Research Article

Statistics in Medicine

Received 2 October 2012, Accepted 26 April 2013 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5855

Graphical tools for model selection in generalized linear models

K. Murray,[a,b*†] S. Heritier[c] and S. Müller[a]

Model selection techniques have existed for many years; however, to date, simple, clear and effective methods of visualising the model building process are sparse. This article describes graphical methods that assist in the selection of models and comparison of many different selection criteria. Specifically, we describe for logistic regression, how to visualize measures of description loss and of model complexity to facilitate the model selection dilemma. We advocate the use of the bootstrap to assess the stability of selected models and to enhance our graphical tools. We demonstrate which variables are important using *variable inclusion plots* and show that these can be invaluable plots for the model building process. We show with two case studies how these proposed tools are useful to learn more about important variables in the data and how these tools can assist the understanding of the model building process. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: model selection curves; Akaike information criterion; graphical methods; Bayesian information criterion; variable selection; model selection; generalized linear models

1. Introduction

Many medical problems involve the collection of data with multiple potential predictor variables. In analysing the data, one usually engages in a process of model building, of which a crucial part is to determine one or more appropriate models. For a general introduction and overview into the topic of model building, we refer to [1]. One of the most commonly used techniques for model selection, which is probably the least advocated by statisticians, is a 'hypothesis test/P-value' stepwise approach, using either forward selection or backward selection or a combination of the two. These approaches have been shown to be inefficient in many situations, and have particular issues such as multiple testing and localization of solutions. For many models, including the vast array of generalized linear models, the information theoretic approach and the use of the log-likelihood to compare models is widespread in general for model selection purposes. For this reason, our article focuses on measuring the descriptive ability of a model via the log-likelihood, but our ideas extend directly to using other loss functions.

To date, in medical research, data analysts have used many different techniques to select models for prediction purposes. In most instances, only one final model is presented, and how such a model is reached is typically not sufficiently described in the statistical methods section of research articles. Reasons for this include space restrictions and a shortage of simple graphical tools that can be shown to explain the reasoning behind the final model. Consequently, future researchers have difficulty obtaining a clear understanding of available model selection techniques in current medical research, what these techniques are really doing, and how to replicate and adapt such techniques.

When using an information theoretic approach to model selection, a somewhat controversial and much debated question is whether to select a model using the AIC [2] or the BIC [3]. The purpose of the analysis drives the model selection. Often, a separation is made between the purposes to describe the data well and to obtain a model that has good predictive qualities. A major difference between AIC and BIC

[a] *School of Mathematics and Statistics, University of Sydney, Carslaw Building (F07), NSW 2006, Australia*
[b] *Centre for Applied Statistics (M019), University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia*
[c] *The George Institute for Global Health, University of Sydney, Sydney, NSW 2050, Australia*
**Correspondence to: K. Murray, School of Mathematics and Statistics, University of Sydney, Carslaw Building (F07), NSW 2006, Australia.*
[†] *E-mail: kevin.murray@uwa.edu.au*

is that AIC attains the minimax rate whereas BIC is consistent (i.e. as n tends to infinity, the probability of selecting the true model tends to 1) and aims to model the dimension of the true model but fails to attain the minimax rate. The concept of a true model is debatable [4], and there are strong arguments in opposition of such assumptions [5]. Put simply, the argument is that all of the predictor variables proposed for a model should have some effect on the response, even if this effect is extremely small. In spite of this, many questions still arise: which variables are important to describe the data well, which components remain when a smaller model is preferred to the AIC model, and what happens if a slightly different penalty is chosen to the one that led to the calculation of the AIC value.

Graphs are powerful tools in statistics. We quote John Tukey who said: 'There is nothing better than a picture for making you think of questions you had forgotten to ask (even mentally)' [6]. When fitting models, it is a standard practice to examine model diagnostics using visual aids (residuals plots, QQ-plots, etc), and we are encouraged to visualise our data throughout the majority of any statistical analysis. In light of this, it seems somewhat perplexing that in carrying out model (or variable) selection, little encouragement is given to employ visualisation approaches. In a general setting, the works of Loftus [7] or more recently with a clinical flavour Krause and O'Connell [8] provide some examples of the benefits of using graphical techniques in data analyses. However, to date, there are relatively few publications that show graphical tools as aids in model selection. The best known is the Mallow's Cp plot [9], which has been proposed for linear regression situations, and some further variants exist [10, 11]. Recently, Müller and Welsh [12] introduced for a $p < n$ setting a variable detection plot, which is similar to the stability paths in Meinshausen and Bühlmann [13] for the $p >> n$ context, where n denotes the sample size and p the total number of variables that are subject to selection. Both graphs are based on slightly different resampling procedures, and among others aim to facilitate the choice of a tuning parameter.

In this article, we take a liberal view of model selection in a medical context, and instead of trying to argue one way or another on issues like AIC versus BIC, or on the advantages of techniques such as model averaging, we aim to provide simple graphical tools for the visualisation of different model selection criteria that facilitate the descriptive process of the data and assist with the model selection problem. It should be noted that we are not advocating the use of our plots for one specific purpose; rather, we aim to demonstrate how they can aid in many different aspects of model building.

The outline of the paper is as follows: In Section 2, we describe the terminology and methodology and introduce our graphical concepts by a first case study with simulated data. Additionally, we show how the methods assist in the model selection process. In Section 3, we present the results of two further case studies, which have more variables and more features than the simulated data. The first is from a retrospective cross-sectional palliative care study, and the second from an ongoing health study analysing the effect of smoking in patients with Crohns disease. We also show additional uses of our model selection plots, in particular, in the context of highly correlated regressors and in the presence of outliers. We conclude with a discussion and comments in Section 4. Further material including R code for reproducing all results and figures shown in this article can be found at http://school.maths.uwa.edu.au/~kev.

2. Terminology and graphical methods

2.1. Generalised linear model framework

Assume that n independent observations $y = (y_1, \ldots, y_n)^T$ and an $n \times p$ design matrix X is available, whose columns are indexed by $1, \ldots, p$. Let α denote any subset of p_α distinct elements from $1, \ldots, p$. Let X_α be the corresponding $n \times p_\alpha$ design matrix and $x_{\alpha_i}^T$ denote the ith row of X_α.

Then, a generalised linear regression model (GLM) α for the relationship between the response y and the design matrix X_α is specified by

$$\mathrm{E}(y_i) = h(\beta_{0,\alpha} + \eta_i), \text{ and } \mathrm{Var}(y_i) = \sigma^2 v^2(\eta_i) \text{ with } \eta_i = x_{\alpha_i}^T \beta_\alpha, \quad i = 1, \ldots, n \tag{1}$$

where $\left(\beta_{0,\alpha}, \beta_\alpha^T\right)^T$ is an unknown $(p_\alpha + 1)$-vector of regression parameters and σ is an unknown scale parameter. Here, h is the inverse of the usual link function, and, for simplicity, we have absorbed h into the variance function v. Both h and v are assumed known. In this article, we focus specifically on the logistic regression model, which has binomial response and logit link function; however, we can extend our methods to any GLM.

