Statistica Sinica **19** (2009), 1155-1170

ROBUST MODEL SELECTION IN GENERALIZED LINEAR MODELS

Samuel Müller and A. H. Welsh

University of Sydney and The Australian National University

Abstract: In this paper, we extend to generalized linear models the robust model selection methodology of Müller and Welsh (2005). As in Müller and Welsh (2005), we combine a robust penalized measure of fit to the sample with a robust measure of out of sample predictive ability that is estimated using a post-stratified m-out-of-n bootstrap. The method can be used to compare different estimators (robust and nonrobust) as well as different models. Specialized to linear models, the present methodology improves on Müller and Welsh (2005): we use a new bias-adjusted bootstrap estimator which avoids the need to include an intercept in every model and we establish an essential monotonicity condition more generally.

Key words and phrases: Bootstrap model selection, generalized linear models, paired bootstrap, robust estimation, robust model selection, stratified bootstrap.

1. Introduction

Model selection is fundamental to the practical application of statistics and there is a substantial literature on the selection of linear regression models. A growing part of this literature is concerned with robust approaches to selecting linear regression models: see Müller and Welsh (2005) for references. The literature on the selection of generalized linear models (GLM; McCullagh and Nelder (1989)) and related marginal models fitted by generalized estimating equations (GEE; Liang and Zeger (1986)) is much smaller and has only recently incorporated robustness considerations. Hurvich and Tsai (1995) and Pan (2001) developed Akaike information criteria (AIC) based on the quasi-likelihood, Cantoni, Mills Flemming and Ronchetti (2005) presented a generalized version of Mallows' C_p, and Pan and Le (2001) and Cantoni, Field, Mills Flemming and Ronchetti (2007) presented approaches based on the bootstrap and cross-validation, respectively. Our purpose in this paper is to generalize the robust bootstrap model selection criterion of Müller and Welsh (2005) to generalized linear models.

The extension of the methodology of Müller and Welsh (2005) from linear regression to generalized linear models is less straightforward than we expected and, as a result, the present paper differs from Müller and Welsh (2005) in two important respects. First, the bias-adjusted m-out-of-n bootstrap estimator

$\widehat{\beta}^{c*}_{\alpha,m} - \mathrm{E}_*(\widehat{\beta}^{c*}_{\alpha,m} - \widehat{\beta}^{c}_{\alpha})$ rather than $\widehat{\beta}^{c*}_{\alpha,m}$ is used in estimating the expected prediction loss $M_n^{(2)}(\alpha)$ (definitions are given in Section 2). As discussed in Section 3.2, this avoids having to include an intercept in every model. Second, we present a simpler, more general method than that used in Müller and Welsh (2005) for showing that the consistency result applies to particular robust estimators of the regression parameter. As discussed in Section 3.3, we use generalized inverse matrices to decompose the asymptotic variance of the estimator into terms that are easier to handle, write the trace as a simple sum, and show that the terms in this sum have the required properties. Both of these changes were necessitated by the more complicated structure of generalized linear models but they also apply to regression models where they represent improvements to the methodology of Müller and Welsh (2005).

Suppose that we have n independent observations $y = (y_1, \ldots, y_n)^T$ and an $n \times p$ matrix X whose columns we index by $\{1, \ldots, p\}$. Let α denote any subset of p_α distinct elements from $\{1, \ldots, p\}$, let X_α denote the $n \times p_\alpha$ matrix with columns given by the columns of X whose indices appear in α and let $x_{\alpha i}^T$ denote the ith row of X_α. Then a generalized linear regression model α for the relationship between the response y and explanatory variables X is specified by

$$\mathrm{E}(y_i) = h(\eta_i), \text{ and } \mathrm{Var}(y_i) = \sigma^2 v^2(\eta_i) \text{ with } \eta_i = x_{\alpha i}^T \beta_\alpha, \quad i = 1, \ldots, n, \tag{1.1}$$

where β_α is an unknown p_α-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. Let $\mathcal{A}$ denote a set of generalized linear regression models (1.1). The purpose of model selection is to choose one or more models α from $\mathcal{A}$ with specified desirable properties.

Our perspective on model selection is that a useful model should (i) parsimoniously describe the relationship between the sample data y and X, and (ii) be able to predict independent new observations. The ability to parsimoniously describe the relationship between the sample data can be measured by applying a penalised loss function to the observed residuals and we use the expected variance-weighted prediction loss to measure the ability to predict new observations. Müller and Welsh (2005) showed there are practical (as well as philosophical) benefits to using both criteria. In addition, we encourage consideration of different types of estimator of each of the models. Possible estimators include the nonrobust maximum likelihood (see Künsch, Stefanski and Carroll (1989), Cantoni and Ronchetti (2001) and Ruckstuhl and Welsh (2001)) and maximum quasi–likelihood estimators (see McCullagh and Nelder (1989)), and the robust estimators of Preisser and Qaqish (1999), Cantoni and Ronchetti (2001), and Cantoni (2004).

We define a class of robust model selection criteria in Section 2, present our theoretical results in Section 3, report the results of a simulation study in Section 4, present a data example in Section 5, and conclude with a short discussion in Section 6. The proof of the main theorem and some additional theoretical results are presented in the online supplement available at `http://www.stat.sinica.edu.tw/statistica`.

2. Robust Model Selection Criterion

Let $\mathcal{C}$ denote a set of estimators, let $\widehat{\beta}_\alpha^c$ denote an estimator of type $c \in \mathcal{C}$ of β_α under (1.1), let ρ be a nonnegative loss function, let δ be a specified function of the sample size n, and let $\widetilde{y}$ be a vector of future observations at X that are independent of y. Then, we choose models α from a set $\mathcal{A}$ fitted by method $c \in \mathcal{C}$ for which the criterion function

$$M(\alpha) = \frac{\sigma^2}{n}\Big\{\mathrm{E}\sum_{i=1}^{n} w_{\alpha i}\rho\Big[\frac{y_i - h(x_{\alpha i}^T\widehat{\beta}_\alpha^c)}{\sigma v(\eta_i)}\Big] + \delta(n)p_\alpha$$
$$+\mathrm{E}\left(\sum_{i=1}^{n} w_{\alpha i}\rho\Big[\frac{\widetilde{y}_i - h(x_{\alpha i}^T\widehat{\beta}_\alpha^c)}{\sigma v(\eta_i)}\Big]\,\Big|\, y, X\right)\Big\} \tag{2.1}$$

is small. In practice, we often supplement this criterion with graphical diagnostic methods that explore the quality of the model in ways that are not amenable to simple mathematical description.

