Biometrika (2017), **104**, 2, *pp.* 251–272
Printed in Great Britain
doi: 10.1093/biomet/asx023
Advance Access publication 9 May 2017

Data integration with high dimensionality

By XIN GAO

Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, Ontario M3J 1P3, Canada
xingao@mathstat.yorku.ca

and RAYMOND J. CARROLL

Department of Statistics, 447 Blocker Building, Texas A&M University, College Station, Texas 77843, U.S.A.
carroll@stat.tamu.edu

Summary

We consider situations where the data consist of a number of responses for each individual, which may include a mix of discrete and continuous variables. The data also include a class of predictors, where the same predictor may have different physical measurements across different experiments depending on how the predictor is measured. The goal is to select which predictors affect any of the responses, where the number of such informative predictors tends to infinity as the sample size increases. There are marginal likelihoods for each experiment; we specify a pseudolikelihood combining the marginal likelihoods, and propose a pseudolikelihood information criterion. Under regularity conditions, we establish selection consistency for this criterion with unbounded true model size. The proposed method includes a Bayesian information criterion with appropriate penalty term as a special case. Simulations indicate that data integration can dramatically improve upon using only one data source.

Some key words: Information criterion; Large deviation; Model misspecification; Pseudolikelihood; Quadratic form.

1. Introduction

Consider the following simple but common examples of data integration.

Example 1. We have a set of individuals whose disease status is observed. We also measure different facets of individual genes, such as mRNA expression, protein expression, RNAseq expression, and so on. The question is: which genes affect the disease in any of the different ways the genes are measured? In this example, the gene is a predictor, which can be assessed in a number of ways through different measurement processes or experiments.

Example 2. Suppose that the individual is assessed through various responses, measurement mechanisms or experiments, while the predictor is the same across these experiments, and we want to examine which predictor affects any of the responses.

We consider a formulation that includes both examples as well as combinations of them. We recognize that the marginal probability densities of the responses will be different among

experiments and the measurements from different experiments can be correlated, as they would be in both of the examples above. Our goal is to combine the various marginal likelihoods in an appropriate manner and perform inference based on a pseudolikelihood and an information criterion that we develop, doing so in such a way as to allow the number of informative predictors to tend to infinity as the sample size increases.

One way to approach Example 1 is to pool the different means of measuring the gene and apply a version of the group lasso, the group being the gene. The group penalty was first formulated in a 1999 Australian National University PhD thesis by S. Bakin and later used to solve group selection problems by Yuan & Lin (2006). The group penalty penalizes the L_2-norm of the grouped parameter vector, so it is able to select predictors based on their overall strengths across experiments. Alternatively, penalty functions such as the smoothly clipped absolute deviation penalty (Fan & Li, 2001) and the minimax concave penalty (Zhang, 2010) can also be employed in the group penalization scheme.

The group penalization of pooled parameters of the same covariates is not appropriate for Example 2, nor is it applicable to combinations of Examples 1 and 2. A joint model is needed for the multiple responses across the experiments. If such a model is difficult to specify, pooling all marginal likelihoods is appropriate for the examples discussed above. However, to the best of our knowledge, the asymptotic properties of group penalized estimation using pseudolikelihood have not been studied in the literature.

If there is one response and one set of predictors, the extended Bayesian information criterion with an appropriate penalty term has been shown to be selection consistent, where the total number of predictors tends to infinity and the number of true predictors is bounded by a constant (Chen & Chen, 2008). Foster & George (1994) proposed a risk inflation criterion for multiple regression. To handle settings where the number of true predictors is unbounded, Zhang & Shen (2010) proposed a corrected risk inflation criterion, and Kim et al. (2012) proposed a generalized information criterion with modified penalty terms. The consistency of both criteria has been established only for the linear regression model. The problem of how to design the penalty term for an information criterion to deal with a varying true model size remains open. We aim to find an appropriate penalty term for the Bayesian information criterion under likelihood settings where the true model size is unbounded. We extend the results to a pseudolikelihood information criterion, thus including both Examples 1 and 2 as well as combinations of them. Model selection consistency requires various assumptions, including uniform boundedness of all predictors, an assumption also made by Kwon & Kim (2012) and Kim et al. (2012).

Pseudolikelihood ratio-type statistics asymptotically follow a weighted chi-squared distribution. This cannot directly provide an upper bound for the tail probability at a given sample size. Non-asymptotic sharp deviation bounds have been computed by Spokoiny & Zhilova (2013) for quadratic forms based on their exact distributions. We use large deviation theory to bound from above the tail probabilities at any given sample size. Our work establishes the consistency of a pseudolikelihood information criterion for divergent true model size.

2. PSEUDOLIKELIHOOD FORMULATION OF DATA INTEGRATION

Consider a setting with predictors $M_1, \ldots, M_P$ contributing to K different experiments. The objective is to integrate the data from all the experiments to perform inference about the effects of the predictors on the responses. Given n independent experimental subjects, the responses from the kth experiment are denoted by $Y_k = (Y_{k1}, \ldots, Y_{kn})$. The parameter vector θ_k consists of $(\theta_{k1}, \ldots, \theta_{kP})$, where θ_{kp} denotes the effect of predictor M_p in experiment k.

Table 1. *Multiple experiments and their parameters: the predictors are $M_1, \ldots, M_P$, and the parameter for predictor M_p in experiment k is θ_{kp}*

	Experiment 1	$\cdots$	Experiment K
Parameters	$\theta_1 = (\theta_{11}, \ldots, \theta_{1P})^{\mathrm{T}}$	$\cdots$	$\theta_K = (\theta_{K1}, \ldots, \theta_{KP})^{\mathrm{T}}$
Densities	$f_1(Y_1; \theta_1)$	$\cdots$	$f_K(Y_K; \theta_K)$
Observation 1	Y_{11}	$\cdots$	Y_{K1}
$\vdots$	$\vdots$		$\vdots$
Observation n	Y_{1n}	$\cdots$	Y_{Kn}
	$Y_1 = (Y_{11}, \ldots, Y_{1n})^{\mathrm{T}}$	$\cdots$	$Y_K = (Y_{K1}, \ldots, Y_{Kn})^{\mathrm{T}}$

Data from the kth experiment have likelihood function $L_k(\theta_k; Y_k) = \prod_{i=1}^{n} f_k(Y_{ki}; \theta_k)$, where f_k denotes the density function. Let $\theta = (\theta_1, \ldots, \theta_k)$. The parameters associated with predictor M_p across different experiments are $\theta^{(p)} = (\theta_{1p}, \ldots, \theta_{Kp})$. The $Y_{(i)} = (Y_{1i}, \ldots, Y_{Ki})$ may be dependent in a way that is hard to specify. Table 1 illustrates the set-up for data integration when all K measurements of $Y_{(i)}$ $(i = 1, \ldots, n)$ are observed. If some of the $Y_{(i)}$ are incomplete, an indicator Z_{ki} can be introduced. If Y_{ki} is observed then $Z_{ki} = 1$; otherwise $Z_{ki} = 0$. In order to integrate the experiments, we propose to describe the overall data using a working-independence pseudo-loglikelihood

$$\ell_I(\theta) = \sum_{k=1}^{K} w_k \ell_k(\theta_k; Y_k) = \sum_{k=1}^{K} w_k \sum_{i=1}^{n} Z_{ki} \log f_k(Y_{ki}; \theta_k),$$

with positive weights w_k $(k = 1, \ldots, K)$. This formulation is similar to composite likelihood (Lindsay, 1988; Cox & Reid, 2004; Varin, 2008), which combines marginal densities from a multivariate distribution. Pseudolikelihood estimation and inference with regard to θ follow standard theory (White, 1982; Lindsay, 1988; Cox & Reid, 2004; Varin, 2008; Ribatet et al., 2012). The maximum pseudolikelihood estimate is denoted by $\hat{\theta}_I = \arg\max_\theta \ell_I(\theta)$, and it is consistent under regularity conditions. The asymptotic covariance matrix of the maximum pseudolikelihood estimator is given by the inverse of the Godambe information matrix $G(\theta) = H(\theta)^{\mathrm{T}} V^{-1}(\theta) H(\theta)$, where $H(\theta) = E\{-\partial^2 \ell_I(\theta)/(\partial\theta\partial\theta^{\mathrm{T}})\}$ and $V(\theta) = \mathrm{cov}\{\partial \ell_I(\theta)/\partial\theta\}$ (Godambe, 1960). For inference about θ, pseudolikelihood ratio statistics and Wald-type statistics can be formed. In the data integration set-up, uniform weights can be assigned to each likelihood. If some experiments are of better quality than others, one could assign them higher weights. In theory, optimal weights can be constructed by projecting the full likelihood score function onto the linear space of the composite score functions. However, such optimal weights are challenging to obtain (Lindsay et al., 2011). Some practical strategies for choosing weights based on the data structure are given in Varin & Vidoni (2006) and Joe & Lee (2009).