2.2. *Measuring description loss and complexity in generalised linear regression model*

The purpose of model selection is to choose one or more models α, from all candidate models $\mathcal{A}$, with specified desirable properties. Many model selection procedures involve the minimisation of an expression, which can take the simple form:

$$\textit{Description Loss} + \lambda \times \textit{Model Complexity} \tag{2}$$

Most prominently, we can measure the 'description loss' of a model by $-2\times$ log-likelihood. Here, we use the term 'description loss', which we regard as any function L_n that measures how well a model fits the data. Model complexity, in its simplest form, is the number of independent regression model parameters. We refer to λ as the penalty multiplier, which often drives the properties of the selection criteria [12]. We will focus on procedures that use the information theoretic choice of description loss, that is, minus twice the log likelihood. The AIC is of form (2) and has penalty multiplier $\lambda = 2$; similarly, the BIC has $\lambda = \log(n)$, or more generally, the generalised information criterion (GIC) [14] has penalty multiplier $\lambda \in \mathbb{R}$.

2.3. *Case study I – simulated logistic regression example*

For ease of presentation, we illustrate our techniques through logistic regression examples. However, these are easily extendable to any GLM. Initially, we describe a simulated logistic regression example in which $n = 250$ responses $\left(y_i's\right)$ are modelled with seven potential predictors $x_1 - x_7$. We simulate the predictors themselves from a multivariate normal distribution, with correlations of 0.7 introduced between predictors x_1 and x_2 and between x_5 and x_6. The data generating model samples responses from Bernoulli random variables with probability of success π_i, such that

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots \beta_7 x_{i7}, i = 1, \ldots, n \tag{3}$$

and the parameter values for the β's (excluding the intercept) are set to be $(2.1, 0, 1.5, 0, 0, 0.9, 1.1)$.

The full model including all seven regressors and intercept, that is, $\alpha_f = \{1, 2, 3, 4, 5, 6, 7, 8\}$, has $-2\times\text{LogLik}(\alpha_f) = 135.9$, whereas the model including only an intercept has $-2\times\text{LogLik}(\{1\}) = 207.7$. Because there are seven explanatory variables, there are a total of $2^7 = 128$ possible logistic GLM's, that is, $\mathcal{A} = \{\{1\}, \ldots, \alpha_f\}$. We visualise the description loss and dimension for all $\alpha \in \mathcal{A}$ next.

2.4. *Case study I – visualising L_n and p_α*

2.4.1. Basic scatter plot. The left panel in Figure 1 shows for all possible $\alpha \in \mathcal{A}$ a simple scatter plot of the number of regression parameters (horizontal axis) and $L_n = -2 \times \text{LogLik}$. As the dimensionality

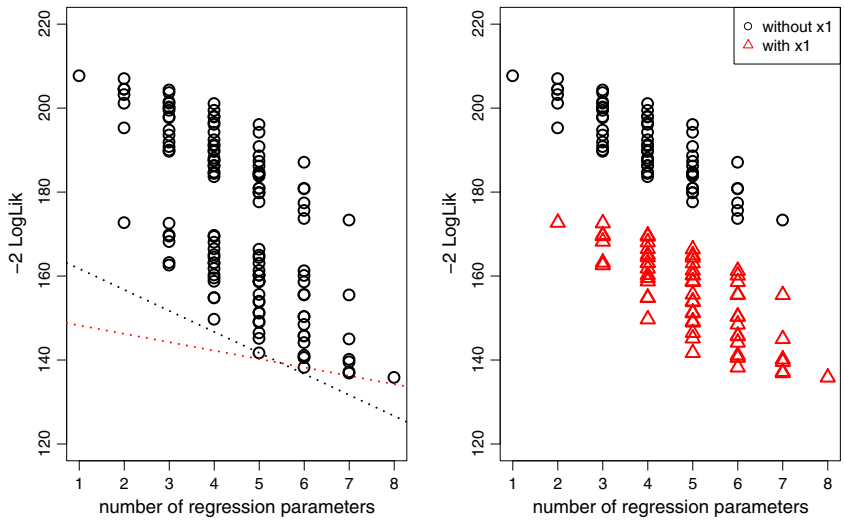

Figure 1. Left panel – L_n versus dimension. Right panel – L_n versus dimension by model type, that is, presence or absence of a highly influential variable (triangle = included, circle = not included).

increases, the general pattern is for the description loss (L_n) to decrease, that is, become better. We can use these simple plots in many different ways. First, we observe for any given model dimensionality that we have a model of rank 1. This is the model that provides minimum L_n value among those models having the same dimension (number of explanatory variables). For example, in Figure 1, with two regression parameters, there is a model with $L_n = 173$, which is the lowest of all seven models at this dimension and hence is defined as the rank 1 model for dimension 2. We have added also to this plot two dashed lines that represent a line of slope -2, which intersects the AIC minimum solution (dimension 6), and a line of slope $-log(n)$, which intersects the BIC minimum solution (dimension 5). From this, one can determine that no other solutions using either a minimum AIC or BIC as model selection criteria exist for this particular data. We consider this in more detail in 2.6.

2.4.2. Enriching the scatter plot. The right panel of Figure 1 is an enriched scatter plot. In this instance, we have labelled models that include the variable x_1 (red triangle) differently to those that do not (black circle). In this example, we observe separation of L_n values over all dimensions, in that all models that include this important variable (x_1) give a much lower L_n than all models that do not include this variable. We can generalise this approach to examine all variables and those interactions that are of interest.

By visual inspection of such graphs, we can determine that this particular variable is 'a must' for the inclusion in the final model (or subset of models), without having to make a choice on the size of the penalty multiplier, or delving into arguments about AIC versus BIC. Mathematically, it is clear that complete separation can only occur when the grouping of models is according to variables present in the best model of a given dimension. For example, by construction of the maximum likelihood approach, adding any variable $k \in 1, \ldots, p$ to a given model $\alpha \in \mathcal{A}$ satisfies $-2 \times \text{LogLik}_{\alpha^*} \leqslant -2 \times \text{LogLik}_{\alpha}$, where $\alpha^* = \alpha \cup \{k\}$. Then, we can use these plots to list and see all models of a given dimension that represent a local minimum in terms of L_n.

2.5. *Case study I – assessing model stability through bootstrapping*

A slight modification to these scatter plots is choosing the plot symbol size to be proportional to a measure of model stability, a concept that was independently introduced by Meinshausen and Bühlmann [13] and Müller and Welsh [12] for different linear regression situations. We can estimate model stability by bootstrapping, which was successfully used for model selection in GLMs by Müller and Welsh [15]. There exist different bootstrapping techniques including the residual bootstrap, the paired bootstrap and the weighted bootstrap as described by Barbe and Bertail [16] and used recently by Minnier *et al.* [17]. We demonstrate how our model selection plots can be modified to give an idea of model stability at each dimension using the weighted bootstrap, which is simple to implement because it is based on repeatedly re-weighting observations. On the other hand, the residual bootstrap can be problematic for generalised linear models because adding bootstrapped residuals to fitted values can result in unobservable responses, whereas the nonparametric bootstrap is known to lead to potential separation problems for logistic regression models.

