Statistica Sinica **28** (2018), 1561-1581
doi:https://doi.org/10.5705/ss.202016.0353

SPARSE ESTIMATION OF GENERALIZED LINEAR MODELS (GLM) VIA APPROXIMATED INFORMATION CRITERIA

Xiaogang Su[1], Juanjuan Fan[2], Richard A. Levine[2], Martha E. Nunn[3] and Chih-Ling Tsai[4]

[1]*University of Texas, El Paso,* [2]*San Diego State University*
[3]*Creighton University and* [4]*University of California, Davis*

Abstract: We propose a sparse estimation method, termed MIC (Minimum approximated Information Criterion), for generalized linear models (GLM) in fixed dimensions. What is essentially involved in MIC is the approximation of the ℓ_0-norm by a continuous unit dent function. A reparameterization step is devised to enforce sparsity in parameter estimates while maintaining the smoothness of the objective function. MIC yields superior performance in sparse estimation by optimizing the approximated information criterion without reducing the search space and is computationally advantageous since no selection of tuning parameters is required. Moreover, the reparameterization tactic leads to valid significance testing results free of post-selection inference. We explore the asymptotic properties of MIC, and illustrate its usage with simulated experiments and empirical examples.

Key words and phrases: BIC, generalized linear models, post-selection inference, regularization, sparse estimation, variable selection.

1. Introduction

Suppose that data $\mathcal{L} := \{(y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$ consist of n i.i.d. copies of $\{y, \mathbf{x}\}$, where y is the response variable and $\mathbf{x} = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$ is the predictor vector. WLOG, we assume that the x_{ij}'s are standardized throughout the paper. Consider the regression models that link the mean response y and covariates $\mathbf{x}$ through its linear predictor $\mathbf{x}^T\boldsymbol{\beta}$ with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, e.g., generalized linear models (GLM; McCullagh and Nelder (1989)). Concerning variable selection, the true $\boldsymbol{\beta}$ can be sparse with some components being zeros. Sparse estimation aims to identify the zero components and estimate the nonzero ones in $\boldsymbol{\beta}$ simultaneously. For simplicity, we assume that either there is no nuisance parameter involved, or that the nuisance parameters (e.g., scale or variance) and $\boldsymbol{\beta}$ are orthogonal (Cox and Reid (1987)). Hence we denote the log-likelihood

function as $L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i, \mathbf{x}_i; \boldsymbol{\beta})$, where $f(y, \mathbf{x}; \boldsymbol{\beta})$ denotes the probability density function of $(y, \mathbf{x})$.

Common variable selection methods can be formulated as the optimization problem

$$\min_{\boldsymbol{\beta} \in \Omega} \quad -2L(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \rho(\beta_j), \tag{1.1}$$

where $\rho(\cdot) \geq 0$ denotes a penalty function applied to each individual component of $\boldsymbol{\beta}$, $\lambda \geq 0$ is the penalty parameter, and $\Omega = \mathbb{R}^p$ is the search space or parameter space for $\boldsymbol{\beta}$. Methods vary in the form of the penalty function and the way of determining the penalty parameter. In the classical best subset selection (BSS), the ℓ_0 norm penalty, or cardinality of $\boldsymbol{\beta}$,

$$\sum_{j=1}^{p} \rho(\beta_j) = \|\boldsymbol{\beta}\|_0 = \operatorname{card}(\boldsymbol{\beta}), \tag{1.2}$$

provides a measure of model complexity with the number of nonzero components in $\boldsymbol{\beta}$; the penalty parameter λ_0 is fixed as 2 in AIC (Akaike (1974)) or $\ln(n)$ in BIC (Schwarz (1978)). We focus more on the use of BIC for its superior empirical performance in variable selection, widely reported in the literature. BSS essentially seeks the best model with minimum BIC. Owing to the discrete nature of the ℓ_0 norm, the optimization is done in two steps: one maximizes the log-likelihood $L(\boldsymbol{\beta})$ for each given sparsity structure or model (2^p in total), then compares across all models. BSS is only feasible for small p, despite the availability of faster algorithms (Furnival and Wilson (1974)).

The second general approach to variable selection is regularization. One basic motivation of regularization is to change the discrete nature of BSS. For this purpose, different continuous penalty functions are proposed. In scenarios where the log-likelihood function is concave, or can be converted so, the ℓ_1 penalty $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ in LASSO (Tibshirani (1996)) helps retain convexity of the optimization problem. LASSO requires strong assumptions in order to ensure selection consistency (Zhao and Yu (2006)) and induces bias in estimating the nonzero parameters. To make improvements, non-convex penalties such as SCAD (Fan and Li (2001)) and MCP (Zhang (2010)) are proposed.

There are several difficulties with regularization. To induce sparsity in the estimated parameters, it is necessary for $\beta = 0$ to be a singular point of $\rho(\beta)$; this makes the optimization in (1.1) non-smooth. Many well-developed smooth optimization routines cannot be used for this and new ones have to be sought. While efficient algorithms such as homotopy (Osborne, Presnell and Turlach (2000)),

the LARS method (Efron et al. (2004)), and coordinate descent (Fu (1998), Friedman, Hastie and Tibshirani (2010), and Breheny and Huang (2011)) have become standard, it is of both methodological and practical interest to see if sparse estimation can be formulated into a smooth optimization problem. Besides, the penalty in regularization no longer corresponds well to model complexity represented by $\|\boldsymbol{\beta}\|_0$. Hence there is no simple rule, as in BIC, for determining the penalty parameter λ and its choice has to be tuned. This leads to the two-step procedure in the practice of regularization: compute the regularization path $\{\widehat{\boldsymbol{\beta}}(\lambda) : \lambda \geq 0\}$, then select the best tuning parameter $\lambda^\star$ by referring to a criterion such as BIC (see, e.g., Wang, Li and Tsai (2007)). Thus, regularization seeks minimum BIC from a much reduced search space, noting that the regularization path is a one-dimensional curve in Ω. Selecting $\lambda^\star$ not only consumes additional computational time, but also causes another statistically awkward issue concerning its inference. Although the best tuning parameter $\hat{\lambda}$ is data-dependent and hence clearly a statistic, no statistical inference is routinely done on λ, at least in the frequentist's approach.

Another problem with both BSS and regularization is the post-selection inference. Conventionally statistical inference is made on the final model with selected variables or nonzero coefficients by ignoring the effect of model selection. This can be problematic, as pointed out by Leeb and Pötscher (2005) and others. One obstacle is that no statistical inference is available for parameters associated with unselected variables in BSS or zero estimates in regularization. How to make valid post-selection inference has been considered in Berk et al. (2013), Efron (2014), and Lockhart et al. (2014).

In this article, we study a sparse estimation method for GLM, termed the Minimum approximated Information Criterion (MIC), first proposed by Su (2015) in linear regression. The exposition in Su (2015) focuses on variable selection only; we expand the use of MIC in sparse estimation. The main idea of MIC is to introduce unit dent functions to approximate the ℓ_0 norm in (1.2). This leads to a smoothed version of BIC that can be directly optimized. A reparameterization step is then devised to enforce sparsity in parameter estimates while maintaining smoothness of the objective function. The formulation results in a non-convex yet smooth programming problem, and many readily available smooth optimization algorithms can be conveniently applied. Moreover, the smoothness of the estimating equation provides leeway in circumventing post-selection inference.

MIC offers several major advantages in sparse estimation. It imitates BSS but extends its capacity to large p. Since MIC seeks optimization of BIC, albeit

approximated, without reducing the search space, it outperforms many regularization methods in the sense of minimum BIC. It is computationally advantageous in avoiding selection of the tuning parameters, and facilitates statistical inference for both zero and non-zero coefficient estimates via the reparameterization trick. Our discussions are restricted to fixed dimensions. The remainder of this article is organized as follows. Section 2 presents the MIC method in detail. In Section 3, we explore its asymptotic properties. Section 4 presents simulation studies and data analysis examples. Section 5 concludes with a brief discussion.