3. Feature selection

Given multiple experiments with high-dimensional predictors, one can perform penalized estimation to select predictors whose parameters are nonzero. If $\theta^{(p)}$ is zero, all corresponding subparameters θ_{kp} $(k = 1, \ldots, K)$ are zero simultaneously. Otherwise, at least one of the parameters θ_{kp} is nonzero. Selecting important features is equivalent to selecting a group of parameters.

We define the overall strength of the predictor M_p as a summarization of the effect sizes in $\theta^{(p)}$, represented by the L_2-norm of $\theta^{(p)}$. We consider the overall objective function

$$Q(\theta) = \ell_I(\theta) - n\sum_{p=1}^{p_n} \Omega_{\lambda_n}(\|\theta^{(p)}\|),$$

with p_n denoting the total number of predictors, Ω_{λ_n} being the penalty function, and $\|\theta^{(p)}\| = (\sum_{k=1}^{K} \theta_{kp}^2)^{1/2}$ denoting the L_2-norm.

As mentioned previously, standard group selection of variables is applicable to Example 1 but not Example 2. Yuan & Lin (2006) considered group selection and proposed the group lasso. Meier et al. (2008) showed that the group lasso for logistic regression yields sparse estimates which are globally consistent in terms of estimation error. Nardi & Rinaldo (2008), Bach (2008) and Zhao et al. (2009) proved selection consistency of the group lasso under regularity conditions. While the group lasso possesses excellent prediction and estimation properties, its variable selection consistency depends on a so-called irrepresentability condition, which requires low correlations between significant and insignificant predictors. This condition is difficult to satisfy when $p_n \gg n$ (Huang et al., 2012). The group lasso tends to overshrink large parameters, because the rate of penalization does not change with the size of the parameters, yielding biased estimates of large parameters (Fan & Li, 2001). Besides the lasso penalty (Tibshirani, 1996), many other types of penalty functions have been proposed, including the smoothly clipped absolute deviation penalty (Fan & Li, 2001) and the minimax concave penalty (Zhang, 2010). These two penalties can achieve both selection consistency and asymptotic unbiasedness. This was extended to the group smoothly clipped absolute deviation penalty and the group minimax concave penalty by Wang et al. (2008), Huang et al. (2012) and Guo et al. (2015). However, no existing work deals with grouped penalization of a pseudolikelihood, which is required for Example 2. In this paper, we focus on the grouped smoothly clipped absolute deviation penalty for pseudolikelihood and establish its properties in high-dimensional models.

The smoothly clipped absolute deviation penalty function satisfies $\Omega_\lambda(0) = 0$, and, with $\theta \geqslant 0$, its first derivative is

$$\Omega'_\lambda(\theta) = \lambda\left\{I(\theta \leqslant \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\right\},$$

where a is a constant that is usually set to 3·7 (Fan & Li, 2001) and $(t)_+ = tI(t > 0)$ is the hinge loss function.

Let $p_n \to \infty$ as $n \to \infty$. We assume that $\|\theta^{(p)}\| > 0$ for $p = 1, \ldots, q_n$ and $\|\theta^{(p)}\| = 0$ for $p = q_n+1, \ldots, p_n$. Define the collections of nonzero and zero parameters as $\theta_a = (\theta^{(1)}, \ldots, \theta^{(q_n)})$ and $\theta_b = (\theta^{(q_n+1)}, \ldots, \theta^{(p_n)})$, respectively. We make the following assumptions.

Assumption 1. The pseudo-loglikelihood admits third derivatives for almost all Y and for all $\theta \in \mathcal{B}$, where the open set $\mathcal{B} \subset \Theta$ contains the true θ^*. The third partial derivatives are bounded by square-integrable functions $W_{jlm}(Y_{(i)})$, so that $|\partial^3 \ell_I(\theta; Y_{(i)})/(\partial\theta_j \partial\theta_l \partial\theta_m)| \leqslant W_{jlm}(Y_{(i)})$ for $\theta \in \mathcal{B}$ and all $j, l, m \in \{vw : v = 1, \ldots, K;\ w = 1, \ldots, p_n\}$, where $E_{\theta^*}\{W_{jlm}(Y_{(i)})\}^{2\kappa} \leqslant M_4$ for an integer $\kappa \geqslant 1$.

Assumption 2. The parameter space $\theta \in \Theta$ is a closed set. Each density $f_k(Y_k; \theta_k)$ is a measurable function of Y_k for any θ_k, and is distinct for different values of θ_k. Let θ^* denote the true

value of θ. We assume that $E_{\theta^*}\{\partial \log f_k(Y_{ki};\theta_k)/\partial\theta_{kj}\} = 0$ and

$$E_{\theta^*}\left\{\frac{\partial^2 \log f_k(Y_{ki};\theta)}{\partial\theta_{kj}\partial\theta_{kl}}\right\} = -E_{\theta^*}\left\{\frac{\partial \log f_k(Y_{ki};\theta_k)}{\partial\theta_{kj}}\,\frac{\partial \log f_k(Y_{ki};\theta_k)}{\partial\theta_{kl}}\right\},$$

for $j, l = 1, \ldots, p_n$ and $k = 1, \ldots, K$.

Assumption 3. Let the submatrices of $H(\theta^*)$ and $V(\theta^*)$ with respect to the parameters in θ_a be denoted by $H^{(1)}(\theta^*)$ and $V^{(1)}(\theta^*)$. Assume that $0 < \lambda_{\min}\{H^{(1)}(\theta^*)\} < \lambda_{\max}\{H^{(1)}(\theta^*)\} < \infty$ and $0 < \lambda_{\min}\{V^{(1)}(\theta^*)\} < \lambda_{\max}\{V^{(1)}(\theta^*)\} < \infty$, where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalues.

Assumption 1–3 are standard in likelihood theory and are analogous to those used in Xu & Reid (2011) and Kwon & Kim (2012).

Assumption 4. Assume θ^* is interior to the parameter space Θ. There exists an integer $\kappa \geqslant 1$ such that for constants (M_1, M_2, M_3),

$$\begin{aligned}
&E_{\theta^*}\{\partial \log f_k(Y_{ki};\theta_k)/\partial\theta_{kj}\}^{2\kappa} \leqslant M_1,\\
&E_{\theta^*}\{\partial^2 \log f_k(Y_{ki};\theta_k)/(\partial\theta_{kj}\partial\theta_{kl})\}^{2\kappa} \leqslant M_2,\\
&E_{\theta^*}[\{\partial \log f_k(Y_{ki};\theta_k)/\partial\theta_{kj}\}\{\partial \log f_k(Y_{ki};\theta_k)/\partial\theta_{kl}\}]^{2\kappa} \leqslant M_3,
\end{aligned}$$

for $j, l = 1, \ldots, p_n$ and $k = 1, \ldots, K$.

Assumption 4 specifies the boundedness of moments of order 2κ for the loglikelihood derivatives, and is used to bound certain tail probabilities. For example, if the density is binomial and $\operatorname{logit}\operatorname{pr}(Y_{ki} = 1 \mid \theta_k) = X_{ki}^{\mathrm{T}}\theta_k$, where $X_{ki} = (X_{ki1}, \ldots, X_{kip_n})^{\mathrm{T}}$ are regression covariates, then

$$\partial \log f_k(Y_{ki};\theta_k)/\partial\theta_k = \big[Y_{ki} - \exp(X_{ki}^{\mathrm{T}}\theta_k)/\{1 + \exp(X_{ki}^{\mathrm{T}}\theta)\}\big]X_{ki}. \tag{1}$$

If the regression covariates are uniformly bounded in absolute value by a constant b, then

$$\max_{1\leqslant j\leqslant p_n} E_{\theta^*}\{\partial \log f_k(Y_{ki};\theta_k)/\partial\theta_{kj}\}^{2\kappa} \leqslant \max_{1\leqslant j\leqslant p_n} X_{kij}^{2\kappa} \leqslant b^{2\kappa}.$$

Similarly it can be verified that in generalized linear models, other densities from exponential families satisfy this assumption if the absolute regression covariates are uniformly bounded.

Assumption 5. There exist constants c_1 and c_2 satisfying $0 < 5c_1 < c_2 < 1$, $q_n = o(n^{c_1})$ and $\min_{1\leqslant j\leqslant q_n} n^{(1-c_2)/2}\|\theta^{*(j)}\| \geqslant M_5$.

