Statistica Sinica: Supplement

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]

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

Supplementary Material

This supplementary material outlines the proofs of Theorems 1 and 2 and provides more details about an R package **glmMIC** that implements MIC, on which basis a comparison study on computing time is also included.

S1 Proofs

The asymptotic properties of MIC can be conveniently derived by following similar arguments of Fan and Li (2001) and utilizing the properties of the hyperbolic tangent penalty.

S1.1 Proof of Theorem 1

We first establish (i) by checking conditions in Theorem 1 of Fan and Li (2001). Note that the quantity $p_{\lambda_n}(|\beta_j|)$ corresponds to

$$p_{\lambda_n}(|\beta_j|) = \frac{\ln(n)}{2n} \cdot w(\gamma_j)$$

in MIC. Some quantities involved in the reparameterization $\beta = \gamma w(\gamma)$ are summarized below:

$$\begin{cases} \dot{w} & = & dw(\gamma)/d\gamma & = & 2a_n\gamma(1-w^2) \\ \ddot{w} & = & d^2w(\gamma)/d\gamma^2 & = & 2a_n(1-w^2)(1-4a_n\gamma^2 w) \\ w & = & \tanh(a_n\gamma^2) & = & \left(e^{a_n\gamma^2} - e^{-a_n\gamma^2}\right) / \left(e^{a_n\gamma^2} + e^{-a_n\gamma^2}\right) \\ 1-w^2 & = & \text{sech}(a_n\gamma^2) & = & 2/\left(e^{a_n\gamma^2} + e^{-a_n\gamma^2}\right) \end{cases}$$

Since $a_n = O(n)$, $\gamma \to \beta$ and $w(\gamma) \to 1$ for $\beta \neq 0$. It follows that, for $\beta \neq 0$,

$$\begin{aligned} \dot{p}_{\lambda_n}(|\beta|) &= \frac{dp_{\lambda_n}(|\beta|)}{d\,\beta} = \frac{\ln(n)}{2n}\frac{\dot{w}}{w+\gamma\dot{w}} \\ &= \frac{\ln(n)}{n}\frac{a_n\gamma(1-w^2)}{w+2a_n\gamma^2(1-w^2)} \\ &= \frac{\ln(n)}{n}\frac{2a_n\gamma}{e^{a_n\gamma^2}+e^{-a_n\gamma^2}+4a_n\gamma^2} \\ &= \frac{\ln(n)}{n}O\left\{a_n e^{-a_n\gamma^2}\right\} \\ &= o(1/\sqrt{n}). \end{aligned}$$

Hence, $\max_j \{\dot{p}_{\lambda_n}(|\beta_{0j}|): \ \beta_{0j} \neq 0\} = o(1/\sqrt{n})$. Similarly, it can be shown that, for $\beta \neq 0$,

$$\ddot{p}_{\lambda_n}(|\beta|) = \frac{d^2 p_{\lambda_n}(|\beta|)}{d\,\beta^2} = \frac{\ln(n)}{2n}\,\frac{w\ddot{w} - 2\dot{w}^2}{(w + \gamma\,\dot{w})^3} \ \xrightarrow{p}\ 0.$$

and so is $\max_j \{\ddot{p}_{\lambda_n}(|\beta_{0j}|): \ \beta_{0j} \neq 0\}$.

Therefore, there exists a local minimizer $\widetilde{\boldsymbol{\beta}}$ of $Q_n(\boldsymbol{\beta})$ such that $\| \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \| = O_p(1/\sqrt{n})$ by Theorem 1 of Fan and Li (2001). □

To establish sparsity of $\widetilde{\boldsymbol{\beta}}_{(0)}$ in (ii), it suffices to show that, for any $\sqrt{n}$-consistent $\boldsymbol{\beta} = (\boldsymbol{\beta}_{(1)}^T, \boldsymbol{\beta}_{(0)}^T)^T$ such that $\| \boldsymbol{\beta}_{(1)} - \boldsymbol{\beta}_{0(1)} \| = O_p(1/\sqrt{n})$ and $\| \boldsymbol{\beta}_{(0)} \| = O_p(1/\sqrt{n})$, we have

$$\frac{\partial Q_n(\boldsymbol{\beta})}{\partial \beta_j} = \begin{cases} > 0 & \text{if } \beta_j > 0 \\ \\ < 0 & \text{if } \beta_j < 0 \end{cases} \tag{S1.1}$$

for any component β_j of $\boldsymbol{\beta}_{(0)}$ with probability tending to 1 as $n \to \infty$.

