Regression model selection via log-likelihood ratio and constrained minimum criterion

Min TSAO

Department of Mathematics & Statistics, University of Victoria, Victoria, British Columbia, Canada

Key words and phrases: Best subset selection; constrained minimum criterion; model selection; regression models; variable selection.

MSC 2020: Primary 62J05; 62J12; secondary 62F03.

Abstract: Although log-likelihood is widely used in model selection, the log-likelihood ratio has had few applications in this area. We develop a log-likelihood ratio based method for selecting regression models by focusing on the set of models deemed plausible by the likelihood ratio test. We show that when the sample size is large and the significance level of the test is small, there is a high probability that the smallest model in this set is the true model; thus, we select this smallest model. The significance level of the test serves as a tuning parameter of this method. We consider three levels of this parameter in a simulation study and compare this method with the Akaike information criterion (AIC) and Bayesian information criterion (BIC) to demonstrate its excellent accuracy and adaptability to different sample sizes. This method is a frequentist alternative and a strong competitor to AIC and BIC for selecting regression models.

Résumé: Bien que la log-vraisemblance soit largement utilisée dans la sélection de modèles, le rapport de vraisemblance logarithmique n’a guère été appliqué dans ce domaine. En se limitant à l’ensemble de modèles jugés plausibles par le test du rapport de vraisemblance, l’auteur du présent travail propose une méthode basée sur le rapport de vraisemblance logarithmique pour sélectionner des modèles de régression. Plus précisément, il montre que lorsque la taille de l’échantillon est grande et que le niveau de signification du test est faible, alors la probabilité que le plus petit modèle de cet ensemble soit le vrai modèle est relativement forte. C’est ainsi que la méthode proposée choisit le plus petit modèle en se servant du niveau de signification du test comme paramètre de réglage. Une étude de simulation avec trois seuils du niveau de signification a permis de montrer que l’approche proposée performe mieux que le critère d’information d’Akaike (AIC) et le critère d’information bayésien (BIC) et ce en termes d’exactitude et d’adaptabilité aux différentes tailles d’échantillon. Enfin, la méthode est présentée comme une alternative fréquentiste et un sérieux concurrent de l’AIC et du BIC pour la sélection de modèles de régression.

1. INTRODUCTION

Regression models are important tools for studying the relationships between response variables and predictor variables. Often, there are many predictor variables available to build a regression model but some of these variables may be inactive in the sense that they have no impact on the response. It is important that we identify the true model containing only and all active variables. To set up the notation, consider a full regression model with $p$ predictor variables:

$$E(Y) = g(X, \beta),$$

where $Y$ is an $n$-vector of independent observations of the response variable, $g$ is a given function, $X = [1, x_1, \ldots, x_p]$ is an $n \times (p + 1)$ design matrix, and $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is a vector
of unknown regression parameters. The marginal distributions of \(Y\) are assumed to be of the same type, usually in the exponential family of distributions, and known except for the values of their parameters. There may be other parameters besides \(\beta\) but they are not of interest in the context of model selection. With the above notation, a variable \(x_i\) is said to be active if its parameter \(\beta_i \neq 0\), and inactive if \(\beta_i = 0\). An important special case of (1) is the generalized linear model, \(E(Y) = g(\eta)\), where \(\eta = X\beta\) is the \(n\)-vector of linear predictors and \(g\) is the inverse of the link function.

Let \(\mathcal{M} = \{M_j\}_{j=1}^{2^p}\) be the collection of \(2^p\) subsets of the \(p\) variables in the full model (1) where each \(M_j\) represents a subset. We call each \(M_j\) a model because it defines a reduced model \(E(Y) = g(X_j, \beta_j)\),

where \(X_j\) is the design matrix containing only variables in \(M_j\), and \(\beta_j\) is the parameter vector for the variables in \(X_j\). Denote by \(M_j'\) the subset containing only and all active variables. We call \(M_j'\) the true model. Throughout this article, we adopt the classical parametric setting where (i) \(p\) is fixed and \(n > p\) and (ii) the full regression model (1) is correctly specified but some of its regression parameters \(\beta_i\) may be zero. Assumption (ii) implies that \(M_j' \subset \mathcal{M}\).

There is a large body of literature on model selection. For a comprehensive review, see Ding, Tarokh & Yang (2018a) and Kadane & Lazar (2004). Here, we only briefly review two commonly used model selection criteria, the Akaike information criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz, 1978), which in part motivated this article. The AIC approach does not assume that the full model is correctly specified. It selects the model from a set of candidate models that minimizes the Kullback–Leibler (KL) divergence between the fitted and the unknown true model that generated the data. Denote by \(\ell(\hat{\beta}_j)\) the maximum log-likelihood of model \(M_j\), where \(\hat{\beta}_j\) is the maximum likelihood estimator of \(\beta_j\), and let \(d_j\) be the number of predictor variables in \(M_j\). The AIC of model \(M_j\) is

\[
AIC(M_j) = -2\ell(\hat{\beta}_j) + 2d_j \quad \text{for } j \in \{1, 2, \ldots, 2^p\},
\]

and the model with the smallest AIC value is the model that minimizes the KL divergence in an asymptotic sense. AIC is a penalized measure of fit of \(M_j\) with the fit measured by its log-likelihood and the penalty given by \(2d_j\). For small sample situations, corrected penalty terms have been proposed by several authors including Hurvich & Tsai (1989) and Broersen (2000). The BIC approach tackles the model selection problem from a Bayesian perspective by assuming a constant prior for the candidate models and an arbitrary prior for the parameter vector of each model. It selects the model with the highest posterior probability, which is asymptotically equivalent to selecting the model with the minimum BIC, where

\[
BIC(M_j) = -2\ell(\hat{\beta}_j) + d_j \log(n) \quad \text{for } j \in \{1, 2, \ldots, 2^p\},
\]

which does not depend on the prior of the parameter vector. BIC is also a penalized measure of fit. When the true model is in the set of candidate models, BIC is consistent (Rao & Wu, 1989) whereas AIC is not. In situations where the number of variables \(p\) is large, the constant prior on the space of candidate models is problematic, and Chen & Chen (2008) proposed an extended BIC based on a different prior that gives higher probabilities to smaller models than the constant prior does. Apart from AIC, BIC, and their variants, there are other criteria based on penalized log-likelihood such as the Hannan and Quinn information criterion (Hannan & Quinn, 1979) and the Bridge criterion (Ding, Tarokh & Yang, 2018b). Indeed, the penalized log-likelihood is the most commonly used approach for balancing the fit and the size of a model. The fact that model
selection methods with different motivations have resulted in criteria with similar penalized log-likelihood forms such as (2) and (3) shows the inherent importance of log-likelihood in model selection. Modern $L_p$ penalized methods such as the lasso (Tibshirani, 1996) have further extended this approach to high-dimensional situations where $p$ may be larger than $n$.

The successful applications of log-likelihood in model selection motivated us to look for a model selection strategy based on the closely related log-likelihood ratio. Let $\lambda(\hat{\beta}_j)$ be the maximum log-likelihood ratio of model $M_j$,

$$\lambda(\hat{\beta}_j) = -2\{\ell(\hat{\beta}_j) - \ell(\hat{\beta})\},$$

where $\hat{\beta}$ is the maximum likelihood estimator of the full model with all $p$ variables, and $\ell(\hat{\beta})$ is the maximum log-likelihood of the full model. The ratio $\lambda(\hat{\beta}_j)$ provides a relative measure of fit of model $M_j$ with respect to the full model. It has an important advantage over the log-likelihood $\ell(\hat{\beta}_j)$ in that its value may be directly used to evaluate the plausibility of model $M_j$ because the asymptotic distribution of $\lambda(\hat{\beta})$ under the null hypothesis is known to be a $\chi^2$ distribution, whereas the value of the log-likelihood of a model alone does not carry information about the plausibility of the model. To use $\lambda(\hat{\beta}_j)$ for model selection, instead of penalizing it with a penalty term proportional to the model size, we take advantage of the known asymptotic distribution of $\lambda(\beta)$ to look for a set of plausible models using the likelihood ratio test at a given significance level $\alpha$. Then, from this set of models we select the smallest model. This amounts to giving the fit of the model (as represented by the log-likelihood ratio) a higher priority over the size of the model, and minimizing the size subject to a lower bound on the fit. We refer to this approach as the constrained minimum criterion (CMC) for model selection.