Assumption 5 specifies the rate at which q_n grows with respect to n and the rate at which the size of the nonzero predictors can approach zero. This means that the proportion of true predictors has to be less than one-fifth the sample size, whereas the number of predictors p_n can exceed n.

Define the oracle estimate $\hat\theta$ to be any local maximizer of the pseudo-loglikelihood $\ell_I(\theta)$ subject to $\|\hat\theta^{(j)}\| = 0$ for $j > q_n$ and $\|\hat\theta - \theta^*\| = O_{\mathrm{p}}\{(q_n/n)^{1/2}\}$. Under Assumptions 1–5 it can be established that such an oracle estimator exists (Fan & Peng, 2004, Theorem 1). Because the penalty function is singular at the origin, we consider the subderivatives of the objective function. The subdifferential of a function is a set-valued mapping and is a generalized derivative

for nondifferentiable functions. Taking the subderivative of $Q(\theta)$ in (4) with respect to the jth grouped parameters $\theta^{(j)}$, we get

$$\frac{\partial Q(\theta)}{\partial \theta^{(j)}} = \begin{cases} \partial \ell_I(\theta)/\partial \theta^{(j)} - n\lambda_n \operatorname{Sign}(\theta^{(j)}), & \|\theta^{(j)}\| \leqslant \lambda_n, \\ \partial \ell_I(\theta)/\partial \theta^{(j)} - n \operatorname{Sign}(\theta^{(j)})\{a\lambda_n - \|\theta^{(j)}\|\}/(a-1), & \lambda_n < \|\theta^{(j)}\| < a\lambda_n, \\ \partial \ell_I(\theta)/\partial \theta^{(j)}, & a\lambda_n \leqslant \|\theta^{(j)}\|, \end{cases} \tag{2}$$

with $\operatorname{Sign}(\cdot)$ denoting a set-valued map for a real vector. Let 0 denote the vector of zeros. When $u \neq 0$, $\operatorname{Sign}(u)$ returns $u/\|u\|$; and for $u = 0$, $\operatorname{Sign}(u)$ returns the set of all possible vectors ω such that $\|\omega\| \leqslant 1$.

THEOREM 1. *Let $S(\lambda_n)$ denote the set of solutions to the subdifferential equation $\partial Q(\theta)/\partial\theta = 0$. Under Assumptions 1–5, $\operatorname{pr}\{\hat{\theta} \in S(\lambda_n)\} \to 1$ provided that $\lambda_n = o\{n^{-(1-c_2+c_1/2)}\}$ and $p_n/(n^{1/2}\lambda_n)^{2\kappa} \to 0$ as $n \to \infty$.*

We emphasize that p_n may be much larger than n, provided that κ defined in Assumption 4 is sufficiently large. If the first, second and third derivatives of the pseudo-loglikelihood have exponentially decaying tails, Theorem 1 holds when $p_n = O\{\exp(n^{c_3})\}$ for some constant $c_3 > 0$ (Kwon & Kim, 2012).

THEOREM 2. *With probability tending to 1, as $n \to \infty$, the root-(n/q_n)-consistent oracle estimator $\hat{\theta} = (\hat{\theta}_a, \hat{\theta}_b)$ in Theorem 1 satisfies*

$$n^{1/2} A_n \{V^{(1)}(\theta^*)\}^{-1/2} H^{(1)}(\theta^*)(\hat{\theta}_a - \theta_a^*) \to N(0, G),$$

where $V^{(1)}(\theta^*)$ and $H^{(1)}(\theta^*)$ are the submatrices of $V(\theta^*)$ and $H(\theta^*)$ with respect to θ_a, $\{V^{(1)}(\theta^*)\}^{1/2}$ is the symmetric square root of $V^{(1)}(\theta^*)$, and A_n is a $m \times q_n^*$ matrix such that $A_n A_n^{\mathrm{T}} \to G$, with G being a $m \times m$ nonnegative-definite symmetric matrix and $q_n^* = K \times q_n$.

In the literature, group penalization has been studied only in true likelihood settings. We establish the oracle property of group penalization in the pseudolikelihood setting. Our results show that group penalization using the smoothly clipped absolute deviation penalty is asymptotically model-selection consistent even when the marginal likelihoods are correlated. For the group lasso to be model-selection consistent, the irrepresentability condition (Meinshausen & Bühlmann, 2006; Zhao & Yu, 2006; Zou, 2006; Bach, 2008) is required. Thus the capacity for the group lasso to be selection consistent is constrained regardless of the strength of the model signals (Fan & Lv, 2010). In contrast, the group smoothly clipped absolute deviation penalty requires less stringent conditions. Provided the coefficient sizes of the nonzero parameters are sufficiently far from zero at the rate specified in Assumption 5, the oracle property of the group smoothly clipped absolute deviation penalty can be established for a local maximizer of the penalized pseudo-loglikelihood. For such a local maximizer to be the unique local maximizer in a certain restricted model space requires the sparse Riesz condition (Zhang, 2010), which is similar to but less stringent than the irrepresentability condition (Kwon & Kim, 2012).

4. PSEUDOLIKELIHOOD INFORMATION CRITERION

Although different data sources have various densities and parameters, in our context they share the same set of predictors. Aggregating different information criteria can boost the power

to select the correct set of predictors. Given all competing models, consistent model selection identifies the smallest correct model with probability tending to 1. Let s be a subset of $(1, \ldots, p_n)$. The model with $\theta^{(p)} = 0$ for all $p \notin s$ is called model s. The sets of underfitting models and overfitting models are denoted by S_- and S_+, respectively. Assume that the largest model size in model space $s \in \mathcal{S}$ is s_n, where $q_n \leqslant s_n \leqslant p_n$.

We propose to aggregate the information in a linear manner. Our proposed pseudolikelihood information criterion is

$$\text{pseu-BIC}(s) = -2\ell_I(\hat{\theta}_I; Y) + d_s^* \gamma_n, \tag{3}$$

where d_s^* is a measure of model complexity and γ_n is a sequence of penalties on the complexity of the model. In (3), the first term is the pseudo-loglikelihood, which reflects the goodness-of-fit for model s jointly assessed among multiple data sources, while the second term is the penalty for model complexity, which enforces sparsity on any model selected.

Let θ_T^* denote the true value of the parameter under the true model T. Under model s, the parameter space is denoted by Θ_s. Define $\theta_s^* = \arg\max_{\theta \in \Theta_s} E_{\theta_T^*}\{\ell_I(\theta)\}$, under the assumption that such a maximizer is unique in the interior of Θ_s. The effective degrees of freedom d_s^* is $\text{tr}\{H_s^{-1}(\theta_s^*) V_s(\theta_s^*)\}$, where $H_s(\theta_s^*)$ and $V_s(\theta_s^*)$ are computed under model s. The term d_s^* has been used to measure model complexity in many pseudolikelihood settings (Varin & Vidoni, 2005).

Most consistency results for model selection criteria have been established for a bounded model T (Chen & Chen, 2008; Gao & Song, 2010) or a divergent true model for linear regression (Zhang & Shen, 2010; Kim et al., 2012). The results were proved based on the exponential decay rate of chi-squared statistics. The exponential decay rate is essential for overall selection consistency, because there are exponentially many $p_n^{s_n}$ competing models. The Bonferroni inequality gives an upper bound on the overall selection error, which is the sum of all the tail probabilities. If the penalty term γ_n is chosen so that the tail probabilities are exponentially small, then the overall selection error will converge to zero.

Unlike in linear regression, pseudolikelihood-type statistics asymptotically follow a weighted chi-squared distribution. It is difficult to bound the tail probability at a given sample size n using the limiting distribution, so we instead obtain the tail probability based on the exact distributions. Our approach consists of two steps: first, showing that differences in pseudo-loglikelihoods between two competing models s and T can be approximated by quadratic forms and that the approximation errors are uniformly bounded across the model space; and second, based on the quadratic forms, applying a large deviation result (Spokoiny & Zhilova, 2013) to quantify the penalty γ_n so that the tail probabilities are exponentially small.

Let ψ denote a random vector, B a matrix, and $\|B\psi\|^2$ a quadratic form. Large deviation results for quadratic forms $\|B\psi\|^2$ were established by Spokoiny & Zhilova (2013) under the exponential moment condition that for $\|t\| \leqslant g$,

$$\log E\{\exp(t^{\mathrm{T}}\psi)\} \leqslant \|t\|^2/2.$$

Here g is a positive constant which differs between Gaussian and non-Gaussian-type deviation bounds. We first prove that such an exponential moment condition can be satisfied asymptotically by taking sample averages, if the original random variables satisfy a cumulant boundedness condition, defined below.

