Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2013)
75, *Part* 3, *pp.* 531–552

Tuning parameter selection in high dimensional penalized likelihood

Yingying Fan

University of Southern California, Los Angeles, USA

and Cheng Yong Tang

University of Colorado, Denver, USA

[Received June 2011. Final revision September 2012]

Summary. Determining how to select the tuning parameter appropriately is essential in penalized likelihood methods for high dimensional data analysis. We examine this problem in the setting of penalized likelihood methods for generalized linear models, where the dimensionality of covariates p is allowed to increase exponentially with the sample size n. We propose to select the tuning parameter by optimizing the generalized information criterion with an appropriate model complexity penalty. To ensure that we consistently identify the true model, a range for the model complexity penalty is identified in the generlized information criterion. We find that this model complexity penalty should diverge at the rate of some power of $\log(p)$ depending on the tail probability behaviour of the response variables. This reveals that using the Akaike information criterion or Bayes information criterion to select the tuning parameter may not be adequate for consistently identifying the true model. On the basis of our theoretical study, we propose a uniform choice of the model complexity penalty and show that the approach proposed consistently identifies the true model among candidate models with asymptotic probability 1. We justify the performance of the procedure proposed by numerical simulations and a gene expression data analysis.

Keywords: Generalized information criterion; Generalized linear model; Penalized likelihood; Tuning parameter selection; Variable selection

1. Introduction

Various types of high dimensional data are encountered in many disciplines when solving practical problems, e.g. gene expression data for disease classifications (Golub *et al.*, 1999), financial market data for portfolio construction and assessment (Jagannathan and Ma, 2003) and spatial earthquake data for geographical analysis (van der Hilst *et al.*, 2007), among many others. To meet the challenges in analysing high dimensional data, penalized likelihood methods have been extensively studied; see Hastie *et al.* (2009) and Fan and Lv (2010) for overviews among a large amount of recent literature.

Though demonstrated effective in analysing high dimensional data, the performance of penalized likelihood methods depends on the choice of the tuning parameters, which control the trade-off between the bias and variance in resulting estimators (Hastie *et al.*, 2009; Fan and Lv, 2010). Generally speaking, the optimal properties of those penalized likelihood methods require certain specifications of the optimal tuning parameters (Fan and Lv, 2010). However,

Address for correspondence: Cheng Yong Tang, Business School, University of Colorado, Denver, PO Box 173364, Denver, CO 80217-3364, USA.
E-mail: chengyong.tang@ucdenver.edu

© 2012 Royal Statistical Society 1369–7412/13/75531

theoretically quantified optimal tuning parameters are not practically feasible, because they are valid only asymptotically and usually depend on unknown nuisance parameters in the true model. Therefore, in practical implementations, penalized likelihood methods are usually applied with a sequence of tuning parameters resulting in a corresponding collection of models. Then, selecting an appropriate model and equivalently the corresponding tuning parameter becomes an important question of interest, both theoretically and practically.

Traditionally in model selection, cross-validation and information criteria—including the Akaike information criterion (AIC) (Akaike, 1973) and Bayes information criterion (BIC) (Schwarz, 1978)—are widely applied. A generalized information criterion (Nishii, 1984) is constructed as follows:

$$\text{measure of model fitting} + a_n \times \text{measure of model complexity}, \tag{1.1}$$

where a_n is some positive sequence that depends only on the sample size n and that controls the penalty on model complexity. The rationale of the information criteria for model selection is that the true model can uniquely optimize the information criterion (1.1) by appropriately choosing a_n. Hence, the choice of a_n becomes crucial for effectively identifying the true model. The minus log-likelihood is commonly used as a measure of the model fitting, and a_n is 2 and $\log(n)$ in the AIC and BIC respectively. It is known that the BIC can identify the true model consistently in linear regression with fixed dimensional covariates, whereas the AIC may fail because of overfitting (Shao, 1997). Meanwhile, cross-validation is shown asymptotically equivalent to the AIC (Yang, 2005) so they behave similarly.

When applying penalized likelihood methods, existing model selection criteria are naturally incorporated to select the tuning parameter. Analogously to those results for model selection, Wang *et al.* (2007) showed that the tuning parameter that is selected by the BIC can identify the true model consistently for the smoothly clipped absolute deviation (SCAD) approach in Fan and Li (2001), whereas the AIC and cross-validation may fail to play such a role (see also Zhang *et al.* (2010)). However, those studies on tuning parameter selection for penalized likelihood methods are mainly for fixed dimensionality. Wang *et al.* (2009) recently considered tuning parameter selection in the setting of linear regression with diverging dimensionality and showed that a modified BIC continues to work for tuning parameter selection. However, their analysis is confined to the penalized least squares method, and the dimensionality p of covariates is not allowed to exceed the sample size n. We also refer to Chen and Chen (2008) for a recent study on an extended BIC and its property for Gaussian linear models, and Wang and Zhu (2011) for tuning parameter selection in high dimensional penalized least squares.

The current trend of high dimensional data analysis poses new challenges for tuning parameter selection. To the best of our knowledge, there is no existing work accommodating tuning parameter selection for general penalized likelihood methods when the dimensionality p grows exponentially with the sample size n, i.e. $\log(p) = O(n^{\kappa})$ for some $\kappa > 0$. The problem is challenging in a few respects. First, note that there are generally no explicit forms of the maximum likelihood estimates for models other than the linear regression model, which makes it more difficult to characterize the asymptotic performance of the first part of criterion (1.1). Second, the exponentially growing dimensionality p induces a huge number of candidate models. We may reasonably conjecture that the true model may be differentiated from a specific candidate model with probability tending to 1 as $n \to \infty$. However, the probability that the true model is not dominated by any of the candidate models may not be straightforward to calculate, and an inappropriate choice of a_n in criterion (1.1) may even cause the model selection consistency to fail.

We explore in this paper tuning parameter selection for penalized generalized linear regression,

with penalized Gaussian linear regression as a special case, in which the dimensionality p is allowed to increase exponentially fast with the sample size n. We systematically examine the generalized information criterion, and our analysis reveals the connections between the model complexity penalty a_n, the data dimensionality p and the tail probability distribution of the response variables, for consistently identifying the true model. Subsequently, we identify a range of a_n such that the tuning parameter that is selected by optimizing the generalized information criterion can achieve model selection consistency. We find that, when p grows polynomially with sample size n, the modified BIC (Wang *et al.*, 2009) can still be successful in tuning parameter selection. But, when p grows exponentially with sample size n, a_n should diverge with some power of $\log(p)$, where the power depends on the tail distribution of response variables. This produces a phase diagram of how the model complexity penalty should adapt to the growth of sample size n and dimensionality p. Our theoretical investigations, numerical implementations by simulations and a data analysis illustrate that the approach proposed can be effectively and conveniently applied in practice. As demonstrated in Fig. 3 in Section 5.2 for analysing a gene expression data set, we find that a single gene identified by the approach proposed can be very informative in predictively differentiating between two types of leukaemia patients.

The rest of this paper is organized as follows. In Section 2, we outline the problem and define the model selection criterion GIC. To study GIC, we first investigate the asymptotic property of a proxy for GIC in Section 3, and we summarize the main result of the paper in Section 4. Section 5 demonstrates the proposed approach via numerical examples of simulations and gene expression data analysis, and Section 6 contains the technical conditions and some intermediate results. The technical proofs are contained in Appendix A.

2. Penalized generalized linear regression and tuning parameter selection

Let $\{(\mathbf{x}_i, Y_i)\}_{i=1}^n$ be independent data with the scalar response variable Y_i and the corresponding p-dimensional covariate vector $\mathbf{x}_i$ for the ith observation. We consider the generalized linear model (McCullagh and Nelder, 1989) with the conditional density function of Y_i given $\mathbf{x}_i$

$$f_i(y_i; \theta_i, \phi) = \exp\{y_i\theta_i - b(\theta_i) + c(y_i, \phi)\}, \tag{2.1}$$

where $\theta_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ is the canonical parameter with $\boldsymbol{\beta}$ a p-dimensional regression coefficient, $b(\cdot)$ and $c(\cdot, \cdot)$ are some suitably chosen known functions, $E[Y_i|\mathbf{x}_i] = \mu_i = b'(\theta_i)$, $g(\mu_i) = \theta_i$ is the link function, and ϕ is a known scale parameter. Thus, the log-likelihood function for $\boldsymbol{\beta}$ is given by

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{Y_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - b(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + c(Y_i, \phi)\}. \tag{2.2}$$

The dimensionality p in our study is allowed to increase with sample size n exponentially fast—i.e. $\log(p) = O(n^{\kappa})$ for some $\kappa > 0$. To enhance the model fitting accuracy and to ensure the model identifiability, it is commonly assumed that the true population parameter $\boldsymbol{\beta}_0$ is sparse, with only a small fraction of non-zeros (Tibshirani, 1996; Fan and Li, 2001). Let $\alpha_0 = \mathrm{supp}(\boldsymbol{\beta}_0)$ be the support of the true model consisting of indices of all non-zero components in $\boldsymbol{\beta}_0$, and let $s_n = |\alpha_0|$ be the number of true covariates, which may increase with n and which satisfies $s_n = o(n)$. To ease the presentation, we suppress the dependence of s_n on n whenever there is no confusion. By using compact notation, we write $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ as the n-vector of response, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_p)$ as the $n \times p$ fixed design matrix and $\boldsymbol{\mu} = \mathbf{b}'(\mathbf{X}\boldsymbol{\beta}) = (b'(\mathbf{x}_1^{\mathrm{T}}\boldsymbol{\beta}), \ldots, b'(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}))^{\mathrm{T}}$ as the mean vector. We standardize each column of $\mathbf{X}$ so that $\|\tilde{\mathbf{x}}_j\|_2 = \sqrt{n}$ for $j = 1, \ldots, p$.

In practice, the true parameter $\boldsymbol{\beta}_0$ is unknown and needs to be estimated from data. Penalized likelihood methods have attracted substantial attention recently for simultaneously selecting and

estimating the unknown parameters. The penalized maximum likelihood estimator (MLE) is broadly defined as

$$\hat{\boldsymbol{\beta}}^{\lambda_n} = \arg\max_{\boldsymbol{\beta}\in\mathbf{R}^p}\Big\{l_n(\boldsymbol{\beta}) - n\sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|)\Big\}, \tag{2.3}$$

where $p_{\lambda_n}(\cdot)$ is some penalty function with tuning parameter $\lambda_n \geqslant 0$. For simplicity, we suppress the dependence of λ_n on n and write it as λ when there is no confusion. Let $\alpha_\lambda = \operatorname{supp}(\hat{\boldsymbol{\beta}}^\lambda)$ be the model that is identified by the penalized likelihood method with tuning parameter λ.

For the penalized likelihood method to identify the underlying true model successfully and to enjoy desirable properties, it is critically important to choose an appropriate tuning parameter λ. Intuitively, a too large or too small tuning parameter imposes respectively an excessive or inadequate penalty on the magnitude of the parameter so the support of $\hat{\boldsymbol{\beta}}^\lambda$ is different from that of the true model α_0. Clearly, a meaningful discussion of tuning parameter selection in equation (2.3) requires the existence of a λ_0 such that $\alpha_{\lambda_0} = \alpha_0$, which has been established in various model settings when different penalty functions are used; see, for example, Zhao and Yu (2006), Lv and Fan (2009) and Fan and Lv (2011).

To identify the λ_0 that leads to the true model α_0, we propose to use the generalized information criterion

$$\mathrm{GIC}_{a_n}(\lambda) = \frac{1}{n}\{D(\hat{\boldsymbol{\mu}}_\lambda; \mathbf{Y}) + a_n|\alpha_\lambda|\}, \tag{2.4}$$

where a_n is a positive sequence depending only on n and $D(\hat{\boldsymbol{\mu}}_\lambda; \mathbf{Y})$ is the scaled deviation measure defined as the scaled log-likelihood ratio of the saturated model and the candidate model with parameter $\hat{\boldsymbol{\beta}}^\lambda$, i.e.

$$D(\hat{\boldsymbol{\mu}}_\lambda; \mathbf{Y}) = 2\{l_n(\mathbf{Y}; \mathbf{Y}) - l_n(\hat{\boldsymbol{\mu}}_\lambda; \mathbf{Y})\} \tag{2.5}$$

with $l_n(\boldsymbol{\mu}; \mathbf{Y})$ the log-likelihood function (2.2) expressed as a function of $\boldsymbol{\mu}$ and $\mathbf{Y}$, and $\hat{\boldsymbol{\mu}}_\lambda = \mathbf{b}'(\mathbf{X}\hat{\boldsymbol{\beta}}^\lambda)$. The scaled deviation measure is used to evaluate the goodness of fit. It reduces to the sum of squared residuals in Gaussian linear regression. The second component in the definition of GIC (2.4) is a penalty on the model complexity. So, intuitively, GIC trades off between the model fitting and the model complexity by appropriately choosing a_n. When $a_n = 2$ and $a_n = \log(n)$, equation (2.4) becomes the classical AIC (Akaike, 1973) and BIC (Schwarz, 1978) respectively. The modified BIC (Wang *et al.*, 2009) corresponds to $a_n = C_n \log(n)$ with a diverging C_n-sequence. The scaled deviation measure and GIC were also studied in Zhang *et al.* (2010) for regularization parameter selection in a fixed dimensional setting.