Tsao (2021) studied CMC for selecting Gaussian linear models under an approximated likelihood ratio test with a significance level depending on $n$ and showed that the method is consistent for selecting Gaussian linear models. The present article uses the exact likelihood ratio test with a fixed significance level $\alpha$ and generalizes CMC to all regression models. When $n$ is large and $\alpha$ is small, we show that there is a high probability that the smallest model in the set of plausible models is the true model. This provides an asymptotic justification for using CMC to select regression models. In real applications, however, $n$ may not be very large and the selection of the $\alpha$ value needs to be guided by the finite sample accuracy of the resulting CMC. We will identify a default $\alpha$ value at which CMC often outperforms AIC and BIC in terms of accuracy as measured by false active and false inactive rates. Having $\alpha$ as a parameter also gives CMC the ability to easily handle special situations such as when $n$ is small. We will discuss how to select the $\alpha$ value for such situations.

The rest of this article is organized as follows. In Section 2, we present CMC based on the likelihood ratio test for selecting regression models. In Section 3, we compare this method to AIC and BIC in terms of selection accuracy in a simulation study with examples of linear, logistic, and Poisson regression models. We also discuss the selection of the significance level $\alpha$. In Section 4, we apply CMC to perform model selection for logistic regression for a South African heart disease dataset. We conclude with a few remarks in Section 5.

2. THE CONSTRAINED MINIMUM CRITERION

For clarity of notation, from now on we will use $\beta'$ to denote the true value of the regression parameter vector for the full model (1), so $\beta'$ is a $(p + 1)$-vector and its elements that correspond to inactive variables are all zero. The notation $\beta$ now denotes a general $(p + 1)$-vector. For simplicity, we make the following three assumptions for all regression models under consideration. The first

DOI: 10.1002/cjs.11756 The Canadian Journal of Statistics / La revue canadienne de statistique
The assumption is that the maximum likelihood estimator for the parameter vector of the full model \( \hat{\beta} \) is consistent, that is,
\[
\hat{\beta} \overset{p}{\rightarrow} \beta' \quad \text{as} \quad n \to \infty,
\]
(5)
where \( \overset{p}{\rightarrow} \) denotes convergence in probability. The second assumption is that the distribution of the log-likelihood ratio evaluated at the true value \( \beta' \) converges to a \( \chi^2_{p+1} \) distribution, that is,
\[
\lambda(\beta') = -2 \left\{ \ell(\beta') - \ell(\hat{\beta}) \right\} \Rightarrow \chi^2_{p+1} \quad \text{as} \quad n \to \infty,
\]
(6)
where \( \Rightarrow \) denotes convergence in distribution. By assumption (6), for any fixed \( \alpha \in (0, 1) \), a 100(1 - \( \alpha \))% asymptotic confidence region for \( \beta' \) is
\[
C_{1-\alpha} = \left\{ \beta \in \mathbb{R}^{p+1} : \lambda(\beta) \leq \chi^2_{1-\alpha,p+1} \right\},
\]
(7)
where \( \chi^2_{1-\alpha,p+1} \) denotes the \((1 - \alpha)\)th quantile of the \( \chi^2_{p+1} \) distribution. The centre of this \((p + 1)\)-dimensional confidence region is \( \hat{\beta} \) because \( \lambda(\hat{\beta}) = 0 \) is the smallest value of \( \lambda(\beta) \). The third assumption is that the size of the confidence region \( C_{1-\alpha} \) goes to zero as \( n \) goes to infinity in the sense that
\[
\max_{\beta \in C_{1-\alpha}} \| \beta - \hat{\beta} \|_2 = o_p(1),
\]
(8)
where \( \| \cdot \|_2 \) denotes the Euclidean norm. For commonly used regression models, regularity conditions for the asymptotic normality of the maximum likelihood estimator \( \hat{\beta} \) of the full model are available in the literature. It may be verified that assumptions (5), (6), and (8) all hold under these conditions. A stronger version of assumption (8),
\[
\max_{\beta \in C_{1-\alpha}} \| \beta - \hat{\beta} \|_2 = O_p\left(n^{-1/2}\right),
\]
also holds under these conditions, but the weaker assumption (8) is already sufficient for our subsequent discussions. As an example of such regularity conditions, for linear regression models, a commonly used pair of such conditions is
\[
\frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \to D,
\]
where \( x_i \) is the \( i \)th row of \( X \), and \( D \) is a fixed \((p + 1) \times (p + 1)\) positive definite matrix, and
\[
\frac{1}{n} \max_{1 \leq i \leq n} x_i^T x_i \to 0.
\]
For generalized linear models, such regularity conditions may be found in Haberman (1977), Gouriéroux & Monfort (1981), and Fahrmeir & Kaufmann (1985). Assuming all of (5), (6), and (8) amounts to assuming that such conditions hold.

The confidence region \( C_{1-\alpha} \) contains the collection of \( \beta \in \mathbb{R}^{p+1} \) not rejected by the likelihood ratio test of \( H_0 : \beta' = \beta' \) at the given \( \alpha \) level. As such, it represents the set of plausible \((p + 1)\)-vectors at the \( \alpha \) level. To extend the notion of plausibility from a \((p + 1)\)-vector to a model \( M_j \), we first find a \((p + 1)\)-vector to represent model \( M_j \). The maximum likelihood estimator \( \hat{\beta}_j \) for \( M_j \) is a vector of dimension \((d_j + 1)\), which is less than \((p + 1)\) when \( M_j \) is not the full model. The estimator is usually a continuous random vector, so with probability 1 all of its elements are
nonzero. We augment the dimension of $\hat{\beta}_j$ by adding $(p - d_j)$ zeros to its elements to represent the $(p - d_j)$ variables not in $M_j$. For example, if $x_1$ is not in $M_j$, then the second element of the augmented $\hat{\beta}_j$ (which corresponds to $x_1$) is a zero. For simplicity, we still use the same notation $\hat{\beta}_j$ for the augmented vector, but it is now a $(p + 1)$-vector representing $M_j$ and its nonzero elements correspond to the intercept and variables in $M_j$. We say that model $M_j$ is plausible at the $\alpha$ level if $\hat{\beta}_j$ is in the confidence region $C_{1-\alpha}$. Alternatively, we may also say that $M_j$ is plausible if $\lambda(\hat{\beta}_j)$ is less than $\chi^2_{1-a,p+1}$. Note that although we need the augmented $(p + 1)$-dimensional version of $\hat{\beta}_j$ to define the plausibility of its corresponding model $M_j$, when computing the maximum log-likelihood ratio of this model $\lambda(\hat{\beta}_j)$, $\hat{\beta}_j$ may be either the augmented $(p + 1)$-dimensional version or the original $(d_j + 1)$-dimensional version, as they both give the same value of $\lambda(\hat{\beta}_j)$.

