Regularization Parameter Selection in the Group Lasso (⋆)

Teppei SHIMAMURA[1], Hiroyuki MINAMI[2], Masahiro MIZUTA[2]

[1] Graduate School of Information Science and Technology, Hokkaido University

e-mail: shima@iic.hokudai.ac.jp

[2] Information Initiative Center, Hokkaido University

Abstract: This article discusses the problem of choosing a regularization parameter in the group Lasso proposed by Yuan and Lin (2006), an l_1-regularization approach for producing a block-wise sparse model that has been attracted a lot of interests in statistics, machine learning, and data mining. It is important to choose an appropriate regularization parameter from a set of candidate values, because it affects the predictive performance of the fitted model. However, traditional model selection criteria, such as AIC and BIC, cover only models estimated by maximum likelihood estimation and can not be directly applied for regularization parameter selection. We propose an information criterion for regularization parameter selection in the group Lasso in the framework of maximum penalized likelihood estimation.

Keywords: L_1-regularization approach, model selection problem, information-theoretic approach

1. Introduction

The group Lasso proposed by Yuan and Lin (2006) produces a block-wise sparse model while keeping high prediction accuracy in linear regression problem with p explanatory variables partitioned into J blocks, which implies that some blocks of the regression coefficients are exactly zero. It achieves variable selection at the group level and is useful for the special case in linear regression when not only continuous but also categorical variables are present or when there are meaningful blocks of variables such as polynomial regression.

This article deals with the problem of choosing a better one for a set of regularization parameter values in the Group Lasso problem. The least regularized group Lasso when the regularization parameter λ is equal to zero corresponds to ordinal least squares estimates, while the most regularized group Lasso using $\lambda = \infty$ yields a constant fit. The choice of the regularization parameter λ affects the predictive performance and sparseness of the

(⋆) Work partially supported by JSPS research fellowships for young scientists

fitted model simultaneously. Thus regularization parameter selection is one of the most important practical issues for the group Lasso.

The purpose of this article is to derive an information criterion (Akaike, 1974; Konishi and Kitagawa, 1996) for regularization parameter selection in the group Lasso. A challenging task of the above derivation is to calculate influence functions for the group Lasso estimators. Like the Lasso, the penalized likelihood function corresponding to the group Lasso is not differentiable at the origin which leads to make it difficult to obtain the influence functions. To avoid this problem, we only consider the nonzero components of the group Lasso estimator satisfying Karush-Kuhn-Tucker conditions and accomplish the above derivation with some assumptions. Section 2 describes the definition of the group Lasso in the context of linear regression models. Section 3 presents information criteria for model selection and regularization parameter selection in the group Lasso.

2. Group Lasso

Consider the regression model with J factors:

$$\boldsymbol{y} = \sum_{j=1}^{J} \boldsymbol{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{y}$ is an $n \times 1$ response vector, $\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, $\boldsymbol{X}_j$ is an $n \times p_j$ design matrix corresponding to the j-th factor, $\boldsymbol{\beta}_j$ is a coefficient vector of size p_j, $j = 1, \ldots, J$, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_J)$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_J')'$. For simplicity of explanation, assume that the response vector and each input variable are centerized, and each $\boldsymbol{X}_j$ is orthonormalized through Gram-Schmidt orthonormalization, that is, $\boldsymbol{X}_j^{\mathrm{T}} \boldsymbol{X}_j = \boldsymbol{I}_{p_j}$. The group Lasso estimator is defined as a solution to minimize the functional

$$\frac{1}{2} \| \boldsymbol{y} - \sum_{j=1}^{J} \boldsymbol{X}_j \boldsymbol{\beta}_j \|^2 + \lambda \sum_{j=1}^{J} (\boldsymbol{\beta}_j^{\mathrm{T}} \boldsymbol{K}_j \boldsymbol{\beta}_j)^{1/2} \tag{2}$$

where $\lambda \geq 0$ is a regularization parameter and $\boldsymbol{K}_1, \ldots, \boldsymbol{K}_J$ are symmetric $p_j \times p_j$ positive definite matrices. In this article, we choose to set $\boldsymbol{K}_j = p_j \boldsymbol{I}_{p_j}$ which is equivalent to the setting of Yuan and Lin (2006).