To obtain our measure of model stability, let w_i be the weight for observation i in our original data, and sample the w_i's from an exponential distribution with expectation 1 as described in Janssen and Pauls [18] and Minnier *et al.* [17]. We refit our original models using these weights and resample B times, where B is the number of bootstrap samples that we set to 1000. For each bootstrap sample, we rank models within each dimension and count the number of times each model is of rank 1. We plot L_n versus dimension with symbol size proportional to the number of times model α has rank 1. This incorporates information on model stability into the scatter plot. Consequently, those models selected a large proportion of times in the weighted bootstrapping have large symbols, and those selected very few times have small symbols. We also highlight the underlying model we simulated from (the data generating model) with the full red circle. Figure 2 shows an example, and we note that the size of the plot symbol for dimension 1 (intercept only) and dimension 8 (α_f) is at its largest because there is only one possible model choice within these dimensions. However, at dimensions 2 through to 7, we can assess the stability of models with lowest description loss visually and determine how reliable any one chosen model may be. For example, at dimension 2, there is one model that stands out as the single best model, whereas at dimensions 3 and 7, there are at least two models that have very similar size symbols at the optimum end of the scale. For information, we include later the figure variables included in the rank 1

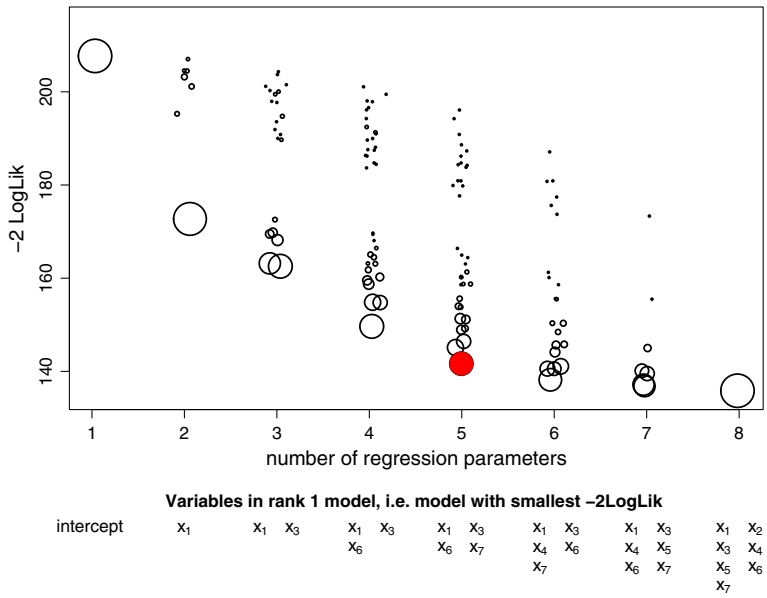

Figure 2. Assessing model stability using weighted bootstrapped probabilities at each dimension with probability proportional to symbol area. Note: Number of regression parameters is jittered to avoid symbols completely overlapping. The data generating model is shown in red. The panel below shows the variables included in the rank 1 model for each dimension.

model for each dimension. For example, at dimension 4, the model with lowest L_n includes the regressor variables x_1, x_3 and x_6.

2.6. *Case study I – the maximum enveloping lower convex curve*

We define and describe the *maximum enveloping lower convex curve* (MELCC), which is developed by the following algorithm. Starting with dimension 1 (or an equivalent appropriate starting point), draw a straight line from the minimum L_n at this point to either the minimum L_n at the next dimension or the minimum L_n at the next dimension that envelops all other minimum L_n/dimension combinations inbetween. Repeat this process until the highest dimension to obtain the curve. Figure 3 depicts this

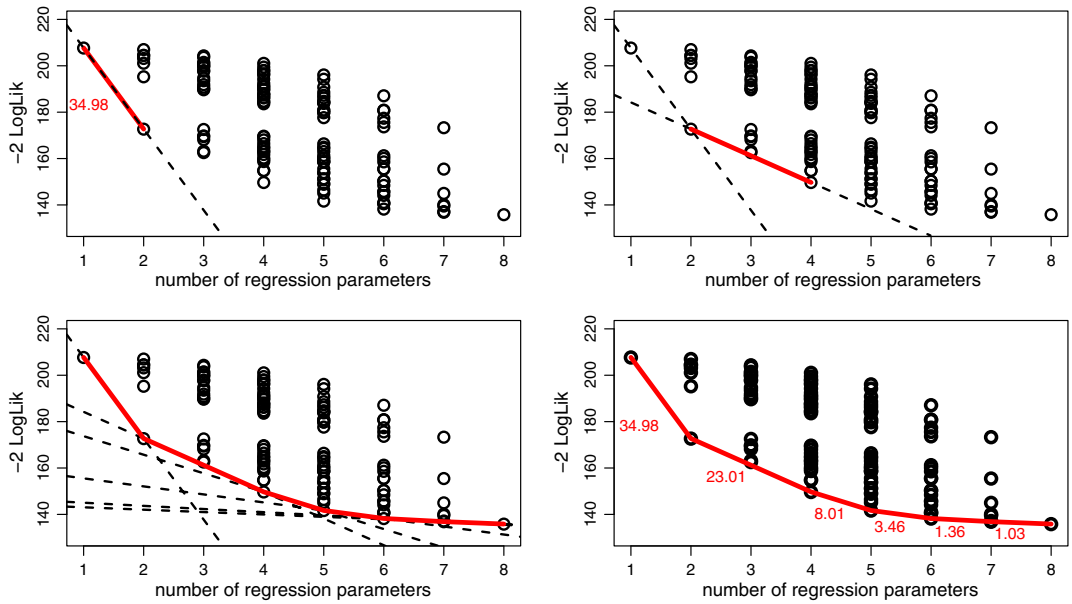

Figure 3. Constructing the maximum enveloping lower convex curve and the decrease in L_n for models lying on this curve.

process, starting in the top left panel where we join the minimum L_n at dimension 1, to the minimum L_n at dimension 2. Continuing, we join the minimum L_n's at subsequent dimensions until the curve is complete. We show these steps with the thick red line indicating the part of the MELCC added. The final curve is shown in the bottom left and right panel of Figure 3 (with all lines and the partial parts of the curve displayed in the bottom left panel). Note, in some instances, the minimum L_n at a given dimension will be omitted from the curve if it is enveloped by a line joining minimum L_n's at dimensions above and below. This is seen in the top right panel of Figure 3 where the minimum L_n at dimension 3 is fully enveloped by the minimum L_n's at dimensions 2 and 4. Intuitively, these model dimensions that do not have an L_n lying on the MELCC indicate that a model of that dimension would never be selected according to a GIC model selection strategy, irrespective of the penalty. The reasoning is simple: We note as follows: (i) that the MELCC has at most p different slopes, a property that follows from the convexity, and (ii) that the GIC solution is found by having a straight line passing through a point such that all other points are not below that line. From (i) and (ii), it follows that only points on the MELCC are GIC solutions.

Visual inspection of this curve shows the lowest L_n for all dimensions effortlessly and identifies the model dimensions that do not appear on the MELCC. For completeness, Figure 3 also shows the finalised MELCC with the value of the decrease for each move along the curve. The first decrease is large (34.98) and subsequent decreases much lower. These decreases determine whether a model with higher dimension would be selected for a given penalty multiplier. For example, with AIC, the penalty multiplier is $\lambda = 2$. Hence, we would select models on the curve that have a decrease in L_n of 2 or greater for each unit increase in dimension. Alternatively, we could use BIC, that is, $\lambda = \log(n)$, which in this example is 5.52, and to accept a larger model, the decrease in L_n would have to be in excess of 5.52 units per unit increase in dimension. Using AIC, we arrive at the solution of the point at minimum L_n on the curve with dimension 6, and in this instance, using BIC gives a solution at dimension 5.