In numerical computations of $\lambda(\hat{\beta}_j)$, we use the $(d_j + 1)$-dimensional version as it appears in (4), which is more convenient. Using the $L_0$ norm, which counts the number of nonzero elements in a vector, the constrained minimum criterion based on the likelihood ratio test chooses the model represented by the solution to the following constrained minimization problem:

$$\min_{\hat{\beta}_j} ||\hat{\beta}_j||_0 \quad \text{subject to} \quad \hat{\beta}_j \in C_{1-\alpha}. \tag{9}$$

We call the solution vector to this minimization problem the CMC solution and its corresponding model the CMC selection. When there are two or more solution vectors, we choose the one with the highest likelihood as the CMC solution.

The $L_p$ penalized methods such as the lasso and ridge regression can also be formulated as constrained minimization problems. To give a geometric comparison between CMC and these methods, Figure 1 illustrates the differences between the CMC solution, the lasso solution, and the ridge solution for an example with a full model containing two predictor variables but no intercept. This figure is based on Figure 2.2 in Hastie, Tibshirani & Wainwright (2015), which

![Figure 1: Geometric comparison between CMC and $L_p$ penalized methods. The red elliptical contours represent the likelihood ratio confidence region $C_{1-\alpha}$ of different confidence levels. For this illustration, the confidence region represented by the outer contour is used. The blue square-shaped region in (a) is the constraint region of the lasso. The blue circular region in (b) is the constraint region of the ridge regression. Green arrows point to CMC solutions. The purple arrow in (a) points to the lasso solution, and the purple arrow in (b) points to the ridge regression solution.](image-url)
compares the lasso with the ridge regression. Figure 1a shows that the lasso solution is the point where the outer contour meets the square-shaped constraint region, which is the top vertex of this region. The CMC solution is the point with the highest likelihood in the intersection between the $\beta_2$-axis and the confidence region represented by the outer contour. Figure 1b shows that the ridge solution is the point where the outer contour meets the circular constraint region. The CMC solution in this case is the point where the outer contour meets the $\beta_2$-axis.

Figure 1 highlights a key difference between how and where CMC and the $L_p$ penalized methods select a model: CMC selects the smallest model from $C_{1-a}$, a region centred at the maximum likelihood estimator $\hat{\beta}$, representing the best set of models (see its properties below); whereas the $L_p$ penalized method selects the best model from a convenient region centred at the origin. Here, “best” means having the highest likelihood, and “convenient” refers to the $L_p$ constraint regions having simple geometric shapes centred at the origin and simple algebraic representations ($\sum |\beta_i| \leq t$ for the lasso and $\sum \beta_i^2 \leq t$ for the ridge regression). These convenient regions also lead to ease of computation. For example, for linear models, the analytic solution of the ridge regression is available, and the lasso solution can be easily computed because of the convexity of its constraint region. The corners of the lasso constraint region also give it an added benefit as an automatic variable selector. Nevertheless, for the low-dimensional problem we study in this article, $C_{1-a}$ is a better constraint region from which to select a model, as it has the following three properties:

(P1) $C_{1-a}$ consists of $\beta$ with the highest likelihood values.

(P2) All $\beta \in C_{1-a}$ converge to $\beta'$ in probability.

(P3) The sensitivity of the region $C_{1-a}$ goes to 1 in that $P$ (elements of $\beta \in C_{1-a}$ corresponding to active variables are all nonzero) approaches 1 as $n \to \infty$. This implies that if an element of $\beta$ in $C_{1-a}$ is zero, then the corresponding variable is likely inactive, which supports CMC as it selects the smallest $\beta$ (i.e., the $\beta$ with the most zeros) in $C_{1-a}$ to minimize the number of selected inactive variables.

Property (P1) comes from the construction of $C_{1-a}$. Properties (P2) and (P3) will be proved later with Theorem 2.1. These properties show that although $C_{1-a}$ does not have a simple shape or algebraic representation, it contains vectors with desirable and well-understood properties that support CMC. In contrast, the $L_p$ constraint regions do not have these properties.

Denote by $\hat{\beta}_j^*$ the maximum likelihood estimator for the unknown true model $M_j^*$. The nonzero elements of this (augmented) $(p+1)$-vector $\hat{\beta}_j^*$ are those corresponding to the active variables and the intercept, and the zero elements of $\hat{\beta}_j^*$ are those corresponding to the inactive variables. Theorem 2.1 below gives the asymptotic properties of the CMC solution and selection.

Theorem 2.1. Suppose Assumptions (5), (6), and (8) hold. For a given $\alpha \in (0, 1)$, let $\hat{\beta}_a$ be the CMC solution of (9) and $\hat{M}_a$ be the corresponding CMC selection. Then, (i) the CMC solution is consistent in that

$$\hat{\beta}_a \xrightarrow{p} \beta' \quad \text{as } n \to \infty, \quad (10)$$

and (ii) the probability that $\hat{M}_a$ is the true model has an asymptotic lower bound

$$\lim_{n \to +\infty} P(\hat{M}_a = M_j^*) \geq 1 - \alpha. \quad (11)$$

The asymptotic lower bound (11) shows that when the sample size $n$ is large, we may choose a small $\alpha$ so that there is a high probability that the CMC selection is the true model. Numerical results show that the lower bound $(1 - \alpha)$ is loose for many $\alpha$ values when $n$ is large in that the
observed probability of the event \( \{ \hat{M}_a = M'_j \} \) is usually much larger than \((1 - \alpha)\). Also, when \( n \) is not large, small \( \alpha \) levels are not appropriate since they may lead to high false inactive rates as can be seen in examples in the next section. We will discuss the selection of the \( \alpha \) level through these examples. We now prove the theorem.

**Proof of Theorem 2.1.** By (5), we have \( \| \hat{\beta} - \beta' \|_2 = o_p(1) \). Since \( \hat{\beta}_a \in C_{1-\alpha} \), by (8) we have \( \| \hat{\beta}_a - \beta \|_2 \leq \max_{\beta \in C_{1-\alpha}} \| \beta - \beta \|_2 = o_p(1) \), which implies \( \| \hat{\beta}_a - \beta \|_2 = o_p(1) \). It follows from these and the triangle inequality that

\[
\| \hat{\beta}_a - \beta' \|_2 \leq \| \hat{\beta}_a - \beta \|_2 + \| \beta - \beta' \|_2 = o_p(1),
\]

which implies the consistency of the CMC solution (10).

To prove the asymptotic lower bound in (11), note that

\[
P(\hat{M}_a = M'_j) = P(\hat{\beta}_a = \hat{\beta}_j)
\]

as events \( \{ \hat{M}_a = M'_j \} \Leftrightarrow \{ \hat{\beta}_a = \hat{\beta}_j \} \), so it suffices to show that \( P(\hat{\beta}_a = \hat{\beta}_j) \) has the asymptotic lower bound in (11). To this end, we first identify the elements of vectors \( \beta \) in \( C_{1-\alpha} \) that may not be zero when \( n \) is large. By definition, elements of \( \beta' \) corresponding to active variables are nonzero. We now show that this also holds for \( \beta \) in \( C_{1-\alpha} \) with a high probability when \( n \) is large. Define an event

\[
A = \{ \text{Elements of } \beta \text{ in } C_{1-\alpha} \text{ corresponding to active variables are nonzero} \}.
\]

Similar to (12), by the triangle inequality and (8), we have \( \| \beta - \beta' \|_2 = o_p(1) \) uniformly for all \( \beta \in C_{1-\alpha} \). This proves property \( (P_2) \) of \( C_{1-\alpha} \) and implies that individual elements of \( \beta \) converge in probability to the corresponding elements of \( \beta' \) uniformly. Since elements of \( \beta' \) corresponding to active variables are nonzero, it follows that elements of \( \beta \) corresponding to active variables converge in probability to nonzero values uniformly, which implies that \( P(A) \to 1 \) as \( n \to \infty \). Incidentally, \( P(A) \to 1 \) implies property \( (P_3) \) of \( C_{1-\alpha} \). It also implies that the probability that all active variables are selected goes to 1. This is because all active variables are selected whenever all elements of \( \hat{\beta}_a \) corresponding to active variables are nonzero, which happens when event \( A \) occurs.