Let a subset of factor indices $\{1, \ldots, J\}$ where the coefficient vectors are nonzero be $\mathcal{G} = \{j \in \{1, \ldots, J\} : \boldsymbol{\beta}_j \neq \boldsymbol{0}\}$. From Karush-Kuhn-Tucker conditions, a necessary and sufficient condition for $\boldsymbol{\beta}$ to be a solution to (2) when $\boldsymbol{K}_j = p_j \boldsymbol{I}_{p_j}$ is

$$-\boldsymbol{X}_j'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sqrt{p_j} \| \boldsymbol{\beta}_j' \boldsymbol{\beta}_j \|^{-1/2} \boldsymbol{\beta}_j = \boldsymbol{0} \quad \forall \boldsymbol{\beta}_j \neq \boldsymbol{0},\ j \in \mathcal{G}, \tag{3}$$

$$\| -\boldsymbol{X}_j'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \| \leq \lambda \sqrt{p_j} \quad \forall \boldsymbol{\beta}_j = \boldsymbol{0},\ j \notin \mathcal{G}. \tag{4}$$

Yuan and Lin (2006) extended the shooting algorithm for the Lasso proposed by Fu (1999) to calculate a solution to (2). We use Yuan and Lin's algorithm directly.

3. Regularization Parameter Selection in the Group Lasso

We start by giving an overview of model selection via information-theoretic approach. Then we move on to discuss the regularization parameter selection problem in the group Lasso and present the derivation of information criteria.

3.1. Information-Theoretic Approach for Model Selection

The problem of model selection is to choose one from a series of r candiate models. Assume that observations $\boldsymbol{y}$ are independently distributed from an unknown "true" distribution function $G(\boldsymbol{y})$ which has the density function $g(\boldsymbol{y})$. One usually models $g(\boldsymbol{y})$ by a member of a parametric family, $\mathcal{F} = \{f(\boldsymbol{y}|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$, where $\boldsymbol{\theta}$ is a p-dimensional unknown parameter vector, and Θ is an open subset of R^p and estimates $\hat{\boldsymbol{\theta}}$ based on observations $\boldsymbol{y}$. We also denote a future observation from the density g by z. Information criteria have been proposed as estimators of the expected log-likelihood given by

$$\text{IC} = -2\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) + 2\hat{b}(G) \tag{5}$$

where $\hat{b}(G)$ is a properly defined approximation to the bias term given by

$$b(G) = E_{G(y)}\left[\int \log f(z|\hat{\boldsymbol{\theta}}) d\hat{G}(z) - \int \log f(z|\hat{\boldsymbol{\theta}}) dG(z)\right] \tag{6}$$

with the expectation taken over the joint distribution of $\boldsymbol{y}$, that is, $\prod_{i=1}^{n} dG(y_i)$.

In general, it is difficult to obtain the bias $b(G)$ in a closed form. Thus the bias estimate $\hat{b}(G)$ is usually given as a consistent estimator of the asymptotic bias, $b_1(G)$, in the expansion

$$b(G) = \frac{1}{n} b_1(G) + O(n^{-2}). \tag{7}$$

For a general statistical functional estimator $\hat{\boldsymbol{\theta}} = \boldsymbol{T}(\hat{G})$ where $\boldsymbol{T}(\cdot)$ is a p-dimensional functional vector on the space of distribution functions, Konishi and Kitagawa (1996) derived the asymptotic bias as a function of the empirical influence function of the estimator $\hat{\boldsymbol{\theta}}$ and the score function of the hypothesized statistical model $f(\cdot|\hat{\boldsymbol{\theta}})$. Then the asymptotic bias in (6) is given by