Our problem of interest now becomes how to choose a_n appropriately such that the tuning parameter λ_0 can be consistently identified by minimizing equation (2.4) with respect to λ— i.e. with probability tending to 1—

$$\inf_{\{\lambda>0:\alpha_\lambda\neq\alpha_0\}}\{\mathrm{GIC}_{a_n}(\lambda) - \mathrm{GIC}_{a_n}(\lambda_0)\} > 0. \tag{2.6}$$

From expressions (2.4) and (2.6), we can see clearly that, to study the choice of a_n, it is essential to investigate the asymptotic properties of $D(\hat{\boldsymbol{\mu}}_\lambda; \mathbf{Y})$ uniformly over a range of λ. Directly studying $D(\hat{\boldsymbol{\mu}}_\lambda; \mathbf{Y})$ is challenging because $\hat{\boldsymbol{\mu}}_\lambda$ depends on $\hat{\boldsymbol{\beta}}^\lambda$, which is the maximizer of a possibly non-concave function (2.3); thus, it takes no explicit form and, more critically, its uniform asymptotic properties are difficult to establish. To overcome these difficulties, we introduce a proxy of $\mathrm{GIC}_{a_n}(\lambda)$, which is defined as

$$\mathrm{GIC}^*_{a_n}(\alpha) = \frac{1}{n}\{D(\hat{\boldsymbol{\mu}}^*_\alpha; \mathbf{Y}) + a_n|\alpha|\} \tag{2.7}$$

for a given model support $\alpha \subset \{1, \ldots, p\}$ that collects indices of all included covariates, and $\hat{\boldsymbol{\mu}}^*_\alpha = \mathbf{b}'\{\mathbf{X}\hat{\boldsymbol{\beta}}^*(\alpha)\}$ with $\hat{\boldsymbol{\beta}}^*(\alpha)$ being the unpenalized MLE restricted to the space $\{\boldsymbol{\beta} \in \mathbf{R}^p : \mathrm{supp}(\boldsymbol{\beta}) = \alpha\}$, i.e.

$$\hat{\boldsymbol{\beta}}^*(\alpha) = \arg \max_{\{\boldsymbol{\beta} \in \mathbf{R}^p : \mathrm{supp}(\boldsymbol{\beta}) = \alpha\}} l_n(\boldsymbol{\beta}). \tag{2.8}$$

The critical difference between equations (2.4) and (2.7) is that $\mathrm{GIC}_{a_n}(\lambda)$ is a function of λ depending on the penalized MLE $\hat{\boldsymbol{\beta}}^\lambda$, whereas $\mathrm{GIC}^*_{a_n}(\alpha)$ is a function of model α depending on the corresponding unpenalized MLE $\hat{\boldsymbol{\beta}}^*(\alpha)$. Under some signal strength assumptions and some regularity conditions, $\hat{\boldsymbol{\beta}}^{\lambda_0}$ and $\hat{\boldsymbol{\beta}}^*(\alpha_0)$ are close to each other asymptotically (Zhang and Huang, 2006; Fan and Li, 2001; Lv and Fan, 2009). As a consequence, $\mathrm{GIC}_{a_n}(\lambda_0)$ and $\mathrm{GIC}^*_{a_n}(\alpha_0)$ are also asymptotically close, as formally presented in the following proposition.

Proposition 1. Under conditions 1, 2 and 4 in Section 6, if $p'_{\lambda_0}(\frac{1}{2}\min_{j \in \alpha_0}|\beta_{0j}|) = o(s^{-1/2}n^{-1/2} \times a_n^{1/2})$, then

$$\mathrm{GIC}_{a_n}(\lambda_0) - \mathrm{GIC}^*_{a_n}(\alpha_0) = o_p(n^{-1}a_n). \tag{2.9}$$

Furthermore, it follows from the definition of $\hat{\boldsymbol{\beta}}^*(\alpha)$ that, for any $\lambda > 0$, $\mathrm{GIC}_{a_n}(\lambda) \geqslant \mathrm{GIC}^*_{a_n}(\alpha_\lambda)$. Therefore, proposition 1 entails

$$\begin{aligned} \mathrm{GIC}_{a_n}(\lambda) - \mathrm{GIC}_{a_n}(\lambda_0) &\geqslant \mathrm{GIC}^*_{a_n}(\alpha_\lambda) - \mathrm{GIC}^*_{a_n}(\alpha_0) + \mathrm{GIC}^*_{a_n}(\alpha_0) - \mathrm{GIC}_{a_n}(\lambda_0) \\ &= \mathrm{GIC}^*_{a_n}(\alpha_\lambda) - \mathrm{GIC}^*_{a_n}(\alpha_0) + o_p(n^{-1}a_n). \end{aligned} \tag{2.10}$$

Hence, the difficulties of directly studying GIC can be overcome by using the proxy GIC* as a bridge, whose properties are elaborated in the next section.

3. Asymptotic properties of the proxy generalized information criterion

3.1. Underfitted models

From definition (2.7), the properties of GIC* depend on the unpenalized MLE $\hat{\boldsymbol{\beta}}^*(\alpha)$ and scaled deviance measure $D(\hat{\boldsymbol{\mu}}^*_\alpha; \mathbf{Y})$. When the truth α_0 is given, it is well known from classical statistical theory that $\hat{\boldsymbol{\beta}}^*(\alpha_0)$ consistently estimates the population parameter $\boldsymbol{\beta}_0$. However, such a result is less intuitive if $\alpha \neq \alpha_0$. In fact, as shown in proposition 2 in Section 6, uniformly for all $|\alpha| \leqslant K$ for some positive integer $K > s$ and $K = o(n)$, $\hat{\boldsymbol{\beta}}^*(\alpha)$ converges in probability to the minimizer $\boldsymbol{\beta}^*(\alpha)$ of the following Kullback–Leibler (KL) divergence:

$$I\{\boldsymbol{\beta}(\alpha)\} = E[\log(f^*/g_\alpha)] = \sum_{i=1}^{n} \{b'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}(\alpha)) - b(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0) + b(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}(\alpha))\}, \tag{3.1}$$

where $\boldsymbol{\beta}(\alpha)$ is a p-dimensional parameter vector with support α, f^* is the density of the underlying true model and g_α is the density of the model with population parameter $\boldsymbol{\beta}(\alpha)$. Intuitively, model α coupled with the population parameter $\boldsymbol{\beta}^*(\alpha)$ has the smallest KL divergence from the truth among all models with support α. Since the KL divergence is non-negative and $I(\boldsymbol{\beta}_0) = 0$, the true parameter $\boldsymbol{\beta}_0$ is a global minimizer of equation (3.1). To ensure identifiability, we assume that equation (3.1) has a unique minimizer $\boldsymbol{\beta}^*(\alpha)$ for every α satisfying $|\alpha| \leqslant K$. This unique minimizer assumption will be further discussed in Section 6. Thus, it follows immediately that $\boldsymbol{\beta}^*(\alpha) = \boldsymbol{\beta}_0$ for all $\alpha \supseteq \alpha_0$ with $|\alpha| \leqslant K$, and consequently $I\{\boldsymbol{\beta}^*(\alpha)\} = 0$. Hereinafter, we refer to

the population model α as the model that is associated with the population parameter $\boldsymbol{\beta}^*(\alpha)$. We refer to α as an overfitted model if $\alpha \supsetneq \alpha_0$, and as an underfitted model if $\alpha \not\supset \alpha_0$.

For an underfitted population model α, the KL divergence $I\{\boldsymbol{\beta}^*(\alpha)\}$ measures the deviance from the truth due to missing at least one true covariate. Therefore, we define

$$\delta_n = \inf_{\substack{\alpha \not\supset \alpha_0 \\ |\alpha| \leqslant K}} \frac{1}{n} I\{\boldsymbol{\beta}^*(\alpha)\} \tag{3.2}$$

as an essential measure of the smallest signal strength of the true covariates, which effectively controls the extent to which the true model can be distinguished from underfitted models.

Let $\boldsymbol{\mu}_\alpha^* = \mathbf{b}'(\mathbf{X}\boldsymbol{\beta}^*(\alpha))$ and $\boldsymbol{\mu}_0 = \mathbf{b}'(\mathbf{X}\boldsymbol{\beta}_0)$ be the population mean vectors corresponding to the parameter $\boldsymbol{\beta}^*(\alpha)$ and the true parameter $\boldsymbol{\beta}_0$ respectively. It can be seen from definition (3.1) that

$$I\{\boldsymbol{\beta}^*(\alpha)\} = \tfrac{1}{2} E[D(\boldsymbol{\mu}_\alpha^*; \mathbf{Y}) - D(\boldsymbol{\mu}_0; \mathbf{Y})].$$

Hence, $2I\{\boldsymbol{\beta}^*(\alpha)\}$ is the population version of the difference between $D(\hat{\boldsymbol{\mu}}_\alpha^*; \mathbf{Y})$ and $D(\hat{\boldsymbol{\mu}}_0^*; \mathbf{Y})$, where $\hat{\boldsymbol{\mu}}_0^* = \hat{\boldsymbol{\mu}}_{\alpha_0}^* = \mathbf{b}'(\mathbf{X}\hat{\boldsymbol{\beta}}^*(\alpha_0))$ is the estimated population mean vector knowing the truth α_0. Therefore, the KL divergence $I(\cdot)$ can be intuitively understood as a population distance between a model α and the truth α_0. The following theorem formally characterizes the uniform convergence result of the difference of scaled deviance measures to its population version $2I\{\boldsymbol{\beta}^*(\alpha)\}$.

Theorem 1. Under conditions 1 and 2 in Section 6, as $n \to \infty$,

$$\sup_{\substack{|\alpha| \leqslant K \\ \alpha \subset \{1,\ldots,p\}}} \frac{1}{n|\alpha|} |D(\hat{\boldsymbol{\mu}}_\alpha^*; \mathbf{Y}) - D(\hat{\boldsymbol{\mu}}_0^*; \mathbf{Y}) - 2I\{\boldsymbol{\beta}^*(\alpha)\}| = O_p(R_n),$$

when either

(a) the Y_is are bounded or Gaussian distributed, $R_n = \sqrt{\{\log(p)/n\}}$, and $\log(p) = o(n)$, or
(b) the Y_is are unbounded and non-Gaussian distributed,

the design matrix satisfies $\max_{ij} |x_{ij}| = O(n^{1/2-\tau})$ with $\tau \in (0, \frac{1}{2}]$, condition 3 holds, $R_n = \sqrt{\{\log(p)/n\} + m_n^2 \log(p)/n}$ and $\log(p) = o[\min\{n^{2\tau}\log(n)^{-1}K^{-2}, nm_n^{-2}\}]$ with m_n defined in condition 3.

Theorem 1 ensures that, for any model α satisfying $|\alpha| \leqslant K$,

$$\mathrm{GIC}_{a_n}^*(\alpha) - \mathrm{GIC}_{a_n}^*(\alpha_0) = \frac{2}{n} I\{\boldsymbol{\beta}^*(\alpha)\} + (|\alpha| - |\alpha_0|)\{a_n n^{-1} - O_p(R_n)\}. \tag{3.3}$$

Hence, it implies that if a model α is far from the truth—i.e. $I\{\boldsymbol{\beta}^*(\alpha)\}$ is large—then this population discrepancy can be detected by looking at the sample value of the proxy $\mathrm{GIC}_{a_n}^*(\alpha)$.

Combining equations (3.2) with (3.3), we immediately find that, if $\delta_n K^{-1} R_n^{-1} \to \infty$ as $n \to \infty$ and a_n is chosen such that $a_n = o(s^{-1} n \delta_n)$, then, for sufficiently large n,

$$\inf_{\alpha \not\supset \alpha_0,\, |\alpha| \leqslant K} \{\mathrm{GIC}_{a_n}^*(\alpha) - \mathrm{GIC}_{a_n}^*(\alpha_0)\} > \delta_n - s a_n n^{-1} - O_p(K R_n) \geqslant \delta_n/2, \tag{3.4}$$

with probability tending to 1. Thus, condition (3.4) indicates that, as long as the signal δ_n is not decaying to 0 too fast, any underfitted model leads to a non-negligible increment in the proxy GIC*. This guarantees that minimizing $\mathrm{GIC}_{a_n}^*(\alpha)$ with respect to α can identify the true model α_0 among all underfitted models asymptotically.

However, for any overfitted model $\alpha \supsetneq \alpha_0$ with $|\alpha| \leqslant K$, $\boldsymbol{\beta}^*(\alpha) = \boldsymbol{\beta}_0$, and thus $I\{\boldsymbol{\beta}^*(\alpha)\} = 0$. Consequently, the true model α_0 cannot be differentiated from an overfitted model α by using the formulation (3.3). In fact, the study of overfitted models is far more difficult in a high dimensional setting, as detailed in the next subsection.

3.2. Overfitted models: the main challenge

It is known that, for an overfitted model α, the difference of scaled deviation measures

$$D(\hat{\boldsymbol{\mu}}_\alpha^*; \mathbf{Y}) - D(\hat{\boldsymbol{\mu}}_0; \mathbf{Y}) = 2\{l_n(\hat{\boldsymbol{\mu}}_0^*; \mathbf{Y}) - l_n(\hat{\boldsymbol{\mu}}_\alpha^*; \mathbf{Y})\} \tag{3.5}$$

follows asymptotically the χ^2-distribution with $|\alpha| - |\alpha_0|$ degrees of freedom when p is fixed. Since there are only a finite number of candidate models for fixed p, a model complexity penalty diverging to ∞ at an appropriate rate with sample size n facilitates an information criterion to identify the true model consistently; see, for example, Shao (1997), Bai *et al.* (1999), Wang *et al.* (2007), Zhang *et al.* (2010) and references therein. However, when p grows with n, the device in traditional model selection theory cannot be carried forward. Substantial challenges arise from two aspects. One is how to characterize the asymptotic probabilistic behaviour of equation (3.5) when $|\alpha| - |\alpha_0|$ itself is diverging. The other is how to deal with so many candidate models, the number of which grows combinatorially fast with p.