Consider

$$\frac{\partial Q_n(\boldsymbol{\beta})}{\partial \beta_j} \ = \ -\ \frac{2}{n}\,\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} + \frac{\ln(n)}{n} \cdot \frac{\partial w(\gamma_j)}{\partial \beta_j} \ = \ I + II$$

for $j = (q+1), \ldots, p$ when evaluated at $\boldsymbol{\beta}$. Note that $\beta_j = O_p(1/\sqrt{n})$ yet $\beta_j \neq 0$ for $\beta_j \in \boldsymbol{\beta}_{(0)}$. By standard arguments (see Fan and Li, 2002) and using the fact that $\| \boldsymbol{\beta} - \boldsymbol{\beta}_0 \| = O_p(1/\sqrt{n})$, it can be shown that the first term I is of order $O_p(1/\sqrt{n})$ under the regularity conditions. For the second term II, the analysis is more subtle, depending on whether $a_n\gamma^2$ goes to 0, a constant, or ∞. Since it is desirable that

$$\begin{aligned} \frac{\partial w(\gamma_j)}{\partial \beta_j} &= \frac{\dot{w}_j}{w_j + \gamma_j \dot{w}_j} = \frac{2a\gamma_j(1-w_j^2)}{w_j + 2a\gamma_j^2(1-w_j^2)} \\ &= \frac{4a_n\gamma}{e^{a_n\gamma^2} + e^{-a_n\gamma^2} + 4a_n\gamma^2} \end{aligned} \tag{S1.2}$$

is $O_p(\sqrt{n})$ or even higher order to have sparsity, neither the choice $a_n\gamma^2 = o(1)$ or $a_n\gamma^2 \to \infty$ is not allowable because in either scenario, $\partial w(\gamma_j)/\partial\beta_j$ is $o(1)$. Now set $a_n\gamma^2 = O_p(1)$. The condition $\gamma w(\gamma) = \gamma \tanh(a_n\gamma^2) = \beta = O_p(1/\sqrt{n})$ leads to the rate $\gamma = 1/\sqrt{n}$ and hence $a_n = O(n)$. Therefore, the $O(n)$ rate for a_n seems to be the unique choice after taking all the side conditions into consideration.

In this case, $\partial w(\gamma_j)/\partial\beta_j = O_p(\sqrt{n})$. The second term becomes $II = O_p\left(\ln(n)n^{1/2}/n\right) = O_p\left(\ln(n)/\sqrt{n}\right)$. Moreover, it can be easily seen that the sign of $\partial w(\gamma_j)/\partial\beta_j$ in (S1.2) is determined by $\dot{w}_j$ and hence γ_j or β_j, because $w_j \geq 0$ and $\gamma_j\dot{w}_j \geq 0$. Put together, $\partial Q_n(\boldsymbol{\beta})/\partial\beta_j$ in (S1.1) is dominated by the second term II and its sign is determined by β_j. Therefore, the desired sparsity of $\widetilde{\boldsymbol{\beta}}$ is established. □

To show asymptotic normality of $\widetilde{\boldsymbol{\beta}}_{(1)}$ in (ii), a close look at the proof of Theorem 2 in Fan and Li (2001) reveals that it suffices to show that the contribution from the penalty term to the estimating equation is negligible relative to the gradient of the log-likelihood function. More specifically, if we can show that

$$\frac{\ln(n)}{n} \left.\frac{\partial w(\gamma_j)}{\partial \beta_j}\right|_{\beta_j=\tilde{\beta}_j} = o_p\left(\frac{1}{\sqrt{n}}\right), \tag{S1.3}$$

for $j = 1, \ldots, q$, then Slutsky's theorem can be applied to complete the proof. Equation (S1.3) holds since, for any non-zero $\beta_j \in \boldsymbol{\beta}_{0(1)}$, we have $\tilde{\beta}_j = \beta_j + O_p(1/\sqrt{n})$ and hence $\tilde{\gamma}_j = \gamma_j + o_p(1)$ by the continuous mapping theorem, where $\tilde{\beta}_j = \tilde{\gamma}_j w(\tilde{\gamma}_j)$ and $\beta_j = \gamma_j w(\gamma_j)$. It follows that $\partial w(\tilde{\gamma}_j)/\partial \beta_j = o_p(1)$ in this case as shown earlier in the proof of (i). Therefore $\dot{\rho}_n(\widetilde{\beta}_j) = o_p\{\ln(n)/n\} = o_p(1/\sqrt{n})$. The proof is completed. □