$$b_1(G) = \text{tr}\left\{\int \boldsymbol{T}^{(1)}(z; G) \left.\frac{\partial \log f(z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\theta=T(G)} dG(z)\right\} \tag{8}$$

where $\boldsymbol{T}^{(1)}(z;G) = (T_1^{(1)}(z;G),\ldots,T_p^{(1)}(z;G))'$, $\partial/\partial\boldsymbol{\theta}' = (\partial/\partial\theta_1,\ldots,\partial/\partial\theta_p)$ and $T_i^{(1)}(z;G)$ is the influence function defined by

$$T_i^{(1)}(z;G) = \lim_{\epsilon\to\infty} \frac{T_i[(1-\epsilon)G + \epsilon\delta_z] - T_i(G)}{\epsilon} \tag{9}$$

with δ_z being a point mass at z. By replacing the asymptotic bias $b_1(G)$ by $b_1(\hat{G})$, we can obtain information criterion. Model selection proceeds by choosing the model which gives the smallest value of information criterion from a set of candidate models $\{f(y|\hat{\boldsymbol{\theta}}_1), f(y|\hat{\boldsymbol{\theta}}_2),\ldots,f(y|\hat{\boldsymbol{\theta}}_r)\}$.

3.2. Proposal for Regularization Parameter Selection

We consider a setting of regularization parameter selection where one chooses a better value among a sequence of regularization parameter values $\{\lambda_1,\lambda_2,\ldots,\lambda_r\}$ in the framework of the group Lasso. Suppose that each model M depends on the regularization parameter λ and is specified by the normal linear model

$$f(y|\hat{\boldsymbol{\theta}}) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}}\exp\left\{-\frac{(y-\hat{\boldsymbol{\beta}}'\boldsymbol{x})^2}{2\hat{\sigma}^2}\right\} \tag{10}$$

where $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1',\ldots,\hat{\boldsymbol{\beta}}_J')'$ and $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}',\hat{\sigma}^2)'$ are given by the maximizer of the penalized log-likelihood

$$l_\lambda(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{i=1}^n \log f(y_i|\boldsymbol{\theta}) - \frac{\lambda}{\sigma^2}\sum_{j=1}^J(\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j)^{1/2} \tag{11}$$

where $\lambda > 0$ is a regularization parameter. Multiplying (11) by σ^2 shows that the problem of maximizing the penalized log-likelihood is precisely equivalent to that maximizing

$$\sigma^2\sum_{i=1}^n \log f(y_i|\boldsymbol{\theta}) - \lambda\sum_{j=1}^J(\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j)^{1/2}. \tag{12}$$

Since, in normal linear case,

$$\sigma^2\sum_{i=1}^n \log f(y_i|\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^n(y_i-\boldsymbol{\beta}'\boldsymbol{x}_i)^2 - \frac{n\sigma^2}{2}\log(2\pi\sigma^2) \tag{13}$$

and σ^2 is independent of the estimation of $\boldsymbol{\beta}$, thus the maximization of (12) in terms of $\boldsymbol{\beta}$ is equivalent to the minimization of

$$\frac{1}{2}\sum_{i=1}^n(y_i-\boldsymbol{\beta}'\boldsymbol{x}_i)^2 + \lambda\sum_{j=1}^J(\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j)^{1/2} \tag{14}$$

eliminating dependence on σ^2, which is exactly the group Lasso criterion used in (2). After $\hat{\boldsymbol{\beta}}$ is obtained, $\hat{\sigma}^2$ is given by the solution of the following equation

$$\frac{\partial}{\partial \sigma^2}\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y-\hat{\boldsymbol{\beta}}'\boldsymbol{x}_i)^2-\frac{n}{2}\log(2\pi\sigma^2)-\frac{\lambda}{\sigma^2}\sum_{j=1}^{J}(\hat{\boldsymbol{\beta}}_j'\boldsymbol{K}_j\hat{\boldsymbol{\beta}}_j)^{1/2}\right\}=0. \tag{15}$$

Our task is to calculate the influence function of the group Lasso estimator $\hat{\boldsymbol{\theta}}$. Note that the penalized log-likelihood (11) is not differentiable in terms of $\boldsymbol{\theta}$ when some block-wise components of $\boldsymbol{\theta}$ are exactly zero in the solution.