When event \( A \) occurs, among the set of vectors \( \{ \hat{\beta}_j \}_{j=1}^{2p} \), only those vectors for models containing all active variables can be in \( C_{1-\alpha} \), so \( \hat{\beta}_j \) of the true model \( M'_j \) is the smallest (in \( L_0 \) norm) member of \( \{ \hat{\beta}_j \}_{j=1}^{2p} \) that can possibly be in \( C_{1-\alpha} \). It follows that \( \{ \hat{\beta}_j \in C_{1-\alpha} \} \cap A \) implies \( \{ \hat{\beta}_a = \hat{\beta}_j \} \), so

\[
P(\hat{\beta}_a = \hat{\beta}_j) \geq P(\{ \hat{\beta}_j \in C_{1-\alpha} \} \cap A) \to P(\hat{\beta}_j \in C_{1-\alpha})
\]

as \( n \) goes to infinity. Also, the event \( \{ \beta' \in C_{1-\alpha} \} \) implies \( \{ \hat{\beta}_j \in C_{1-\alpha} \} \) because \( \hat{\beta}_j \) is the maximum likelihood estimator for \( M'_j \) which has a higher likelihood than \( \beta' \) and thus a smaller log-likelihood ratio than \( \beta' \); that is, \( \lambda(\hat{\beta}_j) < \lambda(\beta') \) and thus \( \{ \beta' \in C_{1-\alpha} \} \) implies \( \{ \hat{\beta}_j \in C_{1-\alpha} \} \). This and (6) imply that

\[
P(\hat{\beta}_j \in C_{1-\alpha}) \geq P(\beta' \in C_{1-\alpha}) \to 1 - \alpha
\]

as \( n \) goes to infinity. Equations (13), (14), and (15) then imply (11).
The above proof follows similar steps as the proof of consistency of CMC for Gaussian linear models in Tsao (2021). However, CMC for Gaussian linear models in that paper is based on an approximated likelihood ratio statistic with a known finite sample distribution. Its \( \alpha \) level is not fixed and goes to zero as \( n \) goes to infinity. The proof of its consistency made use of the availability of the finite sample distribution and a decreasing \( \alpha \). In the present case, \( \alpha \) is fixed, the finite sample distribution of the likelihood ratio statistic is unknown, and we have only the asymptotic distribution (6). Consequently, we are only able to prove the lower bound (11) for the present version of CMC, which is a weaker result than consistency. Nevertheless, this does not affect the accuracy of the present version in the numerical examples considered here; the results show that it is just as accurate as the consistent version for Gaussian linear models (see numerical examples in Section 3). Furthermore, the present version can be applied to all types of regression models satisfying the three assumptions, not just to Gaussian linear models.

Finally, we have presented CMC in the context of the best subset selection from \( \mathcal{M} \) containing all \( 2^p \) models. In practice, we may only need to consider a subset \( \mathcal{M}^* \subset \mathcal{M} \). To use CMC to select a model from \( \mathcal{M}^* \), we simply minimize \( \|\hat{\beta}_j\|_0 \) over \( \mathcal{M}^* \) instead of \( \mathcal{M} \) in (9) with the same constraint, and Theorem 2.1 remains valid so long as the assumption that \( M^*_j \in \mathcal{M}^* \) holds.

3. NUMERICAL COMPARISON OF THE CMC, AIC, AND BIC

We now compare CMC based on the likelihood ratio test (9) with AIC and BIC in terms of false active rate (FAR) and false inactive rate (FIR) through numerical examples. We also discuss the selection of the level \( \alpha \) for CMC. Here, FAR is the number of inactive variables appearing in the selected model divided by the total number of inactive variables in the full model, and FIR is the number of active variables not in the selected model divided by the total number of active variables in the full model. A model selection criterion is accurate when the FIR and FAR of its selected model are both low. To compute the examples, we use the R package “bestglm” by McLeod, Xu & Lai (2020), which performs the best subset selection for generalized linear models. For the best subset selection of Gaussian linear models, “bestglm” uses the “leaps and bounds” algorithm by Furnival & Wilson Jr. (1974), which can handle situations with 40 or fewer predictor variables. For the best subset selection of logistic regression models and Poisson regression models, “bestglm” uses a complete enumeration method by Morgan & Tatar (1972) and has a limit of 15 on the number of predictor variables allowed in the full model. In our examples, we set the number of predictor variables below the limit to a maximum of 30 for linear models and 10 for the two generalized linear models to avoid long simulation times.

3.1. Linear Model Examples

The linear model used for comparison is

\[ y = X\beta + \epsilon, \]  

where \( \epsilon \sim N(0, \sigma^2 I) \) with \( \sigma^2 = 1 \) and \( I \) being the \( n \times n \) identity matrix, \( X = [1, x_1, \ldots, x_p] \), and \( \beta = (1, \beta_1, \ldots, \beta_p^*, 0, \ldots, 0)^T \) with \( \beta_1 = \cdots = \beta_{p^*} = 1 \), so only the first \( p^* \) variables are active. Elements of all \( x_i \) are independent random numbers generated from the standard normal distribution. The performance of CMC depends on the level \( \alpha \). To find the appropriate levels for different sample sizes, we consider three levels, \( \alpha = 0.1, 0.5, \) and 0.9.

Table 1 contains simulated values of the (FIR, FAR) pairs for five model selection criteria, namely AIC, BIC, CMC_0.9, CMC_0.5, and CMC_0.1, at 12 different combinations of \( n, p, \) and \( p^* \). The subscript \( \alpha \) in CMC_\( \alpha \) indicates the level \( \alpha \) used. Each (FIR, FAR) pair in the table is based on 1000 simulation runs. For each run, we first generate an \((X, y)\) pair, and then perform the best subset selection using \((X, y)\) and the five criteria to find the models chosen by these criteria. For
TABLE 1: Model selection accuracy comparison for Gaussian linear models: the entries are simulated (FIR, FAR) values for AIC, BIC, and three CMC criteria in 12 scenarios.

| (n, p, p*) | AIC   | BIC   | CMC0.9 | CMC0.5 | CMC0.1 |
|------------|-------|-------|--------|--------|--------|
| (20, 10, 5)| (0.04, 0.34) | (0.05, 0.24) | (0.05, 0.25) | **(0.09, 0.13)** | (0.21, 0.06) |
| (30, 10, 5)| (0.00, 0.25) | (0.01, 0.12) | (0.01, 0.16) | **(0.02, 0.05)** | (0.09, 0.01) |
| (40, 10, 5)| (0.00, 0.24) | (0.00, 0.09) | (0.00, 0.13) | **(0.00, 0.04)** | (0.03, 0.01) |
| (50, 10, 5)| (0.00, 0.22) | (0.00, 0.08) | (0.00, 0.12) | **(0.00, 0.03)** | (0.01, 0.00) |
| (40, 20, 10)| (0.00, 0.32) | (0.00, 0.15) | (0.01, 0.12) | **(0.02, 0.05)** | (0.06, 0.02) |
| (60, 20, 10)| (0.00, 0.25) | (0.00, 0.09) | (0.00, 0.08) | **(0.00, 0.02)** | (0.01, 0.00) |
| (80, 20, 10)| (0.00, 0.21) | (0.00, 0.06) | (0.00, 0.05) | **(0.00, 0.01)** | (0.00, 0.00) |
| (100, 20, 10)| (0.00, 0.20) | (0.00, 0.05) | (0.00, 0.05) | **(0.00, 0.01)** | (0.00, 0.00) |
| (60, 30, 15)| (0.00, 0.31) | (0.00, 0.12) | (0.00, 0.08) | **(0.00, 0.03)** | (0.02, 0.01) |
| (90, 30, 15)| (0.00, 0.23) | (0.00, 0.07) | (0.00, 0.04) | **(0.01, 0.01)** | (0.01, 0.00) |
| (120, 30, 15)| (0.00, 0.21) | (0.00, 0.05) | (0.00, 0.03) | **(0.00, 0.00)** | (0.00, 0.00) |
| (150, 30, 15)| (0.00, 0.20) | (0.00, 0.04) | (0.00, 0.02) | **(0.00, 0.00)** | (0.00, 0.00) |

Note: The bold CMC results are those at the recommended $\alpha$ level. Results in the table are averages of the two rates for 1000 simulation runs rounded to the second digit after the decimal point.