As in Müller and Welsh (2005) we separate the estimators $\widehat{\beta}_\alpha^c$ and ρ because we want to compare different estimators, and linking ρ to any one of these estimators may favour that estimator. We are interested in fitting the core data and predicting core observations rather than those in the tail of the distribution, so take ρ to be constant for sufficiently large $|x|$. The simplest such function (and the one we use in all our computations) is

$$\rho(z) = \min(z^2, b^2); \tag{2.2}$$

as in Müller and Welsh (2005), we use $b = 2$. Smoother versions of ρ such as are required for theoretical results are easily defined and we can, when appropriate to the problem, use asymmetric ρ functions. The weights $w_{\alpha i}$ are Mallows' type weights which may be included for robustness in the X space, but can and often will be constant. The only restrictions on the function δ are that $\delta(n) \to \infty$ and $\delta(n)/n \to 0$ as $n \to \infty$. A common choice is $\delta(n) = 2\log(n)$ (e.g. Schwarz (1978) and Müller and Welsh (2005)). When we us the penalized loss function alone, δ has to be of higher order than $O(\log\log n)$, as shown in Qian and Field (2002, Thms. 1−3) for logistic regression models.

When needed, σ is estimated from the Pearson residuals $\{y_i - h(x_{\alpha_f i}^T \widehat{\beta}_{\alpha_f}^c)\}/v(x_{\alpha_f i}^T \widehat{\beta}_{\alpha_f}^c)$, $i = 1, \ldots, n$, from a "full" model α_f. A "full" model is a large model (often assumed to be the model $\{1, \ldots, p\}$) which produces a valid measure of residual spread (but hopefully not so large that we incur a high cost from overfitting). We omit the subscript α_f and denote the estimator of σ by $\widehat{\sigma}^c$ for notational simplicity. Then we estimate the penalized in-sample term in the criterion function (2.1) by $\widehat{\sigma}^{c2}\{M_n^{(1)}(\alpha) + n^{-1}\delta(n)p_\alpha\}$, where

$$M_n^{(1)}(\alpha) = n^{-1} \sum_{i=1}^{n} w_{\alpha i}^c \rho\left\{ \frac{y_i - h(x_{\alpha i}^T \widehat{\beta}_\alpha^c)}{\widehat{\sigma}^c v(x_{\alpha_f i}^T \widehat{\beta}_{\alpha_f}^c)} \right\}. \tag{2.3}$$

Next, we implement a proportionally allocated, stratified m-out-of-n bootstrap of rows of (y, X) in which we (i) compute and order the Pearson residuals, (ii) set the number of strata K at between 3 and 8 depending on the sample size n, (iii) set stratum boundaries at the $K^{-1}, 2K^{-1}, \ldots, (K-1)K^{-1}$ quantiles of the Pearson residuals, (iv) allocate observations to the strata in which the Pearson residuals lie, (v) sample $\#$(observations in stratum $k)m/n$ (rounded as necessary) rows of (y, X) independently with replacement from stratum k so that the total sample size is m, (vi) use these data to construct the estimator $\widehat{\beta}_{\alpha,m}^{c*}$, repeat steps (v) and (vi) B independent times and then estimate the conditional expected prediction loss by $\widehat{\sigma}^{c2} M_n^{(2)}(\alpha)$, where

$$M_n^{(2)}(\alpha) = n^{-1} \mathrm{E}_* \sum_{i=1}^{n} w_{\alpha i}^c \rho\left(\frac{y_i - h[x_{\alpha i}^T \{\widehat{\beta}_{\alpha,m}^{c*} - \mathrm{E}_*(\widehat{\beta}_{\alpha,m}^{c*} - \widehat{\beta}_\alpha^c)\}]}{\widehat{\sigma}^c v(x_{\alpha_f i}^T \widehat{\beta}_{\alpha_f}^c)} \right) \tag{2.4}$$

and E_* denotes expectation with respect to the bootstrap distribution. Combining (2.3) and (2.4), we estimate the criterion function (2.1) by

$$M_n(\alpha) = \widehat{\sigma}^{c2}\{M_n^{(1)}(\alpha) + n^{-1}\delta(n)p_\alpha + M_n^{(2)}(\alpha)\}. \tag{2.5}$$

The stratified bootstrap ensures that we obtain bootstrap samples that are similar to the sample data (observations in the tails of the residual distribution and outliers or, with categorical data, groups of categories are represented in each bootstrap sample; this makes computation faster and more stable). The optimal m depends on the true model; as in Müller and Welsh (2005), we suggest using $n/4 \le m \le n/2$ for moderate n ($50 \le n \le 200$). If n is small, m is small and the parameter estimators do not converge for some bootstrap samples (though this problem is reduced by the stratified bootstrap); if n is large, m can be smaller than $n/4$. Choosing $3 \le K \le 8$ is suggested for sample surveys (e.g., Cochran (1977, pp. 132-134)) and seems to work well in practice. The estimated variance function is estimated from a "full" model so does not change with the model α.

This simplifies and makes the procedure more stable. Finally, we use the bias-adjusted bootstrap estimator $\widehat{\beta}^{c*}_{\alpha,m} - \mathrm{E}_*(\widehat{\beta}^{c*}_{\alpha,m} - \widehat{\beta}^{c}_{\alpha})$ rather than the bootstrap estimator $\widehat{\beta}^{c*}_{\alpha,m}$ in $M^{(2)}_n(\alpha)$.

The computational burden of model selection is reduced by using the stratified bootstrap and can be reduced further by limiting the number of different estimators we consider, and by reducing the number of models in $\mathcal{A}$. We use an eclectic mix of methods including robust versions of deviance-tests, search schemes, diagnostics, etc., to produce a relatively small set $\mathcal{A}$ of competing models which we then compare using (2.5). We present a backward search algorithm in Section 3.4 that substantially reduces the number of models to be considered in practice.