Let the subset of factor indices $\{1,\ldots,J\}$ where the nonzero block-wise components of $\boldsymbol{\beta}$ are nonzero for the regularization parameter $\lambda_k,\ k=1,\ldots,r$, be $\mathcal{G}_k=\{j\in\{1,\ldots,J\}:\hat{\boldsymbol{\beta}}_j\neq\mathbf{0}\}$. Also define the group Lasso estimator for λ_k by $\hat{\boldsymbol{\theta}}_k$. Suppose that $\mathcal{G}_k$ is locally constant with respect to $\boldsymbol{y}$, that is, when a small perturbation $\boldsymbol{\varepsilon}$ is imposed on the observations $\boldsymbol{y}$, the zero block-wise components of $\hat{\boldsymbol{\beta}}_k$, that is, $\hat{\boldsymbol{\beta}}_j,\ j\notin\mathcal{G}_k$, stay at $\mathbf{0}$. Therefore the penalized log-likelihood function is twice differentiable in terms of $\boldsymbol{\theta}_{\mathcal{G}_k}$ where $\boldsymbol{\theta}_{\mathcal{G}_k}=(\boldsymbol{\beta}_{\mathcal{G}_k}',\sigma_k^2)'$ and $\boldsymbol{\beta}_{\mathcal{G}_k}$ is the corresponding coefficient vector for $\mathcal{G}_k$. Denote the sub design matrix for $\mathcal{G}_k$ by $\boldsymbol{X}_{\mathcal{G}_k}=[\cdots,\boldsymbol{x}_j,\cdots]_{j\in\mathcal{G}_k}$ where $\boldsymbol{x}_j$ is the j-th column of $\boldsymbol{X}$, and the number of $\hat{\boldsymbol{\theta}}_{\mathcal{G}_k}$ by $p_{\mathcal{G}_k}$.

Let $\boldsymbol{T}_{\mathcal{G}_k}(G)$ be the $p_{\mathcal{G}_k}$-dimensional functional defined by

$$\int \boldsymbol{\psi}_{\mathcal{G}_k}(y,\boldsymbol{\theta}_{\mathcal{G}_k})\Big|_{\boldsymbol{\theta}_{\mathcal{G}}=\boldsymbol{T}_{\mathcal{G}}(G)} dG=0 \tag{16}$$

where

$$\boldsymbol{\psi}_{\mathcal{G}_k}(y,\boldsymbol{\theta}_{\mathcal{G}_k})=\frac{\partial}{\partial\boldsymbol{\theta}_{\mathcal{G}_k}}\left\{-\frac{1}{2\sigma^2}(y-\boldsymbol{\beta}'\boldsymbol{x})^2-\frac{1}{2}\log(2\pi\sigma^2)-\frac{\lambda_k}{\sigma^2}\sum_{j=1}^{J}(\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j)^{1/2}\right\}$$

with G being the true distribution of y. By replacing G by the empirical distribution $\hat{G}$ based on observations $\mathbf{y}$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}_{\mathcal{G}_k}(y_i,\boldsymbol{\theta}_{\mathcal{G}_k})\Big|_{\hat{\boldsymbol{\theta}}_{\mathcal{G}_k}=\boldsymbol{T}_{\mathcal{G}_k}(\hat{G})}=0. \tag{17}$$

This implies that the nonzero components of the group Lasso estimator $\hat{\boldsymbol{\theta}}_{\mathcal{G}_k}$ can be written implicitly as $\hat{\boldsymbol{\theta}}_{\mathcal{G}_k}=\boldsymbol{T}_{\mathcal{G}_k}(\hat{G})$ for the functional $\boldsymbol{T}_{\mathcal{G}_k}(G)$ by (16).

