for the analysis of wide data, especially within the context of data with a correlated outcome and therefore stability selection may provide a viable alternative for such analyses.

4.1. Study limitations

There are a variety of limitations to this empirical study. The

simulated datasets contained normally distributed fixed and random variables which although suitable to meet the assumptions of mixed effect models could be considered idealised. For example, non-standard Gaussian distributions with skewness may be encountered in real data and it is unknown whether the stability selection modelling procedures would perform similarly with such data. No non-linear relationships nor random slopes were tested in the modelling procedures, and this would be a useful subsequent element of research although there is no obvious reason to believe this would result in poorer model performance. The number of true covariates (8 or 20) in the simulated datasets was relatively low so as to align with the 'bet on sparsity principle' (Hastie et al., 2015), which dictates that for causal variables to be identifiable (in any model), a relatively small number of predictors have to be responsible for most of the effect on the outcome of interest. Overall, it is important to note that our results do not extend beyond the scenarios we have tested and further work on stability selection for random effects models with a variety of different data structures and an increasing number of true variables would be of benefit.

4.2. Conclusions

In this research we have proposed a new method of stability selection applicable to wide data with a clustered outcome variable. Empirically, the method appears effective at controlling false discovery rates although with very large numbers of noise variables, this may be accompanied by a slight loss of power (ability to detect the true covariates). The R code and functions to implement examples and analyses in this manuscript are available at https://CRAN.R-project.org/package=stabiliser and https://github.com/cran/stabiliser (also in Supplementary Materials).

Declarations of interest

None.

Acknowledgements

We would like to thank Dr Eliana Lima who played an important role in the original development of the stability selection framework implemented in this manuscript.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.prevetmed.2022.105714.

References

Ambroise, C., McLachlan, G.J., 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. USA 99, 6562–6566. https://doi.org/10.1073/PNAS.102102699.

Barber, R.F., Candés, E.J., 2019. A knockoff filter for high-dimensional selective inference. Ann. Stat. 47, 2504–2537. https://doi.org/10.1214/18-AOS1755.

Barber, R.F., Candès, E.J., 2015. Controlling the false discovery rate via knockoffs. Ann. Stat. 43, 2055–2085. https://doi.org/10.1214/15-AOS1337.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67, 1–48. https://doi.org/10.18637/jss.v067.i01.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Bogdan, M., Ghosh, J.K., Zak-Szatkowska, M., 2008. Selecting explanatory variables with the modified version of the bayesian information criterion. Qual. Reliab. Eng. Int. 627–641. https://doi.org/10.1002/qre.936.

Braithwaite, R.S., Ban, K.F., Stevens, E.R., Caniglia, E.C., 2021. Rounding up the usual suspects: confirmation bias in epidemiological research. Int. J. Epidemiol. 50. https://doi.org/10.1093/ije/dyab091.

Chowdhury, M., Turin, T.C., 2020. Variable selection strategies and its importance in clinical prediction modelling. Fam. Med Com. Health 8, 262. https://doi.org/10.1136/fmch-2019-000262.

R. Core Team, 2020. R Core Team. R: A language and environment for statistical computing.

Curran-Everett, D., 2017. CORP: minimizing the chances of false positives and false negatives. J. Appl. Physiol. https://doi.org/10.1152/japplphysiol.00937.2016.

Dunn, O.J., 1961. Multiple comparisons among means. J. Am. Stat. Assoc. 56, 52–64. https://doi.org/10.1080/01621459.1961.10482090.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7, 1–26. https://doi.org/10.1214/aos/1176344552.

Eliot, M., Ferguson, J., Reilly, M.P., Foulkes, A.S., 2011. Ridge regression for longitudinal biomarker data, 00001022021557467913 53 Int. J. Biostat. 7. https://doi.org/10.2202/1557-4679.1353.

Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. Ann. Stat. 32, 928–961. https://doi.org/10.1214/009053604000000256.

Greenland, S., Pearl, J., Robins, J.M., 1999. Causal diagrams for epidemiologic research. Epidemiology 10. https://doi.org/10.1097/00001648-199901000-00008.