DEFINITION 1. *For a random vector Z of dimension m, let $g(t)$ denote its cumulant generating function, where t is an m-dimensional real vector. Then Z satisfies the cumulant boundedness condition if the first two derivatives of its cumulant generating function satisfy $|\partial g(0)/\partial t_j| \leqslant C_1$*

and $|\partial^2 g(0)/(\partial t_j \partial t_k)| \leqslant C_2$ *and if there exists a constant* δ *such that when* $\|t\| \leqslant \delta$, *the absolute values of all the third derivatives of its cumulant generating function satisfy* $|\partial^3 g(t)/(\partial t_j \partial t_k \partial t_l)| \leqslant C_3$.

LEMMA 1. *Let* $Z_1, \ldots, Z_n$ *be independently distributed random vectors of dimension* m *with zero mean and identity covariance matrices, and let* $\eta = n^{-1/2} \sum_i Z_i$. *If each random vector* Z_i *satisfies the cumulant boundedness condition with the same bounds and* $s_n^4 \log p_n = o(n)$, *then* $\log E\{\exp(t^{\mathrm{T}}\eta)\} \leqslant a^2\|t\|^2/2$ *for* $\|t\| < (s_n^2 \log p_n)^{1/2}$ *and some constant* $a^2 > 1$, *when* n *is sufficiently large.*

This implies that if the conditions in Definition 1 hold, we can apply large deviation results to the pseudolikelihood ratio-type statistics arising in our analysis. Next, we assume the cumulant boundedness conditions for the derivatives of the pseudo-loglikelihood. We also make assumptions about the distances between the true null model and the competing models.

Assumption 6. All the pseudo-loglikelihoods and their first and second derivatives, $\ell_I(\theta_s^*; Y_{(i)})$, $\ell_I^{(1)}(\theta_s^*; Y_{(i)})$ and $\ell_I^{(2)}(\theta_s^*; Y_{(i)})$, satisfy the cumulant boundedness condition in Definition 1 uniformly for all models $s \in \mathcal{S}$. Also, there exists a neighbourhood $\|\theta_s - \theta_s^*\| \leqslant \delta$ such that all the third derivatives of the pseudo-loglikelihoods $\ell_I^{(3)}(\theta_s; Y_{(i)})$ in that neighbourhood satisfy the cumulant boundedness condition in Definition 1 uniformly and $H_s(\theta_s)$ and $V_s(\theta_s)$ have eigenvalues bounded away from zero and infinity uniformly.

Consider generalized linear models with densities from an exponential family. Assume that the link function is three-times continuously differentiable, all the absolute values of the covariates are uniformly bounded and the linear predictors are bounded. Then Assumption 6 is satisfied. For example, if the density is binomial and the canonical link is used, the boundedness of $X_{ki}^{\mathrm{T}}\theta_k$ ensures that $\mu = \exp(X_{ki}^{\mathrm{T}}\theta_k)/\{1 + \exp(X_{ki}^{\mathrm{T}}\theta_k)\}$ is bounded away from 0 and 1. Let $\partial \log f_k(Y_{ki}; \theta_k)/\partial\theta_k$ be formulated as in (1). Then the third derivative of its cumulant generating function $g(t)$ is bounded by $(3/4)\max_{1 \leqslant j \leqslant p_n} |X_{kij}|^3 \max\{1/\mu^3, 1/(1-\mu)^3\} < \infty$.

Assumption 7. Assume that $s_n^4 \log p_n = o(n)$. Define the pseudo-Kullback–Leibler distance between the true model T and the competing model s as $E_{\theta_T}\{\ell_I(\theta_T; Y_{(i)}) - \ell_I(\theta_s^*; Y_{(i)})\}$. Assume that $\liminf_n \min_{s \in S_-} n^{1/2} E_{\theta_T}\{\ell_I(\theta_T; Y_{(i)}) - \ell_I(\theta_s^*; Y_{(i)})\}/(s_n \log p_n)^{1/2} = \infty$.

This assumption regarding the identifiability of the underlying true model allows the pseudo-Kullback–Leibler distance between the true model and the competing models to tend to zero at a certain rate. Similar identifiability conditions were assumed in Chen & Chen (2008) and Fan & Lv (2011). For example, if the limiting minimum distance is a constant, then the assumption is easily satisfied. For nontrivial cases, we allow the minimum distance to approach zero provided that it converges to zero more slowly than $(s_n \log p_n/n)^{1/2}$.

Next we introduce some notation. For any overfitting model s, define a matrix $D_s = (I_{d_T}, 0_{d_T, d_s - d_T})$, where I_{d_T} is the identity matrix of dimension $d_T \times d_T$ and $0_{d_T, d_s - d_T}$ denotes the matrix of zeros of dimension $d_T \times (d_s - d_T)$. For every model s, let the score vector be denoted by $U_n(\theta_s; Y) = \partial \ell_I(\theta_s, Y)/\partial\theta_s$, and construct the quadratic form $Q_s = n^{-1} U_n(\theta_s^*)^{\mathrm{T}} H_s(\theta_s^*)^{-1} U_n(\theta_s^*)$. According to Lemmas A1 and A2 in the Appendix, $2\{\ell_I(\hat{\theta}_s) - \ell_I(\hat{\theta}_T)\} = (Q_s - Q_T)\{1 + o_{\mathrm{p}}(1)\} = Q_{s/T}\{1 + o_{\mathrm{p}}(1)\}$ with $Q_{s/T} = U_s(\theta_s^*)^{\mathrm{T}} M_{s/T} U_s(\theta_s^*)$, where $M_{s/T}$ denotes the difference matrix $H_s(\theta_s^*)^{-1} - D_s^{\mathrm{T}} H_T^{-1}(\theta_T^*) D_s$. Define $B_s = V_s^{1/2}(\theta_s^*) M_{s/T} V_s^{1/2}(\theta_s^*)$. It can be shown that $\mathrm{tr}(B_s) = d_s^* - d_T^*$. Let $\tau = \lambda_{\max}(B_s)$,

$\bar{\tau} = \mathrm{tr}(B_s)/(d_s - d_T)$ and $\omega = \max_{s \in \mathcal{S}} \tau/\bar{\tau}$. For the true loglikelihood, $\omega = 1$. We now establish a consistency result for the pseudolikelihood information criterion for unbounded true model size.

THEOREM 3. *Let* $\gamma_n = 6\omega(1+\gamma)\log p_n$ *for some* $\gamma > 0$, *or* $\gamma_n = 6\omega(\log p_n + \log\log p_n)$. *Under Assumptions 1–7, as* $n \to \infty$,

$$\mathrm{pr}\Big\{\min_{s \in \mathcal{S}} \text{pseu-BIC}(s) > \text{pseu-BIC}(T)\Big\} \to 1.$$

Theorem 3 demonstrates that, with an appropriate penalty term, the BIC-type information criterion based on compounded marginal likelihoods from different sources can be selection consistent, even if the underlying true model size tends to infinity. This result includes the usual BIC based on the true likelihood as a special case with $\omega = 1$ and $\gamma_n = 6(1+\gamma)\log p_n$. Comparing this to the result of Chen & Chen (2008), who proved the consistency of extended BIC with bounded true model size, Theorem 3 establishes the selection consistency of a BIC-type information criterion with unbounded true model size. In § 5 we use the proposed pseudolikelihood information criterion to select the optimal tuning parameter for group penalization.

5. SIMULATIONS

5·1. *Continuous responses*

For our first simulation we generated four different types of experiments, i.e., $K = 4$, each with a continuous response Y_{ki} and associated covariates $X_{ki} = (x_{ki1}, \ldots, x_{kip_n})$. We took the sample size to be $n = 500$ or 1000, and took the number of covariates to be $p_n = 200$ or 1000. For different experiments, the regression covariates were different. The number of true covariates was $q_n = 50$. For $j = 1, \ldots, q_n$, θ_{kj} was drawn from the uniform distribution on (0·05, 0·5), whereas $\theta_{kj} = 0$ for $j = q_n + 1, \ldots, p_n$. The covariates X_{ki} were partitioned into independent blocks of 50 covariates, and within each block the 50 covariates were simulated from the multivariate normal distribution with variances equal to 1 and all off-diagonal covariances equal to 0·2. For each experiment, the mean parameter is $\mu_{ki} = X_{ki}^{\mathrm{T}}\theta_k$. We simulated Y_i from a multivariate normal distribution with mean $\mu_i = (\mu_{1i}, \ldots, \mu_{Ki})$ and covariance matrix Σ. The covariance matrix was compound symmetric with unit variances and off-diagonal covariances 0·7.