By replacing G in (16) by $G_\epsilon=(1-\epsilon)G+\epsilon\delta_y$ with δ_y being a point of mass at y, we obtain

$$\int \psi(y,\boldsymbol{T}_{\mathcal{G}_k}((1-\epsilon)G+\epsilon\delta_y))\Big|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} d\left\{(1-\epsilon)G(y)+\epsilon\delta_y(y)\right\}=0. \tag{18}$$

By differentiating with respect to ϵ and setting $\epsilon = 0$, we have

$$\int \boldsymbol{\psi}_{\mathcal{G}_k}(y, \boldsymbol{T}_{\mathcal{G}_k}(G)) d\left\{\delta_y(y) - G(y)\right\} + \int \frac{\partial}{\partial \boldsymbol{\theta}_{\mathcal{G}_k}} \boldsymbol{\psi}_{\mathcal{G}_k}(y, \boldsymbol{\theta}_{\mathcal{G}_k}) \Big|_{\boldsymbol{\theta}_{\mathcal{G}_k} = \boldsymbol{T}_{\mathcal{G}_k}(G)} dG(y) \cdot \frac{\partial}{\partial \epsilon} \left\{\boldsymbol{T}_{\mathcal{G}_k}((1-\epsilon)G + \epsilon\delta_y)\right\} \Big|_{\epsilon=0} = 0. \tag{19}$$

From the above equation, the influence function of the group Lasso estimator $\hat{\boldsymbol{\theta}}$ is given by

$$\begin{aligned} \boldsymbol{T}^{(1)}_{\mathcal{G}_k}(y, G) \equiv & \frac{\partial}{\partial \epsilon} \left\{\boldsymbol{T}_{\mathcal{G}_k}((1-\epsilon)G + \epsilon\delta_y)\right\} \Big|_{\epsilon=0} \\ = & - \left\{ \int \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\psi}_{\mathcal{G}_k}(y, \boldsymbol{\theta}_{\mathcal{G}_k}) \Big|_{\boldsymbol{\theta}_{\mathcal{G}_k} = \boldsymbol{T}_{\mathcal{G}_k}(G)} dG(y) \right\}^{-1} \boldsymbol{\psi}_{\mathcal{G}_k}(y, \boldsymbol{T}_{\mathcal{G}}(G)) \end{aligned} \tag{20}$$

By replacing G in (20) by $\hat{G}$ and using the result in (8), we have the following information criterion for the group Lasso model $f(y|\hat{\boldsymbol{\theta}})$:

$$\begin{aligned} \text{GLIC} &= -2 \sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) + b_1(\hat{G}) \\ &= -2 \sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) + \frac{2}{n} \sum_{i=1}^{n} \text{tr} \left\{ \boldsymbol{T}_{\mathcal{G}_k}(y_i, \hat{G}) \frac{\partial \log f(y_i|\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right\} \\ &= n \log(2\pi\hat{\sigma}^2) + \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2/\hat{\sigma}^2 + 2\text{tr}\left\{ \boldsymbol{I}(\hat{G}) \boldsymbol{J}(\hat{G})^{-1} \right\} \end{aligned} \tag{21}$$

where

$$\boldsymbol{J}(\hat{G}) = \frac{1}{n\hat{\sigma}^2} \begin{bmatrix} \boldsymbol{X}'_{\mathcal{G}_k} \boldsymbol{X}_{\mathcal{G}_k} + \lambda_k \boldsymbol{D} & \frac{1}{\hat{\sigma}^2} \boldsymbol{X}'_{\mathcal{G}_k} \boldsymbol{\Lambda} \mathbf{1}_n - \frac{\lambda_k}{\hat{\sigma}^2} \boldsymbol{d} \\ \frac{1}{\hat{\sigma}^2} \mathbf{1}'_n \boldsymbol{\Lambda} \boldsymbol{X}_{\mathcal{G}_k} - \frac{\lambda_k}{\hat{\sigma}^2} \boldsymbol{d}' & -\frac{n}{2\hat{\sigma}^2} \end{bmatrix},$$