* DOI: 10.1002/cjs.11756

The Canadian Journal of Statistics / La revue canadienne de statistique
Model (16) was also used to evaluate the consistent CMC for Gaussian linear models in Table 1 of Tsao (2021). The CMC results in Table 1 of Tsao (2021) differ from the CMC results in Table 1 here, especially for cases where \( n = 2p \). These differences are due to the fact that two different tests were used in the formulation of CMC. The tests are asymptotically equivalent, so for large sample sizes \( (n > 3p) \) the CMC results in both tables are very similar. In the examples reported here, we set \( p^* = p/2 \) so that there is an equal number of active and inactive variables, which makes FIR + FAR a meaningful measure of the overall error. For simplicity of presentation, we also set the parameter values of all active variables to 1. We have tried \( p^* \) values not equal to \( p/2 \) and parameter values randomly generated from a uniform distribution on \((0, 1)\) and obtained similar observations concerning the relative performance of the five criteria.

3.2. Logistic Regression Examples

Let \( Y_1, Y_2, \ldots, Y_n \) be \( n \) independent observations of the response variable where \( Y_i \sim \text{Binomial}(m, \pi_i) \), and let \( X = [1, x_1, \ldots, x_p] \) be the corresponding \( n \times (p+1) \) matrix of predictor variables. The logistic regression model is given by

\[
\logit(\pi_i) = x_i \beta,
\]

or alternatively,

\[
\pi_i = \frac{\exp(x_i \beta)}{1 + \exp(x_i \beta)},
\]

where \( x_i \) is the \( i \)th row of \( X \), and \( \beta = (1, \beta_1, \ldots, \beta_p, 0, \ldots, 0)^T \). As in the linear model examples, we set \( \beta_1 = \cdots = \beta_{p^*} = 1 \) and \( p^* = p/2 \), so that only the first half of the variables are active, and elements of all predictor variables \( x_i \) are independent random numbers generated from the standard normal distribution. The sample size here depends on both \( n \) and \( m \), so we used different combinations of \( n \) and \( m \) in the simulation. Table 2 contains the (FIR, FAR) values of the five criteria for 16 combinations of \((n, m, p)\) where each value is the average of 1000 simulated pairs.

We make the following observations about Table 2:

1. AIC has the lowest FIR but the highest FAR for all combinations of \((n, m, p, p^*)\). Because of its high FAR, its overall error rate is, in general, the highest among the five criteria. BIC has much lower FAR than AIC. It is consistent but its FAR converges to zero slowly; it is still about 5% even when \( n \) and \( m \) are at their highest values of 50 and 10, respectively.

2. We had noted that, for Gaussian linear models, the performance of CMC_{0.9} is similar to that of BIC with low FIR. However, for logistic regression models, CMC_{0.9} behaves more like AIC with similar FIR and FAR, especially for cases where \((p, p^*) = (6, 3)\). When \((p, p^*) = (10, 5)\), it has smaller FAR than that of AIC, and behaves more like BIC.

3. For Gaussian linear models, we recommended the 0.5 level as the default \( \alpha \) level for CMC. For logistic regression model selection, both \( n \) and \( m \) affect the accuracy of CMC. Interestingly, however, through exploring a wide range of \( n \) and \( m \) combinations we found that CMC_{0.5} again has stable performance and is usually the most or the second most accurate criterion among the five criteria. We therefore also recommend the 0.5 level as the default level for selecting logistic regression models. When \( n \times m \) is much larger than \( p \), CMC_{0.1} may be used. Table 2 has such cases where \( n \times m \) is 50 times larger than \( p \), and for these cases CMC_{0.1} reached zero error rates, suggesting that CMC with decreasing \( \alpha \) is also consistent for selecting logistic regression models.
Table 2: Model selection accuracy comparison for logistic regression models: the entries are simulated (FIR, FAR) values for AIC, BIC, and three CMC criteria in 16 scenarios.

| (n, m, p, p*) | AIC     | BIC     | CMC<sub>0.9</sub> | CMC<sub>0.5</sub> | CMC<sub>0.1</sub> |
|--------------|---------|---------|-------------------|-------------------|-------------------|
| (20, 5, 6, 3)| (0.06, 0.20) | (0.10, 0.11) | (0.06, 0.20) | (0.14, 0.07) | (0.30, 0.03) |
| (30, 5, 6, 3)| (0.01, 0.17) | (0.02, 0.08) | (0.01, 0.17) | (0.03, 0.05) | (0.12, 0.01) |
| (40, 5, 6, 3)| (0.00, 0.17) | (0.00, 0.07) | (0.00, 0.16) | (0.01, 0.04) | (0.03, 0.00) |
| (50, 5, 6, 3)| (0.00, 0.16) | (0.00, 0.06) | (0.00, 0.16) | (0.00, 0.04) | (0.01, 0.00) |
| (20, 10, 6, 3)| (0.01, 0.17) | (0.01, 0.10) | (0.01, 0.17) | (0.02, 0.05) | (0.08, 0.01) |
| (30, 10, 6, 3)| (0.00, 0.16) | (0.00, 0.07) | (0.00, 0.16) | (0.00, 0.04) | (0.01, 0.00) |
| (40, 10, 6, 3)| (0.00, 0.16) | (0.00, 0.06) | (0.00, 0.16) | (0.00, 0.04) | (0.00, 0.00) |
| (50, 10, 6, 3)| (0.00, 0.16) | (0.00, 0.05) | (0.00, 0.16) | (0.00, 0.03) | (0.00, 0.00) |
| (20, 5, 10, 5)| (0.17, 0.26) | (0.22, 0.17) | (0.19, 0.20) | (0.30, 0.11) | (0.42, 0.06) |
| (30, 5, 10, 5)| (0.04, 0.19) | (0.07, 0.10) | (0.06, 0.13) | (0.12, 0.05) | (0.24, 0.02) |
| (40, 5, 10, 5)| (0.00, 0.18) | (0.02, 0.07) | (0.01, 0.10) | (0.04, 0.03) | (0.13, 0.01) |
| (50, 5, 10, 5)| (0.00, 0.16) | (0.00, 0.05) | (0.00, 0.08) | (0.01, 0.02) | (0.07, 0.00) |
| (20, 10, 10, 5)| (0.06, 0.19) | (0.08, 0.11) | (0.08, 0.13) | (0.15, 0.06) | (0.26, 0.04) |
| (30, 10, 10, 5)| (0.00, 0.17) | (0.01, 0.07) | (0.00, 0.10) | (0.01, 0.02) | (0.06, 0.01) |
| (40, 10, 10, 5)| (0.00, 0.15) | (0.00, 0.06) | (0.00, 0.07) | (0.00, 0.02) | (0.02, 0.00) |
| (50, 10, 10, 5)| (0.00, 0.14) | (0.00, 0.05) | (0.00, 0.07) | (0.00, 0.02) | (0.00, 0.00) |

Note: The bold CMC results are those at the recommended \( \alpha \) level. Results in the table are averages of the two rates for 1000 simulation runs rounded to the second digit after the decimal point.