S1.2 Proof of Theorem 2

According to the definition, $\boldsymbol{\gamma}_0$ is a constant that depends on n via a_n. In view of $\gamma - \beta = \gamma - \gamma w(\gamma) = \gamma\{1 - \tanh(a_n \gamma^2)\} = 2\gamma/\{\exp(2a_n\gamma^2) + 1)\}$, it follows that $|\gamma_{0j} - \beta_{0j}| = O\{\exp(-2a_n\gamma_{0j}^2)\}$ for $\gamma_{0j} \neq 0$ and 0 otherwise. Hence

$$\begin{aligned}
\| \boldsymbol{\gamma}_0 - \boldsymbol{\beta}_0 \|_2 &\leq \| \boldsymbol{\gamma}_0 - \boldsymbol{\beta}_0 \|_1 = \sum_{j=1}^{q} |\gamma_{0j} - \beta_{0j}| \\
&\leq \frac{2q \max_{1\leq j\leq q} \beta_j}{\exp\{2a_n \min_{1\leq j\leq q} \gamma_{0j}^2\} + 1} \\
&= O\left\{\exp\{-2a_n \min_{1\leq j\leq q} \gamma_{0j}^2\}\right\}.
\end{aligned}$$

Moreover, since the function $\beta = \gamma w(\gamma)$ is continuous and so is its inverse, the continuous mapping theorem yields $\parallel \widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 \parallel = o_p(1)$.

To study the asymptotic property of $\widetilde{\boldsymbol{\gamma}}$, we consider $\widetilde{\boldsymbol{\gamma}}$ as a local minimizer of the objective function $Q_n(\cdot) = -2L(\mathbf{W}\boldsymbol{\gamma}) + \ln(n) \cdot \mathrm{tr}(\mathbf{W})$ as stated in (2.1). Since $Q_n(\boldsymbol{\gamma})$ is smooth in $\boldsymbol{\gamma}$, $\tilde{\gamma}$ satisfies the first-order necessary condition $\partial Q_n(\widetilde{\boldsymbol{\gamma}})/\partial\boldsymbol{\gamma} = \mathbf{0}$, which gives

$$\begin{aligned}
& -\frac{2}{n}\frac{\partial L(\widetilde{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\frac{\partial \widetilde{\boldsymbol{\beta}}}{\partial \boldsymbol{\gamma}} + \frac{\ln(n)}{n}\frac{\partial \sum_j w(\tilde{\gamma}_j)}{\partial \boldsymbol{\gamma}} = 0 \\
\Longrightarrow \quad & \nabla L(\widetilde{\boldsymbol{\beta}})\, \mathrm{diag}\,(w_j + \tilde{\gamma}_j \dot{w}_j) = \frac{\ln(n)}{2}\left(\frac{dw_j}{d\gamma_j}\right)_{j=1}^{p} \\
\Longrightarrow \quad & \nabla L(\widetilde{\boldsymbol{\beta}}) = \frac{\ln(n)}{2}\left(\frac{\dot{w}_j}{w_j + \tilde{\gamma}_j \dot{w}_j}\right)_{j=1}^{p}. \qquad \text{(S1.1)}
\end{aligned}$$

Next, applying Taylor's expansion of the LHS $\nabla L(\widetilde{\boldsymbol{\beta}})$ at $\boldsymbol{\gamma}_0$ gives

$$\frac{\ln(n)}{2}\left(\frac{\dot{w}_j}{w_j + \tilde{\gamma}_j \dot{w}_j}\right)_{j=1}^{p} = \nabla L(\boldsymbol{\beta}_0) + \nabla^2 L(\boldsymbol{\beta}_0)\left(\left.\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\gamma}}\right|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0}\right)(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) + \mathbf{r}_n,$$

where $\mathbf{r}_n$ denotes the remainder term. It follows that

$$\begin{aligned}
\left(\mathrm{diag}(w_j + \gamma_j \dot{w}_j)|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0}\right)(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \;=\; & \{-\nabla^2 L(\boldsymbol{\beta}_0)\}^{-1} \cdot \\
& \left\{\nabla L(\boldsymbol{\beta}_0) - \frac{\ln(n)}{2}\left(\frac{\dot{w}_j}{w_j + \tilde{\gamma}_j \dot{w}_j}\right)_{j=1}^{p} + \mathbf{r}_n\right\}.
\end{aligned}$$