3. Theoretical Results

Our procedure is intended to identify useful models that make $M_n(\alpha)$ small whether or not a true model exists. If (i) a true model α_0 exists and (ii) $\alpha_0 \subseteq \{1, \ldots, p\}$, then consistency in the sense that a procedure identifies α_0 with probability tending to one is a desirable property. In this section, we show that choosing the model which minimises $M_n(\alpha)$ is consistent. Specifically, for $c \in \mathcal{C}$, we define

$$\widehat{\alpha}^{c}_{m,n} = \operatorname*{argmin}_{\alpha \in \mathcal{A}} M_n(\alpha), \tag{3.1}$$

and develop conditions under which, for each $c \in \mathcal{C}$,

$$\lim_{n\to\infty} \mathrm{P}\{\widehat{\alpha}^{c}_{m,n} = \alpha_0\} = 1. \tag{3.2}$$

3.1. Conditions

Define the subset of correct models $\mathcal{A}_c$ to be the set of models $\alpha \in \mathcal{A}$ such that $\alpha_0 \subseteq \alpha$; all other models are called incorrect models. For any correct model $\alpha \in \mathcal{A}_c$, the errors $\epsilon_{\alpha i} = y_i - h(x^T_{\alpha i}\beta_\alpha)$ satisfy $\epsilon_{\alpha i} = \epsilon_{\alpha_0 i}$ for $i = 1, \cdots, n$, and the components of β_α corresponding to columns of X_α which are not also in α_0 equal zero. To simplify stating the conditions and the proof of the main result, write

$$\begin{aligned} h_{\alpha i} &= h(x^T_{\alpha i}\beta_\alpha), & h'_{\alpha i} &= h'(x^T_{\alpha i}\beta_\alpha), & h''_{\alpha i} &= h''(x^T_{\alpha i}\beta_\alpha), \\ \sigma_i &= \sigma v(x^T_{\alpha_f i}\beta_{\alpha_f}), & \epsilon_{\alpha_0 i} &= \epsilon_i, & & \\ \psi(x) &= \rho'(x), & \psi_i &= \psi(\epsilon_i/\sigma_i), \text{ and} & \psi'_i &= \psi'(\epsilon_i/\sigma_i). \end{aligned}$$

Then we require the following conditions.

(i) The $p_\alpha \times p_\alpha$ matrix $n^{-1}X^T_\alpha W_{\Gamma_\alpha} X_\alpha \to \Gamma_\alpha$, where $W_{\Gamma_\alpha} = (1/2)\,\mathrm{diag}(\sigma_1^{-2} w_{\alpha 1}(h'^2_{\alpha 1}\mathrm{E}\,\psi'_1 - h''_{\alpha 1}\mathrm{E}\,\psi_1), \ldots, \sigma_n^{-2} w_{\alpha n}(h'^2_{\alpha n}\mathrm{E}\,\psi'_n - h''_{\alpha n}\mathrm{E}\,\psi_n))$ and Γ^c_α is of full rank.

(ii) For all models $\alpha \in \mathcal{A}$ (including the full model), the estimators $\widehat{\beta}_\alpha^c - \beta_\alpha = O_p(n^{-1/2})$, $\widehat{\sigma}^c - \sigma = O_p(n^{-1/2})$ with $\sigma > 0$. For all correct models $\alpha \in \mathcal{A}_c$, $n\mathrm{Var}\,(\widehat{\beta}_\alpha^c) = \Sigma_\alpha + o_p(1)$, where Σ_α is of full rank, and for any two correct models $\alpha_1, \alpha_2 \in \mathcal{A}_c$ such that $\alpha_1 \subset \alpha_2$,

$$\mathrm{trace}(\Sigma_{\alpha_2}\Gamma_{\alpha_2}) - \mathrm{trace}(\Sigma_{\alpha_1}\Gamma_{\alpha_1}) > 0. \tag{3.3}$$

(iii) For all models $\alpha \in \mathcal{A}$, the bootstrap estimator $\widehat{\beta}_{\alpha m}^{c*} \to \beta_\alpha$ in probability. For all correct models $\alpha \in \mathcal{A}_c$, $m\mathrm{Var}_*(\widehat{\beta}_{\alpha m}^{c*}) = n\kappa^c\mathrm{Var}\,(\widehat{\beta}_\alpha^c) + o_p(1)$.

(iv) The sequence $\delta(n) = o(n/m)$ and $m = o(n)$.

(v) The derivatives $\psi = \rho'$ and ψ' exist, are uniformly continuous, bounded, $\mathrm{Var}\,(\epsilon_i\psi_i) < \infty$, and $\mathrm{E}\,\psi'(\epsilon_i) > 0$, $i = 1, \ldots, n$.

(vi) The weights are bounded, h and its first two derivatives are continuous, σ and v are both positive, and v' is bounded.

(vii) The x_i are bounded.

(viii) For any incorrect model α, $\liminf_{n\to\infty} M_n^{(1)}(\alpha) > \lim_{n\to\infty} M_n^{(1)}(\alpha_0)$ a.s..

Condition (i) is a generalization of a standard condition for fitting regression models that we require for generalized linear models. Condition (ii) is satisfied by many estimators; the monotonicity condition (3.3) restricts the estimators we can consider in $\mathcal{C}$ but allows us to include maximum likelihood and other estimators such as the Cantoni and Ronchetti (2001) estimator. Condition (iii) specifies the required properties of the bootstrap parameter estimator. In contrast to Müller and Welsh (2005), we do not have to impose conditions on the asymptotic bias of the bootstrap estimator. Combining (ii) and (iii), we obtain $\mathrm{Var}_*(\widehat{\beta}_{\alpha m}^{c*}) = m^{-1}\kappa^c\Sigma_\alpha + o_p(m^{-1})$. Conditions (v)-(vii) enable us to make various two-term Taylor expansions and to control the remainder terms. We require a higher level of smoothness than exhibited by the ρ-function (2.2), but there are many functions satisfying these properties. We do not require $E\psi_i = 0$ in (v) (and it is not implied by (ii)) because ρ is not linked to any estimator in $\mathcal{C}$. Condition (viii) is a generalisation of Condition (C4) of Shao (1996) to allow a more general choice of $\rho(\cdot)$.

We have specified a simple set of sufficient conditions (particularly in conditions (v)-(vii)) that are appropriate for a robust ρ function and generalized linear models. However, we note that we can specify alternative and simpler conditions for particular cases. For example, we obtain alternative conditions if we allow the x_i to be stochastic; see for example Shao (1996, Condition C3. b.). We can simplify our conditions if we use the nonrobust function $\rho(x) = x^2$; again see Shao (1996, p.661). Even in the robust case, simpler conditions can be given for

homoscedastic linear models because $h(x) = x$, $v(x) = 1$. These possibilities are tangential to our main purpose so we do not pursue them here.

Theorem 3.1. *Under conditions* (i)−(viii), *the consistency result* (3.2) *holds.*

The proof is given in the on-line supplement at `http://www.stat.sinica.edu.tw/statistica`.