2. Minimizing the Approximated BIC

MIC approximates cardinality in the information criteria with a smooth unit dent function and enforces sparsity with reparameterization. In the final form, it solves the unconstrained smooth optimization problem

$$\min_{\boldsymbol{\gamma}} \quad -2\,L(\boldsymbol{\beta}) \;+\; \log(n)\mathrm{tr}(\mathbf{W}), \tag{2.1}$$

where $\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\gamma}$, and $\mathbf{W} = \mathrm{diag}\,(w_j)$ with $w_j = w(\gamma_j) = \tanh(a\,\gamma_j^2)$ for $j = 1, \ldots, p$. The formulation of (2.1) involves a nonnegative parameters a that controls the sharpness of approximation. The empirical performance of MIC is rather stable with respect to the choice of a, hence a is fixed *a priori*. We explain the detailed procedure step-by-step in the ensuing subsections.

2.1. Unit dent functions

In a similar spirit to regularization, we desire to make the discrete BSS process continuous. While most regularization methods are based on optimization considerations, e.g., convex relaxation of the ℓ_0 norm, MIC is mainly motivated by the idea of approximation. Specifically, we seek a continuous or smooth approximation to the cardinality in (1.2).

For convenience, we use β as a generic notation for β_j from time to time. Since the cardinality of $\boldsymbol{\beta}$ is $\|\boldsymbol{\beta}\|_0 = \sum I\{\beta_j \neq 0\}$, we need to approximate the indicator function $I\{\beta \neq 0\}$. To this end, a suitable approximating function $w(\beta)$ must be a unit dent function.

Definition 1. *Let* $\bar{\mathbb{R}} = R \cup \{-\infty, \infty\}$. *A unit dent function is a continuous even function* $w : \bar{\mathbb{R}} \to [0, 1]$ *that is increasing on* $\mathbb{R}_+$ *with* $w(0) = 0$ *and* $\lim_{\beta\to\infty} w(\beta) = 1$.

If $w(\beta)$ is differentiable, then $\dot{w}(\beta) \geq 0$ on $\mathbb{R}+$ and $\dot{w}(\beta) \leq 0$ on $\mathbb{R}_-$. The $[0, 1]$ range requirement ensures that $\sum_j w(\beta_j)$ approximates $\|\boldsymbol{\beta}\|_0$, but it makes $w(\cdot)$

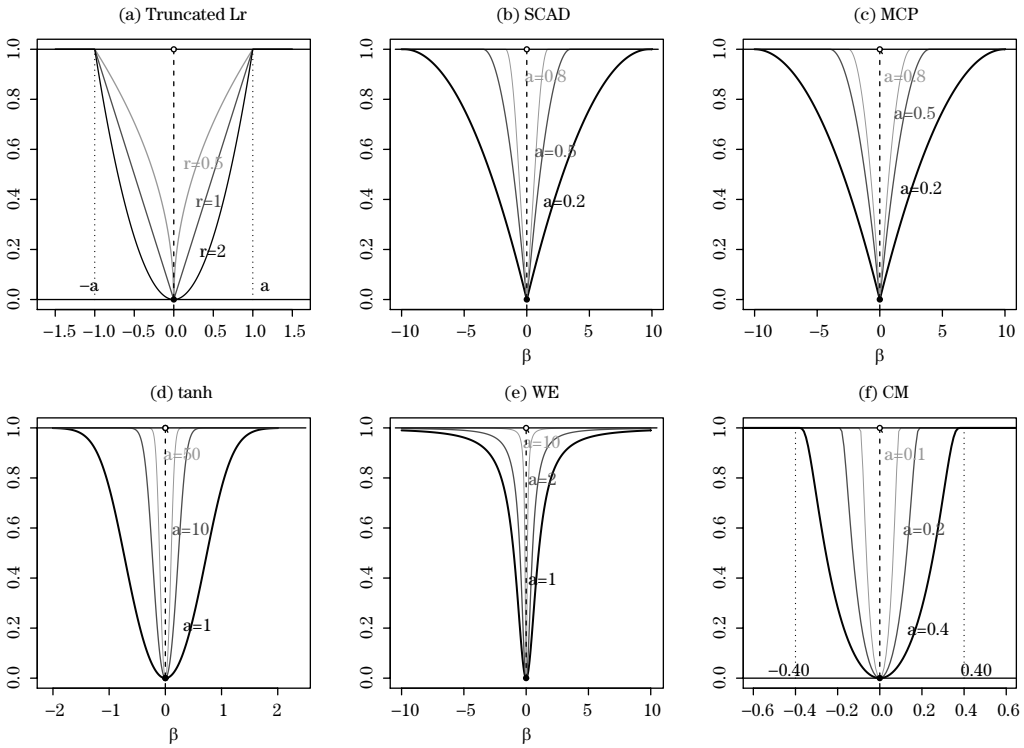

Figure 1. Several unit dent functions for approximating $I(\beta \neq 0)$: (a) truncated L_r; (b) modified SCAD; (c) modified MCP; (d) hyperbolic tangent; (e) weight elimination (WE); (f) converse mollifier (CM).

non-convex. The condition $\lim_{|\beta|\to\infty} w(\beta) = 1$ implies that $w(\beta)$ is approximately a constant function or $\dot{w}(\beta) \approx 0$ when $|\beta|$ is away from 0. As a consequence, the penalty $w(\beta)$ essentially does not alter the related score equations for nonzero β. Motivated by bump functions, we call $w(\cdot)$ a 'dent' function. A special family of bump functions, called mollifiers, are known as smooth approximations to the identity (Friedrichs (1944)). For a mollifier $\phi(\cdot)$ normalized to have the range $[0, 1]$, $1 - \phi(\cdot)$ is a unit dent function.

Let $\mathcal{D}$ denote the family of all unit dent functions. It is easy to see that $\mathcal{D}$ is closed under operations such as composition and product. In particular, if $w(\beta) \in \mathcal{D}$, then $w^k(\beta) \in \mathcal{D}$ for $k \in \mathbb{N}$. Unit dent functions have appeared in the regularization literature, one being the truncated ℓ_r penalty of Shen, Pan and Zhu (2012). The penalty functions SCAD (Fan and Li (2001)) and MCP (Zhang (2010)) can also be modified into unit dent functions. See Figure 1 for graphical illustrations of several unit dent functions: (a) truncated L_r: $w(\beta; a, r) = (|\beta|/a)^r$

if $|\beta| \le a$ and 1 otherwise; (b) modified SCAD: $w(\beta; a) = a|\beta|$ if $|\beta| \le a$; $\{2a(2 - a^2)|\beta| - a^4 - a^2\beta^2\}/\{4(1 - a^2)\}$ if $a < |\beta| < (2 - a^2)/a$; and 1 if $|\beta| > (2 - a^2)/a$ for $0 < a < \sqrt{2/3}$; (c) modified MCP: $w(\beta; a) = a|\beta| - a^2\beta^2/4$ if $|\beta| \le 2/a$ and 1 if $|\beta| > 2/a$ for $0 < a < \sqrt{2}$; (d) hyperbolic tangent $w(\beta; a) = \tanh(a \cdot \beta^2)$; (e) weight elimination (Weigend, Rumelhart and Huberman (1991)) $w(\beta) = (1 + a/\beta^2)^{-1}$ with $a > 0$; (f) converse mollifier $w(\beta) = 1 - \exp\{-\beta^2/(a^2 - \beta^2)\} \cdot I\{|\beta| \le a\}$ for $a > 0$.