Therefore,

$$\sqrt{n}\,[\mathbf{D}(\boldsymbol{\gamma}_0)(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) + \mathbf{b}_n] = \left\{-\frac{\nabla^2 L(\boldsymbol{\beta}_0)}{n}\right\}^{-1}\frac{\nabla L(\boldsymbol{\beta}_0)}{\sqrt{n}} + \mathbf{r}'_n, \qquad \text{(S1.2)}$$

where $\mathbf{D}(\boldsymbol{\gamma}_0)$ and $\mathbf{b}_n$ are defined in (3.3) and (3.4), respectively, and the remainder term is

$$\mathbf{r}'_n = \left\{ -\frac{\nabla^2 L(\boldsymbol{\beta}_0)}{n} \right\}^{-1} \frac{\mathbf{r}_n}{\sqrt{n}}.$$

Under regularity conditions, standard arguments yield $\{-\nabla^2 L(\boldsymbol{\beta}_0)/n\}^{-1} \xrightarrow{p} \mathbf{I}^{-1}(\boldsymbol{\beta}_0)$; $\nabla L(\boldsymbol{\beta}_0)/\sqrt{n} \xrightarrow{d} \mathbf{N}\left\{\mathbf{0}, \mathbf{I}(\boldsymbol{\beta}_0)\right\}$; and $\mathbf{r}'_n = o_p(1)$ as $n \to \infty$. Bringing these results into (S1.2) and an appeal to Slutsky's Theorem give the desired asymptotic normality in (3.2).

Note that the elements D_{jj} of the diagonal matrix $\mathbf{D}(\boldsymbol{\gamma}_0)$ in (3.3) are evaluated at $\boldsymbol{\gamma}_0$. We have

$$D_{jj} = w(\gamma_{0j}) + \gamma_{0j}\, \dot{w}(\gamma_{0j}) = \frac{e^{a_n \gamma_{0j}^2} - e^{-a_n \gamma_{0j}^2} - 4 a_n \gamma_{0j}^2}{e^{a_n \gamma_{0j}^2} + e^{-a_n \gamma_{0j}^2}}.$$

Since $a_n = O(n)$, it can be seen that $\lim_{n\to\infty} D_{jj} = 1$ if $\gamma_{0j} \neq 0$ and 0 otherwise.

To study the limit of bias $\mathbf{b}_n$, we rewrite (3.4) as

$$\mathbf{b}_n = \left\{ -\frac{\nabla^2 L(\boldsymbol{\beta}_0)}{n} \right\}^{-1} \frac{\ln(n)}{2\sqrt{n}} \left(\frac{1}{\sqrt{n}} \frac{\dot{w}_j}{w_j + \tilde{\gamma}_j \dot{w}_j} \right)_{j=1}^{p}. \tag{S1.3}$$

Note that $\{-\nabla^2 L(\boldsymbol{\beta}_0)/n\}^{-1} \xrightarrow{p} \mathbf{I}^{-1}(\boldsymbol{\beta}_0) \succ 0$ is evaluated at the constant $\boldsymbol{\beta}_0$ or $\boldsymbol{\gamma}_0$ while the last term of $\mathbf{b}_n$, with components $\dot{w}/\{\sqrt{n}\,(w + \gamma \dot{w})\}$, is evaluated at $\tilde{\boldsymbol{\gamma}}$. For $\gamma_{0j} \neq 0$, we have $\tilde{\gamma}_j = \gamma_{0j} + o_p(1)$; for $\gamma_{0j} = 0$, we have

$\tilde{\gamma}_j = O_p(1/\sqrt{n})$. Consider the term

$$\begin{aligned}\frac{\dot{w}}{w+\gamma\dot{w}} &= \frac{4a_n\gamma}{\exp(a_n\gamma^2)-\exp(-a_n\gamma^2)+4a_n\gamma^2} \\ &= \begin{cases} O_p(a_n\, e^{-a_n\gamma_{0j}^2}) = o_p(1) & \text{if} \quad \tilde{\gamma}_j = \gamma_{0j} + o_p(1), \\ O_p(1/\tilde{\gamma}_j) = O_p(\sqrt{n}) & \text{if} \quad \tilde{\gamma}_j = O_p(1/\sqrt{n}). \end{cases}\end{aligned}$$

Hence the last term of $\mathbf{b}_n$ is $O_p(1)$ in both cases. As a result, $\mathbf{b}_n = o_p(1)$ as $n \to \infty$. Its componentwise convergence rates are exponential for estimates of nonzero γ_{0j}'s and $O_p\{\ln(n)/\sqrt{n}\}$ for estimates of zero coefficients. This completes the proof. □

S2 MIC Implementation

We have made available an R package **glmMIC** that implements the proposed MIC method, for which some details are provided in Section S2.1. Section S2.2 presents a comparison study on computing time.