We used the group smoothly clipped absolute deviation penalty function to perform feature selection and used the pseudolikelihood information criterion to select the tuning parameters. For group penalized estimation, we used the group descent algorithm proposed by Breheny & Huang (2015). With regard to the penalty term, Theorem 3 provides a theoretical value of $6\omega(1+\gamma)d_s^* \log p_n$. Here the effective degrees of freedom is $\hat{d}_s^* = \mathrm{tr}(\hat{H}_s^{-1}\hat{V}_s)$, where $\hat{H}_s$ is the observed Hessian matrix and $\hat{V}_s$ is the sample covariance matrix of the composite scores. We set the penalty term to be $c\hat{d}_s^* \log p_n$, where c is constant. This penalty term therefore has the same asymptotic order as the theoretical penalty term. We set $c = 1$ or $c = 6$ and examined how the sensitivity and selectivity of our method changed. Table 2 reports the positive selection rates and false discovery rates of our data integration method and the single-experiment analysis based on the first experiment only. When c changes from 1 to 6, our method's positive selection rate and false discovery rate decrease slightly. A large improvement in its performance is observed compared with single-experiment analysis. For example, when $c = 1$, $n = 500$ and $p_n = 1000$, the positive selection rate and false discovery rate of the data integration method are 1·00 and 0·02, whereas those of single-experiment analysis are 0·81 and 0·35.

Table 2. *Positive selection rates* (%) *and false discovery rates* (%) *of the data integration method compared with single-experiment analysis for multivariate normal responses as described in* § 5·1*; the reported numbers are average rates obtained from* 100 *simulated datasets*

		$c = 1$				$c = 6$			
p	n	DI		SA		DI		SA	
		PSR	FDR	PSR	FDR	PSR	FDR	PSR	FDR
200	500	100	2	91	28	99	0	73	3
	STD	1	2	5	8	1	0	13	4
200	1000	100	0	96	27	100	0	90	4
	STD	0	1	3	8	1	0	6	3
1000	500	100	7	81	35	99	0	57	2
	STD	1	7	7	10	1	1	13	3
1000	1000	100	0	91	29	100	0	81	4
	STD	0	1	4	8	1	0	7	3

DI, data integration method; SA, single-experiment analysis; PSR, positive selection rate; FDR, false discovery rate; STD, sample standard deviation of PSR and FDR from 100 simulations; c, the free multiplicative constant for the penalty.

5·2. *Continuous responses with correlations between predictors and nonpredictors*

We investigated the performance of the proposed method with the group smoothly clipped absolute deviation penalty and the group lasso penalty when there are correlations between predictors and nonpredictors. This setting violates the strong irrepresentability condition and thus affects the performance of the lasso penalty. We conducted four different experiments. The covariates X_{ki} were partitioned into independent blocks of 200 covariates. The first block contains 50 true predictors and 150 nonpredictors. These 200 covariates were simulated from multivariate normal distributions with unit variances and off-diagonal covariances 0·2 or 0·5. The remaining nonpredictors were simulated from independent normal distributions with unit variances. All other parameter settings are the same as in § 5·1. We chose $n = 1000$ and $p = 1000$. Table 3 shows that in the presence of correlation between predictors and nonpredictors, the group smoothly clipped absolute deviation outperforms the group lasso.

5·3. *Multiple experiments with varying quality of information*

Our third simulation examines the performance of our method when experiments contain different amounts of information. The sizes of the nonzero parameters are different across four different experiments. In the first case, all experiments provide information relating the predictors and the responses. The nonzero parameters θ_{1j}, θ_{2j}, θ_{3j} and θ_{4j} were drawn from uniform distributions on $(0{\cdot}05, 0{\cdot}5)$, $(0{\cdot}05, 0{\cdot}4)$, $(0{\cdot}05, 0{\cdot}3)$ and $(0{\cdot}05, 0{\cdot}15)$. In the second case, the other three experiments provide almost no information, with θ_{2j}, θ_{3j} and θ_{4j} drawn from uniform distributions on $(0, 0{\cdot}05)$ for $j = 1, \ldots, q_n$. All other parameter settings are the same as in § 5·1. We chose $n = 500$ and $p = 1000$. Table 4 shows that if the other three experiments contain information, even if the amount is much smaller than that in the first experiment, combining all four experiments outperforms any single-experiment analysis. However, if the other three experiments contain no information at all, then combining all four experiments provides worse results than the single-experiment analysis on the first experiment.

Table 3. *Positive selection rates* (%)*, false discovery rates* (%) *and estimation errors of the data integration method using group lasso and using the group smoothly clipped absolute deviation penalty in the presence of correlated covariates with* $n = 1000$ *and* $p = 1000$*, as described in* § 5·2*; the reported numbers are average values from* 100 *simulated datasets*

	Lasso			SCAD		
r	PSR	FDR	SSE	PSR	FDR	SSE
0·20	100	1	224	100	0	47
STD	1	1	82	1	1	7
0·50	99	3	464	99	1	282
STD	2	5	106	2	3	118

SCAD, smoothly clipped absolute deviation penalty; PSR, positive selection rate; FDR, false discovery rate; SSE, sum of squared errors of the penalized estimate $\|\hat{\theta} - \theta\|_2^2$, multiplied by 100; r, the correlation between true predictors and false predictors; STD, sample standard deviation of PSR, FDR and SSE computed from 100 simulations.

Table 4. *Positive selection rates* (%) *and false discovery rates* (%) *of the data integration method and single-experiment analysis with four experiments containing different amounts of information, as described in* § 5·3

	DI		SA_1		SA_2		SA_3		SA_4	
Information	PSR	FDR	PSR	FDR	PSR	FDR	PSR	FDR	PSR	FDR
YES	98	6	81	35	76	31	66	24	34	9
STD	2	6	7	10	7	10	9	9	12	9
NO	51	5	81	35	0	0	0	3	0	1
STD	24	6	7	10	1	0	1	16	1	11

DI, data integration method; SA_1–SA_4, single-experiment analysis on four different platforms; PSR, positive selection rate; FDR, false discovery rate; STD, sample standard deviation of PSR and FDR computed from 100 simulations; YES, the 2nd to 4th experiments have information relating the predictors and the responses; NO, the 2nd to 4th experiments have no information relating the predictors and the responses.

5·4. *Mixtures of continuous and binary responses*

Our fourth simulation examines the performance of our method on data with correlated continuous and binary responses and $K = 4$. All experiments share the same set of covariates $X_i = (x_{i1}, \ldots, x_{ip_n})$. We took $n = 1000, 1500$ and $p_n = 200, 1000$. For different experiments, the regression covariates were different. The number of true covariates was $q_n = 50$. For $j = 1, \ldots, q_n$, θ_{kj} was drawn from the uniform distribution on $(0{\cdot}05, 0{\cdot}5)$, whereas $\theta_{kj} = 0$ for $j = q_n + 1, \ldots, p_n$. The covariates X_i were standard normal. For each experiment, the mean parameter is $\mu_{ki} = X_i^{\mathrm{T}}\theta_k$. We simulated Y_i^* from a multivariate normal distribution with mean $\mu_i = (\mu_{1i}, \ldots, \mu_{Ki})$ and covariance matrix Σ. The covariance matrix was compound symmetric with unit variances and off-diagonal covariances 0·7. For the first two experiments, the observed responses were continuous values $Y_{ki} = Y_{ki}^*$ $(k = 1, 2)$; for the third and fourth experiments, the binary data were $Y_{ki} = I(Y_{ki}^* > 0)$ $(k = 3, 4)$. Table 5 shows the performance of the data integration method when $c = 1$ and 6. The results are consistent with Table 2, where the data integration method outperforms the single-experiment analysis based on the first experiment

Table 5. *Positive selection rates* (%) *and false discovery rates* (%) *of the data integration method compared with single-experiment analysis for multivariate mixed binary and continuous responses, in the case where the binary and continuous responses are correlated, as described in* § 5·4*; the reported numbers are average rates obtained from* 100 *simulated datasets*

		$c = 1$				$c = 6$			
p	n	DI		SA		DI		SA	
		PSR	FDR	PSR	FDR	PSR	FDR	PSR	FDR
200	1000	100	1	96	30	93	0	89	5
	STD	1	2	3	8	13	0	6	3
200	1500	100	1	99	29	97	0	94	5
	STD	0	1	2	7	3	0	4	4
1000	1000	99	1	90	34	83	0	81	5
	STD	1	1	5	7	26	0	7	4
1000	1500	100	1	95	32	96	0	88	4
	STD	1	3	4	7	3	0	6	4

c, the free multiplicative constant for the penalty; DI, data integration method; SA, single-experiment analysis; PSR, positive selection rate; FDR, false discovery rate; STD, sample standard deviation of PSR and FDR from 100 simulations.

only. For example, when $n = 1000$, $p_n = 1000$ and $c = 1$, the positive selection rate and false discovery rate of our data integration method are 0·99 and 0·01, respectively, whereas those of the single-experiment analysis are 0·90 and 0·34.