$$\boldsymbol{I}(\hat{G}) = \frac{1}{n\hat{\sigma}^4} \begin{pmatrix} \boldsymbol{X}'_{\mathcal{G}_k} \boldsymbol{\Lambda} - \lambda_k \boldsymbol{d} \\ \frac{1}{2\hat{\sigma}^2} \mathbf{1}'_n \boldsymbol{\Lambda}^2 - \frac{1}{2} \mathbf{1}'_n + \frac{\lambda_k}{\hat{\sigma}^2} (\hat{\boldsymbol{\beta}}' \boldsymbol{K} \hat{\boldsymbol{\beta}})^{1/2} \mathbf{1}'_n \end{pmatrix} \begin{pmatrix} \boldsymbol{\Lambda} \boldsymbol{X}_{\mathcal{G}_k}, & \frac{1}{2\hat{\sigma}^2} \boldsymbol{\Lambda}^2 \mathbf{1}_n - \frac{1}{2} \mathbf{1}_n \end{pmatrix}$$

with

$$
\begin{aligned}
\mathbf{1}_n &= (1, 1, \ldots, 1)' \quad (n \times 1 \text{ vector}), \\
\hat{\sigma}^2 &= \{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + 2\lambda(\hat{\boldsymbol{\beta}}'\boldsymbol{K}\hat{\boldsymbol{\beta}})^{1/2}\}/n \\
\boldsymbol{D} &= \text{bdiag}[\boldsymbol{D}_j]_{j \in \mathcal{G}_k}, \\
\boldsymbol{D}_j &= \|\hat{\boldsymbol{\beta}}_j'\hat{\boldsymbol{\beta}}_j\|^{-1/2}\sqrt{p_j}\boldsymbol{I}_j - \|\hat{\boldsymbol{\beta}}_j'\hat{\boldsymbol{\beta}}_j\|^{-3/2}\sqrt{p_j}\hat{\boldsymbol{\beta}}_j\hat{\boldsymbol{\beta}}_j', \\
\boldsymbol{\Lambda} &= \text{diag}[y_1 - \hat{\boldsymbol{\beta}}'\boldsymbol{x}_1, y_2 - \hat{\boldsymbol{\beta}}'\boldsymbol{x}_2, \ldots, y_n - \hat{\boldsymbol{\beta}}'\boldsymbol{x}_n] \\
\boldsymbol{d} &= (\boldsymbol{d}_j)_{j \in \mathcal{G}_k}, \\
\boldsymbol{d}_j &= \|\hat{\boldsymbol{\beta}}_j'\hat{\boldsymbol{\beta}}_j\|^{-1/2}\sqrt{p_j}\hat{\boldsymbol{\beta}}_j.
\end{aligned}
$$

We choose the regularization parameter value which gives the minimizer of GLIC among a sequence of regularization parameter values $\{\lambda_1, \lambda_2, \ldots, \lambda_r\}$.

4. Conclusion

In this article, we discussed the problem of choosing a better regularization parameter from a series of candidate values in the group Lasso. We derived the influence function of the group Lasso estimator and proposed an information criterion for regularization parameter selection in the normal linear regression problem with the group Lasso. Further works remains to be done towards extending the derived criterion in the framework of the penalized likelihood-based models from exponential families.

References

[1] Akaike, H. (1974), A new look at the statistical model identification. *IEEE Trans. Automatic Control*, **AC-19**, 716-723.

[2] Konishi, S. and Kitagawa, G. (1996), Generalised information criteria in model selection, *Biometrika*, **83**, 875-890.

[3] Yuan, M. and Lin Y. (2006), Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society, Series B*, **68**, 49-67.