3.2. The elimination of bias

One of the main difficulties in constructing model selection criteria like $M_n(\alpha)$ is removing the bias (equivalently the linear term) in the expansion of $M_n^{(2)}(\alpha)$. Suppose that instead of the bias-adjusted bootstrap estimator $\widehat{\beta}^{c*}_{\alpha,m} - \mathrm{E}_*(\widehat{\beta}^{c*}_{\alpha,m} - \widehat{\beta}^{c}_{\alpha})$, we use the bootstrap estimator $\widehat{\beta}^{c*}_{\alpha,m}$ in $M_n^{(2)}(\alpha)$. Then when we expand $M_n^{(2)}(\alpha)$ as in Shao (1996), Müller and Welsh (2005), or the proof of Theorem 3.1, we obtain the linear term

$$\mathrm{E}_*(\widehat{\beta}^{*}_{\alpha,m} - \widehat{\beta}_{\alpha})^T \frac{1}{n} \sum_{i=1}^{n} \widehat{\sigma}_i^{-1} w_{\alpha i} x_{\alpha i} h'(x_{\alpha i}^T \widehat{\beta}_{\alpha}) \psi\Big(\frac{y_i - h(x_{\alpha i}^T \widehat{\beta}_{\alpha})}{\widehat{\sigma}_i}\Big). \tag{3.4}$$

As shown in Müller and Welsh (2005), the bias term $\mathrm{E}_*(\widehat{\beta}^{*}_{\alpha,m} - \widehat{\beta}_{\alpha})$ is typically a function of α with leading term $O_p(m^{-1})$, the same as the quadratic term in the expansion. Since the quadratic term governs the selection of correct models, it is crucial that the linear term be at least of smaller order.

There are various ways to make (3.4) of order $o_p(m^{-1})$. Ordinarily, the mean in (3.4) is asymptotic to $n^{-1}\sum_{i=1}^{n} \sigma_i^{-1} w_{\alpha i} x_{\alpha i} h'_{\alpha i} \mathrm{E}\,\psi_i$ which is $O(1)$. However, if $\mathrm{E}\,\psi_i = 0$ (as in Shao (1996)), then it can be $O_p(n^{-1/2})$ which can be made $o_p(m^{-1})$. It holds when $\psi(x) = x$, but this is a nonrobust choice and hence unappealing in general. Müller and Welsh (2005) made each model contain an intercept and centered the explanatory variables to have mean zero so the bias would be forced into the intercept. The intercept can be eliminated by replacing the intercept of the bootstrap estimator by that of the estimator $\widehat{\beta}_{\alpha}$, or by fixing the intercept at that estimated under a "full" model. This approach is much less attractive here because the centering vector has to include estimates of σ_i, $\mathrm{E}\,\psi_i$ and $h'_{\alpha i}$ so is stochastic, and the centered explanatory variables cannot be simply conditioned on. Even if we overcome these difficulties, the arguments in the next subsection do not apply unless the model is fitted with the same covariates as the model selection criterion uses, so this approach is not attractive.

A different approach would be to require as in Müller and Welsh (2005) that $\mathrm{E}_*(\widehat{\beta}^{*}_{\alpha,m} - \widehat{\beta}_{\alpha}) = m^{-1}B_{\alpha} + o_p(m^{-1})$, estimate B_{α}, and then adjust the criterion by subtracting off an estimate of (3.4). Although this would remove the bias, it would contribute to the quadratic term and affect the consistency proof.

Moreover, the new criterion would lack natural interpretability. It is better to adjust the bootstrap estimator $\widehat{\beta}^*_{\alpha,m}$ for bias. We only need the leading term but we would need to derive and estimate B_α for each estimator we consider. Fortunately, we have the bias in the natural form $\mathrm{E}_*(\widehat{\beta}^*_{\alpha,m} - \widehat{\beta}_\alpha)$ so we can remove it entirely. This is the solution we have adopted in $M_n^{(2)}(\alpha)$.

3.3. The monotonicity of trace($\Sigma_\alpha\Gamma_\alpha$)

The assumption (ii) that trace($\Sigma_\alpha\Gamma_\alpha$) is monotone in p_α does not hold in general for arbitrary positive semi–definite matrices Σ_α and Γ_α. However Müller and Welsh (2005) proved that, for linear regression models, the condition holds for the class of Mallows type M–estimators or one–step Mallows type M–estimators etc., because of the relationship between $\mathrm{Var}\,(\widehat{\beta}_\alpha)$ and Γ_α. For generalized linear models, the maximum likelihood estimator $\widehat{\beta}_\alpha$ satisfies

$$n\mathrm{Var}\,(\widehat{\beta}_\alpha) = (X_\alpha^T W_{\Sigma_\alpha} X_\alpha)^{-1} + o_p(1),$$

where $W_{\Sigma_\alpha} = \mathrm{diag}(h'^2_{\alpha 1}/\sigma_1^2, \ldots, h'^2_{\alpha n}/\sigma_n^2)$ (McCullagh and Nelder (1989, p.43)) so to establish (3.3) we have to show that

$$\mathrm{trace}\left\{(X_\alpha^T W_{\Sigma_\alpha} X_\alpha)^{-1} X_\alpha^T W_{\Gamma_\alpha} X_\alpha\right\}$$

is strictly monotone increasing in p_α. Reorder the rows of X_α if necessary so that the top $p_\alpha \times p_\alpha$ submatrix C_α is nonsingular. Then the $p_\alpha \times n$ matrix $X_\alpha^- = (C_\alpha^{-1}, 0)$ is a generalized inverse of X_α, so $X_\alpha X_\alpha^- = \mathrm{blockdiag}(I_{p_\alpha}, 0)$ and

$$\begin{aligned}\mathrm{trace}\left\{(X_\alpha^T W_{\Sigma_\alpha} X_\alpha)^{-1} X_\alpha^T W_{\Gamma_\alpha} X_\alpha\right\} &= \mathrm{trace}\left(X_\alpha X_\alpha^- W_{\Sigma_\alpha}^{-1} X_\alpha X_\alpha^- W_{\Gamma_\alpha}\right) \\ &= \frac{1}{2}\sum_{i=1}^{p_\alpha} w_{\alpha i} \frac{h'^2_{\alpha i}\mathrm{E}\,\psi_i' - h''_{\alpha i}\mathrm{E}\,\psi_i}{h_i'^2}.\end{aligned}$$

Since $h'^2_{\alpha i}\mathrm{E}\,\psi_i' > 0$, the simplest sufficient condition for monotonicity is

$$h''_{\alpha i}\mathrm{E}\,\psi_i \le 0, \quad i = 1, \ldots, n. \tag{3.5}$$

We show in the supplementary on-line material (`http:www.stat.sinica.edu.tw/statistica`) that (3.5) is also a sufficient condition for monotonicity of the Mallows quasi–likelihood estimator of Cantoni and Ronchetti (2001, Sec. 2.2).