6. DATA ANALYSIS

First we applied our method to Example 1 discussed in § 1. The data consist of two different microarray experiments on breast cancer cells (Wang et al., 2005; Iwamoto et al., 2011). In the first experiment, the gene expression profiles from total RNA were obtained from frozen tumour samples from lymph-node-negative patients who had not received adjuvant systemic treatment. In the second experiment, pretreatment fine-needle aspirations from primary tumours were obtained and RNA was extracted and hybridized to microarrays. Because of the different experimental protocols, the two sets of gene expression profiles are globally different. Both experiments were conducted to study the difference in gene expression profiles between estrogen-receptor-positive and estrogen-receptor-negative patients. The training dataset consists of a total of 170 samples, with 35 samples from the estrogen-receptor-positive patients and 50 samples from the estrogen-receptor-negative patients. In Figs. 1 (a) and (b), the heatmaps of the two experiments are shown. The objective of the analysis is to combine the data from the two experiments and find a common set of candidate genes to classify the estrogen-receptor-positive and estrogen-receptor-negative cases. For each of the experiments, we constructed a logistic model with the two subclasses as the binary responses and the expression levels of all the genes as the covariates. We applied our method and used the group smoothly clipped absolute deviation penalty to penalize the regression coefficients. With increasing penalty size, we obtained a shorter list of genes. Figures 1 (c) and (d) show the selected genes when the candidate list has been reduced to four candidates. The selected top candidates exhibit consistent significant differential behaviour in the two experiments. The logistic models based on the selected four covariates were used to classify the subclasses of a

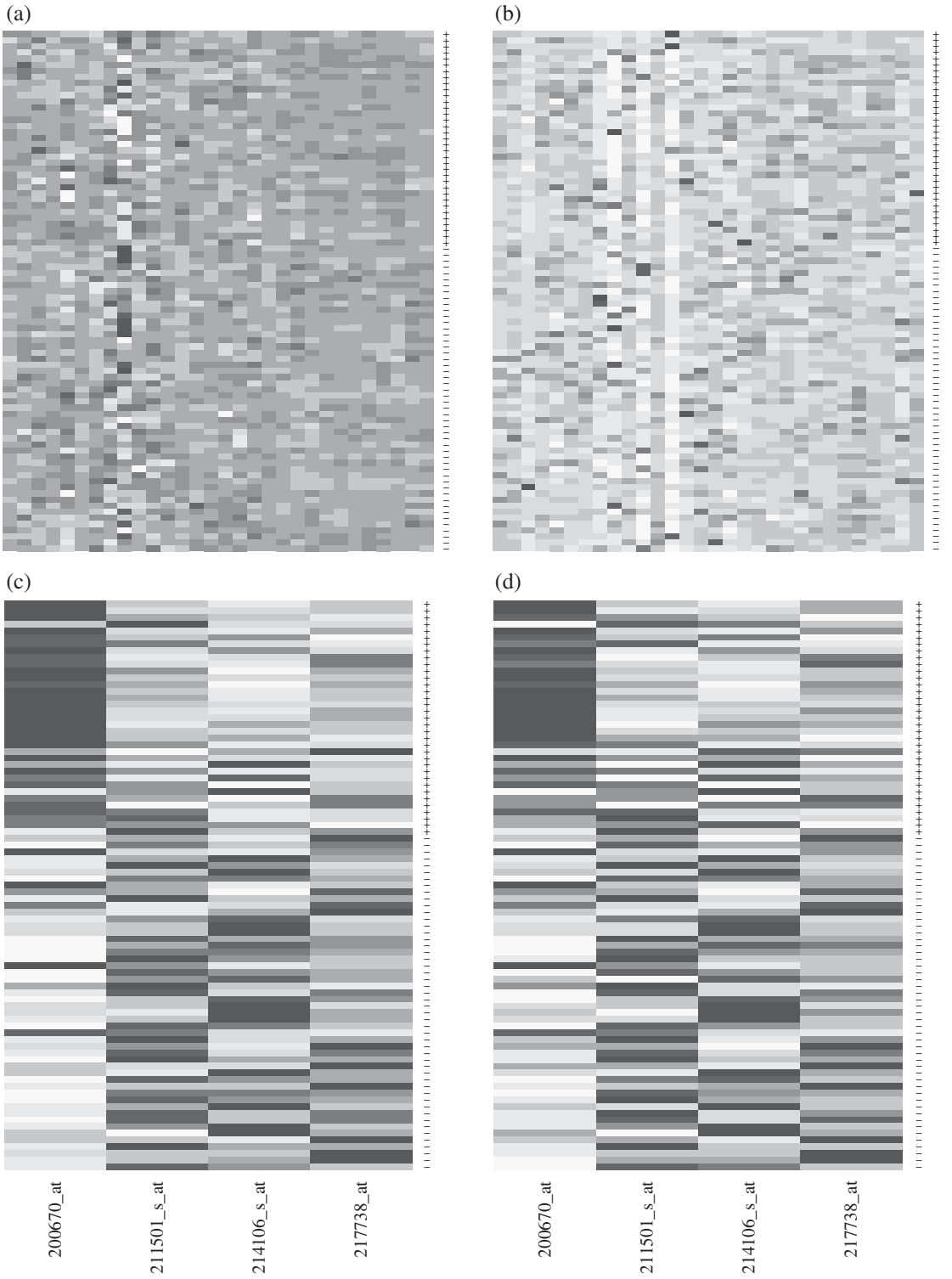

Fig. 1. Panels (a) and (b) show heatmaps of the microarray gene expression profiles from the two experiments in Wang et al. (2005) and Iwamoto et al. (2011), respectively; only the first 30 genes are depicted in the heatmaps. Panels (c) and (d) display the heatmaps of the gene expression levels of the four selected genes from both experiments; + signs denote estrogen-receptor-positive samples and − signs represent estrogen-receptor-negative samples.

Table 6. *Sum of squared prediction errors of selected subset models and the full model for three responses in the financial market indices data*

Model	VIX	S&P 500	Dow Jones
SSPE of submodel by group SCAD	4930·00	2·10	15·27
SSPE of submodel by group lasso	5746·81	2·57	14·81
SSPE of full model	6375·95	1·62	17·41
Total variation in the response	12539·58	165·24	160·83

SSPE, sum of squared prediction errors; SCAD, smoothly clipped absolute deviation.

different validation dataset, which contains 13 samples from the first experiment and 54 samples from the second experiment. Among the 67 validation samples, 16 samples were misclassified. The overall accuracy rate of the classification on the validation data was 76%.

Second, we applied our method to Example 2 discussed in § 1. The dataset consists of financial market indices. We are interested in three indices, the S&P 500 index, the Dow Jones index and the VIX index. The VIX is a measurement of implied volatility of the S&P 500 index and is highly negatively correlated with it. The S&P 500 and Dow Jones are positively correlated. The 46 covariates are the major international equity indices, the North American bond indices, and the major commodities indices. The goal of the analysis is to select a subset of covariates to model the S&P 500, Dow Jones and VIX indices. The training dataset consists of three-year market performances of the S&P 500 index, the Dow Jones index and the VIX index along with the 46 covariates. For each index, the value used in the analysis is log(today's value/yesterday's value) $\times 100$. There are a total of 232 records, with three-day spacing between the values. The values are not autocorrelated at a 5% significance level.

For each response, we constructed a linear regression model based on the same set of covariates. We used both the group lasso penalty and the group smoothly clipped absolute deviation penalty. The subset selected by the group lasso contains 37 covariates, while the subset obtained by the group smoothly clipped absolute deviation penalty contains 34 covariates. The two methods had 31 covariates in common. In order to validate the submodels, we used the model built from the training dataset to perform prediction on a different validation dataset of 232 records. Table 6 shows the sum of squared prediction errors for the submodels selected by the group smoothly clipped absolute deviation penalty and the group lasso penalty and for the full model, together with the total sum of squared variation in the responses. Both selected submodels have small prediction errors compared with the total variation in the responses in the validation dataset. The submodel selected by the group smoothly clipped absolute deviation penalty has smaller prediction errors than that selected by the group lasso in two out of the three responses.