Green, M., Lima, E., Hyde, R.M., 2021. Selection stability in high dimensional statistical modelling: defining a threshold for robust model inference. Preprint. (Version 1) available Res. Sq.. https://doi.org/10.21203/rs.3.rs-738092/v1.

Groll, A., 2017. glmmLasso: Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation.

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman and Hall/CRC.

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied Logistic Regression: Third Edition. Applied Logistic Regression: Third Edition 1–510. https://doi.org/10.1002/9781118548387.

Ignatiadis, N., Klaus, B., Zaugg, J.B., Huber, W., 2016. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nat. Methods 13, 577–580. https://doi.org/10.1038/nmeth.3885.

Korthauer, K., Kimes, P.K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E.J., Hicks, S.C., 2019. A practical guide to methods controlling false discoveries in computational biology. Genome Biol. 20, 118. https://doi.org/10.1186/s13059-019-1716-1.

Kuhn, M., Johnson, K., 2013. Applied predictive modeling. Appl. Predict. Model. 1–600. https://doi.org/10.1007/978-1-4614-6849-3.

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2017. lmerTest Package: Tests in Linear Mixed Effects Models. J. Stat. Softw. 82, 1–26. https://doi.org/10.18637/jss.v082.i13.

Lei, L., Fithian, W., 2018. AdaPT: an interactive procedure for multiple testing with side information. J. R. Stat. Soc. Ser. B (Stat. Method. 80, 649–679. https://doi.org/10.1111/rssb.12274.

Lima, E., Green, M., Lovatt, F., Davies, P., King, L., Kaler, J., 2020. Use of bootstrapped, regularised regression to identify factors associated with lamb-derived revenue on commercial sheep farms. Prev. Vet. Med. 174, 104851 https://doi.org/10.1016/j.prevetmed.2019.104851.

Lima, E., Hyde, R., Green, M., 2021a. Model selection for inferential models with high dimensional data: synthesis and graphical representation of multiple techniques. Sci. Rep. 11, 412. https://doi.org/10.1038/s41598-020-79317-8.

Liu, J.Y., Zhong, W., Li, R.Z., 2015. A selective overview of feature screening for ultrahigh-dimensional data. Sci. China Math. 58, 2033–2054. https://doi.org/10.1007/s11425-015-5062-9.

Li, Y., Wang, S., Song, P.X.-K., Wang, N., Zhou, L., Zhu, J., 2018. Doubly regularized estimation and selection in linear mixed-effects models for high-dimensional longitudinal data. Stat. its Interface 11, 721–737. https://doi.org/10.4310/SII.2018.v11.n4.a15.

Meinshausen, N., Bühlmann, P., 2010. Stability selection. J. R. Stat. Soc.: Ser. B 72, 417–473. https://doi.org/10.1111/j.1467-9868.2010.00740.x.

Royston, P., Moons, K.G.M., Altman, D.G., Vergouwe, Y., 2009. Prognosis and prognostic research: Developing a prognostic model. BMJ (Online) 338. https://doi.org/10.1136/bmj.b604.

Schelldorfer, J., Meier, L., Bühlmann, P., 2014. GLMMLasso: an algorithm for high-dimensional generalized linear mixed models using ℓ1-penalization. J. Comput. Graph. Stat. 23, 460–477. https://doi.org/10.1080/10618600.2013.773239.

Simas, R., Maestri, F., Normando, D., 2014. Controlling false positive rates in research and its clinical implications. Dent. Press J. Orthod. https://doi.org/10.1590/2176-9451.19.3.024-025.ebo.

Sirimongkolkasem, T., Drikvandi, R., 2019. On regularisation methods for analysis of high dimensional data. Ann. Data Sci. 6, 737–763. https://doi.org/10.1007/s40745-019-00209-4.

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. 58, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Wasserman, L., Roeder, K., 2009. High Dimensional Variable Selection. Ann. Stat. 1, 2178–2201. https://doi.org/10.1038/jid.2014.371.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. 67, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.