To enforce sparsity, the penalty function must have $\beta = 0$ as a singular point (Fan and Li (2001)). However, MIC advocates the use of smooth unit dent functions since the smoothness property allows us to capitalize on well-developed theories and methods in optimization and statistical inference. We achieve sparsity in a different way.

While many smooth unit dent functions can be considered, we use the hyperbolic tangent function in MIC for its simple form:

$$w(\beta) = \tanh(a\beta^2) = \frac{\exp(2a\beta^2) - 1}{\exp(2a\beta^2) + 1} = 2\,\text{logistic}(2a\beta^2) - 1. \tag{2.2}$$

Its derivatives are easily available, with the first two given by $\dot{w}(\beta) = 2a\beta(1-w^2)$ and $\ddot{w}(\beta) = 2a(1-w^2)(1-4a\beta^2 w)$. In addition, the $\tanh(\cdot)$ function is associated with the logistic or expit function that is widely used in statistics. A plot of $w(\beta)$ versus β for different a values is provided in Figure 1(d). It can be seen that a larger a yields a sharper approximation to the indicator function $I\{\beta \ne 0\}$.

With $w(\beta) = \tanh(a\beta^2)$, one seeks to solve

$$\min_{\boldsymbol{\beta}} \quad -2L(\boldsymbol{\beta}) + \lambda_0 \sum_{j=1}^{p} w(\beta_j). \tag{2.3}$$

Expanding $L(\boldsymbol{\beta})$ at the MLE $\widehat{\boldsymbol{\beta}}$ and using the fact that $\nabla L(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$, we have

$$L(\boldsymbol{\beta}) \approx L(\widehat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^T \left\{ \frac{\nabla^2 L(\widehat{\boldsymbol{\beta}})}{2} \right\} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}),$$

where $\nabla L(\widehat{\boldsymbol{\beta}})$ and $\nabla^2 L(\widehat{\boldsymbol{\beta}})$ are the gradient vector and Hessian matrix of $L(\boldsymbol{\beta})$ evaluated at $\widehat{\boldsymbol{\beta}}$, respectively. Thus, the penalized optimization form in (2.3) can be viewed as the Lagrangian that corresponds to the constrained optimization problem

$$\min_{\boldsymbol{\beta}} \quad (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^T \left\{ -\nabla^2 L(\widehat{\boldsymbol{\beta}}) \right\} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \quad \text{subject to} \quad \sum_{j=1}^{p} w(\beta_j) \le t_0, \tag{2.4}$$

for some $t_0 \ge 0$. Figure 2(a) presents a graphical illustration of the optimization

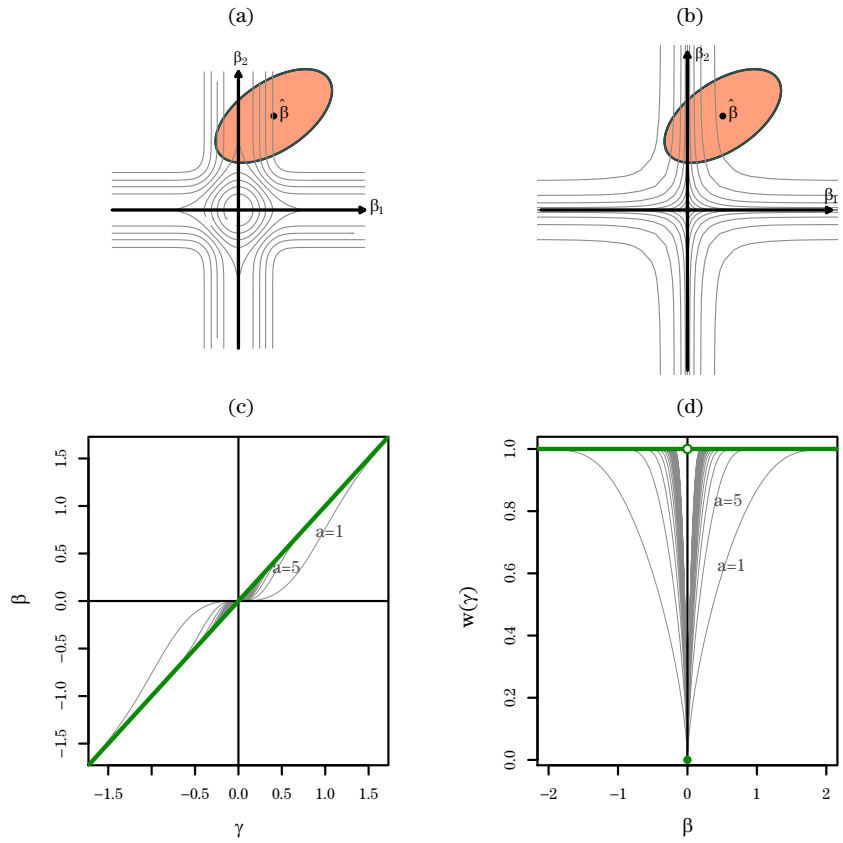

Figure 2. Illustration of the reparameterization step: (a) the contour plot of (2.4); (b) the contour plot of (2.7); (c) $\beta = \gamma w(\gamma)$ vs. γ; and (d) $w(\gamma)$ as a function of β with different a values.

problem (2.4) in the two-dimensional case. The objective function is an ellipsoid centered at MLE $\widehat{\boldsymbol{\beta}}$. As shown as the contour plots of Figure 2(a), the feasible sets for the constraint $w(\beta_1) + w(\beta_2) \le t_0$ contain both sharpened diamonds for large t_0 and discs for small t_0, resembling the grouped LASSO penalty (Bakin (1999)) as pointed out by a referee. By the Taylor expansion, $w(\beta) = a\beta^2 + O(\beta^6)$ for $\beta \to 0$, implying that sparsity may not be enforced. We address this issue in the next section. Hereafter, $w(\beta)$ is referred to the hyperbolic tangent penalty, unless otherwise stated.

2.2. Reparameterization

To enforce sparsity, we consider a reparameterization procedure originally motivated by the nonnegative garrotte (NG) of Breiman (1995). NG is a sign-

constrained regularization based on the decomposition $\beta = \text{sgn}(\beta)\,|\beta|$. Supposing that the sign of each β_j can be correctly specified by the MLE $\widehat{\boldsymbol{\beta}}$, it remains to estimate $|\boldsymbol{\beta}_j|$. Reparameterizing $\boldsymbol{\beta} = \text{diag}\{\text{sgn}(\widehat{\boldsymbol{\beta}})\}\,\boldsymbol{\gamma}$, for some nonnegative vector $\boldsymbol{\gamma}$ such that $\gamma_j = |\beta_j|$, leads to the NG formulation

$$\min_{\boldsymbol{\gamma}} -2L(\boldsymbol{\beta}) \;\text{ s.t. }\; \sum_{j=1}^{p} \gamma_j \le t \text{ and } \gamma_j \ge 0$$

with tuning parameter t. A fundamental problem with sign-constrained regularization is that if any sign is wrongly specified by the initial estimator $\widehat{\boldsymbol{\beta}}$, which occurs often in data owing to multicollinearity or other complexities, then it cannot make a correction.

Our immediate aim is to introduce singularity to the penalty function at 0. For this purpose, we consider the decomposition $\beta = \beta\, I\{\beta \neq 0\}$. Set $\gamma = \beta$ and approximate $I\{\gamma \neq 0\}$ by $w(\gamma)$. This motivates the reparameterization $\beta_j = \gamma_j w(\gamma_j)$ for $j = 1, \ldots, p$. In matrix form, $\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\gamma}$, where matrix $\mathbf{W}$ is defined in (2.1). As shown in Figure 2(c), β is a strictly increasing function of γ and $\beta = \gamma$ except for a small neighborhood of 0, in which a shrinkage on $|\beta|$ is imposed.