3.3. Poisson Regression Examples

Let \( Y_1, Y_2, \ldots, Y_n \) be \( n \) independent observations of the response variable where \( Y_i \sim \text{Poisson}(\mu_i) \). The Poisson regression model with log link is

\[
\ln(\mu_i) = x_i \beta, \tag{18}
\]

where \( x_i \) is the \( i \)th row of the \( n \times (p + 1) \) matrix \( X = [1, x_1, \ldots, x_p] \) of predictor variables, and \( \beta \) is the vector of regression parameters. We set \( \beta = (1, \beta_1, \ldots, \beta_p, 0, \ldots, 0)^T \) with \( p^* = p/2 \) and \( \beta_1 = \cdots = \beta_{p^*} = 0.5 \). Table 3 contains the simulated (FIR, FAR) results for the five criteria. In the case of the logistic regression models in Table 2, the performance of CMC<sub>0.9</sub> is similar in most cases to that of AIC which has the lowest FIR but the highest FAR. BIC has slightly higher FIR than AIC and CMC<sub>0.9</sub> but lower FAR. On the relative performance of the three CMC criteria, CMC<sub>0.9</sub> has lower overall error when the sample size \( n \) is small (\( n \leq 2p \)). When \( 2p < n \leq 4p \), CMC<sub>0.5</sub> is usually the most accurate. When \( n > 4p \), CMC<sub>0.1</sub> is usually the most accurate but CMC<sub>0.5</sub> is a close second. Based on these findings and for simplicity, we again recommend the 0.5 level as the default level of \( \alpha \). For small sample sizes, the 0.9 level may be used. For very large sample sizes, the 0.1 level may be used.

To summarize the numerical comparison, recommendations on the \( \alpha \) level of CMC that we have made above are based on the objective of minimizing the overall error. For fixed \( n \) and \( p \), the FIR of CMC decreases and the FAR increases when \( \alpha \) increases. This gives users of CMC control of the balance between these two rates through the choice of the \( \alpha \) level. If a
Table 3: Model selection accuracy comparison for Poisson regression models: the entries are simulated (FIR, FAR) values for AIC, BIC, and three CMC criteria in 10 scenarios.

| (n, p, p*) | AIC   | BIC   | CMC₀.₉ | CMC₀.₅ | CMC₀.₁ |
|------------|-------|-------|--------|--------|--------|
| (20, 6, 3) | (0.06, 0.19) | (0.09, 0.10) | (0.05, 0.20) | (0.12, 0.06) | (0.28, 0.03) |
| (30, 6, 3) | (0.01, 0.16) | (0.02, 0.07) | (0.01, 0.16) | (0.02, 0.03) | (0.08, 0.01) |
| (40, 6, 3) | (0.00, 0.16) | (0.01, 0.06) | (0.00, 0.17) | (0.01, 0.03) | (0.03, 0.00) |
| (50, 6, 3) | (0.00, 0.15) | (0.00, 0.05) | (0.00, 0.16) | (0.00, 0.03) | (0.01, 0.00) |
| (100, 6, 3) | (0.00, 0.15) | (0.00, 0.03) | (0.00, 0.15) | (0.00, 0.02) | (0.00, 0.00) |
| (20, 10, 5) | (0.13, 0.20) | (0.17, 0.14) | (0.16, 0.15) | (0.26, 0.08) | (0.39, 0.06) |
| (30, 10, 5) | (0.01, 0.16) | (0.03, 0.07) | (0.02, 0.09) | (0.06, 0.03) | (0.15, 0.02) |
| (40, 10, 5) | (0.00, 0.16) | (0.00, 0.06) | (0.00, 0.08) | (0.01, 0.02) | (0.05, 0.00) |
| (50, 10, 5) | (0.00, 0.16) | (0.00, 0.05) | (0.00, 0.08) | (0.00, 0.01) | (0.01, 0.00) |
| (100, 10, 5) | (0.00, 0.16) | (0.00, 0.03) | (0.00, 0.07) | (0.00, 0.01) | (0.00, 0.00) |

Note: The bold CMC results are those at the recommended 𝛼 level. Results in the table are averages of the two rates for 1000 simulation runs rounded to the second digit after the decimal point.

Low FAR is the priority instead of a lower overall error, one can set 𝛼 to 0.1 regardless of the sample size n and dimension p. If a low FIR is the priority, one can set 𝛼 to 0.9. We have only considered three 𝛼 levels here, but other levels may also be used. For example, in Table 1 for the linear model (16), the lowest 𝛼 level is 0.1. When n ≫ p such as (n, p) = (200, 20), even smaller 𝛼 levels such as 0.05 may be used; we tried CMC₀.₀₅ for this case and obtained error rates of zero. Finally, we note that the predictor variables in the above examples have low correlations as they are independently generated. When there are strongly correlated predictor variables, simulation results (not included here) show that CMC₀.₅ may be more accurate than CMC₀.₉ and CMC₀.₁ for small and moderate sample sizes. Nevertheless, CMC₀.₅ is still often the most or second most accurate, and it is often more accurate than AIC and BIC. Because of these, we recommend the 0.5 level as the default regardless of the type of regression model, the sample size, and the correlation situation of the predictor variables. This makes the application of CMC straightforward, as the user does not have to spend time deciding which level of 𝛼 to use. However, to optimize CMC, one may consider a different level depending on the sample size, the correlation situation, and the signal-to-noise ratio.

3.4. Robustness Against Departures from Model Assumptions

AIC, BIC, and CMC all select from a set of parametric candidate models. It is important to investigate their robustness against departures from the parametric assumptions. We now examine their robustness against misspecification of the error distribution for the linear model (16) through a small simulation study. We apply the three criteria to select models from the set of 2^p Gaussian linear models when the true error distributions are standardized t_5, x_5^2, and EXP(1), respectively, representing error distributions that are symmetric non-normal or skewed with the same mean and variance as N(0, 1) in the Gaussian model (16). Table 4 contains the FAR and FIR results for these three error distributions at p = 20 and p* = 10. Interestingly, results in these three cases are almost the same. Comparing these results to those for the normal error in the middle block of Table 1, we see that they are also nearly the same. These findings suggest that the three criteria are robust against departures from the normal error assumption in the linear model in that their performance in terms of FAR and FIR is little affected by such departures.
Table 4: Model selection accuracy for Gaussian linear models when the true error distributions are the standardized $t_5$, $\chi^2_5$, and EXP(1) distributions. The entries are the simulated (FIR, FAR) values for AIC, BIC, and three CMC criteria.