Let $\mathbf{H}_0 = \mathrm{diag}\{\mathbf{b}''(\mathbf{X}\boldsymbol{\beta}_0)\}$ be the diagonal matrix of the variance of $\mathbf{Y}$, and $\mathbf{X}_\alpha$ be a submatrix of $\mathbf{X}$ formed by columns whose indices are in α. For any overfitted model α, we define the associated projection matrix as

$$\mathbf{B}_\alpha = \mathbf{H}_0^{1/2}\mathbf{X}_\alpha(\mathbf{X}_\alpha^{\mathrm{T}}\mathbf{H}_0\mathbf{X}_\alpha)^{-1}\mathbf{X}_\alpha^{\mathrm{T}}\mathbf{H}_0^{1/2}. \tag{3.6}$$

When the Y_is are Gaussian, $\hat{\boldsymbol{\beta}}^*(\alpha)$ is the least squares estimate and admits an explicit form so that direct calculations yield

$$D(\hat{\boldsymbol{\mu}}_\alpha^*; \mathbf{Y}) - D(\hat{\boldsymbol{\mu}}_0; \mathbf{Y}) = -(\mathbf{Y} - \boldsymbol{\mu}_0)^{\mathrm{T}}\mathbf{H}_0^{-1/2}(\mathbf{B}_\alpha - \mathbf{B}_{\alpha_0})\mathbf{H}_0^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}_0). \tag{3.7}$$

When the Y_is are non-Gaussian, this result still holds, but only approximately. In fact, as formally shown in proposition 3 in Section 6,

$$\begin{aligned} D(\hat{\boldsymbol{\mu}}_\alpha^*; \mathbf{Y}) - D(\hat{\boldsymbol{\mu}}_0; \mathbf{Y}) = &-(\mathbf{Y} - \boldsymbol{\mu}_0)^{\mathrm{T}}\mathbf{H}_0^{-1/2}(\mathbf{B}_\alpha - \mathbf{B}_{\alpha_0})\mathbf{H}_0^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}_0) \\ &+ (|\alpha| - |\alpha_0|)\,(\text{uniformly small term}). \end{aligned} \tag{3.8}$$

The interim result (3.8) facilitates characterizing the deviation result for the scaled deviance measures by concentrating on the asymptotic property of

$$Z_\alpha = (\mathbf{Y} - \boldsymbol{\mu}_0)^{\mathrm{T}}\mathbf{H}_0^{-1/2}(\mathbf{B}_\alpha - \mathbf{B}_{\alpha_0})\mathbf{H}_0^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}_0).$$

When the Y_is are Gaussian, it can be seen that $Z_\alpha \sim \chi^2_{|\alpha|-|\alpha_0|}$ for each fixed α. Thus, the deviation result on $\max_{\alpha \supset \alpha_0,\, |\alpha| \leqslant K} Z_\alpha$ can be obtained by explicitly calculating the tail probabilities of χ^2 random variables. However, if the Y_is are non-Gaussian, it is challenging to study the asymptotic property of Z_α, not to mention the uniform result across all overfitted models. To overcome this difficulty, we use the decoupling inequality (De La Peña and Montgomery-Smith, 1994) to study Z_α. The main results for overfitted models are given in the following theorem.

Theorem 2. Suppose that the design matrix satisfies $\max_{ij} |x_{ij}| = O(n^{1/2-\tau})$ with $\tau \in (\frac{1}{3}, \frac{1}{2}]$ and $\log(p) = O(n^\kappa)$ for some $0 < \kappa < 1$. Under conditions 1 and 2 in Section 6, as $n \to \infty$,

$$\frac{1}{|\alpha| - |\alpha_0|}\{D(\hat{\boldsymbol{\mu}}_\alpha^*; \mathbf{Y}) - D(\hat{\boldsymbol{\mu}}_0; \mathbf{Y})\} = O_p(\psi_n)$$

uniformly for all $\alpha \supsetneq \alpha_0$ with $|\alpha| \leqslant K$, and ψ_n is specified in the following two situations:

(a) $\psi_n = \sqrt{\log(p)}$ when the Y_is are bounded, $K = O(\min\{n^{(3\tau-\kappa-1)/3}, n^{(4\tau-1-3\kappa)/8}\})$ and $\kappa \leqslant 3\tau - 1$;
(b) $\psi_n = \log(p)$ when the Y_is are Gaussian distributed, or when the Y_is are unbounded non-Gaussian distributed, additional condition 3 holds, $K = O[n^{(6\tau-2-\kappa)/6}\{\sqrt{\log(n)} + m_n\}^{-1}]$, $\kappa \leqslant 6\tau - 2$, and $m_n = o(n^{(6\tau-2-\kappa)/6})$.

The results in theorem 2 hold for all overfitted models, which provides an insight into a high dimensional scenario beyond the asymptotic result characterized by a χ^2-distribution when p is fixed. Theorem 2 entails that, when a_n is chosen such that $a_n \psi_n^{-1} \to \infty$, uniformly for any overfitted model $\alpha \supsetneq \alpha_0$,

$$\mathrm{GIC}^*_{a_n}(\alpha) - \mathrm{GIC}^*_{a_n}(\alpha_0) = \frac{|\alpha| - |\alpha_0|}{n}\left\{a_n - O_p(\psi_n)\right\} > \frac{a_n}{2n} \tag{3.9}$$

with asymptotic probability 1. Thus, we can now differentiate overfitted models from the truth by examining the values of the proxy $\mathrm{GIC}^*_{a_n}(\alpha)$.

4. Consistent tuning parameter selection with the generalized information criterion

Now, we are ready to study the appropriate choice of a_n such that the tuning parameter λ_0 can be selected consistently by minimizing GIC as defined in equation (2.4). In practical implementation, the tuning parameter λ is considered over a range and, correspondingly, a collection of models is produced. Let $\lambda_{\max}$ and $\lambda_{\min}$ be respectively the upper and lower limits of the regularization parameter, where $\lambda_{\max}$ can be easily chosen such that $\alpha_{\lambda_{\max}}$ is empty and $\lambda_{\min}$ can be chosen such that $\hat{\boldsymbol{\beta}}^{\lambda_{\min}}$ is sparse, and the corresponding model size $K = |\alpha_{\lambda_{\min}}|$ satisfies conditions in theorem 3 below. Using the same notation as in Zhang *et al.* (2010), we partition the interval $[\lambda_{\min}, \lambda_{\max}]$ into subsets

$$\Omega_- = \{\lambda \in [\lambda_{\min}, \lambda_{\max}] : \alpha_\lambda \not\supset \alpha_0\},$$
$$\Omega_+ = \{\lambda \in [\lambda_{\min}, \lambda_{\max}] : \alpha_\lambda \supset \alpha_0 \text{ and } \alpha_\lambda \neq \alpha_0\}.$$

Thus, Ω_- is the set of λs that result in underfitted models, and Ω_+ is the set of λs that produce overfitted models.

We now present the main result of the paper. Combining expressions (2.10), (3.4) and (3.9) with proposition 1, we have the following theorem.

Theorem 3. Under the same conditions in proposition 1, theorem 1 and theorem 2, if $\delta_n K^{-1} R_n^{-1} \to \infty$, a_n satisfies $n\delta_n s^{-1} a_n^{-1} \to \infty$ and $a_n \psi_n^{-1} \to \infty$, where R_n and ψ_n are specified in theorems 1 and 2, then, as $n \to \infty$,

$$P\{\inf_{\lambda \in \Omega_- \cup \Omega_+} \mathrm{GIC}_{a_n}(\lambda) > \mathrm{GIC}_{a_n}(\lambda_0)\} \to 1,$$

where λ_0 is the tuning parameter in condition 4 in Section 6 that consistently identifies the true model.

The two requirements on a_n specify a range such that GIC is consistent in model selection for penalized MLEs. They reveal the synthetic impacts due to the signal strength, tail probability behaviour of the response and the dimensionality. Specifically, $a_n \psi_n^{-1} \to \infty$ means that a_n should diverge to ∞ adequately fast so that the true model is not dominated by overfitted models. In contrast, $n\delta_n s^{-1} a_n^{-1} \to \infty$ restricts the diverging rate of a_n, which can be viewed as constraints due to the signal strength quantified by δ_n in equation (2) and the size s of the true model.

Note that ψ_n in theorem 2 is a power of $\log(p)$. The condition $a_n\psi_n^{-1}$ in theorem 3 clearly demonstrates the effect of dimensionality p so the penalty on the model complexity should incorporate $\log(p)$. From this perspective, the AIC and even the BIC may fail to identify the true model consistently when p grows exponentially fast with n. As can be seen from the technical proofs in Appendix A, the huge number of overfitted candidate models leads to the model complexity penalty involving $\log(p)$. Moreover, theorem 3 actually accommodates the existing results—e.g. the modified BIC as in Wang *et al.* (2009). If dimensionality p is only of polynomial order of sample size n (i.e. $p=n^c$ for some $c \geqslant 0$), then $\log(p)=O\{\log(n)\}$, and thus the modified BIC with $a_n=\log\{\log(n)\}\log(n)$ can consistently select the true model in Gaussian linear models. As mentioned in Section 1, theorem 3 produces a phase diagram of how the model complexity penalty should adapt to the growth of sample size n and dimensionality p.

Theorem 3 specifies a range of a_n for consistent model selection:

$$n\delta_n s^{-1}a_n^{-1}\to\infty \quad \text{and} \quad a_n\psi_n^{-1}\to\infty.$$

For practical implementation, we propose to use a uniform choice $a_n=\log\{\log(n)\}\log(p)$ in GIC. The diverging part $\log\{\log(n)\}$ ensures $a_n\psi_n^{-1}\to\infty$ for all situations in theorem 3, and the slow diverging rate can ideally avoid underfitting. As a direct consequence of theorem 3, we have the following corollary for the validity of the choice of a_n.

Corollary 1. Under the same conditions in theorem 3 and letting $a_n=\log\{\log(n)\}\log(p)$, as $n\to\infty$

$$P\{\inf_{\lambda\in\Omega_-\cup\Omega_+}\mathrm{GIC}_{a_n}(\lambda)>\mathrm{GIC}_{a_n}(\lambda_0)\}\to 1$$

if $\delta_n K^{-1}R_n^{-1}\to\infty$ and $n\delta_n s^{-1}\log\{\log(n)\}^{-1}\log(p)^{-1}\to\infty$, where R_n is as specified in theorem 1.

When a_n is chosen appropriately as in theorem 3 and corollary 1, minimizing equation (2.4) identifies the tuning parameter λ_0 with probability tending to 1. This concludes a valid tuning parameter selection approach for identifying the true model for penalized likelihood methods.

5. Numerical examples

5.1. Simulations

We implement the proposed tuning parameter selection procedure using GIC with $a_n=\log\{\log(n)\}\log(p)$ as proposed in corollary 1. We compare its performance with those obtained by using the AIC ($a_n=2$) and BIC ($a_n=\log(n)$). In addition $a_n=\log(p)$ is also assessed, which is one of the possible criteria proposed in Wang and Zhu (2011). Throughout the simulations, the number of replications is 1000. In the numerical studies, the performance of the AIC is substantially worse than that of other tuning parameter selection methods, especially when p is much larger than n, so we omit the corresponding results for ease of presentation.

We first consider the Gaussian linear regression where continuous response variables are generated from the model

$$Y_i=\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\varepsilon_i, \qquad i=1,\ldots,n. \tag{5.1}$$

The row vectors $\mathbf{x}_i$ of the design matrix $\mathbf{X}$ are generated independently from a p-dimensional multivariate standard Gaussian distribution, and the ε_is are independant and identically distributed $N(0,\sigma^2)$ with $\sigma=3.0$ corresponding to the noise level. In our simulations, p is taken to be the integer part of $\exp\{(n-20)^{0.37}\}$. We let n increase from 100 to 500 with p ranging from 157

to 18376. The number of true covariates s grows with n in the following manner. Initially, $s=3$, and the first five elements of the true coefficient vector $\boldsymbol{\beta}_0$ are set to be $(3.0, 1.5, 0.0, 0.0, 2.0)^{\mathrm{T}}$ and all remaining elements are 0. Afterward, s increases by 1 for every 40-unit increment in n and the new element takes the value 2.5. For each simulated data set, we calculate the penalized MLE $\hat{\boldsymbol{\beta}}^{\lambda}$ by using equation (2.3) with $l_n(\boldsymbol{\beta})$ being the log-likelihood function for the linear regression model (5.1).

We then consider logistic regression where binary response variables are generated from the model

$$P(Y_i=1|\mathbf{x}_i)=\frac{1}{1+\exp(-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}, \qquad i=1,\ldots,n. \tag{5.2}$$

The design matrix $\mathbf{X}=(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)^{\mathrm{T}}$, dimensionality p and sample size n are similarly specified as in the linear regression example. The first five elements of $\boldsymbol{\beta}_0$ are set to be $(-3.0, 1.5, 0.0, 0.0, -2.0)^{\mathrm{T}}$ and the remaining components are all 0s. Afterwards, the number of non-zero parameters s increases by 1 for every 80-unit increment in n with the value being 2.0 and -2.0 alternately. The penalized MLE $\hat{\boldsymbol{\beta}}^{\lambda}$ is computed for each simulated data set based on equation (2.3) with $l_n(\boldsymbol{\beta})$ being the log-likelihood function for the logistic regression model (5.2).