7. DISCUSSION

A missing data problem arises if some predictors have not been measured in some of the experiments. To see the possible difficulties, for simplicity let us consider linear models with observations Y_{ki} of the form $Y_{ki} = \alpha_k + \sum_{p=1}^{P} X_{kpi}\theta_{kp} + \epsilon_{ki}$, where the ϵ_{ki} may be correlated within subject i but are independent of the covariates. This means that $E(Y_{ki} \mid X_{k1i}, \ldots, X_{kPi}) = \alpha_k + \sum_{p=1}^{P} X_{kpi}\theta_{kp}$. Now suppose that for experiment $k = 2$, X_{2pi} for $p = 1$ is missing for all subjects. Then, for experiment $k = 2$, the regression mean is $E(Y_{2i} \mid X_{22i}, \ldots, X_{2Pi}) =$

$\alpha_2 + E(X_{21i} \mid X_{22i}, \ldots, X_{2Pi})\theta_{21} + \sum_{p=2}^{P} X_{2pi}\theta_{2p}$. Unless $E(X_{21i} \mid X_{22i}, \ldots, X_{2Pi})\theta_{21}$ is constant, θ_{21} contaminates all the other θ_{2p} for $p \geqslant 2$. If $E(X_{21i} \mid X_{22i}, \ldots, X_{2Pi})\theta_{21}$ is constant, we can ignore θ_{21} and remove it from $\theta^{(1)}$. Then the length of $\theta^{(1)}$ is less than K. In general, let K_p denote the length of $\theta^{(p)}$, which can be smaller than K. We can consider the objective function

$$Q(\theta) = \ell_I(\theta) - n \sum_{p=1}^{p_n} \Omega_{\lambda_n}\{(K/K_p)^{1/2} \|\theta^{(p)}\|\}, \tag{4}$$

with $(K/K_p)^{1/2}$ adjusting for the different lengths of $\theta^{(p)}$. In this case, our results will hold for criterion (4).

In contrast, if $E(X_{21i} \mid X_{22i}, \ldots, X_{2Pi})$ is not constant and X_{21i} is not missing for all subjects, we could model the covariate missing mechanism under a missingness at random assumption. In Claeskens & Consentino (2008), a missing data AIC-type criterion was proposed with the observed-data likelihood being replaced by the expected complete-data likelihood. Garcia et al. (2010) investigated the smoothly clipped absolute deviation penalty and adaptive lasso and proposed a model selection and estimation procedure for use when there are missing data. Both methods were established for finite p. Further research is needed to consider the large-p scenario in this case.

Acknowledgement

We are grateful to the referees and the associate editor for their insightful comments, and to the editor for his great patience throughout the revision process. Gao's research was supported by the Natural Sciences and Engineering Research Council of Canada. Carroll's research was supported by the National Cancer Institute. Carroll is also Distinguished Professor at the School of Mathematical and Physical Sciences, University of Technology, Sydney, Australia.

Appendix

Proof of Theorem 1. By taking the first derivative of the objective function $Q(\theta)$ with respect to the jth grouped parameters $\theta^{(j)}$, we show that $\hat{\theta}$ satisfies the Karush–Kuhn–Tucker conditions. By the definition of an oracle estimate, for $1 \leqslant j \leqslant q_n$, $\partial \ell_I(\theta)/\partial\theta^{(j)}|_{\hat{\theta}} = 0$. It can be shown that $\mathrm{pr}(\min_{1\leqslant j\leqslant q_n} \|\hat{\theta}^{(j)}\| \geqslant a\lambda_n) \to 1$, because $\min_{1\leqslant j\leqslant q_n} \|\hat{\theta}^{(j)}\| \geqslant \min_{1\leqslant j\leqslant q_n} \|\theta^{*(j)}\| - \max_{1\leqslant j\leqslant q_n} \|\theta^{*(j)} - \hat{\theta}^{(j)}\|$, $\min_{1\leqslant j\leqslant q_n} \|\theta^{*(j)}\| > M_5 n^{-(1-c_2)/2}$, $\max_{1\leqslant j\leqslant q_n} \|\theta^{*(j)} - \hat{\theta}^{(j)}\| = O_{\mathrm{p}}(n^{-(1-c_1)/2})$ and $\lambda_n = o(n^{-(1-c_2+c_1)/2})$. Thus $\hat{\theta}^{(j)}$ belongs to the third case in formula (2) and $\partial Q(\theta)/\partial\theta^{(j)}|_{\hat{\theta}} = 0$.

Let $S_j(\theta) = \partial \ell_I(\theta)/\partial\theta^{(j)}$. For the remaining parameters, we prove that $\mathrm{pr}(\max_{q_n<j\leqslant p_n} \|S_j(\hat{\theta})\| \leqslant n\lambda_n) \to 1$. For each k, $\hat{\theta}_k$ is an oracle estimate for θ_k. Therefore, by formula (A.7) in Kwon & Kim (2012), it can be shown that $\mathrm{pr}\{\max_{q_n<j\leqslant p_n} |\partial \ell_I(\hat{\theta})/\theta_{kj}| > n\lambda_n/K^{1/2}\} \to 0$. Hence $\mathrm{pr}\{\max_{q_n<j\leqslant p_n} \|S_j(\hat{\theta})\| > n\lambda_n\} \leqslant \sum_{k=1}^{K} \mathrm{pr}\{\max_{q_n<j\leqslant p_n} |\partial \ell_I(\hat{\theta})/\theta_{kj}| > n\lambda_n/K^{1/2}\} \to 0$. Therefore $\hat{\theta}^{(j)}$ belongs to the first case in formula (2) and $\partial Q(\theta)/\partial\theta^{(j)}|_{\hat{\theta}} = 0$. □

Proof of Theorem 2. Let $\nabla \ell_I(\theta) = \partial \ell_I(\theta)/\partial\theta$ denote the score vector of the pseudolikelihood. Let ∇_1 denote partial differentiation with respect to θ_a. Let $\nabla^2 \ell_I(\theta)$ denote the matrix of second derivatives $\partial^2 \ell_I(\theta)/(\partial\theta\partial\theta^{\mathrm{T}})$. We expand $\nabla_1 \ell_I(\hat{\theta})$ about θ^*, knowing that $\nabla_1 \ell_I(\hat{\theta}) = 0$, as $\nabla_1 \ell_I(\hat{\theta}) = \nabla_1 \ell_I(\theta^*) + \nabla_1^2 \ell_I(\theta^*)(\hat{\theta}_a - \theta_a^*) + R$, where R is a $q_n^* \times 1$ vector of remainder terms with $R_i =$

$(1/2)\sum_{j,l}\partial^3\ell_I(\theta)/(\partial\theta_i\partial\theta_j\partial\theta_l)\big|_{\tilde\theta}(\hat\theta_j-\theta_j^*)(\hat\theta_k-\theta_k^*)$ for $i,j,k\in\{st : s=1,\ldots,K;\ t=1,\ldots,q_n\}$ and $\tilde\theta$ between θ^* and $\hat\theta$. This leads to $n^{-1}\nabla_1^2\ell_I(\theta^*)(\hat\theta_a-\theta_a^*) = -n^{-1}\{\nabla_1\ell_I(\theta^*)+R\}$. By Assumption 1, $|\partial^3\ell_I(\theta^*)/(\partial\theta_i\partial\theta_j\partial\theta_k)|\leqslant\sum_{l=1}^n W_{ijk}(Y_{(l)})$. Thus

$$\begin{aligned}|R_i/n| &\leqslant n^{-1}\sum_l\sum_j\sum_k W_{ijk}(Y_{(l)})(\hat\theta_j-\theta_j^*)(\hat\theta_k-\theta_k^*)\\ &= n^{-1}\sum_l\sum_j\sum_k\big[W_{ijk}(Y_{(l)})-E\{W_{ijk}(Y_{(l)})\}\big](\hat\theta_j-\theta_j^*)(\hat\theta_k-\theta_k^*)\\ &\quad+n^{-1}\sum_l\sum_j\sum_k E\{W_{ijk}(Y_{(l)})\}(\hat\theta_j-\theta_j^*)(\hat\theta_k-\theta_k^*)=I_1+I_2,\end{aligned}$$

where $I_2\leqslant Mq_n\|\tilde\theta_1-\theta_1^*\|^2=O_{\mathrm{p}}(q_n^2/n)$ for some constant M by the Cauchy–Schwarz inequality. Let W_{ijkl}^* denote the centred random variable $W_{ijk}(Y_{(l)})-E\{W_{ijk}(Y_{(l)})\}$. By the Rosenthal inequality, $E\{(\sum_l W_{ijkl}^*)^2\}=O(n)$. Using the Markov inequality, we have $(\sum_l W_{ijkl}^*)^2=O_{\mathrm{p}}(n)$. For I_1, by the Cauchy–Schwarz inequality,

$$\begin{aligned}I_1 &\leqslant n^{-1}\left\{\sum_j(\hat\theta_j-\theta_j^*)^2\right\}^{1/2}\left\{\sum_k(\hat\theta_k-\theta_k^*)^2\right\}^{1/2}\left\{\sum_j\sum_k\left(\sum_l W_{ijkl}^*\right)^2\right\}^{1/2}\\ &= O_{\mathrm{p}}(q_n/n^2)\left\{\sum_j\sum_k\left(\sum_l W_{ijkl}^*\right)^2\right\}^{1/2}=O_{\mathrm{p}}(q_n^2n^{3/2}).\end{aligned}$$