Condition (3.5) holds if $\mathrm{E}\,\psi_i = 0$ or $h''_{\alpha i} = 0$. The first case occurs when (i) $\rho(x) = x^2$ or (ii) the $\epsilon_i = y_i - h_{\alpha i}$ have a distribution which is symmetric about zero and ψ is antisymmetric, and the second when we use the identity link so $h(x) = x$. Shao (1996) exploited (i), but this choice favours least squares estimation and is non-robust so we prefer not to use it; (ii) applies to Gaussian

models but not to models with asymmetric distributions. Similarly, the identity link is widely used in Gaussian models and may be used in gamma models, but is not useful in binomial and Poisson models. In these cases, we need to examine (3.5) more carefully. We show in the on-line supplementary material that (3.5) holds for the binomial distribution with the logistic link, and for right skewed distributions with the log or reciprocal link provided b in (2.2) is large enough.

3.4. The reduction of models

For any incorrect model $\alpha \in \mathcal{A} \setminus \mathcal{A}_c$ it follows from condition (vi) in Section 3.1 and the proof of the theorem, that for fixed p_{α_f} we have

$$\liminf_{n\to\infty} \min_{\alpha\in\mathcal{A}\setminus\mathcal{A}_c} M_n(\alpha) > \lim_{n\to\infty} \max_{\alpha\in\mathcal{A}_c} M_n(\alpha_0) \quad \text{a.s..} \tag{3.6}$$

Equation (3.6) ensures that backward model selection schemes based on $M_n(\alpha)$ are consistent for the true model if $\mathcal{A}$ is the set of all possible $2^{p_{\alpha_f}}$ submodels. We therefore suggest using the following backward selection algorithm if p_{α_f} is large.

1. Calculate $M_n(\alpha)$ for the full model $\alpha_f = \{1, \ldots, p_{\alpha_f}\}$ and $\alpha_{f,-i} = \{1, \ldots, p_{\alpha_f}\} \setminus \{i\}$, $i = 1, \ldots, p_{\alpha_f}$, resulting in $\{M_n(\alpha) : \#\alpha \geq p_{\alpha_f} - 1\}$.
2. Set $\alpha_f = \text{argmin}_{\{\#\alpha \geq p_{\alpha_f}-1\}} M_n(\alpha)$ and repeat 1. if $\alpha_f \geq 2$.
3. Estimate α by argmin $M_n(\alpha)$ over all $1 + \sum_{i=1}^{p_{\alpha_f}} i = 1 + k(k+1)/2$ considered models.

An example of the solution paths of all submodels and of the backward selected submodels is given in Figure 1 in Section 5.

4. Simulation Study

In this section we present simulation results for Poisson regression models

$$y_i \sim Poi(\mu_i),\ \eta_i = \log \mu_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i},\ i = 1, \ldots, n, \tag{4.1}$$

which compared the proposed robust model selection criterion with the maximum likelihood ($\widehat{\beta}_{ML}$) and Cantoni and Ronchetti (2001) estimators ($\widehat{\beta}_{CR}$) computed using the R functions `glm.fit` and `glmrob`, respectively. All the simulations used $b = 2$ in (2.2), $\delta(n) = 2\log(n)$, and were based on 500 simulation runs so the standard errors of the estimated model selection probabilities were less than 0.023.

We first considered 4-parameter cases with $n = 64$ observations generated with parameter vectors $(1, 0, 0, 0)$, $(-1, 2, 0, 0)$, and $(-1, 1, 1, 0)$. The explanatory variables were generated from the multivariate normal distribution with mean

Table 1. Estimated selection probabilities with no outlying points.

			$\widehat{\beta}_{ML}$			$\widehat{\beta}_{CR}$
true β^T	model	type	AIC	BIC	$\widehat{\alpha}^{ss}_{m,n}$	$\widehat{\alpha}^{ss}_{m,n}$
$(1,0,0,0)$	$(\beta_1,0,0,0)$	α_0	0.58	0.60	0.90	0.89
	$(\beta_1,\beta_2,0,0)$	$\mathcal{A}_c$	0.10	0.10	0.02	0.03
	$(\beta_1,0,\beta_3,0)$	$\mathcal{A}_c$	0.13	0.13	0.04	0.05
	$(\beta_1,0,0,\beta_4)$	$\mathcal{A}_c$	0.13	0.12	0.03	0.03
	$(\beta_1,\beta_2,\beta_3,0)$	$\mathcal{A}_c$	0.03	0.02	0.00	0.00
	$(\beta_1,\beta_2,0,\beta_4)$	$\mathcal{A}_c$	0.02	0.02	0.00	0.00
	$(\beta_1,0,\beta_3,\beta_4)$	$\mathcal{A}_c$	0.02	0.02	0.00	0.00
	$(\beta_1,\beta_2,\beta_3,\beta_4)$	$\mathcal{A}_c$	0.00	0.00	0.00	0.00
$(-1,2,0,0)$	$(\beta_1,0,0,0)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,\beta_2,0,0)$	α_0	0.65	0.67	0.94	0.93
	$(\beta_1,0,\beta_3,0)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,0,0,\beta_4)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,\beta_2,\beta_3,0)$	$\mathcal{A}_c$	0.15	0.15	0.03	0.03
	$(\beta_1,\beta_2,0,\beta_4)$	$\mathcal{A}_c$	0.17	0.16	0.03	0.03
	$(\beta_1,0,\beta_3,\beta_4)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,\beta_2,\beta_3,\beta_4)$	$\mathcal{A}_c$	0.03	0.02	0.00	0.00
$(-1,1,1,0)$	$(\beta_1,0,0,0)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,\beta_2,0,0)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,0,\beta_3,0)$	–	0.00	0.00	0.05	0.07
	$(\beta_1,0,0,\beta_4)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,\beta_2,\beta_3,0)$	α_0	0.81	0.82	0.91	0.89
	$(\beta_1,\beta_2,0,\beta_4)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,0,\beta_3,\beta_4)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,\beta_2,\beta_3,\beta_4)$	$\mathcal{A}_c$	0.19	0.18	0.03	0.04

$(1,1,1)$ and identity covariance matrix. For the criterion $\widehat{\alpha}$, we used $m = 24$ with $K = 8$ equal-sized strata based on the Pearson residuals from the full model and $B = 50$ bootstrap samples. Selection probabilities are presented in Table 1. Without outliers, the overall performance of the selection criterion $\widehat{\alpha}$ was superior to classical criteria such as the AIC and BIC criterion independently of the chosen estimation procedure. For example, for $(1,0,0,0)$, the selection probabilities of the true model using $\widehat{\beta}_{ML}$ are 0.58 for AIC, 0.60 for BIC, 0.90 for $\widehat{\alpha}$, and using $\widehat{\beta}_{CR}$ the estimated probability was 0.89 for $\widehat{\alpha}$.

Although the results are not reported here, we repeated the experiment with random explanatory variables (i.e. different in each simulation run) and obtained essentially indistinguishable results from those in Table 1. This suggests that our procedure works as well with random explanatory variables as with fixed ones. In both cases, we also repeated the simulation with $m = 16$ and $m = 32$, and confirmed that the results were insensitive to the choice of m in the recommended

Table 2. Estimated selection probabilities with eight moderately outlying points.