2.7. *Case study I – the variable inclusion plot*

In addition to being able to simply visualise the loss and dimensionality, we show here how a 'variable inclusion plot' (VIP) can provide insightful information. Müller and Welsh [12] introduced the VIP for linear regression models. Here, we introduce the VIP for GLMs using the weighted bootstrap. The VIP visualises 'inclusion probabilities' as a function of the penalty multiplier λ. For each variable x_j subject to selection, the proportion of times this variable is retained in the B final selected bootstrapped model is plotted for a range of λ values, for example, $\lambda \in [0, 2\log(n)]$. More specifically, we calculate for bootstrap sample $b = 1, \ldots, B$ and for each considered λ multiplier value that model $\hat{\alpha}_\lambda^{(b)} \in \mathcal{A}$ which has smallest $\text{GIC}(\alpha; \lambda) = -2 \times \text{LogLik}(\alpha) + \lambda p_\alpha$ value. Thus, the inclusion probability for variable x_j is estimated by

$$\frac{1}{B} \sum_{b=1}^{B} 1\left\{ j \in \hat{\alpha}_\lambda^b \right\}$$

where the indicator function $1\left\{ j \in \hat{\alpha}_\lambda^{(b)} \right\}$ is one if variable x_j is in the final model and zero otherwise.

Figure 4 shows the VIP for all seven variables in our simulated data, and to help identifying what happens when using the AIC and BIC methods, we show vertical lines at the AIC penalty multiplier value $\lambda = 2$ and for BIC at $\lambda = \log(n)$, respectively.

From this chart, we can immediately see two things: First, regardless of the value of the penalty multiplier, variable x_1 is always going to be included in a final model with a bootstrapped probability of 1 for penalty parameter values up to 10. Similarly, for relatively large λ values, variable x_3 seems to maintain a high inclusion probability. Beyond this, we can examine each curve individually. Variable x_6 appears to have a reasonably high inclusion probability regardless of the penalty as does variable x_7, but the latter tails off more steeply as the penalty is increased. On the other hand, because we know the data generating model, redundant variable x_4 has a relatively high chance of being included with the AIC, only a fair chance with BIC and beyond that is not likely to be selected. Redundant variables x_2 and x_5 both have a fair chance of inclusion with AIC, but beyond that, the bootstrapped probabilities diminish, indicating that a higher penalty would result in neither of these variables being included. One of the big advantages of the VIP is to visualise differences between AIC and BIC and any other GIC as a model selector, which reveals much more information than when 'blindly' choosing a fixed λ value.

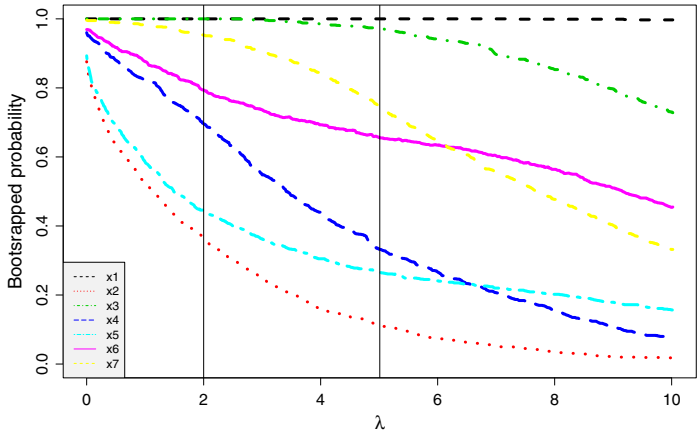

Figure 4. Variable inclusion plot for simulated data.

3. Results and further applications

We further investigate the ideas described in the previous section through two further case studies and show additional ways in which combinations of our graphical displays can be used.

3.1. Case study II – cross-sectional palliative care study

We analyse data from a retrospective cross-sectional study on how home-based palliative care services affect hospitalisations of people in a cohort identified systematically through death registrations as described in [19] and [20]. The cohort comprised 1071 people who died in Western Australia between 1 August 2005 and 30 June 2006, had an informal primary carer at the time of death, did not reside in a residential aged care facility and died of one of 10 conditions amenable to palliative care.

Using a logistic GLM with logit link function and the terminology described previously, we model the dichotomous response variable 'did' or 'did not receive community-based palliative care'. It is known that cancer patients are much more likely to receive such care because of the nature of their condition, so we considered the dichotomous variable underlying cause of death (cancer or noncancer) as a vitally important explanatory variable; we label this x_1 for simplicity. Additionally, we considered variables age at death (x_2) and age squared (x_3), gender (x_4), number of days spent in hospital in final year of life (x_5), number of emergency department visits in final year of life (x_6) and usual place of residence (metropolitan or rural, based on postcode) (x_7). Therefore, we include $p = 7$ potential parameters in the full model, resulting in $\#\mathcal{A} = 2^p = 128$ possible submodels that include an intercept. However, if the marginality principle is obeyed, we only include the age squared term in the model together with the linear term. This reduces the number of models to $\#\mathcal{A} = 96$. In Table I, we show parameter estimates and standard errors for the full model, the model selected using both AIC and BIC as selection criteria, along

Table I. Full model, AIC and BIC best model, forward/backward best model, parameter estimates and standard errors.

	Full model		AIC best		BIC best		Forward selection		Backward selection	
Variable	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	−2.127	1.514	−2.408	1.496	−1.392	0.153	−0.220	0.507	−1.392	0.153
x_1	2.118	0.185	2.086	0.182	2.250	0.176	2.105	0.182	2.250	0.176
x_2	0.057	0.045	0.057	0.045			−0.012	0.006		
x_3	−0.001	0.000	−0.001	0.000						
x_4	−0.087	0.150								
x_5	−0.084	0.069								
x_6	−0.193	0.115	−0.243	0.108			−0.244	0.108		
x_7	−1.918	0.185	−1.924	0.185	−1.884	0.182	−1.902	0.184	−1.884	0.182

SE, standard errors.

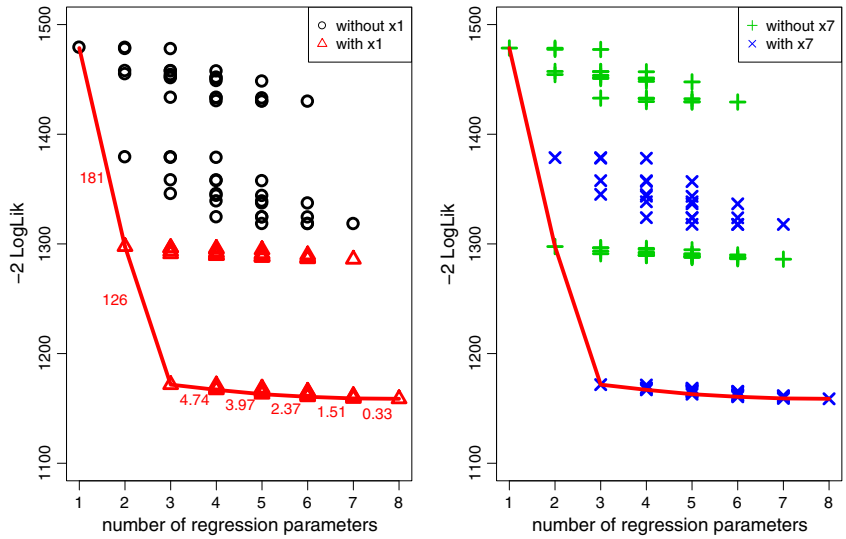

Figure 5. Basic and enriched scatter plot with MELCC and reduction in L_n for models on the MELCC for case study II.

with a p-value based forward and backward hypothesis testing approach to select a final model. Note the substantial differences in the models selected between these methods, even with our relatively small p. For example, the AIC model includes five explanatory variables, whereas both BIC and backward selection only include x_1 and x_7.