| ERROR   | (n, p, $p^*$) | AIC     | BIC     | CMC0.9   | CMC0.5   | CMC0.1   |
|---------|---------------|---------|---------|----------|----------|----------|
| $t_5$   | (40, 20, 10)  | (0.00, 0.31) | (0.01, 0.15) | (0.01, 0.12) | **(0.02, 0.05)** | (0.06, 0.02) |
|         | (60, 20, 10)  | (0.00, 0.25) | (0.00, 0.08) | (0.00, 0.07) | **(0.00, 0.02)** | (0.01, 0.00) |
|         | (80, 20, 10)  | (0.00, 0.21) | (0.00, 0.06) | (0.00, 0.06) | **(0.00, 0.01)** | (0.00, 0.00) |
|         | (100, 20, 10) | (0.00, 0.20) | (0.00, 0.05) | (0.00, 0.05) | **(0.00, 0.01)** | (0.00, 0.00) |
| $\chi^2_5$ | (40, 20, 10)  | (0.00, 0.32) | (0.01, 0.16) | (0.01, 0.14) | **(0.02, 0.06)** | (0.05, 0.02) |
|         | (60, 20, 10)  | (0.00, 0.24) | (0.00, 0.08) | (0.00, 0.07) | **(0.00, 0.02)** | (0.02, 0.00) |
|         | (80, 20, 10)  | (0.00, 0.22) | (0.00, 0.06) | (0.00, 0.05) | **(0.00, 0.01)** | (0.00, 0.00) |
|         | (100, 20, 10) | (0.00, 0.20) | (0.00, 0.05) | (0.00, 0.05) | **(0.00, 0.01)** | (0.00, 0.00) |
| EXP(1)  | (40, 20, 10)  | (0.00, 0.31) | (0.01, 0.15) | (0.01, 0.12) | **(0.02, 0.05)** | (0.06, 0.02) |
|         | (60, 20, 10)  | (0.00, 0.24) | (0.00, 0.08) | (0.00, 0.07) | **(0.00, 0.02)** | (0.02, 0.00) |
|         | (80, 20, 10)  | (0.00, 0.21) | (0.00, 0.06) | (0.00, 0.05) | **(0.00, 0.01)** | (0.00, 0.00) |
|         | (100, 20, 10) | (0.00, 0.20) | (0.00, 0.05) | (0.00, 0.05) | **(0.00, 0.01)** | (0.00, 0.00) |

Note: The bold CMC results are those at the recommended $\alpha$ level. Results in the table are averages of the two rates for 1000 simulation runs.

AIC does not assume that the true model is in the set of candidate models. It finds the member in the set that is closest to the true model, so it is expected to be robust so long as there are members in the set that are close to the true model. Presently, it is not clear why BIC and CMC also showed good robustness in the above example. Robust selection of linear models against the impact of outliers has been studied by Müller & Welsh (2005) and Güney, Bozdogan & Arslan (2021), among others. As part of our future research on CMC, we plan to further study its robustness against violations of model assumptions and also explore its robustness and robust modifications against outliers.

4. SOUTH AFRICAN HEART DISEASE DATA ANALYSIS

To illustrate CMC with a real example, we now apply it to perform model selection for a dataset from a heart disease study conducted by Rousseauw et al. (1983). The dataset can be found in various publicly available sources such as the R package “bestglm” by McLeod, Xu & Lai (2020) and the online resource for the book Elements of Statistical Learning by Hastie, Tibshirani & Friedman (2009). The response variable in the dataset is the coronary heart disease status (chd), a binary variable recording the presence (chd = 1) or absence (chd = 0) of coronary heart disease for a sample of 462 males from a heart disease high-risk region of the Western Cape, South Africa. There are nine predictor variables: systolic blood pressure (sbp), tobacco use (tob), low-density lipoprotein cholesterol (ldl), adiposity (adi), family history of heart disease (fhd), type-A behaviour (typ), obesity (obe), alcohol consumption (alc), and age at onset (age). Fitting the full logistic regression model to chd using all nine predictor variables yields the output in Table 5. Five variables have small $p$-values, and in ascending order of their $p$-values these five variables are fhd, age, typ, tob, and ldl.

Using “bestglm”, we obtain the 10 models with the highest likelihood among the models with the same number of predictor variables. These 10 models, their AIC values, BIC values, and maximum log-likelihood ratio values (LogLR) are shown in Table 6. The model with the

DOI: 10.1002/cjs.11756 The Canadian Journal of Statistics / La revue canadienne de statistique
### TABLE 5: Estimated full logistic regression model for the South Africa heart disease data.

| Variable | Estimate | Std. error | z-value | p-value |
|----------|----------|------------|---------|---------|
| (Intercept) | -6.15072 | 1.30826 | -4.70145 | 2.583e-06 |
| sbp | 0.00650 | 0.00573 | 1.13500 | 2.563e-01 |
| tob | 0.07937 | 0.02660 | 2.98375 | 2.857e-03 |
| ldl | 0.17392 | 0.05966 | 2.91516 | 3.555e-03 |
| adi | 0.01858 | 0.02928 | 0.63458 | 5.257e-01 |
| fhd | 0.92537 | 0.22789 | 4.06052 | 4.896e-05 |
| typ | 0.03959 | 0.01232 | 3.21382 | 1.309e-03 |
| obe | -0.06290 | 0.04424 | -1.42176 | 1.550e-01 |
| alc | 0.00012 | 0.00448 | 0.02713 | 9.783e-01 |
| age | 0.04522 | 0.01212 | 3.72846 | 1.926e-04 |

### TABLE 6: Models with the highest likelihood.

| sdp | tob | ldl | adi | fhd | typ | obe | alc | age | AIC | BIC | LogLR |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------|
| 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 596.1084 | 596.1084 | 123.96 |
| 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 527.5623 | 531.6979 | 53.422 |
| 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 1   | 510.6582 | 518.9293 | 34.518 |
| 0   | 1   | 0   | 0   | 1   | 0   | 0   | 0   | 1   | 501.3854 | 513.7921 | 23.245 |
| 0   | 1   | 0   | 0   | 1   | 1   | 0   | 0   | 1   | 492.7143 | 509.2566 | 12.574† |
| 0   | 1   | 1   | 0   | 1   | 1   | 0   | 0   | 1   | 485.6856‡ | 506.3634‡ | 3.5455‡ |
| 0   | 1   | 1   | 0   | 1   | 1   | 1   | 0   | 1   | 485.9799 | 510.7933 | 1.8398 |
| 1   | 1   | 1   | 0   | 1   | 1   | 1   | 0   | 1   | 486.5490 | 515.4979 | 0.4089 |
| 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 488.1408 | 521.2253 | 0.0001 |
| 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 490.1400 | 527.3601 | 0.0000 |

**Note:** Each row represents the model with the highest likelihood among models with the same number of variables. “1” indicates the variable in the column heading is in the model, and “0” means it is not in the model. The “LogLR” column gives the maximum log-likelihood ratios of the models.

† Indicates the model chosen by CMC0.9.
‡ Indicates the model chosen by all of AIC, BIC, CMC0.9, and CMC0.5.

The smallest AIC value is the five-variable model containing the five variables with the smallest p-values, fhd + age + typ + tob + ldl. The model with the smallest BIC value is also this five-variable model. Since there are p = 9 variables in the full model, the degrees of freedom of the $\chi^2$ distribution for calibrating the log-likelihood ratio is 10. The $\chi^2_{1-0.10}$ quantiles defining the confidence regions (7) associated with CMC0.9, CMC0.5, and CMC0.1 are, respectively, 4.865, 9.341, and 15.987. From the LogLR column in Table 6, we see that models with log-likelihood ratios below 4.865 (or in $\chi^2_{1-0.9}$) are the last five models with 5–9 variables, so CMC0.9 chooses the smallest model in this set, which is the model with five variables chosen by AIC and BIC. Similarly, CMC0.5 also chooses the same five-variable model. On the other hand, models with
log-likelihood ratios below 15.987 are the last six models with 4–9 variables, so CMC_{0.1} chooses the smallest model in this set of six models, which is the four-variable model consisting of the four variables with the smallest p-values, fhd + age + typ + tob. Although this model is different from the choice of the other four criteria, it is worth considering. This is because for this dataset the sample size is n = 462, which is 50 times larger than the number of variables p = 9, and CMC_{0.1} has been very accurate in our simulation study when the sample size is this large.

McLeod & Xu (2020) analyzed this dataset and obtained the above five-variable model and four-variable model, respectively, under two different BIC_{q} criteria discussed in that paper. In Chapter 4 of their book, Hastie, Tibshirani & Friedman (2009) also analyzed this dataset. They obtained a different four-variable model containing fhd + age + tob + ldl using a backward selection method. Different model selection criteria may lead to different selections. The CMC criteria at different α levels are no exception, but CMC provides a simple and unified framework to view the different selections through their log-likelihood ratios and associated α levels. This gives users more information about the best models.