			$\widehat{\beta}_{ML}$			$\widehat{\beta}_{CR}$
true β^T	model	type	AIC	BIC	$\widehat{\alpha}^{ss}_{m,n}$	$\widehat{\alpha}^{ss}_{m,n}$
$(1,0,0,0)$	$(\beta_1,0,0,0)$	α_0	0.41	0.42	0.94	0.94
	$(\beta_1,\beta_2,0,0)$	$\mathcal{A}_c$	0.12	0.12	0.02	0.02
	$(\beta_1,0,\beta_3,0)$	$\mathcal{A}_c$	0.07	0.07	0.02	0.02
	$(\beta_1,0,0,\beta_4)$	$\mathcal{A}_c$	0.25	0.24	0.03	0.02
	$(\beta_1,\beta_2,\beta_3,0)$	$\mathcal{A}_c$	0.04	0.04	0.00	0.00
	$(\beta_1,\beta_2,0,\beta_4)$	$\mathcal{A}_c$	0.06	0.05	0.00	0.00
	$(\beta_1,0,\beta_3,\beta_4)$	$\mathcal{A}_c$	0.05	0.05	0.00	0.00
	$(\beta_1,\beta_2,\beta_3,\beta_4)$	$\mathcal{A}_c$	0.01	0.01	0.00	0.00
$(-1,2,0,0)$	$(\beta_1,0,0,0)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,\beta_2,0,0)$	α_0	0.01	0.01	0.66	0.78
	$(\beta_1,0,\beta_3,0)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,0,0,\beta_4)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,\beta_2,\beta_3,0)$	$\mathcal{A}_c$	0.00	0.00	0.01	0.02
	$(\beta_1,\beta_2,0,\beta_4)$	$\mathcal{A}_c$	0.79	0.80	0.33	0.20
	$(\beta_1,0,\beta_3,\beta_4)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,\beta_2,\beta_3,\beta_4)$	$\mathcal{A}_c$	0.20	0.19	0.01	0.01
$(-1,1,1,0)$	$(\beta_1,0,0,0)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,\beta_2,0,0)$	–	0.00	0.00	0.01	0.01
	$(\beta_1,0,\beta_3,0)$	–	0.00	0.00	0.02	0.07
	$(\beta_1,0,0,\beta_4)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,\beta_2,\beta_3,0)$	α_0	0.00	0.00	0.55	0.73
	$(\beta_1,\beta_2,0,\beta_4)$	–	0.00	0.00	0.04	0.02
	$(\beta_1,0,\beta_3,\beta_4)$	–	0.00	0.00	0.05	0.03
	$(\beta_1,\beta_2,\beta_3,\beta_4)$	$\mathcal{A}_c$	0.99	0.99	0.34	0.13

range.

Next we generated data from the same model (4.1), but added moderate outliers in the response to the eight observations with largest explanatory variable x_4. That is, if $\text{rank}(x_{4i}) := \sum_{k=1}^{n} \mathbf{1}(x_{4k} \leq x_{4i}) \geq 57$, then $y_i \sim Poi(10)$, $i = 1, \ldots, n$. The selection probabilities are presented in Table 2. We see that the proposed selection criterion used with the robust estimator from Cantoni and Ronchetti (2001) performed outstandingly well. Used with the maximum likelihood estimator, it still performed very well compared to AIC and BIC. For example, for $(-1,2,0,0)$, the selection probabilities of the true model using $\widehat{\beta}_{ML}$ were 0.01 for AIC, 0.01 for BIC, 0.66 for $\widehat{\alpha}$, and using $\widehat{\beta}_{CR}$ the selection probability was 0.78 for $\widehat{\alpha}$.

For a more severe test, we generated data from (4.1) but added two influential outliers to the response variable according to the condition that if $\text{rank}(x_{4i}) \leq 2$ then $y_i \sim Poi(100)$, $i = 1, \ldots, n$. The selection probabilities presented in Table 3

Table 3. Estimated selection probabilities with two strongly outlying points.

			$\widehat{\beta}_{ML}$			$\widehat{\beta}_{CR}$
true β^T	model	type	AIC	BIC	$\widehat{\alpha}^{ss}_{m,n}$	$\widehat{\alpha}^{ss}_{m,n}$
$(1,0,0,0)$	$(\beta_1,0,0,0)$	α_0	0.00	0.00	0.03	0.97
	$(\beta_1,\beta_2,0,0)$	$\mathcal{A}_c$	0.00	0.00	0.00	0.01
	$(\beta_1,0,\beta_3,0)$	$\mathcal{A}_c$	0.00	0.00	0.02	0.01
	$(\beta_1,0,0,\beta_4)$	$\mathcal{A}_c$	0.00	0.00	0.04	0.00
	$(\beta_1,\beta_2,\beta_3,0)$	$\mathcal{A}_c$	0.00	0.00	0.00	0.00
	$(\beta_1,\beta_2,0,\beta_4)$	$\mathcal{A}_c$	0.00	0.00	0.00	0.00
	$(\beta_1,0,\beta_3,\beta_4)$	$\mathcal{A}_c$	0.05	0.05	0.88	0.00
	$(\beta_1,\beta_2,\beta_3,\beta_4)$	$\mathcal{A}_c$	0.95	0.95	0.04	0.00
$(-1,2,0,0)$	$(\beta_1,0,0,0)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,\beta_2,0,0)$	α_0	0.00	0.00	0.17	0.99
	$(\beta_1,0,\beta_3,0)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,0,0,\beta_4)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,\beta_2,\beta_3,0)$	$\mathcal{A}_c$	0.00	0.00	0.04	0.01
	$(\beta_1,\beta_2,0,\beta_4)$	$\mathcal{A}_c$	0.00	0.00	0.15	0.00
	$(\beta_1,0,\beta_3,\beta_4)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,\beta_2,\beta_3,\beta_4)$	$\mathcal{A}_c$	1.00	1.00	0.63	0.00
$(-1,1,1,0)$	$(\beta_1,0,0,0)$	–	0.00	0.00	0.00	0.01
	$(\beta_1,\beta_2,0,0)$	–	0.00	0.00	0.05	0.06
	$(\beta_1,0,\beta_3,0)$	–	0.00	0.00	0.00	0.22
	$(\beta_1,0,0,\beta_4)$	–	0.00	0.00	0.04	0.00
	$(\beta_1,\beta_2,\beta_3,0)$	α_0	0.00	0.00	0.00	0.71
	$(\beta_1,\beta_2,0,\beta_4)$	–	0.02	0.02	0.88	0.00
	$(\beta_1,0,\beta_3,\beta_4)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,\beta_2,\beta_3,\beta_4)$	$\mathcal{A}_c$	0.98	0.98	0.03	0.00