We apply regularization methods with the lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001) and the minimax concave penalty (Zhang, 2010), and co-ordinate descent algorithms (Breheny and Huang, 2010; Friedman *et al.*, 2010) are carried out in optimizing the objective functions. The results of the minimax concave penalty are very similar to those of the SCAD penalty, and they have been omitted. We also compare with a reweighted adaptive lasso method, whose adaptive weight for β_j is chosen as $p'_{\lambda}(\hat{\beta}^{\lambda}_{j,\mathrm{lasso}})$ with $p'_{\lambda}(\cdot)$ being the derivative of the SCAD penalty, and $\hat{\boldsymbol{\beta}}^{\lambda}_{\mathrm{lasso}}=(\hat{\beta}^{\lambda}_{1,\mathrm{lasso}},\ldots,\hat{\beta}^{\lambda}_{p,\mathrm{lasso}})^{\mathrm{T}}$ being the lasso estimator. We remark that this reweighted adaptive lasso method shares the same spirit as the original SCAD-regularized estimate. In fact, a similar method, the local linear approximation method, has been proposed and studied in Zou and Li (2008). They showed that, under some conditions of the initial estimator $\hat{\boldsymbol{\beta}}_{\mathrm{lasso}}$, the reweighted adaptive lasso estimator discussed above enjoys the same oracle property as the original SCAD-regularized estimator. The similarities between these two estimates can also be seen in Figs 1 and 2.

For each regularization method—say, the SCAD method—when carrying out the tuning parameter selection procedure, we first calculate partly the solution path by choosing $\lambda_{\min}$ and $\lambda_{\max}$. Here, $\lambda_{\max}$ is chosen in such a way that no covariate is selected by the SCAD method in the corresponding model, whereas $\lambda_{\min}$ is the value where $[3\sqrt{n}]$ covariates are selected. Subsequently, for a grid of 200 values of λ equally spaced on the log-scale over $[\lambda_{\min},\lambda_{\max}]$, we calculate the SCAD-regularized estimates. This results in a sequence of candidate models. Then we apply each of the aforementioned tuning parameter selection methods to select the best model from the sequence. We repeat the same procedure for other regularization methods.

To evaluate the tuning parameter selection methods, we calculate the percentage of correctly specified models, the average number of false 0s identified and the median model error $E[\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0]^2$ for each selected model. We remark that the median and mean of model errors are qualitatively similar, and we use the median just to make results comparable with those in Wang *et al.* (2009). The comparison results are summarized in Figs 1 and 2. We clearly see that, for SCAD and the adaptive lasso, higher percentages of correctly specified models are achieved when $a_n=\log\{\log(n)\}\log(p)$ is used. The lasso method performs relative poorly, owing to its bias issue (Fan and Lv, 2010). In fact, our GIC aims at selecting the true model, whereas it is known that the lasso tends to overselect many variables. Thus GIC selects larger values of tuning parameter λ for the lasso than for other regularization methods to enforce the model

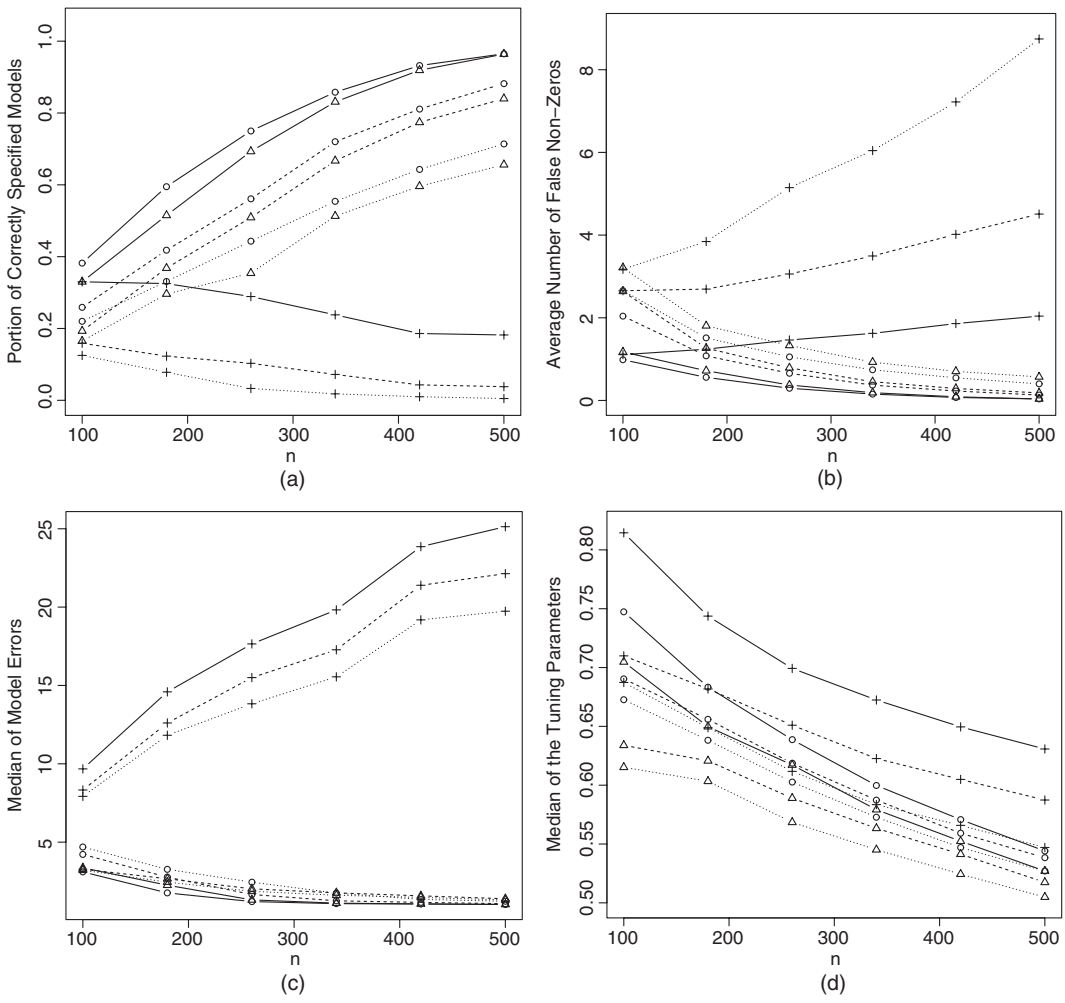

Fig. 1. Results for the linear model and Gaussian errors and $c_n = \log\{\log(n)\}$ (○, SCAD; △, adaptive lasso; +, lasso; ———, $\mathrm{GIC}_1 - c_n \log(p)$; - - - - - - -, $\mathrm{GIC}_2 - \log(p)$; · · · · · · · , $\mathrm{BIC} - \log(n)$): (a) correctly specified models; (b) average number of false non-zeros; (c) median of model errors; (d) median of the selected tuning parameters

sparsity, as shown in Figs 1(d) and 2(d). This larger thresholding level λ results in an even more severe bias issue as well as missing true weak covariates for the lasso method, which in turn cause larger model errors (see Figs 1(c) and 2(c)).

As expected and seen from Figs 1(b) and 2(b), $a_n = \log\{\log(n)\}\log(p)$ in combination with SCAD and the adaptive lasso has much fewer false positive results, which is the main reason for the substantial improvements in model selection. This demonstrates the need for applying an appropriate value of a_n in ultrahigh dimensions. In Figs 1(c) and 2(c), we report the median of relative model errors of the refitted unpenalized estimates for each selected model. We use the oracle model error from the fitted true model as the baseline, and we report the ratios of model errors for selected models to the oracle errors. From Figs 1(c) and 2(c), we can see that the median relative model errors corresponding to $\log\{\log(n)\}\log(p)$ decrease to 1 very fast and are consistently smaller than those by using the BIC, for both SCAD and the adaptive

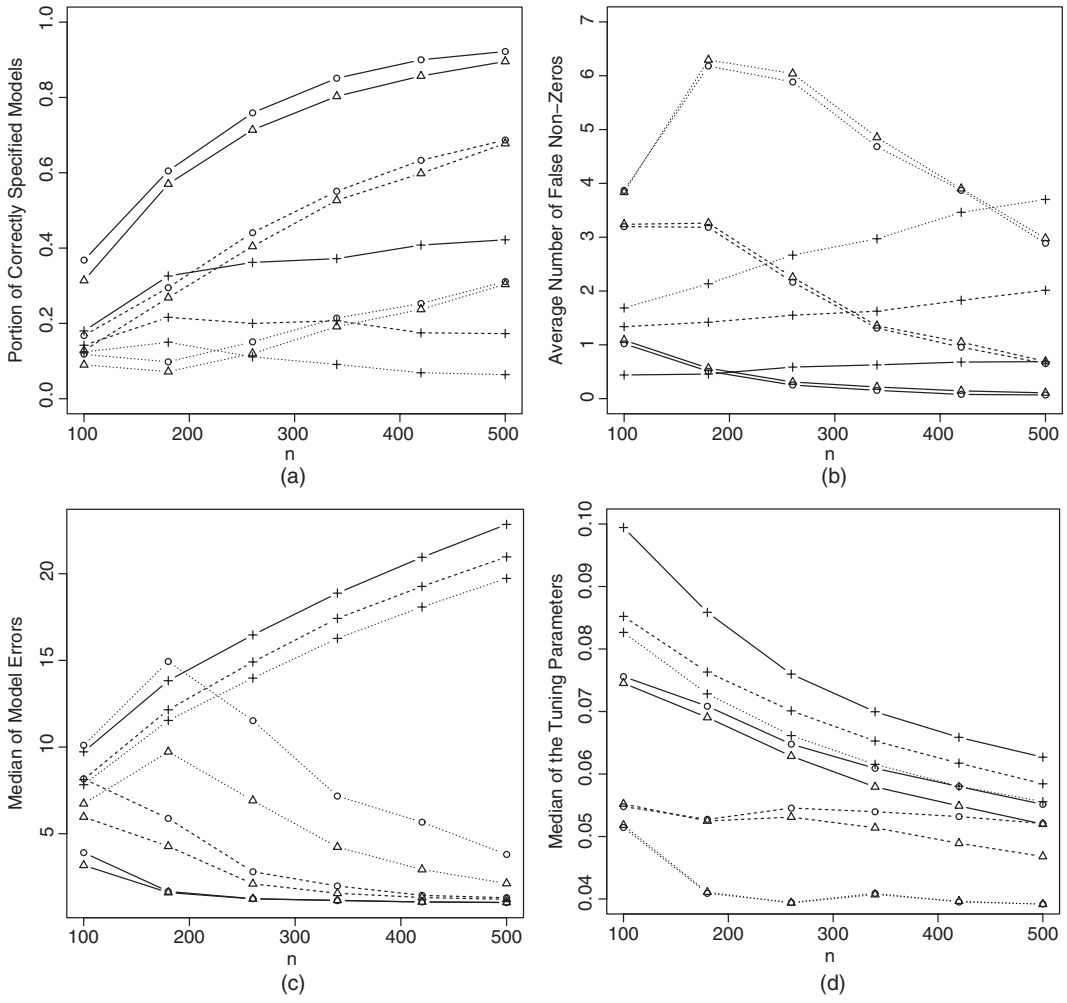

Fig. 2. Results for logistic regressions and $c_n = \log\{\log(n)\}$ (○, SCAD); △, adaptive lasso; +, lasso; ———, $\mathrm{GIC}_1 - c_n \log(p)$; - - - - - - -, $\mathrm{GIC}_2 - \log(p)$; · · · · · · ·, BIC$-\log(n)$): (a) correctly specified models; (b) average number of false non-zeros; (c) median of model errors; (d) median of the selected tuning parameters

lasso. This demonstrates the improvement by using a more accurate model selection procedure in an ultrahigh dimensional setting. As the sample size n increases, the chosen tuning parameter decreases as shown in Figs 1(d) and 2(d). We also observe from Figs 1(d) and 2(d) that $a_n = \log\{\log(n)\}\log(p)$ results in relatively larger values of selected λ. Since λ controls the level of sparsity of the model, Figs 1(d) and 2(d) reflect the extra model complexity penalty made by our GIC to select the true model from a huge collection of candidate models, as theoretically demonstrated in previous sections.