Figure 5 shows the basic scatter plot. We show the impact of including or excluding the two most important main effects by enriching this plot by choosing different symbols for models with and without x_1 and with and without x_7 as well as drawing the MELCC. We observe the following: First, in the left hand plot, we observe a stark separation of the model L_n for all dimensions, when the variable of interest (in this case, cause of death) is either included or not included in the model. Such separation does not exist for any of the other main effects (not shown here), and we conclude that cause of death is clearly the single most important variable irrespective of model dimension or concepts such as p-values. Second, the right hand panel of Figure 5 shows the impact of including a second important variable (x_7) in the model, subsequent to fixing either level of cause of death in the model. We observe that the model with both Cause of Death and Usual Place of Residence gives us consistently lower L_n regardless of dimension. Again, such a pattern is not apparent for other variables when combined with Cause of Death. We conclude that these two variables are important as a pair. We could continue in this fashion, considering other effects including interactions and higher order terms.

Figure 5 also shows the MELCC for this data. We note the dramatically steep decline as we move from the intercept only model, to the models at dimensions 2 and 3, further highlighting the importance of the two variables x_1 and x_7. From the left hand plot, we see the reduction from dimension 3 to 4 is 4.74, which is less than $\log(n) = 6.98$. Here, the BIC solution has dimension 3. Furthermore, when continuing down the curve until a reduction of less than 2 is observed, we obtain the AIC model. We attain this at dimension 6 as also seen in Figure 5. For this case study, the change in the L_n values as we move from dimension to dimension is monotone decreasing, which is not the case in general.

The model stability plot shown in Figure 6 indicates again the stability at each of dimension 2 and dimension 3 with large symbols indicating one model is selected the majority of the time at these dimensions. However, in the right hand panel of Figure 6, we show the zoomed in version of this plot focussing on dimensions 4 through 7 and identify several models that have similar size symbol to that of the optimal model at each of these dimensions.

The VIP shown in Figure 7 gives an indication of which variables would and should be included, dependent on the value of the penalty parameter. First, we note again the importance of variables Cause of Death (x_1) and Usual Place of Residence (x_7), which are always included regardless of the penalty multiplier used, as shown by the two overlaying horizontal lines with probability of 1 for all λ values in the range given. Second, we see that with lower values of λ, a large proportion of the time variables x_2 and x_6 would be included in the 'best model', and to some extent, variables x_3 and x_5 could also be considered. At higher penalty values, it would be difficult to argue for the inclusion of anything

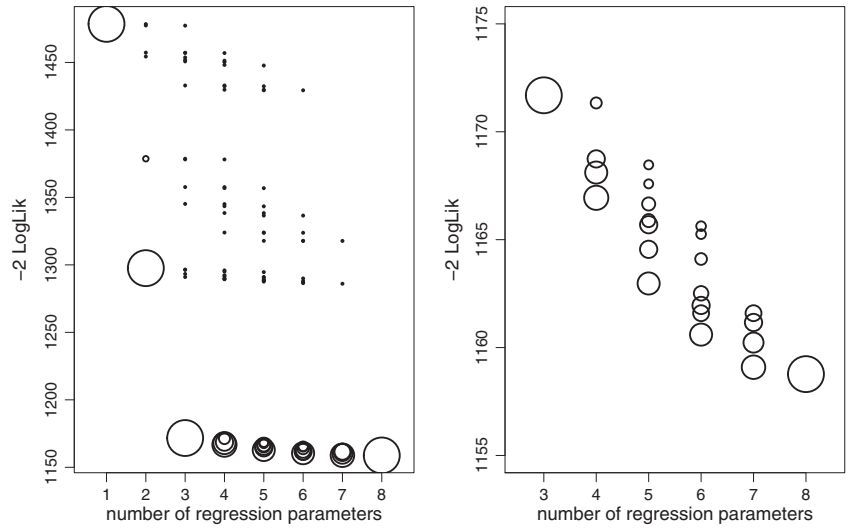

Figure 6. Model stability for case study II data with weighted bootstrapped proportions conditional on dimension proportional to symbol area. Full plot (left panel) and zoomed in plot (right panel).

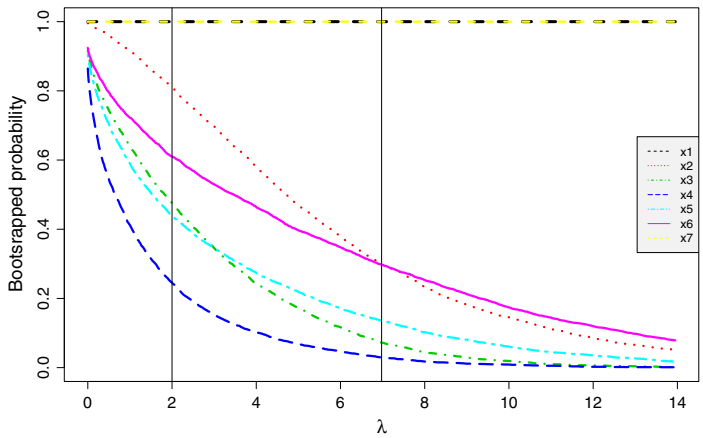

Figure 7. Variable inclusion plot for case study II.

more than the two main variables, which is reflected in the BIC model for this data. The VIP clearly shows that variables x_2 and x_6 are important to describe the data but are not informative to predict the response variable.