To see how the reparameterization helps enforce sparsity, consider the resulting optimization problem

$$\min_{\boldsymbol{\beta}} \quad -2L(\boldsymbol{\beta}) + \ln(n) \sum_{j=1}^{p} w(\gamma_j). \tag{2.5}$$

Compared to (2.3), the only change is that the penalty function $w(\cdot)$ is now applied to the reparameterized γ_j instead of β_j. The $w(\gamma_j)$ in (2.5) is an implicit function of β_j. Figure 2(d) plots $w(\gamma)$ as a penalty function of β for different values of a, which shows a similar pattern to the non-convex SCAD or MCP penalty with a cusp at $\beta = 0$. It can be verified that $w(\gamma)$ is a unit dent function of β that approximates $I(\beta \neq 0)$.

The singularity at 0 can be further confirmed by calculating the derivatives of $w(\gamma)$ at β. Applying the chain rule gives

$$\frac{d\,w(\gamma)}{d\,\beta} = \frac{d\,w(\gamma)}{d\,\gamma}\frac{d\,\gamma}{d\,\beta} = \frac{d\,w(\gamma)}{d\,\gamma}\left(\frac{d\,\beta}{d\,\gamma}\right)^{-1} = \frac{\dot{w}}{w + \gamma\dot{w}}, \tag{2.6}$$

where $w = w(\gamma)$ and $\dot{w} = \dot{w}(\gamma) = 2a\gamma(1 - w^2)$, and it follows $d\,\beta/d\,\gamma = w + \gamma\dot{w}$. The first derivative in (2.6) is expressed in terms of γ via implicit differentiation since the explicit formula of γ in terms of β is unavailable. The validity of (2.6), however, requires $d\,\beta/d\,\gamma \;\neq\; 0$, which holds everywhere except at $\beta = 0$.

Similar arguments can be used to derive the form of higher-order derivatives. For example, the second-derivative is given by

$$\frac{d^2\, w(\gamma)}{d\,\beta^2} = \frac{w\,\ddot{w} - 2\,\dot{w}^2}{(w+\gamma\dot{w})^3}$$

with $\ddot{w} = \ddot{w}(\gamma) = 2a(1-w^2)(1-4a\gamma^2 w)$, which again does not exist at $\beta = 0$. It can be verified that $w(\gamma)$ is a smooth function of β except at $\beta = 0$.

The reparameterization $\beta = \gamma w(\gamma)$ to enforce singularity at 0 holds for any smooth function in $\mathcal{D}$. We have utilized the differentiation of the inverse function to achieve this. Accordingly, the derivatives of $w(\gamma)$ as a function of β exist everywhere except at $\beta = 0$.

Figure 2(b) provides a two-dimensional illustration of the constrained optimization problem that corresponds to (2.5):

$$\min_{\boldsymbol{\beta}} \quad (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^T \left\{ -\nabla^2 L(\widehat{\boldsymbol{\beta}}) \right\} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \;\; \text{s.t.} \;\; \text{tr}(\mathbf{W}) \le t_0 \text{ with } \boldsymbol{\beta} = \mathbf{W}\boldsymbol{\gamma}. \tag{2.7}$$

The contour lines of the constraint $w(\gamma_1) + w(\gamma_2) \le t$ (as a function of β_1 and β_2) are sharpened diamonds, which serve better for the variable selection purpose.

The smooth formulation facilitated by reparameterization allows us to utilize available results in optimization theory and statistical inference, and leads to some important advantages. For computation, we estimate γ instead by solving (2.1). Compared to (2.5) where the objective function is nonsmooth in $\boldsymbol{\beta}$, we have switched the decision vector to $\boldsymbol{\gamma}$. Solving (2.1) is a smooth optimization problem and many standard algorithms apply. Estimation of $\boldsymbol{\gamma}$ is meaningful in its own right. The fact that the correspondence between β and γ is one-to-one with $\beta_j = 0$ iff $\gamma_j = 0$ allows us to derive significance testing for $\boldsymbol{\beta}$ through $\boldsymbol{\gamma}$ that is free of post-selection inference. The objective function in (2.1) is smooth for estimating $\boldsymbol{\gamma}$. Thus standard arguments in M-estimators can be applied for making inference on $\boldsymbol{\gamma}$. The procedure is given next.

3. Asymptotic Properties

In this section, we study the asymptotic oracle properties of the MIC estimator $\widetilde{\boldsymbol{\beta}}$, including its $\sqrt{n}$-consistency, selection consistency, and the asymptotic normality of its nonzero components. We then present significance testing on $\boldsymbol{\beta}$ via $\boldsymbol{\gamma}$ that is free of post-selection inference. The proofs are in the supplementary materials.

3.1. Oracle properties of the MIC estimator $\widetilde{\boldsymbol{\beta}}$

We consider the MIC estimator $\widetilde{\boldsymbol{\beta}}$ obtained by minimizing the objective function in (2.5),

$$Q_n(\boldsymbol{\beta}) = -2\,\frac{L(\boldsymbol{\beta})}{n} + \frac{\ln(n)}{n}\,\sum_{j=1}^{p} w(\gamma_j), \tag{3.1}$$

where $L(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta})$ with $l_i(\boldsymbol{\beta}) = \log f(\mathbf{X}_i, Y_i; \boldsymbol{\beta})$. We denote a as a_n so that $\beta_j = \gamma_j w(\gamma_j) = \gamma_j \tanh(a_n \beta_j^2)$, and assume $a_n = O(n)$; this rate for a_n will be manifested in the derivation.

Denote the true parameter as $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0(1)}^T, \boldsymbol{\beta}_{0(0)}^T)^T$, where $\boldsymbol{\beta}_{0(1)} \in \mathbb{R}^q$ consists of all q nonzero components and $\boldsymbol{\beta}_{0(0)} = \mathbf{0}$ consists of all the $(p-q)$ zero components. For simplicity, we use $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ to denote the MIC and MLE estimators, respectively. Let $\mathbf{I} = \mathbf{I}(\boldsymbol{\beta}_0)$ and $\mathbf{I}_1$ be the Fisher information matrix for the whole and reduced true model with $\boldsymbol{\beta}_{0(0)} = \mathbf{0}$, respectively. It is well known that $\mathbf{I}_1$ is the q-th principal submatrix of $\mathbf{I}$.

Theorem 1. *Let* $\{(\mathbf{X}_i, Y_i) : i = 1, \ldots, n\}$ *be* n *i.i.d. copies from a density* $f(\mathbf{X}, Y; \boldsymbol{\beta}_0)$. *Under the regularity conditions* (A)–(C) *in* Fan and Li (2001), *we have*

(i). ($\sqrt{n}$*-Consistency) there exists a local minimizer* $\widetilde{\boldsymbol{\beta}}$ *of* $Q_n(\boldsymbol{\beta})$ *that is* $\sqrt{n}$*-consistent for* $\boldsymbol{\beta}_0$ *in the sense that* $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$.

(ii). (Sparsity and Asymptotic Normality) Partition $\widetilde{\boldsymbol{\beta}}$ *in (i) as* $(\widetilde{\boldsymbol{\beta}}_{(1)}^T, \widetilde{\boldsymbol{\beta}}_{(0)}^T)^T$ *in a similar manner to* $\boldsymbol{\beta}_0$. *With probability tending to 1 as* $n \to \infty$, $\widetilde{\boldsymbol{\beta}}_{(0)} = \mathbf{0}$ *and* $\sqrt{n}(\widetilde{\boldsymbol{\beta}}_{(1)} - \boldsymbol{\beta}_{0(1)}) \to N\left(\mathbf{0}, \mathbf{I}_1^{-1}\right)$.