5. CONCLUDING REMARKS

Unlike AIC, which is founded on information theory, or BIC, which originated from Bayesian inference, CMC has its roots in the classical frequentist inference, as it advocates selecting the smallest model in the likelihood ratio confidence region for the parameter vector of the full model. Although it is a simple idea, it has proven to be very effective in finite sample applications. Indeed, at the default level α = 0.5, CMC often outperforms AIC and BIC in terms of selection accuracy. With the tuning parameter α, it is also easy to adapt CMC to special situations. There have been various efforts in finding finite sample adjustments for AIC and BIC in order to improve their small-sample performances; see, for example, Hurvich & Tsai (1989), Broersen (2000), and Sclove (1987). CMC does not need such adjustments. When the sample size is small or when there are strongly correlated predictor variables, we may simply use a larger α level, say α = 0.9, to handle such special situations.

We have used the likelihood ratio test to define the set of plausible models. The score test and Wald test are asymptotically equivalent to the likelihood ratio test, and in principle they may also be used to define the set of plausible models for constructing CMC. However, these tests are computationally more complicated than the likelihood ratio test. Further, one of the key arguments used in establishing the lower bound (11) for CMC based on the likelihood ratio test is that the event \{ \hat{\beta} \in C_{1-\alpha} \} implies \{ \hat{\beta} \in C_{1-\alpha} \}. This argument would be invalid if other tests were used, which will make the theoretical investigation of the CMC selection more difficult. Nevertheless, we plan to study a Wald test based CMC to determine whether it has theoretical advantages over the log-likelihood ratio based CMC. In particular, by letting C_{W} be the Wald test induced confidence region for \beta, we hope to find a sequence \alpha_{n} \to 0 such that \P( \hat{\beta} \in C_{\alpha_{n}}^{W} ) \to 1 and

\[
\max_{\beta \in C_{\alpha_{n}}^{W}} \| \beta - \hat{\beta} \| _{2} = o_{p}(1)
\]

uniformly for all \alpha_{n} as n \to \infty. If such a sequence of \alpha_{n} can be found, then we can show that the CMC defined by \alpha_{n} is consistent, so it may have superior large-sample accuracy than CMC_{0.1}. The Wald test induced confidence region has an analytic expression which should be helpful in looking for such a sequence. Confidence regions based on other tests do not have this advantage.

CMC has a computational advantage over AIC and BIC for the best subset selection. While AIC and BIC both require the computation of the maximum likelihood of all 2^{p} models, CMC may need far fewer. To see this, denote by \M_{-i} the model containing all variables except x_{i}. Denote by \hat{\beta}_{-i} the maximum likelihood estimator and by \lambda(\hat{\beta}_{-i}) the maximum log-likelihood...
ratio of this model. We first compute \( \hat{\beta}_i \) and \( \lambda(\hat{\beta}_i) \) for \( i = 1, 2, \ldots, p \). Since excluding the variable \( x_i \) is the same as setting its parameter \( \beta_i \) to zero, the (augmented) \( \hat{\beta}_i \) is the point on the hyperplane \( \beta_i = 0 \) with the highest likelihood or lowest log-likelihood ratio. Suppose \( \lambda(\hat{\beta}_i) > \chi^2_{1-\alpha, p+1} \). Then, \( \hat{\beta}_i \notin C_{1-\alpha} \). This implies that other points on the hyperplane \( \beta_i = 0 \) are also not in \( C_{1-\alpha} \); that is, the hyperplane and the confidence region \( C_{1-\alpha} \) do not intersect. Figure 1a is an example where the hyperplane \( \beta_2 = 0 \) (i.e., \( \beta_1 \)-axis) does not intersect with the confidence region. This means for any \( \beta \) in \( C_{1-\alpha} \), its element corresponding to \( x_i \) is nonzero. The CMC solution is in \( C_{1-\alpha} \), so its element corresponding to \( x_i \) is nonzero. It follows that CMC will select \( x_i \). Suppose there are \( p' \) such variables. CMC will select them all, so we only need to determine whether CMC will select the remaining \( p - p' \) variables. The number of models that need to be computed for the CMC-based best subset selection is thus \( p + 2^{p-p'} \), which could be substantially smaller than the \( 2^p \) required by the AIC and BIC.

Finally, recent advances in computational statistics have made it possible to carry out the best subset selection for regression models with a large number of variables. For linear models, Bertsimas, King & Mazumder (2016) proposed a mixed integer optimization algorithm, which can solve the best subset selection problem with hundreds of variables in minutes. In situations where \( p - p' \) is still too large for the best subset selection, we may use the lasso with cross-validation or a dedicated variable screening method such as the sure independence screening method of Fan & Lv (2008) to reduce the total number of variables to a manageable level first, and then apply CMC to select a model.

ACKNOWLEDGEMENTS

We thank two anonymous reviewers and the editor, Professor Johanna Nešlehová, for their helpful comments, which have led to many improvements to this article. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, 716–723.
Bertsimas, D., King, A., & Mazumder, R. (2016). Best subset selection via a modern optimization lens. Annal of Statistics, 44, 813–852.
Broersen, P. M. T. (2000). Finite sample criteria for autoregressive order selection. IEEE Transactions on Signal Processing, 48, 3550–3558.
Chen, J. & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. Biometrika, 95, 759–771.
Ding, J., Tarokh, V., & Yang, Y. (2018a). Model selection techniques: An overview. IEEE Signal Processing Magazine, 35, 16–34.
Ding, J., Tarokh, V., & Yang, Y. (2018b). Bridging AIC and BIC: A new criterion for autoregression. IEEE Transactions on Information Theory, 64, 4024–4043.
Fahrmeir, L. & Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. Journal of Royal Statistical Society, Series B, 70, 849–911.
Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. Journal of Royal Statistical Society, Series B, 70, 849–911.
Gouriéroux, C. & Monfort, A. (1981). Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. Journal of Econometrics, 17, 83–97.
Güney, Y., Bozdogan, H., & Arslan, O. (2021). Robust model selection in linear regression models using information complexity. Journal of Computational and Applied Mathematics, 398, 113679.
Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. Annals of Statistics, 5, 815–841.
Hannan, E. J. & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of Royal Statistical Society, Series B*, 41, 190–195.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Elements of Statistical Learning: Data Mining, Inference and Predictions*, 2nd ed., Springer Verlag, New York.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical Learning with Sparsity*, 2nd ed., CRC Press, Boca Raton, FL.

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

Kadane, J. B. & Lazar, N. A. (2004). Methods and criteria for model selection. *Journal of the American Statistical Association*, 99, 279–290.

McLeod, A. I. & Xu, C. (2020). bestglm: Best Subset GLM. Vignette for R package ‘bestglm’. Available at: http://www2.uaem.mx/r-mirror

McLeod, A. I., Xu, C. & Lai, Y. (2020). Package ‘bestglm’. An R Package. Available at: https://cran.r-project.org

Morgan, J. A. & Tatar, J. F. (1972). Calculation of the residual sum of squares for all possible regressions. *Technometrics*, 14, 317–325.

Müller, S. & Welsh, A. H. (2005). Outlier robust model selection in linear regression. *Journal of the American Statistical Association*, 100, 1297–1310.

Rao, C. R. & Wu, Y. H. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika*, 76(2), 369–374.

Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J., Jooste, P., & Ferreira, J. (1983). Coronary risk factor screening in three rural communities. *South African Medical Journal*, 64, 430–436.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333–343.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Tsao, M. (2021). A constrained minimum method for model selection. *Stat*, 10(1), e387.

Received 25 January 2022
Accepted 26 August 2022