show that the robust model selection criterion can break down if it is used with $\widehat{\beta}_{ML}$ but still perform well with robust parameter estimators. For example, for $(-1,1,1,0)$, the selection probabilities of the true model using $\widehat{\beta}_{ML}$ were 0 for AIC, BIC, and $\widehat{\alpha}$, but using $\widehat{\beta}_{CR}$ the estimated probability was 0.71 for $\widehat{\alpha}$. The methods AIC, BIC, and $\widehat{\alpha}$ using $\widehat{\beta}_{ML}$ that use the nonrobust maximum likelihood estimator all break down in this case. The fact that AIC and BIC essentially always choose the full model (regardless of the true model) is not desirable and implies that no model selection is occurring. The poor performance of $\widehat{\alpha}$ using $\widehat{\beta}_{ML}$ shows that robust model selection requires both a robust criterion and a robust estimator.

Finally, to explore the effect of correlation in the explanatory variables in an example more like that considered in the next Section (the largest models selected have 6 parameters with low correlation between the explanatory variables), we considered a 6-parameter case with $n = 128$ observations generated with parameter vectors $(-1,1,1,0,0,0)$, $(-2,1,1,1,0,0)$, and $(-1,1,1,-1,1,0)$, and the explanatory variables assigned correlation 0.5. For the criterion $\widehat{\alpha}$, we used $m = 32$ with $K = 8$ equal-sized strata and $B = 50$ bootstrap samples. We considered all 32 possible models with an intercept but, to save space, report

Table 4. Estimated selection probabilities with correlated explanatory variables.

			$\widehat{\beta}_{ML}$			$\widehat{\beta}_{CR}$
true β^T	model	type	AIC	BIC	$\widehat{\alpha}^{ss}_{m,n}$	$\widehat{\alpha}^{ss}_{m,n}$
$(-1,1,1,0,0,0)$	$(\beta_1,\beta_2,\beta_3,0,0,0)$	α_0	0.63	0.72	0.98	0.98
	$(\beta_1,\beta_2,\beta_3,\beta_4,0,0)$	$\mathcal{A}_c$	0.13	0.10	0.00	0.01
	$(\beta_1,\beta_2,\beta_3,0,\beta_5,0)$	$\mathcal{A}_c$	0.10	0.09	0.01	0.01
	$(\beta_1,\beta_2,\beta_3,0,0,\beta_6)$	$\mathcal{A}_c$	0.09	0.06	0.01	0.01
	$(\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,0)$	$\mathcal{A}_c$	0.02	0.02	0.00	0.00
	$(\beta_1,\beta_2,\beta_3,\beta_4,0,\beta_6)$	$\mathcal{A}_c$	0.02	0.06	0.00	0.00
	$(\beta_1,\beta_2,\beta_3,0,\beta_5,\beta_6)$	$\mathcal{A}_c$	0.01	0.01	0.00	0.00
$(-2,1,1,1,0,0)$	$(\beta_1,\beta_2,\beta_3,0,0,0)$	–	0.00	0.00	0.00	0.01
	$(\beta_1,0,\beta_3,\beta_4,0,0)$	–	0.00	0.00	0.00	0.00
	$(\beta_1,\beta_2,\beta_3,\beta_4,0,0)$	α_0	0.69	0.76	0.97	0.97
	$(\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,0)$	$\mathcal{A}_c$	0.13	0.12	0.01	0.01
	$(\beta_1,\beta_2,\beta_3,\beta_4,0,\beta_6)$	$\mathcal{A}_c$	0.15	0.11	0.01	0.01
	$(\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6)$	$\mathcal{A}_c$	0.03	0.02	0.00	0.00
$(-1,1,1,-1,1,0)$	$(\beta_1,0,\beta_3,0,\beta_5,0)$	–	0.00	0.00	0.02	0.02
	$(\beta_1,\beta_2,\beta_3,\beta_4,0,0)$	–	0.00	0.00	0.02	0.03
	$(\beta_1,\beta_2,\beta_3,0,\beta_5,0)$	–	0.00	0.00	0.01	0.02
	$(\beta_1,0,\beta_3,\beta_4,\beta_5,0)$	–	0.00	0.00	0.01	0.01
	$(\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,0)$	α_0	0.84	0.88	0.93	0.91
	$(\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6)$	$\mathcal{A}_c$	0.16	0.12	0.01	0.01

only those models which at least one method selected three or more times (out of 500 simulations). The results in Table 4 indicate that the performance of $\widehat{\alpha}$ is not affected by this level of correlation.

In further simulations (not reported here), we compared the performance of (2.3) and (2.4) applied separately with that of (2.5). The results confirmed the finding of Müller and Welsh (2005) that the combined criterion captures the strengths of the two separate criteria and, over a range of examples, leads to a better criterion.

5. Data Example

In this section, we analyse data on the diversity of arboreal marsupials (possums) in montane ash forest (Australia). The dataset described by Lindenmayer et al. (1991, 1990) is part of the `robustbase` package in R (`possumDiv.rda`). The response is the number of different species (`diversity`) observed on $n = 151$ sites. The explanatory variables describe the sites in terms of the number of shrubs (`shrubs`), number of cut stumps from past logging operations (`stumps`), the number of stags (`stags`), a bark index (`bark`, 30 levels), the basal area of acacia species (`acacia`, 11 levels), a habitat score (`habitat`, 40 levels), the species of

Table 5. Selected best model for the Lindenmayer et al. (1990, 1991) data using a range of model selection procedures.

selection criterion	$\widehat{\beta}$	selected variables in the best model
$\widehat{\alpha}$	$\widehat{\beta}_{CR}$	stags, habitat
$\widehat{\alpha}$	$\widehat{\beta}_{ML}$	stags, habitat
AIC	$\widehat{\beta}_{ML}$	stags, bark, acacia, habitat, aspect
BIC	$\widehat{\beta}_{ML}$	stags, bark, acacia, aspect
p-value forward stepwise	$\widehat{\beta}_{CR}$	stags, bark, acacia, habitat, aspect
p-value forward stepwise	$\widehat{\beta}_{ML}$	stags, bark, acacia, habitat, aspect

eucalypt with the greatest stand basal area (`eucalypt`, 3 nominal levels), and the aspect of the site (`aspect`, 4 nominal levels). We calculate $\widehat{\alpha}$ based on $\widehat{\beta}_{CR}$ with the same specifications as in the simulation study, but because n is larger than 64 we use $m = 40 \approx 0.26n$. Table 5 presents a summary of selected best models which includes also the results of Cantoni and Ronchetti (2001, Sec. 5.2). The best model according to our criterion $M_n(\alpha)$ includes `stags` and `habitat`, which are also selected if the backward selection algorithm in Section 3.4 is applied. The solution path of $M_n(\alpha)$ is given in Figure 1 which shows the minimal value of $M_n(\alpha)$ for all considered models with the same number of variables. Cantoni and Ronchetti (2001) found four potentially influential data points, namely observations 59, 110, 133, and 139. Based on the construction of our criterion, its consistency and the results of our simulation study, we consider $\widehat{\alpha}$ with $\widehat{\beta}_{CR}$ to be superior to AIC, BIC, and $\widehat{\alpha}$ with $\widehat{\beta}_{ML}$.