3.2. Case study III – smoking in patients with Crohns disease

The final case study consists of data from an ongoing study described in Lawrance *et al*. (in preparation) that investigates the impact of smoking in patients with Crohns disease, specifically focussing on those patients who have already had one surgery relating to their condition since their initial diagnosis. We examine the dichotomous response of having a repeat surgery (yes/no) and the impact of potential explanatory variables: age group (young, middle aged and old), sex, whether the patient had ever been a smoker (yes/no), the amount of time they were followed up (measured in years post their first surgery), the location of the condition (broken down into one of three locations) and an additional indicator location variable depicting severity of location. In addition, whether the patient was using steroids (yes/no), had a perianal condition (yes/no), was classified as immunosuppressed (yes/no) or had received Anti TNF α agents (yes/no) were included as potential explanatory variables. In total, we examined 10 variables (seven binary indicator variables, one continuous variable and two factors each with three levels). Changing the factors to dummy variables (two additional parameters each), we arrive at a total of 12

Table II. Best models by dimension showing decrease in L_n for Case Study III.

p_α+1	Model	$-2 \times$ LogLik	Decrease	
1	intercept only	711.73	NA	
2	$+ x_5$	634.26	77.47	
3	$+ x_5 + x_{11}$	614.35	19.91	
4	$+ x_5 + x_7 + x_{11}$	601.34	13.01	BIC
5	$+ x_5 + x_7 + x_{10} + x_{11}$	595.65	5.69	bw and fw
6	$+ x_5 + x_6 + x_7 + x_{10} + x_{11}$	593.35	2.30	AIC
7	$+ x_5 + x_6 + x_7 + x_{10} + x_{11} + x_{12}$	592.28	1.07	
8	$+ x_5 + x_6 + x_7 + x_8 + x_{10} + x_{11} + x_{12}$	591.48	0.80	
9	$+ x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{11} + x_{12}$	591.09	0.39	
10	$+ x_1 + x_2 + x_5 + x_6 + x_7 + x_8 + x_{10} + x_{11} + x_{12}$	590.61	0.48	
11	$+ x_1 + x_2 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{11} + x_{12}$	590.25	0.36	
12	$+ x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_{10} + x_{11} + x_{12}$	590.32	−0.07	
13	$+ x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{11} + x_{12}$	589.90	0.42	

possible explanatory variables. We retained the factors as a whole in the model; hence, the two parameters associated with each of the two factors would be regarded as a grouped variable, with levels that are either simultaneously included or excluded.

We use logistic GLMs with logit link function to model the occurrence of a second surgery (yes/no) response and demonstrate further our ideas. With $p = 12$ potential variables in the final model, we have $\#\mathcal{A} = 2^{12} = 4096$ possible submodels that include an intercept. Taking into consideration the two factors and ensuring that the corresponding dummy variables are either simultaneously excluded or included, the number of potential models reduces to $\#\mathcal{A} = 2^{10} = 1024$.

In Table II, we report the model selection for this data by showing for each dimension what model is optimal (has smallest L_n value). Note that both backward and forward variable selection using a p-value approach would give us a model of dimension 5, the minimum AIC model is a model of dimension 6 and the minimum BIC model is a model of dimension 4. Again, this confirms that there is often lack of agreement in model selection depending on what the objective is and what methodology is used.

We show the basic scatter plot in the left panel of Figure 8. There is clear separation between those models that include the variable 'time post first surgery' and those that do not, indicating this to be an important variable. Intuitively, this makes sense, given patients are more likely to have recorded a second surgery if their follow-up time is longer.

The middle panel of Figure 8 shows a zoomed version of the scatter plot and the MELCC for this data. As before, we can make several 'simple' observations from this plot: For example, dimensions 9 and 12 do not lie on the MELCC. Hence, regardless of the choice of penalty multiplier, one can rule these out of the subset of *optimal* models, which is regardless of the choice of λ, no GIC method will select a final model with 8 or 11 explanatory variables.

As expected, examination of model stability indicates that the model including only the variable time is not only the best model at dimension 2, but extremely stable. However, as seen in the right hand panel of Figure 8, beyond dimension 3 through to dimension 12, there are many models that are identified almost as frequently as the optimal models lying on the MELCC. This should prompt consideration of models that lie within close proximity of the optimal model at these dimensions.

The VIP shown in Figure 9 gives an indication of which variables would be included, dependent on the value of the penalty parameter. Notice that the variables x_1 and x_2 have trajectories, which are exactly the same. This is true of variables x_3 and x_4 and is due to the criteria we set for these variables, which represent the dummies for age and location, respectively. The horizontal line at probability 1 corresponds to the variable time (x_5) and again demonstrates the importance of this variable, which would be included in any final model. We note also the high line corresponding to x_7, which for both AIC and BIC criteria, has probability in excess of 0.8. Beyond this, one could argue for the inclusion of variables x_{10} and x_{11} using criteria with penalties at or below the BIC penalty. However, only lower penalties would include x_6, and it would be difficult to argue for the inclusion of other variables regardless of penalty parameter.

3.3. Additional uses for model selection plots

In this section, we consider two additional challenges that are (potentially) problematic in any model selection problem. First, we examine how our plots perform in the presence of (highly) correlated

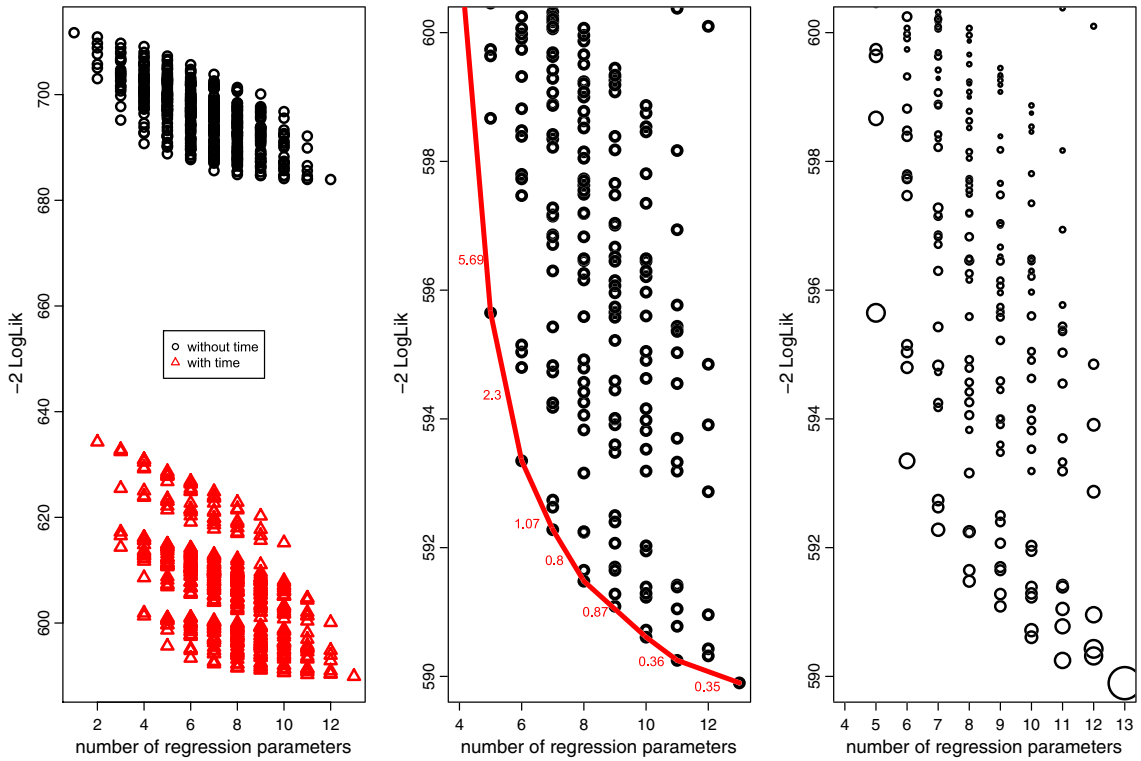

Figure 8. Case study III – L_n versus dimension: enriched scatter plot (left panel), zoomed in MELCC with reduction in L_n (middle panel) and zoomed in model stability plot (right panel).