The results in Theorem 1 are analogous to Theorems 1 and 2 in Fan and Li (2001). It establishes that $\widetilde{\boldsymbol{\beta}}_{(0)}$ is selection consistent and $\widetilde{\boldsymbol{\beta}}_{(1)}$ is a best asymptotic normal (BAN; see, e.g., Serfling (1980)) estimator of $\boldsymbol{\beta}_{0(1)}$. The standard errors (SE) for nonzero components in $\widetilde{\boldsymbol{\beta}}$ can be computed by replacing $\mathbf{I}_1$ in Theorem 1(ii) with the observed Fisher information matrix (Efron and Hinkley (1978)) and plugging in $\widetilde{\boldsymbol{\beta}}$. Since $\widetilde{\boldsymbol{\beta}}$ is essentially an M-estimator, alternative sandwich SE formulas (Stefanski and Boos (2002)) are available. However, as part of the post-selection inferences, all these SE formulas are only available for nonzero components in $\widetilde{\boldsymbol{\beta}}$ and hence caution should be exercised.

3.2. Inference on $\boldsymbol{\beta}$ via $\boldsymbol{\gamma}$

MIC completes sparse estimation in a single optimization step. This brings

about a unique opportunity to address the fundamental post-selection inference problem. Inference on zero components in $\boldsymbol{\beta}$ is unavailable in MIC because the asymptotic normality of M-estimators often entails a condition that the expected objective function $E\{Q_n(\boldsymbol{\beta})\}$ admits a second-order Taylor expansion at $\boldsymbol{\beta}_0$ whereas sparsity requires singularity of the penalty function $w(\gamma)$ at $\beta = 0$. However, the reparameterisation helps us to circumvent this non-smoothness issue. The transformation $\beta = \gamma w(\gamma)$ is a bijection and $\beta = 0$ iff $\gamma = 0$. Hence testing $H_0 : \beta_j = 0$ is equivalent to testing $H_0 : \gamma_j = 0$. As the objective function of $\boldsymbol{\gamma}$, $Q_n(\boldsymbol{\gamma})$ in (3.1) is smooth in $\boldsymbol{\gamma}$. Therefore, the statistical properties of $\widetilde{\boldsymbol{\gamma}}$ are readily available following standard M-estimation arguments.

Theorem 2. *If $\boldsymbol{\gamma}_0$ is the reparameterized parameter vector associated with $\boldsymbol{\beta}_0$ such that $\beta_{0j} = \gamma_{0j} w(\gamma_{0j})$, then*

$$\|\boldsymbol{\gamma}_0 - \boldsymbol{\beta}_0\|_2 = O\{\exp(-2a_n \min_{1\le j\le q} \gamma_{0j}^2)\}.$$

Under the regularity conditions (A)–(C) *in* Fan and Li (2001)*, we have*

$$\sqrt{n}\,\{\mathbf{D}(\boldsymbol{\gamma}_0)(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) + \mathbf{b}_n\} \xrightarrow{d} N\left\{\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\beta}_0)\right\}. \tag{3.2}$$

where

$$\mathbf{D}(\boldsymbol{\gamma}_0) = diag(w_j + \gamma_j \dot{w}_j)|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0} = diag\,(D_{jj}) \tag{3.3}$$

and the asymptotic bias

$$\mathbf{b}_n = \left\{-\nabla^2 L(\boldsymbol{\beta}_0)\right\}^{-1} \frac{\ln(n)}{2} \left(\frac{\dot{w}_j}{w_j + \tilde{\gamma}_j \dot{w}_j}\right)_{j=1}^{p} = (b_{nj})_{j=1}^{p} \tag{3.4}$$

satisfies (i) $\lim_{n\to\infty} D_{jj} = I\{\beta_{0j} \neq 0\}$ *and (ii)* $\mathbf{b}_n = o_p(1)$.

A practical implication of Theorem 2 is that both $\mathbf{D}(\boldsymbol{\gamma}_0)$ and $\mathbf{b}_n$ may be ignored in computing the standard errors of $\widetilde{\boldsymbol{\gamma}}$. Furthermore, since $\|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\beta}_0\| \le \|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| + \|\boldsymbol{\gamma}_0 - \boldsymbol{\beta}_0\| = o_p(1)$, $\widetilde{\boldsymbol{\gamma}}$ is a consistent estimator of $\boldsymbol{\beta}_0$ and can be used to replace $\boldsymbol{\beta}_0$ in estimating the Fisher information matrix. Thus, an asymptotic $(1-\alpha)\times 100\%$ confidence interval for γ_{0j} is

$$\tilde{\gamma}_j \pm z_{1-\alpha/2} \sqrt{\left\{\frac{\mathbf{I}_n^{-1}(\widetilde{\boldsymbol{\gamma}})}{n}\right\}_{jj}}, \tag{3.5}$$

where $\mathbf{I}_n$ denotes the observed Fisher information matrix and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-th percentile of $N(0,1)$. Significance testing on γ_{0j} can be done accordingly.

4. Numerical Results

In this section, we present simulation experiments and data examples to

illustrate MIC in comparison with other methods.

4.1. Computational issues

MIC solves for $\widetilde{\boldsymbol{\gamma}}$ by optimizing (2.1). Considering its nonconvex nature, a global optimization method is desirable. Mullen (2014) provides a comprehensive comparison of global optimization algorithms currently available in R (R Core Team (2017)). According to her recommendations, we have chosen the `GenSA` package (Xiang et al. (2013)) that implements the generalized simulation annealing of Tsallis and Stariolo (1996), because of its superior performance in both identification of the true optimal point and computing speed. With estimated $\widetilde{\boldsymbol{\gamma}}$, the MIC estimator $\widetilde{\boldsymbol{\beta}}$ can be obtained via the transformation $\widetilde{\boldsymbol{\beta}} = \widetilde{\mathbf{W}}\widetilde{\boldsymbol{\gamma}}$, where $\widetilde{\mathbf{W}} = \mathrm{diag}(\tilde{w}_j)$ with $\tilde{w}_j = w(\tilde{\gamma}_j)$. Because of the shrinkage effect of the reparameterization around 0, estimates $\tilde{\gamma}_j$ close to 0 yield small values of $|\tilde{\beta}_j|$, which can be virtually taken as 0.

Implementation of MIC involves the choice of a or a_n. In theory, the asymptotic results in Section 3.1 entail $a_n = O(n)$. To apply the arguments of Fan and Li (2001), this $O(n)$ rate seems unique. In practice, the empirical performance of MIC is quite stable with respect to the choice of a_n, as demonstrated in Su (2015) for linear regression. In MIC, a_n is a shape or scale parameter in the unit dent function that modifies the sharpness of its approximation to the indicator function. Its role is largely similar to that of the parameter a in SCAD (Fan and Li (2001)), where a is fixed at $a = 3.7$. In general, a larger a_n enforces a better approximation of the indicator function with the hyperbolic tangent function, while a smaller a_n is appealing for optimization purposes, by introducing more smoothness. Based on our numerical experience, applying an a_n value smaller than 1 is not advisable owing to poor approximation. The performance of MIC stabilizes substantially when a_n gets large, especially when it is 10 or above. On this basis, we recommend fixing a to any value in [10, 50].

Four known methods are included for comparison with MIC: the best subset selection (BSS) with BIC, LASSO, SCAD, and MCP. The oracle estimate is added as a benchmark. All the computations are done in R (R Core Team (2017)). Specifically, we have used the R package `bestglm` for BSS, `lars` and `glmnet` for LASSO, and `ncvreg` and `SIS` for SCAD and MCP, with their default settings.

4.2. Simulated experiments

We generated data sets by using the simulation settings of Zou and Li (2008).