5.2. Gene expression data analysis

We now examine the tuning parameter selection procedures on the data from a gene expression study of leukaemia patients. The study is described in Golub *et al.* (1999) and the data set is available from `http://www.genome.wi.mit.edu/MPR`. The training set contains gene

expression levels of two types of acute leukaemias: 27 patients with acute lymphoblastic leukaemia and 11 patients with acute myeloid leukaemia (AML). Gene expression levels for another 34 patients are available in a test set. We applied the preprocessing steps as in Dudoit *et al.* (2002), which resulted in $p = 3051$ genes. We create a binary response variable based on the types of leukaemias by letting $Y_i = 1$ (or $Y_i = 0$) if the corresponding patient has acute lymphoblastic leukaemia (or AML). By using the gene expression levels as covariates in $\mathbf{x}_i$, we fit the data to the penalized logistic regression model (5.2) using the SCAD penalty for a sequence of tuning parameters. Applying the AIC, seven genes were selected, which is close to the results by the cross-validation procedure applied in Breheny and Huang (2011). Applying the BIC, four genes were selected. When applying GIC with $a_n = \log\{\log(n)\}\log(p)$, only one gene, CST3 Cystatin C (amyloid angiopathy and cerebral haemorrhage), was selected. We note that this gene was included in those selected by the AIC and BIC. Given the small sample size ($n = 38$) and extremely high dimensionality ($p > 3000$), the variable selection result is not surprising. By further examining the gene expression level of CST3 Cystatin C, we can find that it is actually highly informative in differentiating between the two types acute lymphoblastic leukaemia and AML even by using only one gene. To assess the out-of-sample performance, we generated the accuracy profile by first ordering the patients according to the gene expression level of CST3 Cystatin C and then plotting the top $x\%$ patients against the $y\%$ of actual AML cases among them. By looking at the accuracy profile in Fig. 3 according to the ranking using the gene expression level of CST3 Cystatin C, we can see that the profile is very close to the oracle profile that knows the truth. For comparison, we also plot the accuracy profiles based on genes selected by the AIC and BIC. As remarked in Dudoit *et al.* (2002), the out-of-sample test set

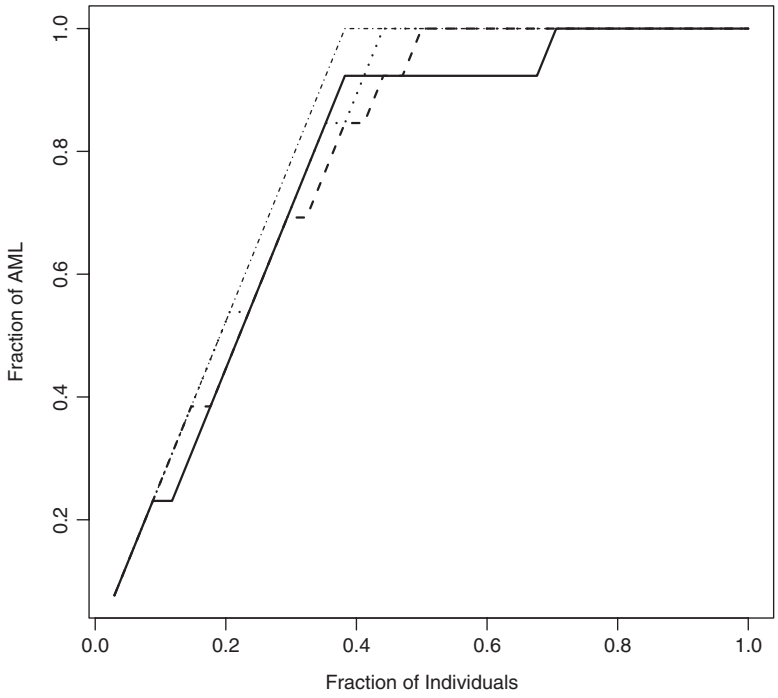

Fig. 3. Accuracy profiles of the selected gene expression level for discriminating between the types of leukaemia: ·-·-·-, oracle; ———, GIC; – – –, AIC; · · · · · · · , BIC

is more heterogeneous because of a broader range of samples, including those from peripheral blood and bone marrow, from childhood AML patients, and even from laboratories that used different sample preparation protocols. In this case, the accuracy profile is instead an informative indication in telling the predicting power of gene expression levels.

6. Technical conditions and intermediate results

For model identifiability, we assume in Section 3.1 that equation (3.1) has a unique minimizer $\hat{\boldsymbol{\beta}}^*(\alpha)$ for all α satisfying $|\alpha| \leqslant K$. By optimization theory, $\boldsymbol{\beta}^*(\alpha)$ is the unique solution to the first-order equation

$$\mathbf{X}_\alpha^{\mathrm{T}}\{\mathbf{b}'(\mathbf{X}\boldsymbol{\beta}_0) - \mathbf{b}'(\mathbf{X}\boldsymbol{\beta}(\alpha))\} = \mathbf{0}, \tag{6.1}$$

where $\mathbf{X}_\alpha$ is the design matrix corresponding to model α. It has been discussed in Lv and Liu (2010) that a sufficient condition for the uniqueness of the solution to equation (6.1) is the combination of condition 1 below and the assumption that $\mathbf{X}_\alpha$ has full rank. In practice, requiring the design matrix $\mathbf{X}_\alpha$ to be full rank is not stringent because a violation means that some explanatory variables can be expressed as linear combinations of other variables, and thus they can always be eliminated to make the design matrix non-singular.

For theoretical analysis, we assume that the true parameter $\boldsymbol{\beta}_0$ is in some sufficiently large, convex, compact set $\mathcal{B}$ in $\mathbf{R}^p$, and that $\|\boldsymbol{\beta}^*(\alpha)\|_\infty$ is uniformly bounded by some positive constant for all models α with $|\alpha| \leqslant K$. Denote by $\mathbf{W} = (W_1, \ldots, W_n)^{\mathrm{T}}$ where $W_i = Y_i - E[Y_i]$ is the model error for the ith observation. The following conditions are imposed in the theoretical developments of results in this paper.

Condition 1. The function $b(\theta)$ is three times differentiable with $c_0 \leqslant b''(\theta) \leqslant c_0^{-1}$ and $|b'''(\theta)| \leqslant c_0^{-1}$ in its domain for some constant $c_0 > 0$.

Condition 2. For any $\alpha \subset \{1, \ldots, p\}$ such that $|\alpha| \leqslant K$, $n^{-1}\mathbf{X}_\alpha^{\mathrm{T}}\mathbf{X}_\alpha$ has the smallest and largest eigenvalues bounded from below and above by c_1 and $1/c_1$ for some $c_1 > 0$, where K is some positive integer satisfying $K > s$ and $K = o(n)$.

Condition 3. For unbounded and non-Gaussian distributed Y_i, there is a diverging sequence $m_n = o(\sqrt{n})$ such that

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}_1} \max_{1 \leqslant i \leqslant n} |b'(|\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}|)| \leqslant m_n, \tag{6.2}$$

where $\mathcal{B}_1 = \{\boldsymbol{\beta} \in \mathcal{B} : |\mathrm{supp}(\boldsymbol{\beta})| \leqslant K\}$. Additionally W_is follow the uniform sub-Gaussian distribution—i.e. there are constants $c_2, c_3 > 0$ such that, uniformly for all $i = 1, \ldots, n$,

$$P(|W_i| \geqslant t) \leqslant c_2 \exp(-c_3 t^2) \qquad \text{for any } t > 0. \tag{6.3}$$

Condition 4. There is a $\lambda_0 \in [\lambda_{\min}, \lambda_{\max}]$ such that $\alpha_{\lambda_0} = \alpha_0$ and $\|\hat{\boldsymbol{\beta}}^{\lambda_0} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-\pi})$ for $0 < \pi < \frac{1}{2}$. Moreover, for each fixed λ, $p'_\lambda(t)$ is non-increasing over $t \in (0, \infty)$. Also, $n^\pi \min_{j \in \alpha_0} |\beta_{0j}| \to \infty$ as $n \to \infty$.

Condition 1 implies that the generalized linear model (2.2) has smooth and bounded variance function. It ensures the existence of the Fisher information for statistical inference with model (2.2). For commonly used generalized linear models, condition 1 is satisfied. These include the Gaussian linear model, the logistic regression model and the Poisson regression model with bounded variance function. Thus, all models fitted in Section 5 satisfy this condition. Condition 2 on the design matrix is important for ensuring the uniqueness of the population parameter

$\boldsymbol{\beta}^*(\alpha)$. If a random-design matrix is considered, Wang (2009) showed that condition 2 holds with probability tending to 1 under appropriate assumptions on the distribution of the predictor vector $\mathbf{x}_i$, true model size s and the dimensionality p, which are satisfied by the settings in our simulation examples.

Condition 3 is a technical condition that is used to control the tail behaviour of unbounded non-Gaussian response Y_i. It is imposed to ensure a general and broad applicability of the method. For many practically applied models such as those in Section 5, this condition is not required. Inequality (6.2) is on the mean function of the response variable, whereas condition (6.3) is on the tail probability distribution of the model error. The combination of conditions (6.2) and (6.3) controls the magnitude of the response variable Y_i in probability uniformly. If we further have $\|\boldsymbol{\beta}\|_\infty \leqslant C$ with some constant $C > 0$ for any $\boldsymbol{\beta} \in \mathcal{B}_1$, with $\mathcal{B}_1$ defined in condition 3, then

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}_1} \max_{1\leqslant i\leqslant n} |\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}| \leqslant \sup_{\boldsymbol{\beta}\in\mathcal{B}_1} \|\mathbf{X}_{\mathrm{supp}(\boldsymbol{\beta})}\|_\infty \|\boldsymbol{\beta}\|_\infty \leqslant CK \max_{ij} |x_{ij}|.$$

Hence, condition (6.2) holds if $|b'(t)|$ is bounded by m_n for all $|t| \leqslant CK \max_{ij} |x_{ij}|$. Conditions that are analogous to condition (6.2) were made in Fan and Song (2010) and Bühlmann and van de Geer (2011) for studying high dimensional penalized likelihood methods.

Since our interest is in tuning parameter selection, we impose condition 4 to ensure that the true model can be recovered by regularization methods. No requirements in this condition are restrictive from the practical perspective, and their validity and applicability can be supported by existing results in the literature on variable selection via regularization methods. Specifically, the first part of condition 4 is satisfied automatically if the penalized likelihood method maximizing equation (2.3) has the oracle property (Fan and Li, 2001). Meanwhile, the desirable oracle property and selection consistency for various penalized likelihood methods have been extensively studied recently. For example, Zhao and Yu (2006) proved that, in the linear model setting, the lasso method with l_1-penalty $p_\lambda(t) = \lambda t$ has model selection consistency under the strong irrepresentable condition. Zhang and Huang (2006) studied the sparsity and bias of the lasso estimator and established the consistency rate, and Lv and Fan (2009) established the weak oracle property of the regularized least squares estimator with general concave penalty functions. For generalized linear models, Fan and Lv (2011) proved that the penalized likelihood methods with folded concave penalty functions enjoy the oracle property in the setting of non-polynomial dimensionality. The second part of condition 4 is a mild assumption on $p_\lambda(t)$ to avoid excessive bias, which is satisfied by commonly used penalty functions in practice, including those in our numerical examples—i.e. the lasso, SCAD and minimax concave penalty. The last part of condition 4, $n^\pi \min_{j\in\alpha_0} |\beta_{0j}| \to \infty$, is a general and reasonable specification on the signal strength for ensuring the model selection sign consistency—i.e. $\mathrm{sgn}(\hat{\boldsymbol{\beta}}^{\lambda_0}) = \mathrm{sgn}(\boldsymbol{\beta}_0)$, of the estimator $\hat{\boldsymbol{\beta}}^{\lambda_0}$. This, together with the technical condition $p'_{\lambda_0}(\frac{1}{2}\min_{j\in\alpha_0} |\beta_{0j}|) = o(s^{-1/2}n^{-1/2}a_n^{1/2})$ in proposition 1, is used to show that $\|\hat{\boldsymbol{\beta}}^{\lambda_0} - \hat{\boldsymbol{\beta}}^*(\alpha_0)\|_2 = o_p\{\log(p)^{\xi/2}/\sqrt{n}\}$ with ξ defined in proposition 3. For a more specific data model and penalty function, alternative weaker conditions may replace condition 4 as long as the same result holds.

We now establish the uniform convergence of the MLE $\hat{\boldsymbol{\beta}}^*(\alpha)$ to the population parameter $\boldsymbol{\beta}^*(\alpha)$ over all models α with $|\alpha| \leqslant K$. This intermediate result plays a pivotal role in measuring the goodness of fit of underfitted and overfitted models in Section 3.

Proposition 2. Under conditions 1 and 2, as $n \to \infty$,

$$\sup_{\substack{|\alpha|\leqslant K \\ \alpha\subset\{1,\ldots,p\}}} \frac{1}{\sqrt{|\alpha|}} \|\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha)\|_2 = O_p\left[L_n \sqrt{\left\{\frac{\log(p)}{n}\right\}}\right],$$

when either

(a) the Y_is are bounded or Gaussian distributed, $L_n = O(1)$ and $\log(p) = o(n)$ or
(b) the Y_is are unbounded non-Gaussian distributed, additional condition 3 holds, $L_n = O\{\sqrt{\log(n)} + m_n\}$ and $\log(p) = o(n/L_n^2)$.

Proposition 2 extends the consistency result of $\hat{\boldsymbol{\beta}}^*(\alpha_0)$ to $\boldsymbol{\beta}_0$ to the uniform setting over all candidate models with model size less than K, where there are $\binom{p}{k} \sim p^K$ such models in total. The large amount of candidate models causes the extra term $\log(p)$ in the rate of convergence.

On the basis of proposition 2, we have the following result on the log-likelihood ratio for non-Gaussian generalized linear model response. It parallels result (3.7) in the Gaussian response setting.