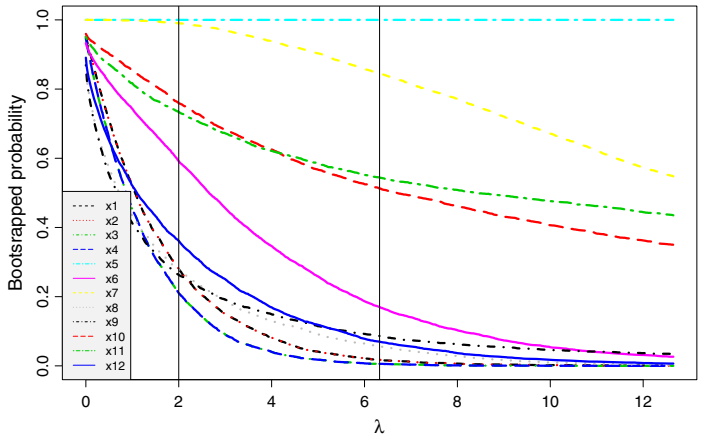

Figure 9. Variable inclusion plot for case study III.

predictors. Second, we consider extensions of our plots when outliers are present in our data, and propose some robust extensions to our methodology.

3.3.1. Correlated predictors. We investigate the effect of correlated predictors through a simple example. Let us again consider a similar logistic regression example to that described in Section 2.3. For this example, we chose $n = 150$ and have four potential predictors with the parameter values for the β's (excluding the intercept) set to be $(1.5, 0, 0, 2.5)$. We set the correlation between x_1 and x_2 to be 0.95, 0.85, 0.7 or 0.5 in four separate simulations.

Figure 10 shows the impact of the correlated variables on the model stability plots. Superimposed are the results from the simulation with correlation of 0.5 (red circles) and the correlation 0.95 (black circles). By using the smaller correlation (of 0.5), the distinction between competing models at dimension 3 (the underlying true dimension) is more pronounced, with only one prominent large red circle.

However, even though the results are not as clear with the high correlation, there is still a subtle distinction at dimension 3 between the model that includes x_1 and not x_2 (largest black circle) and the model that includes x_2 and not x_1 (smaller black circle). This indicates that even in the presence of a high correlation, we are still able in this instance to obtain some reasonable results, implying that our graphs are useful in this context.

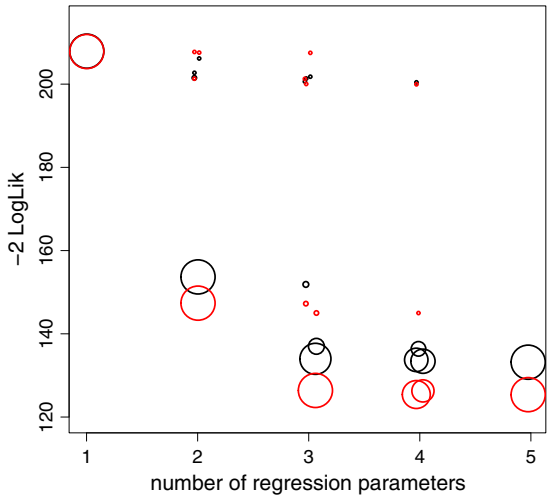

Figure 10. Model stability plot for simulated correlated data. Correlations between x_1 and x_2 are 0.95 (black circles) and 0.5 (red circles).

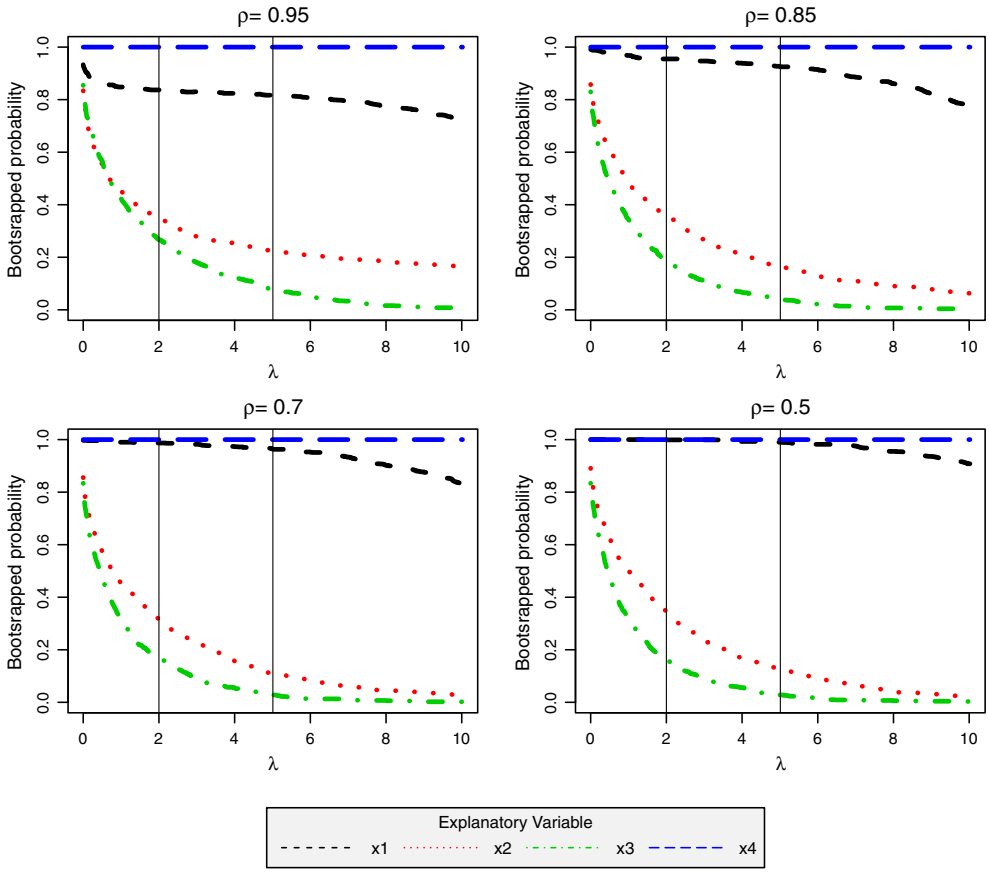

Figure 11. Variable inclusion plot for simulated correlated data. Correlations between x_1 and x_2 are 0.95 and 0.85, (top row) and 0.7 and 0.5 (bottom row).

Figure 11, which shows the VIP for the four different correlations, is more informative. Clearly, the bootstrapped probability of selection of the true variable x_1 from which the data was simulated is much higher than x_2 for a large range of λ, even for high correlation values like 0.95. As expected, the trade-off between selecting x_1 and x_2 becomes more apparent when the correlation becomes very close to one.

3.3.2. Outliers in data. As with any applications in statistics, outliers can cause problems. Given our methods depend on the likelihood as a loss function, our plots are influenced by one or more outliers in the data whenever the likelihood is influenced as well. A simple solution to this is preprocessing the data and exclusion of outliers before estimation and selection. However, in the robustness literature, deletion of outliers is often not the preferred way of dealing with this challenge. Therefore, we prefer to extend our ideas by replacing the likelihood by robust alternatives of the loss function. These, as well as alternative choices of the penalty term, were recently reviewed in [12]. In particular, one could consider log quasi-likelihoods, L1 and other loss functions for the robust estimation and selection of parameters [14, 15, 21–23]. Generally speaking, as long as one can quantify the loss function and penalty in some sensible and robust fashion, then we can produce slightly modified versions of our plots for robust model selection. A step in this direction was suggested for GLMs in [24, p 159].