The models are

$$\begin{cases} \text{Model A: } y|\mathbf{x} \sim N\{\mu(\mathbf{x}), 1\} \text{ with } \mu(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}, \\ \text{Model B: } y|\mathbf{x} \sim \text{Bernoulli}\{\mu(\mathbf{x})\} \text{ with } \mu(\mathbf{x}) = \text{expit}(\mathbf{x}^T\boldsymbol{\beta}), \\ \text{Model C: } y|\mathbf{x} \sim \text{Poisson}\{\mu(\mathbf{x})\} \text{ with } \mu(\mathbf{x}) = \exp(\mathbf{x}^T\boldsymbol{\beta}), \end{cases} \tag{4.1}$$

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0)^T$ in Models A and B, and $(1.2, 0.6, 0, 0, 0.8, 0, 0, 0, 0, 0, 0, 0)^T$ in Model C. Each data set involves $p = 12$ predictors that follow a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\sigma_{jj'})$ and $\sigma_{jj'} = 0.5^{|j-j'|}$ for $j, j' = 1, \ldots, p$. In Model B, six binary predictors are created by setting $x_{2j-1} := I\{x_{2j-1} < 0\}$ for $j = 1, \ldots, 6$. Thus, there are six continuous and six binary predictors in Model B. We consider sample sizes $n = 100$ and $n = 200$, and 500 simulation runs were taken for each model configuration.

To apply the MIC method, we fixed $\lambda_0 = \ln(n)$ and $a_n = 10$. Five performance measures were used for comparison. The first is the empirical model error (ME), $\sum_{i=1}^n (\mu_i - \hat{\mu}_i)^2/n$, where μ_i is given in (4.1) and $\hat{\mu}_i$ is obtained by plugging in the estimate of $\boldsymbol{\beta}$. We computed ME based on an independent test sample of size $n = 500$ and report the averaged ME over 500 realizations. The other measures were the average model Size, the number of nonzero parameter estimates; FP, the number of nonzero estimates for zero parameters; FN, the number of zero estimates for nonzero parameters; and the proportion of correct selections, C.

Table 1 indicates that MIC performs similarly to BSS across all three models. All performance measures of MIC improve as the sample size increases. By comparing MIC against the other regularization methods, we find that MIC outperforms them in general, except for the Gaussian linear regression case where its performance is only comparable. We think this is mainly because the objective function of MIC involves the Gaussian profile likelihood $n \ln \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$, which is nonconvex, while regularization methods can work with the convex least squares problem $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ directly. Nevertheless, they all have to deal with the same log-likelihood function in Models B and C. No implementation of MCP is available for the log-linear regression, hence it is not presented for Model C. In sum, MIC not only enjoys computational efficiency, but also demonstrates an excellent finite sample performance.

We evaluated the standard error formula for nonzero parameter estimates. Table 2 presents the median absolute deviation (MAD) value of $\widetilde{\boldsymbol{\beta}}_{(1)}$ out of 500 runs, which provides a more robust estimate of its standard deviation. This MAD value matches reasonably well with the median of standard errors of $\widetilde{\boldsymbol{\beta}}_{(1)}$. Also

Table 1. Simulation results on MIC (with $\lambda_0 = \ln(n)$ and $a = 10$) in comparison with other methods. Reported quantities include ME, Size, FP, FN, and C, all based on 500 realizations.

(a) Model A – Linear Regression

	$n = 100$					$n = 200$				
Method	ME	Size	FP	FN	C	ME	Size	FP	FN	C
MIC	0.054	3.47	0.47	0.00	0.640	0.021	3.25	0.25	0.00	0.790
Oracle	0.034	3.00	0.00	0.00	1.000	0.015	3.00	0.00	0.00	1.000
BIC	0.055	3.35	0.35	0.00	0.710	0.022	3.19	0.19	0.00	0.834
LASSO	0.085	6.09	3.09	0.00	0.092	0.039	6.23	3.23	0.00	0.102
SCAD	0.045	3.58	0.58	0.00	0.752	0.022	3.71	0.71	0.00	0.752
MCP	0.047	3.57	0.57	0.00	0.750	0.020	3.41	0.41	0.00	0.814

(b) Model B – Logistic Regression

	$n = 100$					$n = 200$				
Method	ME	Size	FP	FN	C	ME	Size	FP	FN	C
MIC	0.017	3.74	1.03	0.29	0.354	0.005	3.42	0.49	0.07	0.624
Oracle	0.005	3.00	0.00	0.00	1.000	0.002	3.00	0.00	0.00	1.000
BIC	0.015	3.40	0.67	0.27	0.514	0.005	3.21	0.28	0.06	0.766
LASSO	0.023	6.54	3.79	0.25	0.012	0.012	7.32	4.37	0.05	0.018
SCAD	0.019	3.69	1.09	0.41	0.206	0.012	3.92	1.11	0.19	0.278
MCP	0.019	3.12	0.65	0.53	0.236	0.011	3.39	0.64	0.24	0.420

(c) Model C – Log-Linear Regression

	$n = 100$					$n = 200$				
Method	ME	Size	FP	FN	C	ME	Size	FP	FN	C
MIC	12.310	3.34	0.35	0.00	0.712	4.367	3.23	0.23	0.00	0.828
Oracle	9.289	3.00	0.00	0.00	1.000	3.555	3.00	0.00	0.00	1.000
BIC	25.884	3.39	0.39	0.00	0.714	4.897	3.23	0.23	0.00	0.826
LASSO	600.821	1.55	0.37	1.81	0.184	348.182	1.46	0.18	1.72	0.282
SCAD	40.753	4.08	1.08	0.00	0.336	12.843	3.64	0.64	0.00	0.528
MCP	80.931	3.48	0.59	0.11	0.698	18.979	3.50	0.56	0.05	0.745

presented is the MAD of standard errors. Table 3 presents the empirical size and power results in testing $H_0 : \gamma_j = 0$ at the significance level $\alpha = 0.05$, together with the coverage of 95% confidence intervals, over 1,000 simulation runs. The coverage proportion of 95% confidence intervals is presented only for each nonzero estimate; the coverage for a zero-valued γ_j estimate equals 1 minus the empirical size in this case and hence has been omitted. Sample sizes $n \in \{50, 200\}$ were considered. It can be seen that the proposed testing procedure has empirical sizes close to the nominal level 0.05 while showing exceptional empirical powers and coverage probabilities.

Table 2. Simulation results on standard errors of nonzero $\widehat{\boldsymbol{\beta}}$ with $n = 200$ over 500 simulation runs. Reported quantities are MAD of the parameter estimates, Median of the standard errors, and MAD of the standard errors.

(a) Model A – Gaussian Linear Regression

	oracle			MIC		
	MAD	Median SE	MAD SE	MAD	Median SE	MAD SE
β_1	0.083	0.082	0.006	0.083	0.082	0.006
β_2	0.084	0.082	0.006	0.087	0.082	0.006
β_5	0.072	0.072	0.005	0.073	0.072	0.005

(b) Model B – Logistic Regression

	oracle			MIC		
	MAD	Median SE	MAD SE	MAD	Median SE	MAD SE
β_1	0.528	0.475	0.086	0.529	0.492	0.094
β_2	0.399	0.389	0.048	0.448	0.407	0.064
β_5	0.380	0.356	0.059	0.405	0.367	0.061

(c) Model C – Loglinear Regression

	oracle			MIC		
	MAD	Median SE	MAD SE	MAD	Median SE	MAD SE
β_1	0.037	0.036	0.007	0.037	0.036	0.007
β_2	0.039	0.039	0.007	0.040	0.039	0.007
β_5	0.032	0.032	0.006	0.033	0.033	0.006

Table 3. Hypothesis testing on $\boldsymbol{\gamma}_0$ in MIC. Empirical size (ES), empirical power (EP), and the coverage proportion of the 95% CI are obtained at $\alpha = 0.05$ based on 1,000 simulation runs. The stronger signals correspond to Model A, B, and C in (4.1) while the case of the weaker signals resets $\boldsymbol{\beta} := \boldsymbol{\beta}/3$.