Proposition 3. Suppose that the design matrix satisfies $\max_{ij} |x_{ij}| = O(n^{1/2-\tau})$ with $\tau \in (0, \frac{1}{2}]$. Then, under conditions 1 and 2, uniformly for all models $\alpha \supseteq \alpha_0$ with $|\alpha| \leqslant K$, as $n \to \infty$,

$$\begin{aligned} l_n\{\hat{\boldsymbol{\beta}}^*(\alpha)\} - l_n\{\boldsymbol{\beta}^*(\alpha)\} = {} & \tfrac{1}{2}(\mathbf{Y} - \boldsymbol{\mu}_0)^{\mathrm{T}} \mathbf{H}_0^{-1/2} \mathbf{B}_\alpha \mathbf{H}_0^{-1/2} (\mathbf{Y} - \boldsymbol{\mu}_0) \\ & + |\alpha|^{5/2} O_p\{L_n^2 n^{1/2-2\tau} \log(p)^{1+\xi/2}\} + |\alpha|^4 O_p\{n^{1-4\tau} \log(p)^2\} \\ & + |\alpha|^3 O_p\{L_n^3 n^{1-3\tau} \log(p)^{3/2}\} \end{aligned}$$

when

(a) the Y_is are bounded, $\xi = \frac{1}{2}$ and $L_n = O(1)$ or
(b) the Y_is are unbounded non-Gaussian distributed, additional condition 3 holds, $\xi = 1$ and $L_n = \sqrt{\log(n)} + m_n$.

Acknowledgements

We thank the Editor, the Associate Editor and two referees for their insightful comments and constructive suggestions that have greatly improved the presentation of the paper. Fan's work was partially supported by National Science Foundation 'Career' award DMS-1150318 and grant DMS-0906784 and the 2010 Zumberge Individual Award from the University of Southern California's James H. Zumberge Faculty Research and Innovation Fund. Tang is affiliated also with the National University of Singapore and acknowledges support from National University of Singapore academic research grants and a research grant from the Risk Management Institute, National University of Singapore.

Appendix A

A.1. Lemmas

We first present a few lemmas whose proofs are given in the on-line supplementary material.

Lemma 1. Assume that $W_1, \ldots, W_n$ are independent and have uniform sub-Gaussian distribution (6.3). Then, with probability at least $1 - o(1)$,

$$\|\mathbf{W}\|_\infty \leqslant C_1 \sqrt{\log(n)}$$

with some constant $C_1 > 0$. Moreover, for any positive sequence $\tilde{L}_n \to \infty$, if n is sufficiently large, there is some constant $C_2 > 0$ such that

$$n^{-1} \sum_{i=1}^{n} E[W_i | \Omega_n]^2 \leqslant C_2 \tilde{L}_n \exp(-C_2 \tilde{L}_n^2).$$

Lemma 2. If the Y_is are unbounded non-Gaussian distributed and conditions 1 and 2 hold, then, for any diverging sequence $\gamma_n \to \infty$ satisfying $\gamma_n L_n \sqrt{\{K \log(p)/n\}} \to 0$,

$$\sup_{|\alpha|\leqslant K}\frac{1}{|\alpha|}Z_\alpha\left[\gamma_n L_n\sqrt{\left\{|\alpha|\frac{\log(p)}{n}\right\}}\right]=O_p\{L_n^2 n^{-1}\log(p)\},\tag{A.1}$$

where $L_n=2m_n+C_1\sqrt{\log(n)}$ with C_1 defined in lemma 1. If the Y_is are bounded and conditions 1 and 2 hold, then the same result holds with L_n replaced with 1.

Lemma 3. Let $\tilde{\mathbf{Y}}\equiv(\tilde{Y}_1,\ldots,\tilde{Y}_n)^{\mathrm{T}}=\mathbf{H}_0^{-1/2}(\mathbf{Y}-\boldsymbol{\mu}_0)$. For any $K=o(n)$,

$$\sup_{\alpha\supset\alpha_0,|\alpha|\leqslant K}\frac{1}{|\alpha|-|\alpha_0|}\tilde{\mathbf{Y}}^{\mathrm{T}}(\mathbf{B}_\alpha-\mathbf{B}_{\alpha_0})\tilde{\mathbf{Y}}=O_p\{\log(p)^\xi\},$$

where

(a) $\xi=\frac{1}{2}$ when the $\tilde{Y}_i$s are bounded and
(b) $\xi=1$ when the $\tilde{Y}_i$s are uniform sub-Gaussian random variables.

We use empirical process techniques to prove the main results. We first introduce some notation. For a given model α with $|\alpha|\leqslant K$ and a given $N>0$, define the set

$$\mathcal{B}_\alpha(N)=\{\boldsymbol{\beta}\in\mathbf{R}^p:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*(\alpha)\|_2\leqslant N,\operatorname{supp}(\boldsymbol{\beta})=\alpha\}\cup\{\boldsymbol{\beta}^*(\alpha)\}.$$

Consider the negative log-likelihood loss function $\rho(s,Y_i)=-Y_i s+b(s)-c(Y_i,\phi)$ for $s\in\mathbf{R}$. Then

$$l_n(\boldsymbol{\beta})=-\sum_{i=1}^n\rho(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta},Y_i).$$

Further, define $Z_\alpha(N)$ as

$$Z_\alpha(N)=\sup_{\boldsymbol{\beta}\in\mathcal{B}_\alpha(N)}n^{-1}|l_n(\boldsymbol{\beta})-l_n\{\boldsymbol{\beta}^*(\alpha)\}-E[l_n(\boldsymbol{\beta})-l_n\{\boldsymbol{\beta}^*(\alpha)\}]|.\tag{A.2}$$

It is seen that $Z_\alpha(N)$ is the supremum of the absolute value of an empirical process indexed by $\boldsymbol{\beta}\in\mathcal{B}_\alpha(N)$. Define the event $\Omega_n=\{\|\mathbf{W}\|_\infty\leqslant\tilde{L}_n\}$ with $\mathbf{W}=\mathbf{Y}-E[\mathbf{Y}]$ being the error vector and $\tilde{L}_n$ some positive sequence that may diverge with n. Then, for bounded responses, $P(\Omega_n)=1$ if $\tilde{L}_n$ is chosen as a sufficiently large constant; for unbounded and non-Gaussian responses, by lemma 1, $P(\Omega_n)=1-o(1)$ if $\tilde{L}_n=C_1\sqrt{\log(n)}$ with $C_1>0$ a sufficiently large constant. On the event Ω_n, $\|\mathbf{Y}\|_\infty\leqslant m_n+C_1\sqrt{\log(n)}$. Throughout, we use C to denote a generic positive constant, and we slightly abuse the notation by using $\boldsymbol{\beta}(\alpha)$ to denote either the p-vector or its subvector on the support α when there is no confusion.

A.2. Proof of proposition 1

First note that $\hat{\boldsymbol{\beta}}_0\equiv\hat{\boldsymbol{\beta}}^*(\alpha_0)$ maximizes the log-likelihood $l_n(\boldsymbol{\beta})$ restricted to model α_0. Thus, $\partial l_n(\hat{\boldsymbol{\beta}}_0)/\partial\boldsymbol{\beta}=0$. Moreover, it follows from condition 1 that

$$\frac{\partial}{\partial^2\boldsymbol{\beta}}l_n(\boldsymbol{\beta})=\mathbf{X}^{\mathrm{T}}\mathbf{H}(\boldsymbol{\beta})\mathbf{X}.$$

Thus, by Taylor's expansion and condition 2 we obtain

$$\begin{aligned}0&\geqslant\mathrm{GIC}^*_{a_n}(\alpha_0)-\mathrm{GIC}_{a_n}(\lambda_0)\\&=\frac{1}{n}\{l(\hat{\boldsymbol{\beta}}^{\lambda_0})-l(\hat{\boldsymbol{\beta}}_0)\}\\&=-\frac{1}{n}(\hat{\boldsymbol{\beta}}^{\lambda_0}-\hat{\boldsymbol{\beta}}_0)^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{H}(\tilde{\boldsymbol{\beta}})\mathbf{X}(\hat{\boldsymbol{\beta}}^{\lambda_0}-\hat{\boldsymbol{\beta}}_0)\geqslant-C\|\hat{\boldsymbol{\beta}}^{\lambda_0}-\hat{\boldsymbol{\beta}}_0\|_2^2,\end{aligned}\tag{A.3}$$

where $\tilde{\boldsymbol{\beta}}$ lie on the line segment connecting $\hat{\boldsymbol{\beta}}^{\lambda_0}$ and $\hat{\boldsymbol{\beta}}_0$, and we have used $\operatorname{supp}(\hat{\boldsymbol{\beta}}^{\lambda_0})=\operatorname{supp}(\hat{\boldsymbol{\beta}}_0)=\alpha_0$ for the last inequality. It remains to prove that $\|\hat{\boldsymbol{\beta}}^{\lambda_0}-\hat{\boldsymbol{\beta}}_0\|_2$ is small.

Let $\hat{\boldsymbol{\beta}}^{\lambda_0}_{\alpha_0}$ and $\hat{\boldsymbol{\beta}}_{0,\alpha_0}$ be the subvectors of $\hat{\boldsymbol{\beta}}^{\lambda_0}$ and $\hat{\boldsymbol{\beta}}_0$ on the support α_0, correspondingly. Since $\hat{\boldsymbol{\beta}}^{\lambda_0}$ minimizes $l_n(\boldsymbol{\beta})+n\sum_{j=1}^n p_{\lambda_0}(|\beta_j|)$, it follows from classical optimization theory that $\hat{\boldsymbol{\beta}}^{\lambda_0}_{\alpha_0}$ is a critical value, and thus

$$\mathbf{X}_0^{\mathrm{T}}\{\mathbf{Y}-\mathbf{b}'(\mathbf{X}_0\hat{\boldsymbol{\beta}}^{\lambda_0}_{\alpha_0})\}+n\,\bar{p}'_{\lambda_n}(\hat{\boldsymbol{\beta}}^{\lambda_0}_{\alpha_0})=0,$$

where $\mathbf{X}_0$ is the design matrix of the true model, and $\bar{p}'_{\lambda_n}(\hat{\boldsymbol{\beta}}^{\lambda_0}_{\alpha_0})$ is a vector with components $\operatorname{sgn}(\hat{\beta}_j^{\lambda_0})\,p'_{\lambda_0}(|\hat{\beta}_j^{\lambda_0}|)$ and $j\in\alpha_0$. Since $\hat{\boldsymbol{\beta}}_0$ is the MLE when restricted to the support α_0, $\mathbf{X}_0^{\mathrm{T}}\{\mathbf{Y}-\mathbf{b}'(\mathbf{X}_0\hat{\boldsymbol{\beta}}_{0,\alpha_0})\}=0$. Thus, the above equation can be rewritten as

$$\mathbf{X}_0^{\mathrm{T}}\{\mathbf{b}'(\mathbf{X}_0\hat{\boldsymbol{\beta}}_{0,\alpha_0}) - \mathbf{b}'(\mathbf{X}_0\hat{\boldsymbol{\beta}}_{\alpha_0}^{\lambda_0})\} + n\,\bar{p}'_{\lambda_n}(\hat{\boldsymbol{\beta}}_{\alpha_0}^{\lambda_0}) = 0, \tag{A.4}$$

Now, applying Taylor's expansion to equation (A.4) we obtain that $\mathbf{X}_0^{\mathrm{T}}\mathbf{H}(\mathbf{X}_0\bar{\boldsymbol{\beta}})\mathbf{X}_0(\hat{\boldsymbol{\beta}}_{\alpha_0}^{\lambda_0} - \hat{\boldsymbol{\beta}}_{0,\alpha_0}) = n\,\bar{p}_{\lambda_0}(|\hat{\boldsymbol{\beta}}_{\alpha_0}^{\lambda_0}|)$, where $\bar{\boldsymbol{\beta}}$ lies between the line segment connecting $\hat{\boldsymbol{\beta}}_{\alpha_0}^{\lambda_0}$ and $\hat{\boldsymbol{\beta}}_{0,\alpha_0}$. Therefore,

$$\hat{\boldsymbol{\beta}}_{\alpha_0}^{\lambda_0} - \hat{\boldsymbol{\beta}}_{0,\alpha_0} = n\{\mathbf{X}_0^{\mathrm{T}}\mathbf{H}(\mathbf{X}_0\bar{\boldsymbol{\beta}})\mathbf{X}_0\}^{-1}\,\bar{p}_{\lambda_0}(\hat{\boldsymbol{\beta}}_{\alpha_0}^{\lambda_0}).$$

This together with conditions 1 and 2 ensures that

$$\|\hat{\boldsymbol{\beta}}_{\alpha_0}^{\lambda_0} - \hat{\boldsymbol{\beta}}_{0,\alpha_0}\|_2 \leqslant C\|\bar{p}_{\lambda_0}(\hat{\boldsymbol{\beta}}_{\alpha_0}^{\lambda_0})\|_2. \tag{A.5}$$

Since we have assumed that $\|\hat{\boldsymbol{\beta}}^{\lambda_0} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-\pi})$, it follows that, for sufficiently large n, $\min_{j\in\alpha_0}|\hat{\boldsymbol{\beta}}_j^{\lambda_0}| \geqslant \min_{j\in\alpha_0}|\beta_{0j}| - n^{-\pi} \geqslant 2^{-1}\min_{j\in\alpha_0}|\beta_{0j}|$. Thus, by theorem assumptions,

$$\|\bar{p}'_{\lambda_0}(\hat{\boldsymbol{\beta}}_{\alpha_0}^{\lambda_0})\|_2 \leqslant \sqrt{s}\,p'_{\lambda_0}(\tfrac{1}{2}\min_{j\in\alpha_0}|\beta_{0j}|) = o\{\sqrt{(n^{-1}a_n)}\}.$$

Combining the above inequality with inequality (A.5) yields

$$\|\hat{\boldsymbol{\beta}}^{\lambda_0} - \hat{\boldsymbol{\beta}}_0\|_2 = \|\hat{\boldsymbol{\beta}}_{\alpha_0}^{\lambda_0} - \hat{\boldsymbol{\beta}}_{0,\alpha_0}\|_2 \leqslant o(n^{-1/2}a_n^{1/2}).$$

This, together with expression (A.3), completes the proof of result (2.9).