3.3.3. Other uses for model selection plots. We can use our model selection plots, or modifications thereof, in many additional ways to those described in this article. To date, we have examined the logistic regression model; however, we can easily extend these techniques to any generalised linear model or even all models in which an information theoretic approach can be taken for model selection. Adapting the model stability plots to show symbol size proportional to bootstrap probabilities across all dimensions, rather than dimension specific, is one alternative to what we have proposed as is the inclusion of plots to focus on models included in model averaging. Given the simplicity of the graphs and the ease at which they can be expanded upon, they can be easily applied and/or adapted with the sample code as made available on http://school.maths.uwa.edu.au/~kev.

4. Discussion

We have shown, in the context of information criteria, that by using simple graphical techniques, there are many ways to assist in the process of model selection and model building. The presented charts visualise how the choice of penalty multiplier can impact on the models selected and how volatile selected models are. The techniques show for each possible penalty multiplier the corresponding minimum description loss, separately for each possible model dimension. This demonstrates that any model selection procedure, having the simple form of Description Loss + $\lambda\times$ Model Complexity can be visualised. A particular strength of the graphs are that they help to address the questions of which variables to include, model stability and potentially model averaging. We propose that this prevails regardless of whether the aim is inference or prediction, because varying the penalty multiplier allows examination of both.

Our techniques imply that L_n, a measure of description loss, can be calculated for all possible models. In situations where the number of variables is between 4 and 20, that is, the maximum number of models to be considered is between 16 and about 1 000 000, our techniques will provide invaluable information. With only a few models, hypothesis testing seems more appropriate than automated all subset procedures. On the other hand, with a growing number of variables, the number of models becomes eventually infeasible to fully evaluate. In situations where p is too large, say $p > 20$, even with huge graphical displays, it will become increasingly difficult to ascertain exactly what advantage can be gained when visualising all possible models. However, in these situations, our techniques are still useful when the number of models is substantially reduced by preprocessing of the data and initial screening of models. How to best preprocess goes beyond the scope of this article. A simple and fast way for logistic regression models would be to use the logistic LASSO [25] repeatedly (through bootstrapping) and to focus only on those models that appear on one of the repeated LASSO paths (see Müller *et al.* [26, Section 4] for a more detailed explanation). Investigating interactions is more challenging, and future work is needed. The group LASSO [27] is one possible way to investigate interactions.

Our aim was not to suggest that using the shown techniques is the only way to go about model selection in the future. Rather, we suggest that our charts together with their nice geometric properties assist with model selection and help the understanding of what is actually achieved when using some of the well-documented techniques, such as AIC or BIC. In this respect, we regard our diagnostic tools as very helpful, especially considering it is still widespread to use model selection techniques as 'black boxes',

or at best a recipe that has to be followed despite the consequences. Graphically displaying loss functions will greatly assist in understanding model stability, AIC, BIC and model selection.

Acknowledgements

Samuel Müller was supported by a grant from the Australian Research Council (DP110101998).

References

1. Miller A. *Subset Selection in Regression*, 2nd edn, Monographs on Statistics and Applied Probability, Vol. 95. Chapman & Hall/CRC, Boca Raton, FL, 2002.
2. Akaike H. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium of Information Theory*, Petrov BN, Csáki F (eds). Akadémiai Kiadó: Budapest, 1973; 267–281.
3. Schwarz G. Estimating the dimension of a model. *Annals of Statistics* 1978; **6**(2):461–464.
4. Shao J. An asymptotic theory for linear model selection. *Statistica Sinica* 1997; **7**(2):221–264.
5. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference – A Practical Information-Theoretic Approach*, 2nd edn. Springer-Verlag: New York, 2002.
6. Tukey JW, Tukey PA. Computer graphics and explaoratory data analysis: An introduction. *In Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics'*, National Computer Graphics Association, Fairfax, VA., 1985; **85**(3): 773–785.
7. Loftus GR. A picture is worth a thousand p values: on the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods & Instrumentation* 1983; **25**(2):250–256.
8. Krause A, O'Connell M. *A Picture is Worth a Thousand Tables. Graphics in Life Sciences*. Springer: London, 2012.
9. Mallows CL. Some comments on Cp. *Technometrics* 1973; **8**:661–675.
10. Spjotvoll E. Alternatives to plotting Cp in multiple regression. *Biometrika* 1977; **64**(1):1–8.
11. Siniksaran E. A geometric interpretation of Mallows' C_p statistic and an alternative plot in variable selection. *Computational Statistics & Data Analysis* 2008; **52**(7):3459–3467.
12. Müller S, Welsh AH. On model selection curves. *International Statistical Review* 2010; **78**(2):240–256.
13. Meinshausen N, Bühlmann P. Stability selection. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 2010; **72**(4):417–473.
14. Konishi S, Kitagawa G. Generalised information criteria in model selection. *Biometrika* 1996; **83**(4):875–890.
15. Müller S, Welsh AH. Robust model selection in generalized linear models. *Statistica Sinica* 2009; **19**(3):1155–1170.
16. Barbe P, Bertail P. *The Weighted Bootstrap*, Lecture Notes in Statistics, Vol. 98. Springer-Verlag: New York, 1995.
17. Minnier J, Tian L, Cai T. A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association* 2011; **106**(496):1371–1382.
18. Janssen A, Pauls T. How do bootstrap and permutation tests work? *Annals of Statistics* 2003; **31**(3):768–806.
19. Rosenwax LK, McNamara BA, Murray K, McCabe RJ, Aoun SM, Currow DC. Hospital and emergency department use in the last year of life: a baseline for future modifications to end-of-life care. *Medical Journal of Australia* 2011; **194**(11):570–573.
20. McNamara BA, Rosenwax LK. Which carers of family members at the end of life need more support from health services and why? *Social Science & Medicine* 2010; **70**(7):1035–1041.
21. Ronchetti E, Staudte RG. A robust version of Mallows' C_P. *Journal of the American Statistical Association* 1994; **89**(426):550–559.
22. Müller S, Welsh AH. Outlier robust model selection in linear regression. *Journal of the American Statistical Association* 2005; **100**(472):1297–1310.
23. Ronchetti E. Robustness aspects of model choice. *Statistica Sinica* 1997; **7**(2):327–338.
24. Heritier S, Cantoni E, Copt S, Victoria-Feser MP. *Robust Methods in Biostatistics*, Wiley Series in Probability and Statistics. John Wiley & Sons Ltd.: Chichester, 2009.
25. Park MY, Hastie T. Penalized logistic regression for detecting gene interactions. *Biostatistics* 2008; **9**(1):30–50.
26. Müller S, Scealy JL, Welsh AH. Model selection in linear mixed models. *Statistical Science* 2013; **28**(2):135–167. DOI: 10.1214/12-STS410.
27. Meier L, van de Geer S, Bühlmann P. The group Lasso for logistic regression. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 2008; **70**(1):53–71.