6. Discussion and Conclusions

We have proposed a bootstrap criterion for robustly selecting generalized linear models. The criterion is a generalization of that developed for regression models by Müller and Welsh (2005) and has its strengths while still improving on that criterion. In particular, the criterion (i) combines a robust penalised criterion (which reflects goodness-of-fit to the data) with an estimate of a robust measure of the conditional expected prediction error (which measures the ability to predict as yet unobserved observations), (ii) separates the comparison of models from any particular method of estimating them, and (iii) uses the stratified bootstrap to make the criterion more stable. The improvement is achieved by using the bootstrap to estimate the bias of the bootstrap estimator of the regression parameter, and then using the bias-adjusted bootstrap estimator instead of the raw bootstrap estimator in the criterion. This step widens the applicability of the method by removing the requirement of Müller and Welsh (2005) that the models under consideration include an intercept. We have also developed a more

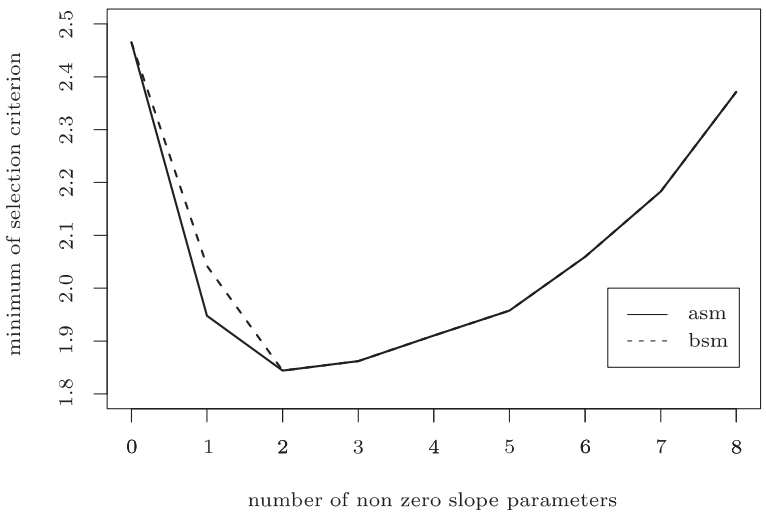

Figure 1. Solution path for $\operatorname{argmin} M_n(\alpha)$ given a fixed number of non zero slope parameters for all (asm) and backward selected (bsm) submodels.

widely applicable method than that given in Müller and Welsh (2005) for establishing that the criterion can be applied with particular robust estimators of the regression parameters. Our main theoretical result establishes the asymptotic consistency of the method and a simulation study shows that the model selection method works very well in finite samples.

Acknowledgement

This research is supported by the Australian Research Council DP. We acknowledge the helpful suggestions of the Editors, and associate editor and the referee.

References

Cantoni, E. (2004). Analysis of robust quasi-deviances for generalized linear models. *J. Statist. Software* **10**, Issue 4.

Cantoni, E., Field, C., Mills Flemming, J. and Ronchetti, E. (2007). Longitudinal variable selection by cross-validation in the case of many covariates. *Statist. Medicine* **26**, 919-930.

Cantoni, E., Mills Flemming, J. and Ronchetti, E. (2005). Variable selection for marginal longitudinal generalized linear models. *Biometrics* **61**, 507-513.

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *J. Amer. Statist. Assoc.* **96**, 1022-1030.

Cochran, W. G. (1977). *Sampling Techniques*. 3rd edition. Wiley, New York.

Hurvich, C. M. and Tsai, C.-L. (1995). Model selection for extended quasi-likelihood models in small samples. *Biometrics* **51**, 1077-1084.

Künsch, H. R., Stefanski, L. A. and Carroll, R. J. (1989). Conditionally unbiased bounded–influence estimation in general regression models, with applications to generalized linear models. *J. Amer. Statist. Assoc.* **84**, 460-466.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Lindenmayer, D. B., Cunningham, R. B., Tanton, M. T., Nix, H. A. and Smith, A. P. (1991). The conservation of arboreal marsupials in the Montane ash forests of the central highlands of Victoria, South-East Australia: III. The habitat requirements of Leadbeater's possum *Gymnobelideus leadbeateri* and models of the diversity and abundance of arboreal marsupials. *Biological Conservation* **56**, 295-315.

Lindenmayer, D. B., Cunningham, R. B., Tanton, M. T., Smith, A. P. and Nix, H. A. (1990). The conservation of arboreal marsupials in the Montane ash forests of the central highlands of Victoria, South-East Australia: I. Factors influencing the occupancy of trees with hollows. *Biological Conservation* **54**, 111-131.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd edition. Chapman and Hall/CRC, London.

Müller, S. and Welsh, A. H. (2005). Outlier robust model selection in linear regression. *J. Amer. Statist. Assoc.* **100**, 1297-1310.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics* **57**, 120-125.

Pan, W. and Le, C.T. (2001). Bootstrap selection in generalized linear models. *J. Agricultural, Biological, and Environmental Statistics* **6**, 49-61.

Preisser, J. S. and Qaqish, B. F. (1999). Robust regression for clustered data with application to binary responses. *Biometrics* **55**, 574-579.

Qian, G. and Field, C. (2002). Law of iterated logarithm and consistent model selection criterion in logistic regression. *Statist. Probab. Lett.* **56**, 101-112.

Ruckstuhl, A. F. and Welsh, A. H. (2001). Robust fitting of the binomial model. *Ann. Statist.* **29**, 1117-1136.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Shao, J. (1996). Bootstrap model selection. *J. Amer. Statist. Assoc.* **91**, 655-665.

School of Mathematics and Statistics F07, University of Sydney, NSW 2006, Australia.

E-mail: mueller@maths.usyd.edu.au

Centre for Mathematics and its Applications, The Australian National University, Canberra ACT 0200, Australia.

E-mail: Alan.Welsh@anu.edu.au

(Received November 2007; accepted June 2008)