A.3. Proof of proposition 2

We first consider non-Gaussian responses. Using the similar idea in van de Geer (2002), for a given $N > 0$, define a convex combination $\hat{\boldsymbol{\beta}}_u(\alpha) = u\,\hat{\boldsymbol{\beta}}^*(\alpha) + (1-u)\,\boldsymbol{\beta}^*(\alpha)$ with $u = \{1 + \|\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha)\|_2/N\}^{-1}$. Then, by definition, $\|\hat{\boldsymbol{\beta}}_u(\alpha) - \boldsymbol{\beta}^*(\alpha)\|_2 = u\|\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha)\|_2 \leqslant N$. If $\mathrm{supp}(\hat{\boldsymbol{\beta}}_u) \neq \alpha$, then modify the definition of u a little by slightly increasing N to make $\mathrm{supp}(\hat{\boldsymbol{\beta}}_u) = \alpha$. So we assume implicitly that $\mathrm{supp}(\hat{\boldsymbol{\beta}}_u) = \alpha$ and thus that $\hat{\boldsymbol{\beta}}_u \in \mathcal{B}_\alpha(N)$. The key is to prove

$$\sup_{|\alpha|\leqslant K}\frac{1}{\sqrt{|\alpha|}}\|\hat{\boldsymbol{\beta}}_u(\alpha) - \boldsymbol{\beta}^*(\alpha)\|_2 = O_p\Big[L_n\sqrt{\Big\{\frac{\log(p)}{n}\Big\}}\Big]. \tag{A.6}$$

Then, by noting that $\|\hat{\boldsymbol{\beta}}_u(\alpha) - \boldsymbol{\beta}^*(\alpha)\|_2 \leqslant N/2$ implies that $\|\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha)\|_2 \leqslant N$, the result in proposition 2 is proved.

Now, we proceed to prove result (A.6). By the concavity of the log-likelihood function,

$$l_n\{\hat{\boldsymbol{\beta}}_u(\alpha)\} \geqslant u\,l_n\{\hat{\boldsymbol{\beta}}^*(\alpha)\} + (1-u)\,l_n\{\boldsymbol{\beta}^*(\alpha)\}.$$

Since $\hat{\boldsymbol{\beta}}^*(\alpha)$ maximizes $l_n(\boldsymbol{\beta})$ over all models with support α, the above inequality can further be written as $l_n\{\boldsymbol{\beta}^*(\alpha)\} \leqslant l_n\{\hat{\boldsymbol{\beta}}_u(\alpha)\}$. However, since $\boldsymbol{\beta}^*(\alpha)$ minimizes the KL divergence $I\{\boldsymbol{\beta}(\alpha)\}$ in equation (3.1), we obtain

$$E[l_n\{\boldsymbol{\beta}^*(\alpha)\} - l_n\{\hat{\boldsymbol{\beta}}_u(\alpha)\}] = I\{\hat{\boldsymbol{\beta}}_u(\alpha)\} - I\{\boldsymbol{\beta}^*(\alpha)\} \geqslant 0,$$

where $E[l_n\{\hat{\boldsymbol{\beta}}_u(\alpha)\}] = -\Sigma_{i=1}^n E[\rho(\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_u, Y_i)]$ should be understood as $E[\rho(\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_u, Y_i)] = \int \rho(\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_u, y)\,\mathrm{d}F_i(y)$ with $F_i(\cdot)$ being the distribution function of Y_i. Combining these two results yields

$$\begin{aligned} 0 &\leqslant E[l_n\{\boldsymbol{\beta}^*(\alpha)\} - l_n\{\hat{\boldsymbol{\beta}}_u(\alpha)\}] \\ &\leqslant l_n\{\hat{\boldsymbol{\beta}}_u(\alpha)\} - E[l_n\{\hat{\boldsymbol{\beta}}_u(\alpha)\}] - (l_n\{\boldsymbol{\beta}^*(\alpha)\} - E[l_n\{\boldsymbol{\beta}^*(\alpha)\}]) \leqslant n\,Z_\alpha(N), \end{aligned} \tag{A.7}$$

where $Z_\alpha(N)$ is defined in equation (A.2). However, by equation (6.1), for any $\boldsymbol{\beta}(\alpha) \in \mathcal{B}_\alpha(N)$,

$$\begin{aligned} E[l_n\{\boldsymbol{\beta}(\alpha)\} - l_n\{\boldsymbol{\beta}^*(\alpha)\}] &= \mathbf{b}'(\mathbf{X}\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{X}\{\boldsymbol{\beta}(\alpha) - \boldsymbol{\beta}^*(\alpha)\} - \mathbf{1}^{\mathrm{T}}\{\mathbf{b}(\mathbf{X}\,\boldsymbol{\beta}(\alpha)) - \mathbf{b}(\mathbf{X}\,\boldsymbol{\beta}^*(\alpha))\} \\ &= \mathbf{b}'(\mathbf{X}\boldsymbol{\beta}^*(\alpha))^{\mathrm{T}}\mathbf{X}\{\boldsymbol{\beta}(\alpha) - \boldsymbol{\beta}^*(\alpha)\} - \mathbf{1}^{\mathrm{T}}\{\mathbf{b}(\mathbf{X}\,\boldsymbol{\beta}(\alpha)) - \mathbf{b}(\mathbf{X}\,\boldsymbol{\beta}^*(\alpha))\} \\ &= -\tfrac{1}{2}(\boldsymbol{\beta}(\alpha) - \boldsymbol{\beta}^*(\alpha))^{\mathrm{T}}\mathbf{X}_\alpha^{\mathrm{T}}\tilde{\mathbf{H}}\mathbf{X}_\alpha(\boldsymbol{\beta}(\alpha) - \boldsymbol{\beta}^*(\alpha)), \end{aligned}$$

where $\tilde{\mathbf{H}} = \mathrm{diag}\{\mathbf{b}''(\mathbf{X}\bar{\boldsymbol{\beta}}(\alpha))\}$ and $\bar{\boldsymbol{\beta}}(\alpha)$ lies on the segment connecting $\boldsymbol{\beta}^*(\alpha)$ and $\boldsymbol{\beta}(\alpha)$. Thus, it follows from conditions 1 and 3 that, for any $\boldsymbol{\beta}(\alpha) \in \mathcal{B}_\alpha(N)$,

$$E[l_n\{\boldsymbol{\beta}(\alpha)\} - l_n\{\boldsymbol{\beta}^*(\alpha)\}] \leqslant -\tfrac{1}{2}c_0c_1n\|\boldsymbol{\beta}(\alpha) - \boldsymbol{\beta}^*(\alpha)\|_2^2.$$

This, together with inequality (A.7), entails that, for any $\boldsymbol{\beta}(\alpha) \in \mathcal{B}_\alpha(N)$,

$$\|\boldsymbol{\beta}(\alpha) - \boldsymbol{\beta}^*(\alpha)\|_2^2 \leqslant 2(c_0 c_1)^{-1} Z_\alpha(N).$$

Since $\hat{\boldsymbol{\beta}}_u \in \mathcal{B}_\alpha(N)$, taking $N = N_n \equiv \gamma_n L_n \sqrt{\{|\alpha| \log(p)/n\}}$ and by lemma 2, we have

$$\sup_{|\alpha| \leqslant K} \frac{1}{\sqrt{|\alpha|}} \|\hat{\boldsymbol{\beta}}_u(\alpha) - \boldsymbol{\beta}^*(\alpha)\|_2 \leqslant 2(c_0 c_1)^{-1} \left\{ \sup_{|\alpha| \leqslant K} \frac{1}{|\alpha|} Z_\alpha(N_n) \right\}^{1/2} = O_p\left[L_n \sqrt{\left\{ \frac{\log(p)}{n} \right\}} \right],$$

where $L_n = 2m_n + O\{\sqrt{\log(n)}\}$ when the Y_is are unbounded non-Gaussian, and $L_n = O(1)$ when the Y_is are bounded. This completes the proof of result (A.6).

Now consider the Gaussian response. For a given model α, we have the explicit form that $\hat{\boldsymbol{\beta}}^*(\alpha) = (\mathbf{X}_\alpha^{\mathrm{T}} \mathbf{X}_\alpha)^{-1} \mathbf{X}_\alpha^{\mathrm{T}} \mathbf{Y}$. Since $\mathbf{X}_\alpha^{\mathrm{T}} \{\mathbf{X} \boldsymbol{\beta}^*(\alpha) - \mathbf{X} \boldsymbol{\beta}_0\} = 0$, direct calculation yields

$$\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha) = (\mathbf{X}_\alpha^{\mathrm{T}} \mathbf{X}_\alpha)^{-1} \mathbf{X}_\alpha^{\mathrm{T}} \mathbf{W}.$$

Since $\mathbf{W} \sim N(0, \sigma^2 I_n)$, it follows that $\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha) \sim N(0, \sigma^2 I_{|\alpha|})$. So $\sigma^{-2} \|\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha)\|_2^2 \sim \chi^2_{|\alpha|}$. Thus, for $t > 0$, there exists $C > 0$:

$$P\{\|\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha)\|_2^2 \geqslant |\alpha| t\} \leqslant C \exp(-C|\alpha| t).$$

Using a similar method to that before, we obtain that

$$\sup_{|\alpha| \leqslant K} |\alpha|^{-1/2} \|\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha)\|_2 = O_p[\sqrt{\{\log(p)/n\}}].$$

This completes the proof.

A.4. Proof of proposition 3

By Taylor series expansion, $l_n\{\hat{\boldsymbol{\beta}}^*(\alpha)\} - l_n\{\boldsymbol{\beta}^*(\alpha)\}$ can be written as

$$l_n\{\hat{\boldsymbol{\beta}}^*(\alpha)\} - l_n\{\boldsymbol{\beta}^*(\alpha)\} = I_1(\alpha) - I_2(\alpha) + I_3(\alpha), \tag{A.8}$$

where

$$I_1(\alpha) = (\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha))^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \{\mathbf{Y} - \mathbf{b}''(\mathbf{X} \boldsymbol{\beta}^*(\alpha))\}, \tag{A.9}$$

$$I_2(\alpha) = \tfrac{1}{2} (\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha))^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{H}_0 \mathbf{X} (\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha)), \tag{A.10}$$

and $I_3(\alpha)$ is the remainder term. We shall study them one by one.

We first consider $I_1(\alpha)$. Since $\hat{\boldsymbol{\beta}}^*(\alpha)$ is the MLE, it satisfies the first-order equation $\mathbf{X}_\alpha^{\mathrm{T}} \{\mathbf{Y} - \mathbf{b}'(\mathbf{X} \hat{\boldsymbol{\beta}}^*(\alpha))\} = 0$. Applying the Taylor series expansion to $\mathbf{b}'(\mathbf{X} \hat{\boldsymbol{\beta}}^*(\alpha))$ yields

$$\mathbf{X}_\alpha^{\mathrm{T}} \mathbf{Y} = \mathbf{X}_\alpha^{\mathrm{T}} \{\mathbf{b}'(\mathbf{X} \boldsymbol{\beta}^*(\alpha)) + \mathbf{H}_0 \mathbf{X} (\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha)) + \mathbf{v}_\alpha\},$$

where

$$\mathbf{v}_\alpha = (v_1, \ldots, v_n)^{\mathrm{T}} \qquad v_i = \tfrac{1}{2} b'''\{\mathbf{x}_i^{\mathrm{T}} \tilde{\boldsymbol{\beta}}^*(\alpha)\} \{\mathbf{x}_i^{\mathrm{T}} (\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha))\}^2 \tag{A.11}$$

and $\tilde{\boldsymbol{\beta}}^*(\alpha)$ lying on the line segment connecting $\boldsymbol{\beta}^*(\alpha)$ and $\hat{\boldsymbol{\beta}}^*(\alpha)$. Since equation (6.1) ensures that $\mathbf{X}_\alpha^{\mathrm{T}} \mathbf{b}'(\mathbf{X} \boldsymbol{\beta}_0) = \mathbf{X}_\alpha^{\mathrm{T}} \mathbf{b}'(\mathbf{X} \boldsymbol{\beta}^*(\alpha))$, thus we have

$$\begin{aligned} \hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha) &= (\mathbf{X}_\alpha^{\mathrm{T}} \mathbf{H}_0 \mathbf{X}_\alpha)^{-1} \mathbf{X}_\alpha^{\mathrm{T}} \{\mathbf{Y} - \mathbf{b}'(\mathbf{X} \boldsymbol{\beta}^*(\alpha)) - \mathbf{v}_\alpha\} \\ &= (\mathbf{X}_\alpha^{\mathrm{T}} \mathbf{H}_0 \mathbf{X}_\alpha)^{-1} \mathbf{X}_\alpha^{\mathrm{T}} (\mathbf{Y} - \boldsymbol{\mu}_0 - \mathbf{v}_\alpha). \end{aligned} \tag{A.12}$$