			Empirical Size									Empirical Power			Coverage		
Model	Signal	n	γ_3	γ_4	γ_6	γ_7	γ_8	γ_9	γ_{10}	γ_{11}	γ_{12}	γ_1	γ_2	γ_5	γ_1	γ_2	γ_5
A	Stronger	50	0.051	0.049	0.036	0.040	0.037	0.017	0.036	0.028	0.035	1.000	1.000	1.000	0.957	0.956	0.959
		200	0.037	0.041	0.043	0.023	0.036	0.024	0.030	0.033	0.042	1.000	1.000	1.000	0.960	0.963	0.973
	Weaker	50	0.062	0.054	0.038	0.043	0.041	0.018	0.037	0.031	0.037	1.000	0.681	0.912	0.942	0.867	0.944
		200	0.036	0.040	0.045	0.022	0.035	0.023	0.031	0.034	0.043	1.000	1.000	1.000	0.959	0.961	0.972
B	Stronger	50	0.012	0.011	0.007	0.004	0.002	0.005	0.008	0.009	0.010	0.509	0.202	0.327	0.997	0.988	0.995
		200	0.065	0.045	0.053	0.043	0.059	0.036	0.053	0.035	0.062	1.000	1.000	1.000	0.924	0.934	0.935
	Weaker	50	0.051	0.061	0.074	0.044	0.057	0.067	0.078	0.071	0.062	0.712	0.242	0.350	0.914	0.952	0.944
		200	0.047	0.061	0.049	0.045	0.048	0.038	0.046	0.052	0.051	1.000	0.759	0.932	0.941	0.926	0.946
C	Stronger	50	0.023	0.017	0.017	0.016	0.013	0.019	0.019	0.014	0.022	1.000	0.977	0.997	0.984	0.985	0.986
		200	0.071	0.077	0.066	0.060	0.058	0.060	0.052	0.051	0.085	1.000	1.000	1.000	0.990	0.970	0.994
	Weaker	50	0.045	0.033	0.033	0.039	0.028	0.022	0.024	0.033	0.044	0.650	0.185	0.220	0.817	0.964	0.987
		200	0.041	0.030	0.016	0.014	0.009	0.008	0.008	0.008	0.017	0.999	0.528	0.878	0.917	0.702	0.943

Table 4. Illustration with real data examples.

(a) Linear Regression with Diabetes Data

	Best Subset		MIC					
	$\hat{\beta}_j$	SE	p-value*	$\hat{\beta}_j$	SE	LASSO	SCAD	MCP
age			1.00					
sex	−0.15	0.04	0.00	−0.14	0.04	−0.12	−0.15	−0.14
bmi	0.32	0.04	0.00	0.33	0.04	0.32	0.32	0.33
map	0.20	0.04	0.00	0.20	0.04	0.18	0.20	0.20
tc			1.00			−0.06	−0.38	
ldl			1.00				0.22	−0.07
hdl	−0.18	0.04	0.01	−0.17	0.04	−0.14		−0.18
tch			1.00				0.08	
ltg	0.29	0.04	0.00	0.29	0.04	0.32	0.43	0.30
glu			1.00			0.03	0.04	0.03

(b) Logistic Regression with Heart Data

	Best Subset		MIC					
	$\hat{\beta}_j$	SE	p-value*	$\hat{\beta}_j$	SE	LASSO	SCAD	MCP
intercept	−0.85	0.12	0.00	−0.84	0.12	−0.79	−0.85	−0.84
sbp			1.00			0.04		0.06
tobacco	0.37	0.12	0.00	0.35	0.12	0.30	0.37	0.37
ldl	0.35	0.11	0.00	0.33	0.11	0.27	0.35	0.37
famhist	0.46	0.11	0.00	0.45	0.11	0.37	0.46	0.46
obesity			1.00			−0.01	−0.09	
alcohol			1.00					
age	0.64	0.14	0.00	0.66	0.14	0.54	0.65	0.63

(c) Log-Linear Regression with Fish Data

	Best Subset		MIC					
	$\hat{\beta}_j$	SE	p-value*	$\hat{\beta}_j$	SE	LASSO	SCAD	MCP
intercept	−0.31	0.07	0.00	−0.30	0.07	0.36	−0.01	
nofish			1.00				0.03	
livebait			1.00					
camper			1.00					
persons			1.00					
child	−0.64	0.10	0.00	−0.64	0.10		−0.65	−0.65
xb	1.47	0.03	0.00	1.46	0.03	0.33	1.46	1.46
zg	0.60	0.07	0.00	0.60	0.07		0.60	0.60
xb:zg			1.00			0.18		

4.3. Data examples

We consider the diabetes data (Efron et al. (2004)), the heart data (Hastie, Tisshirani and Friedman (2009)), and the fish count data (available from `http://www.ats.ucla.edu/stat/data/fish.csv`) to illustrate linear regression, lo-

gistic regression, and log-linear regression models, respectively. The results are presented in Table 4, where the p-values in MIC are based on testing $H_0 : \gamma_j = 0$.

Table 4 shows that MIC provides the similar selection as the BIC-based best subset selection across all three examples. In addition, the resulting MIC estimates and their standard errors are quite close to these of the BIC model, indicating that MIC approximates the best subset selection method well. This, together with MIC's computational efficacy, allows us to employ MIC on data with large numbers of covariates, even when BSS is infeasible. In the diabetes data, it is interesting that the sign of the parameter estimate on `hdl` is positive under the full model fitting, but is negative in the MIC model and others. This sign change could be problematic for sign-constrained methods such as NG (Breiman (1995)), but it comes out naturally in MIC. Furthermore, MIC is computationally much advantageous by design. See Table 1 in the Supplementary Materials for a comparison study on computing time.

To illustrate the stability of MIC with respect to the value of a, we obtained the MIC estimates for $a \in \{1, 5, 10, 15, \ldots, 100\}$ and plot them in Figure 3. While there are some minor variations mainly owning to the non-convex optimization nature, almost all the estimated coefficients are quite steady in all three examples, suggesting that the MIC estimation is robust to the choice of a.

5. Discussion

MIC is the first method that does sparse estimation by explicitly approximating BIC. BIC is optimal in two aspects: it approximates the posterior distribution of candidate models besides being selection-consistent. This is why BIC has been used as an ultimate yardstick in many variable selection and regularization methods. MIC extends the best subset selection (BSS) to scenarios with large p by optimizing an approximated BIC. Formulated as a smooth optimization problem, MIC is computationally advantageous to the discrete-natured BSS and enjoys the additional benefit in avoiding the post-selection inference. Moreover, the search space in MIC remains to be the entire parameter space. This explains why we expect MIC to outperform many regularization methods that have a much reduced search space for the minimum BIC. By borrowing the knowledge of the fixed penalty parameter for model complexity in BIC, MIC circumvents the tuning parameter selection problem and hence is also computationally advantageous to regularization methods.