S2.1 R Package glmMIC

We put together an R package **glmMIC** based on our implementation of MIC, which is now available from GitHub: `https://github.com/xgsu/glmMIC`. The package can be easily installed with the facility from R package **devtools** (Wickham et al., 2016) as follows:

```
devtools::install_github("xgsu/glmMIC")

library(glmMIC)
```

The main function `glmMIC` has the following arguments:

```
glmMIC(formula, preselection=NULL, family=c("gaussian",
    "binomial", "poisson"), data, beta0=NULL,
    select.intercept=T, criterion="BIC", lambda0=NULL,
    a0=NULL, rounding.digits=4, use.GenSA=F,
    lower=NULL, upper=NULL, maxit.global=100,
    maxit.local=50, epsilon=1e-06, se.gamma = T,
    CI.gamma=T, conf.level=0.95,
    se.beta=T, fit.ML=FALSE, details=F)
```

Details about the options can be obtained with the help file `?glmMIC`, where several real examples (i.e., results presented in Table 4) can also be found. In particular, we have made two optimization methods available for MIC. The first is to use the simulated annealing algorithm implemented by the `method="SANN"` option in R function `optim` followed by a local optimization BFGS algorithm `method="BFGS"`. The quasi Newton BFGS algorithm makes sure the procedure stops at a local minimum. This is an approach found quite useful in practice. The second is to use the generalized simulated annealing as implemented in R package `GenSA`. There is no need of additional iterations from a local optimization algorithm with this approach. In our experiences, `GenSA` is slightly slower but it often converges to the same solution with different runs. Inspired by the comments from an anonymous reviewer, the `glmMIC` function also includes a `preselection` option that allows one to pre-select variables into the model. The output of `glmMIC` is an object of S3 class `glmMIC`, for which generic functions `print` and `plot` can be used.

S2.2 Comparison in Computing Time

To compare computing time, we generate data from the same models (A, B, C) given by (4.1) and change the dimensions by adding predictors with 0 regression coefficients so that $p \in \{12, 50\}$ and varying sample sizes so that $\in \{100, 1000\}$. Since the best subset selection implemented by R package **bestglm** (McLeod and Xu, 2014) breaks down for $p \geq 50$, we have used one of its stepwise surrogates – the backward deletion implemented by R function `step()`. Computation of LASSO, SCAD, and MCP is done via the updated package **ncvreg** (Breheny and Lee, 2016).

Table 1 presents the average CUP time in seconds from six methods: the whole model with all predictors included (Whole), the backward deletion, MIC, LASSO, SCAD, and MCP. The default settings of these implementations are essentially used. However, to make them comparable, we have applied the same tolerance level $\epsilon = 0.0001$ for convergence, which is the default in **ncvreg**. All the computations are done on a Lenovo T450s laptop computer with CPU 2.60 GHz and 12GB RAM. It is no surprise to see that MIC outperforms other selection methods for a great deal in all the scenarios considered here. In fact, the time consumed by MIC is quite comparable with that used in fitting the whole model via OLS or ML.

References

Breheny, P. and Lee, S. (2016). R Package **ncvreg**. URL `https://cran.r-project.org/web/packages/ncvreg/`.

McLeod, A. I. and Xu, C. (2014). R Package **bestglm**. URL `https://cran.r-project.org/web/packages/bestglm/`.

Wickham, H., Chang, W., et al. (2016). R Package **devtools**. URL `https://cran.r-project.org/web/packages/devtools/`.

Table 1: Computing time (in seconds) comparison among six methods: the whole model without variable selection (Whole), backward deletion as a surrogate of best subset selection (Backward), MIC, LASSO, SCAD, and MCP. The computing time is averaged over five simulation runs.

			Methods					
Model	p	n	Whole	Backward	MIC	LASSO	SCAD	MCP
A	12	100	0.000	0.042	0.006	0.040	0.046	0.034
		1000	0.002	0.074	0.012	0.152	0.154	0.170
	50	100	0.000	0.656	0.094	0.190	0.292	0.274
		1000	0.010	2.886	0.080	0.408	0.590	0.662
B	12	100	0.004	0.130	0.012	0.204	0.180	0.172
		1000	0.008	0.548	0.060	1.074	1.080	1.028
	50	100	0.016	8.614	0.028	1.152	0.684	0.436
		1000	0.032	24.268	0.598	2.534	3.618	3.960
C	12	100	0.006	0.126	0.010	0.218	0.166	0.136
		1000	0.014	0.512	0.022	1.032	0.894	0.878
	50	100	0.006	2.966	0.146	1.782	1.328	0.952
		1000	0.036	20.386	0.094	2.420	3.752	4.090