Combining equations (A.9) and (A.12), we obtain

$$I_1(\alpha) = (\mathbf{Y} - \boldsymbol{\mu}_0)^{\mathrm{T}} \mathbf{H}_0^{-1/2} \mathbf{B}_\alpha \mathbf{H}_0^{-1/2} (\mathbf{Y} - \boldsymbol{\mu}_0) + R_{1,\alpha}, \tag{A.13}$$

where $\mathbf{B}_\alpha$ is defined in equation (3.6) and $R_{1,\alpha} = -\mathbf{v}_\alpha^{\mathrm{T}} \mathbf{H}_0^{-1/2} \mathbf{B}_\alpha \mathbf{H}_0^{-1/2} \mathbf{W}$. We only need to study $R_{1,\alpha}$. By the Cauchy–Schwartz inequality, we have

$$|R_{1,\alpha}| \leqslant \|\mathbf{B}_\alpha \mathbf{H}_0^{-1/2}\mathbf{W}\|_2 \|\mathbf{H}_0^{-1/2}\mathbf{v}_\alpha\|_2 \leqslant (\|\mathbf{B}_{\alpha_0}\mathbf{H}_0^{-1/2}\mathbf{W}\|_2 + \|\tilde{R}_{1,\alpha}\|_2)\|\mathbf{H}_0^{-1/2}\mathbf{v}_\alpha\|_2, \tag{A.14}$$

where $\tilde{R}_{1,\alpha} = (\mathbf{B}_\alpha - \mathbf{B}_{\alpha_0})\mathbf{H}_0^{-1/2}\mathbf{W}$. We consider the terms on the very right-hand side of inequality (A.14) one by one. By the Markov inequality, and noting that $\mathbf{H}_0 = E[\mathbf{W}\mathbf{W}^{\mathrm{T}}]$ and $\mathrm{tr}(\mathbf{B}_{\alpha_0}\mathbf{B}_{\alpha_0}) = |\alpha_0|$, we can derive that for any $\gamma_n \to \infty$

$$\begin{aligned} P\{\|\mathbf{B}_{\alpha_0}\mathbf{H}_0^{-1/2}\mathbf{W}\|_2 \geqslant \sqrt{(|\alpha_0|\gamma_n)}\} &\leqslant \frac{1}{|\alpha_0|\gamma_n} E[\|\mathbf{B}_{\alpha_0}\mathbf{H}_0^{-1/2}\mathbf{W}\|_2^2] \\ &= \frac{1}{|\alpha_0|\gamma_n}\mathrm{tr}(\mathbf{B}_{\alpha_0}\mathbf{H}_0^{-1/2}E[\mathbf{W}\mathbf{W}^{\mathrm{T}}]\mathbf{H}_0^{-1/2}\mathbf{B}_{\alpha_0}) = \frac{1}{\gamma_n} \to 0. \end{aligned}$$

Therefore,

$$\|\mathbf{B}_{\alpha_0}\mathbf{H}_0^{-1/2}\mathbf{W}\|_2 = O_p(\sqrt{|\alpha_0|}). \tag{A.15}$$

Next, by lemma 3 we obtain that uniformly for all α

$$(|\alpha| - |\alpha_0|)^{-1/2}\|\tilde{R}_{1,\alpha}\|_2 = O_p\{\log(p)^{\xi/2}\}, \tag{A.16}$$

where ξ is defined therein. Finally we consider $\|\mathbf{H}_0^{-1/2}\mathbf{v}_\alpha\|_2$. Since $b'''(\cdot)$ is bounded, $\max_{ij}|x_{ij}| = O(n^{1/2-\tau})$ and $\mathrm{supp}\{\hat{\boldsymbol{\beta}}^*(\alpha)\} = \mathrm{supp}\{\boldsymbol{\beta}^*(\alpha)\} = \alpha$, by equation (A.11) and condition 2,

$$\begin{aligned} \|\mathbf{H}_0^{-1/2}\mathbf{v}_\alpha\|_2 &\leqslant C\|\mathbf{v}_\alpha\|_2 \leqslant C\left[\sum_{i=1}^n |\mathbf{x}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha))|^4\right]^{1/2} \\ &\leqslant C|\alpha|n^{3/2-2\tau}\|\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha)\|_2^2 = |\alpha|^2 O_p\{L_n^2 n^{1/2-2\tau}\log(p)\}. \end{aligned} \tag{A.17}$$

Combining expressions (A.14)–(A.17), and in view of equation (A.13), we obtain that

$$I_1(\alpha) = (\mathbf{Y} - \boldsymbol{\mu}_0)^{\mathrm{T}}\mathbf{H}_0^{-1/2}\mathbf{B}_\alpha\mathbf{H}_0^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}_0) + R_{1,\alpha}, \tag{A.18}$$

where, uniformly for all overfitted models α,

$$R_{1,\alpha} = |\alpha|^{5/2} O_p\{L_n^2 n^{1/2-2\tau}\log(p)^{1+\xi/2}\}. \tag{A.19}$$

Next, we consider $I_2(\alpha)$ defined in equation (A.10). By equation (7.12) we have the decomposition

$$\begin{aligned} I_2(\alpha) &= \tfrac{1}{2}(\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha))^{\mathrm{T}}\mathbf{X}_\alpha^{\mathrm{T}}\mathbf{H}_0\mathbf{X}_\alpha(\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha)) \\ &= \tfrac{1}{2}(\mathbf{Y} - \boldsymbol{\mu}_0)^{\mathrm{T}}\mathbf{H}_0^{-1/2}\mathbf{B}_\alpha\mathbf{H}_0^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}_0) + \tfrac{1}{2}R_{2,\alpha} - R_{1,\alpha}, \end{aligned} \tag{A.20}$$

where $R_{2,\alpha} = \mathbf{v}_\alpha^{\mathrm{T}}\mathbf{H}_0^{-1/2}\mathbf{B}_\alpha\mathbf{H}_0^{-1/2}\mathbf{v}_\alpha$, and $R_{1,\alpha}$ is defined in equations (A.18) and (A.19). We only need to study $R_{2,\alpha}$. Since $b''(\cdot)$ is bounded, $\max_{i,j}|x_{ij}| = O(n^{1/2-\tau})$ and $\mathbf{B}_\alpha$ is a projection matrix, it is easy to derive that

$$R_{2,\alpha} = \mathbf{v}_\alpha^{\mathrm{T}}\mathbf{H}_0^{-1/2}\mathbf{B}_\alpha\mathbf{H}_0^{-1/2}\mathbf{v}_\alpha \leqslant \mathbf{v}_\alpha^{\mathrm{T}}\mathbf{H}_0^{-1}\mathbf{v}_\alpha \leqslant C\|\mathbf{v}_\alpha\|_2^2 = |\alpha|^4 O_p\{n^{1-4\tau}\log(p)^2\}, \tag{A.21}$$

where the last step is because of theorem 4 and expression (A.11). The above result is uniformly over all α with $|\alpha| \leqslant K$. This, together with expressions (A.10) and (A.19)–(A.21), ensures that, uniformly for all overfitted models α,

$$I_2(\alpha) = \tfrac{1}{2}(\mathbf{Y} - \boldsymbol{\mu}_0)^{\mathrm{T}}\mathbf{H}_0^{-1/2}\mathbf{B}_\alpha\mathbf{H}_0^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}_0) + |\alpha|^{5/2}O_p\{L_n^2 n^{1/2-2\tau}\log(p)^{1+\xi/2}\} + |\alpha|^4 O_p\{n^{1-4\tau}\log(p)^2\}. \tag{A.22}$$

Finally, we consider $I_3(\alpha)$ in equation (A.8). Since $b'''(\cdot)$ is bounded, by theorem 4 we have

$$|I_3(\alpha)| \leqslant Cn^{5/2-3\tau}|\alpha|^{3/2}\|\hat{\boldsymbol{\beta}}^*(\alpha) - \boldsymbol{\beta}^*(\alpha)\|_2^3 = |\alpha|^3 O_p\{n^{1-3\tau}L_n^3\log(p)^{3/2}\},$$

where this result is uniformly over all α with $|\alpha| \leqslant K$. This result, together with equations (A.18), (A.22) and (A.8), completes the proof of proposition 3.

A.5. Proof of theorem 1

We note that theorem 1 is a direct consequence of the following two propositions, the proofs of which are given in the on-line supplementary material.

Proposition 4. In either situation (a) or (b) in proposition 2, and under the same conditions, as $n \to \infty$,

$$\sup_{\substack{|\alpha| \leqslant K \\ \alpha \subset \{1,\ldots,p\}}} \frac{1}{n|\alpha|} \{l_n(\hat{\boldsymbol{\mu}}^*_\alpha; \mathbf{Y}) - l_n(\boldsymbol{\mu}^*_\alpha; \mathbf{Y})\} = O_p\{n^{-1} L_n^2 \log(p)\}.$$

Proposition 5. Under conditions 1 and 2, as $n \to \infty$,

$$\sup_{\substack{|\alpha| \leqslant K \\ \alpha \subset \{1,\ldots,p\}}} \frac{1}{n|\alpha|} |l_n(\boldsymbol{\mu}^*_\alpha; \mathbf{Y}) - E[l_n(\boldsymbol{\mu}^*_\alpha; \mathbf{Y})]| = O_p\left[\sqrt{\left\{\frac{\log(p)}{n}\right\}}\right]$$

when either

(a) the Y_is are bounded or Gaussian distributed and $\log(p) = o(n)$ or
(b) the Y_is are unbounded and non-Gaussian distributed, additional condition 3 holds, $\max_{ij} |x_{ij}| = O(n^{1/2-\tau})$ with $\tau \in (0, \frac{1}{2}]$ and $K^2 \log(p) = o(n^{2\tau})$.

A.6. Proof of theorem 2

Theorem 2 follows directly from lemma 3 and proposition 3.

A.7. Proof of theorem 3

Combining equations (3.4) with (3.9), and in view of theorems 4 and 2, we obtain that, if $\delta_n K^{-1} R_n^{-1} \to \infty$, a_n satisfies $n\delta_n s^{-1} a_n^{-1} \to \infty$, and $a_n \psi_n^{-1} \to \infty$, then

$$P\left[\inf_{\alpha \supsetneq \alpha_0} \{\mathrm{GIC}^*_{a_n}(\alpha) - \mathrm{GIC}^*_{a_n}(\alpha_0)\} > \frac{\delta_n}{2} \quad \text{and} \quad \inf_{\alpha \not\supset \alpha_0} \{\mathrm{GIC}^*_{a_n}(\alpha) - \mathrm{GIC}^*_{a_n}(\alpha_0)\} > \frac{a_n}{2n}\right] \to 1, \tag{A.23}$$

where R_n and ψ_n are specified in theorems 1 and 2. This, together with proposition 1 and equation (2.10), completes the proof of the theorem.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B. N. Petrov and F. Csàki). Budapest: Akademiai Kiado.

Bai, Z. D., Rao, C. R. and Wu, Y. (1999) Model selection with data-oriented penalty. *J. Statist. Planng Inf.*, **77**, 103–117.

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

Bühlmann, P. and van de Geer, S. (2011) *Statistics for High-dimensional Data: Methods Theory and Applications*. New York: Springer.

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

De La Peña, V. H. and Montgomery-Smith, S. J. (1994) Bounds on the tail probability of U-statistics and quadratic forms. *Bull. Am. Math. Soc.*, **31**, 223–227.

Dudoit, S., Fridlyand, J. and Speed, T. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Statist. Ass.*, **97**, 77–87.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.

Fan, J. and Lv, J. (2010) A selective overview of variable selection in high dimensional feature space. *Statist. Sin.*, **20**, 101–148.

Fan, J. and Lv, J. (2011) Non-concave penalized likelihood with np-dimensionality. *IEEE Trans. Inform. Theor.*, **57**, 5467–5484.

Fan, J. and Song, R. (2010) Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.*, **38**, 3567–3604.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Statist. Softwr.*, **33**, 1–22.

van de Geer, S. (2002) M-estimation using penalties or sieves. *J. Statist. Planng Inf.*, **108**, 55–69.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning—Data Mining, Inference and Prediction*, 2nd edn. New York: Springer.
van der Hilst, R., de Hoop, M., Wang, P., Shim, S.-H., Ma, P. and Tenorio, L. (2007) Seismo-stratigraphy and thermal structure of earth's core-mantle boundary region. *Science*, **315**, 1813–1817.
Jagannathan, R. and Ma, T. (2003) Risk reduction in large portfolios: why imposing the wrong constraints helps. *J. Finan.*, **58**, 1651–1683.
Lv, J. and Fan, Y. (2009) A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.*, **37**, 3498–3528.
Lv, J. and Liu, J. (2010) Model selection principles in misspecified models. *Manuscript*. University of Southern California, Los Angeles.
McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
Nishii, R. (1984) Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758–765.
Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Shao, J. (1997) An asymptotic theory for linear model selection. *Statist. Sin.*, **7**, 221–264.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Wang, H. (2009) Forward regression for ultra-high dimensional variable screening. *J. Am. Statist. Ass.*, **104**, 1512–1524.
Wang, H., Li, B. and Leng, C. (2009) Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Statist. Soc.* B, **71**, 671–683.
Wang, H., Li, R. and Tsai, C. L. (2007) Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553–568.
Wang, T. and Zhu, L. (2011) Consistent tuning parameter selection in high dimensional sparse linear regression. *J. Multiv. Anal.*, **102**, 1141–1151.
Yang, Y. (2005) Can the strengths of aic and bic be shared?: a conflict between model identification and regression estimation. *Biometrika*, **92**, 937–950.
Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942.
Zhang, C. H. and Huang, J. (2006) The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.*, **36**, 1567–1594.
Zhang, Y., Li, R. and Tsai, C. L. (2010) Regularization parameter selections via generalized information criterion. *J. Am. Statist. Ass.*, **105**, 312–323.
Zhao, P. and Yu, B. (2006) On model selection consistency of Lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.
Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.*, **36**, 1509–1533.

Supporting information
Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material to "Tuning parameter selection in high-dimensional penalized likelihood" '.