The hyperbolic tangent function has been used to approximate the cardinal-

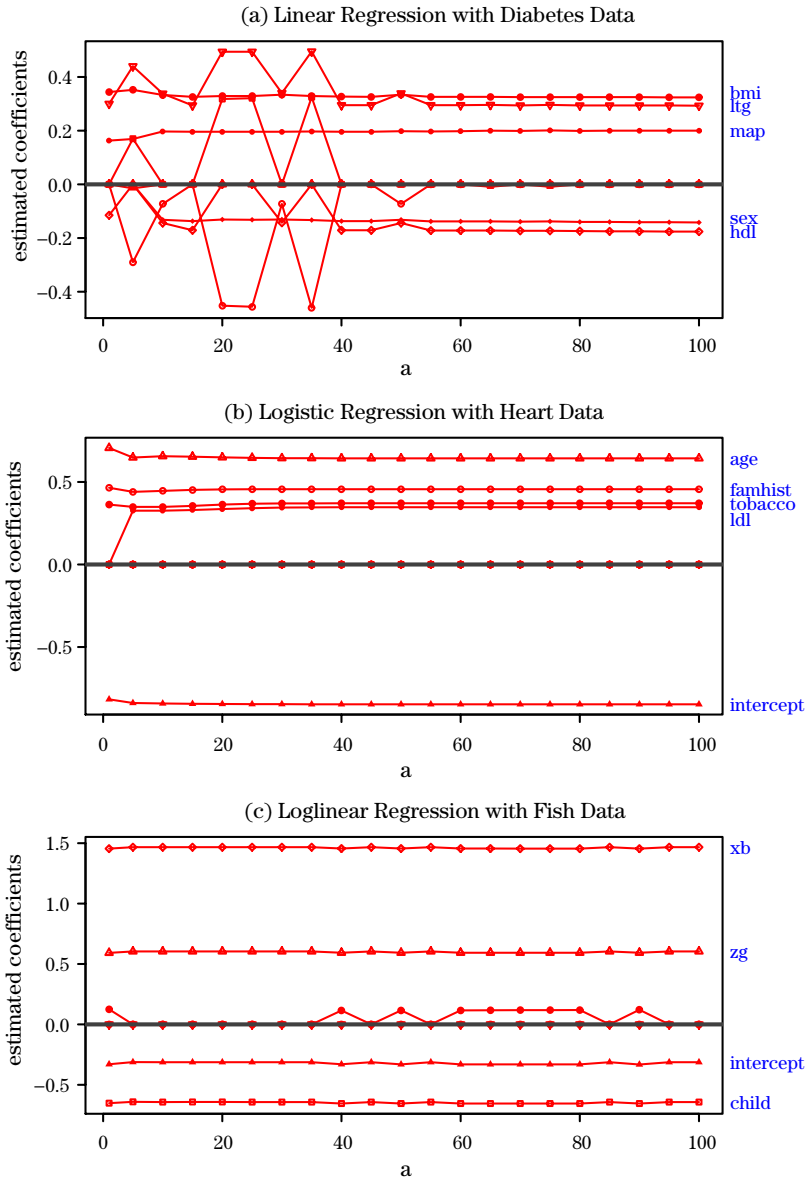

Figure 3. Illustrating the robustness of MIC with respect to the choice of a in these examples. The values of a considered are $\{1, 5, 10, 15, \ldots, 100\}$.

ity in MIC, but can be replaced by other unit dent functions. Since one focus of this paper is on the variable selection consistency, we have adopted BIC by taking $\lambda_0 = \ln(n)$. If the aim is on the model selection efficiency or predictive accuracy, then we can adopt AIC by setting $\lambda_0 = 2$. It can be shown that the

resulting MIC is selection-efficient by applying techniques similar to those used in Zhang, Li and Tsai (2010). In sum, we can obtain variants of MIC by changing its penalty function w and penalty parameter λ_0 to meet practical needs.

To broaden the usefulness of MIC, we would like to generalize MIC by accommodating grouped or structured sparsity (see, e.g., Huang, Breheny and Ma (2012)), and extend MIC to other complex model or dependence structures, such as finite mixture models, longitudinal data, and structural equation modelings (SEM). Similar ideas can be applied to approximate the effective degrees of freedom as well. In these settings, MIC can be particularly useful because the log-likelihood function is not concave and having convex penalties does not help with the optimization problem. We would like to also develop the MIC method for diverging $p \to \infty$ with $p/n \to 0$ (Fan and Peng (2004)) and ultra-high dimensions with $p \gg n$ (Fan and Lv (2008)) by approximating the extended or generalized BIC as pioneered by Chen and Chen (2008).

Supplementary Materials

In the Supplementary Materials, we outline the proofs of Theorems 1 and 2 and we provide more details about an R package **glmMIC** that implements MIC, on which basis a comparison study on computing time is also included.

Acknowledgment

The authors would like to thank the Editor, an associate editor, and two anonymous referees, whose insightful comments and constructive suggestions have greatly improved an earlier version of this manuscript.

References

Akaike, H. (1974). A new look at model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Bakin, S. (1999). *Adaptive Regression and Model Selection in Data Mining Problems*. PhD Thesis, Australian National University, Canberra, Australia.

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics* **41**, 802–837.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics* **5**, 232–253.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373–384.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model spaces. *Biometrika* **95**, 759–771.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society. Series B Statistical Methodology* **49**, 1–18.

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association* **109**, 991–1007.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion). *The Annals of Statistics* **32**, 407–499.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika* **65**, 457–482.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society. Series B Statistical Methodology* **70**, 849–911.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32**, 928–961.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1).

Friedrichs, K. O. (1944). The identity of weak and strong extensions of differential operators. *Transactions of the American Mathematical Society* **55**, 132–151.

Furnival, G. M. and Wilson, R. W. (1974). Regression by leaps and bounds. *Technometrics* **16**, 499–511.

Fu, W. (1998). Penalized regressions: the Bridge versus the Lasso. *Journal of Computational and Graphical Statistics* **7**, 397–416.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*, 2nd Edition. Springer, New York.

Huang, J., Breheny, P. and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science* **4**, 481–499.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: facts and fiction. *Econometric Theory* **21**, 21–59.

Lockhart, R., Taylor, J., Tibshirani, R. and Tibshirani, R. (2014). A significance test for the LASSO. *The Annals of Statistics* **42**, 413–468.

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

Mullen, K. M. (2014). Continuous global optimization in R. *Journal of Statistical Software* **60**(6).

Osborne, M., Presnell, B. and Turlach, B. (2000). On the Lasso and its dual. *Journal of Computational and Graphical Statistics* **9**, 319–337.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

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

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons. New York, NY.

Shen, X., Pan, W. and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* **107**, 223–232.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician* **56**, 29–38.

Su, X. (2015). Variable selection via subtle uprooting. *Journal of Computational and Graphical Statistics* **24**, 1092–1113.

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

Tsallis, C. and Stariolo, D. A. (1996). Generalized simulated annealing. *Physica A* **233**, 395–406.

Wang, H., Li, R. and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568.

Weigend, A. S., Rumelhart, D. E. and Huberman, B. A. (1991). Generalization by weight-elimination with application to forecasting. In *Advances in Neural Information Processing Systems 3* (Denver 1990), R. P. Lippmann, J. E. Moody, and D. S. Touretzky, Editors, 875–882. Morgan Kaufmann. San Mateo, CA.

Xiang, Y., Gubian, S., Suomela, B. and Hoeng, J. (2013). Generalized simulated annealing for global optimization: The GenSA package. *The R Journal* **5**(1).

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894–942.

Zhang, Y., Li, R. and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association* **105**, 312–323.

Zou, H. and Li, Y. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* **36**, 1509–1533.

Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. *Journal of Machine Learning Research* **7**, 2541–2563.

Department of Mathematical Sciences, University of Texas, El Paso, TX 79968, USA.

E-mail: xsu@utep.edu

Department of Mathematics and Statistics, San Diego State University, CA 92182, USA.

E-mail: jjfan@mail.sdsu.edu

Department of Mathematics and Statistics, San Diego State University, CA 92182, USA.

E-mail: ralevine@sciences.sdsu.edu

Department of Periodontology, Creighton University, Omaha, NE 68178, USA.

E-mail: MarthaNunn@creighton.edu

Graduate School of Management, University of California, Davis, CA 95616, USA.

E-mail: cltsai@ucdavis.edu

(Received July 2016; accepted April 2017)