Upon combining these results for I_1 and I_2, $|R_i|=O_{\mathrm{p}}(q_n^2)$. Let A_{nr} denote the rth row of A_n. It then follows that

$$\big|n^{-1/2}A_{nr}\{V^{(1)}(\theta^*)\}^{-1/2}R\big|\leqslant n^{-1/2}\|A_{nr}\|\lambda_{\max}\big[\{V^{(1)}(\theta^*)\}^{-1/2}\big]\|R\|=O_{\mathrm{p}}\{(q_n^5/n)^{1/2}\}=o_{\mathrm{p}}(1).$$

Therefore the vector $n^{-1/2}A_n\{V^{(1)}(\theta^*)\}^{-1/2}R$ converges to zero in probability. By Lemma 8 in Fan & Peng (2004), $\|\{H^{(1)}(\theta^*)+n^{-1}\nabla_1^2(\theta^*)\}(\hat\theta_a-\theta_a^*)\|\leqslant o_{\mathrm{p}}(q_n^{-1})O_{\mathrm{p}}\{(q_n/n)^{1/2}\}$, so

$$\begin{aligned}&\big|n^{1/2}A_{nr}\{V^{(1)}(\theta^*)\}^{-1/2}\big\{H^{(1)}(\theta^*)+n^{-1}\nabla_1^2(\theta^*)\big\}(\hat\theta_a-\theta_a^*)\big|\\ &\quad\leqslant n^{1/2}\|A_{nr}\|\lambda_{\max}\{V^{(1)}(\theta^*)\}^{-1/2}\big\|\big\{H^{(1)}(\theta^*)+n^{-1}\nabla_1^2(\theta^*)\big\}(\hat\theta_a-\theta_a^*)\big\|=o_{\mathrm{p}}(1).\end{aligned}$$

It follows that the vector $n^{1/2}A_n\{V^{(1)}(\theta^*)\}^{-1/2}\{H^{(1)}(\theta^*)+n^{-1}\nabla_1^2(\theta^*)\}(\hat\theta_a-\theta_a^*)$ converges to zero in probability. This means that $n^{1/2}A_n\{V^{(1)}(\theta^*)\}^{-1/2}H^{(1)}(\theta^*)(\hat\theta_a-\theta_a^*)=n^{-1/2}A_n\{V^{(1)}(\theta^*)\}^{-1/2}\nabla_1\ell_I(\theta^*)+o_{\mathrm{p}}(1)$. Next, let $Z_l=n^{-1/2}A_n\{V^{(1)}(\theta^*)\}^{-1/2}\nabla_1\ell_I(\theta^*,Y_{(l)})$. By the argument in the proof of Theorem 2 of Fan & Peng (2004), $\sum_{l=1}^n E\|Z_l\|^2I\{Z_l\geqslant\epsilon\}=o(1)$ and $\lim_n\sum_{l=1}^n\mathrm{cov}(Z_l)=G$. According to the Lindeberg–Feller central limit theorem, this means that $n^{-1/2}A_n\{V^{(1)}(\theta^*)\}^{-1/2}\nabla_1\ell_I(\theta^*)\to N(0,G)$ in distribution, completing the proof. □

Proof of Theorem 3. According to Lemma A3 below, $\max_{s\in S_-}2\{\ell_I(\hat\theta_s)-\ell_I(\theta_s^*)\}=O_{\mathrm{p}}(s_n\log p_n)$. For the true model T, $2\{\ell_I(\hat\theta_T)-\ell_I(\theta_T^*)\}=O_{\mathrm{p}}(1)$. Define $\lambda_{T|s}(Y)=\ell_I(\theta_T^*;Y)-\ell_I(\theta_s^*;Y)$. Based on Lemma A6, $\max_{s\in S_-}\lambda_{T|s}(Y)-E_{\theta_T^*}\{\lambda_{T|s}(Y)\}=O_{\mathrm{p}}\{(ns_n\log p_n)^{1/2}\}$. Therefore, for an underfitting model,

$$\begin{aligned}-2\{\ell_I(\hat\theta_s)-\ell_I(\hat\theta_T)\} &\geqslant -2\Big[\max_{s\in S_-}\{\ell_I(\hat\theta_s)-\ell_I(\theta_s^*)\}\Big]+2\{\ell_I(\hat\theta_T)-\ell_I(\theta_T^*)\}\\ &\quad+2\big[\lambda_{T|s}(Y)-E_{\theta_T^*}\{\lambda_{T|s}(Y)\}\big]+2E_{\theta_T^*}\{\lambda_{T|s}(Y)\}\\ &=O_{\mathrm{p}}(s_n\log p_n)+O_{\mathrm{p}}\{(ns_n\log p_n)^{1/2}\}+2E_{\theta_T^*}\{\lambda_{T|s}(Y)\}.\end{aligned}$$

This means that

$$\min_{s\in S_-} \text{pseu-BIC}(s) - \text{pseu-BIC}(T) \geqslant \min_{s\in S_-} -2\{\ell_I(\hat{\theta}_s) - \ell_I(\hat{\theta}_T)\} + \gamma_n(d_s^* - d_T^*).$$

Because $|\gamma_n(d_s^* - d_T^*)| = O(s_n \log p_n)$, $\liminf_{n\to\infty} \min_{s\in S_-} E_{\theta_T^*}\{\lambda_{T|s}(Y)\}/(ns_n \log p_n)^{1/2} = \infty$ and $s_n^4 \log p_n = o(n)$, we have that $\text{pr}_{\theta_T^*}\{\text{pseu-BIC}(T) < \min_{s\in S_-} \text{pseu-BIC}(s)\} \to 1$.

For an overfitting marginal model s,

$$\begin{aligned}\text{pseu-BIC}(s) - \text{pseu-BIC}(T) &= -2\{\ell_I(\hat{\theta}_s) - \ell_I(\hat{\theta}_T)\} + (d_s^* - d_T^*)\gamma_n \\ &\geqslant -\max_{s\in S+} Q_{s/T} + (d_s^* - d_T^*)\gamma_n + o_{\text{p}}(1).\end{aligned}$$

By Lemma A4, $\text{pr}_{\theta_T^*}\{\max_{s\in S+} Q_{s/T} < (d_s^* - d_T^*)\gamma_n\} \to 1$, completing the proof. □

Proof of Lemma 1. By Taylor expansion, for $\|t\| \leqslant \delta$, the cumulant generating function for Z_i is $g_i(t) = t^{\text{T}}t/2 + (1/6)\sum_{j,k,l=1}^m \partial^3 g_i(t^*)/(\partial t_j \partial t_k \partial t_l)\, t_j t_k t_l$ for some $0 \leqslant \|t^*\| \leqslant \|t\| \leqslant \delta$. Let $\partial^3 \bar{g}(t)/(\partial t_j \partial t_k \partial t_l) = n^{-1}\sum_{i=1}^n \partial^3 g_i(t)/(\partial t_j \partial t_k \partial t_l)$. Because each third-order partial derivative is uniformly bounded, so too is the average third-order partial derivative. For any $\|t\|/n^{1/2} \leqslant \delta$, the moment generating function of $\eta = n^{-1/2}\sum_{i=1}^n Z_i$ is

$$\phi_\eta(t) = \exp\left\{t^{\text{T}}t/2 + (1/6)\sum_{i,j,k=1}^m n^{-1/2} t_i t_j t_k \partial^3 \bar{g}(t^*/n^{1/2})/(\partial t_i \partial t_j \partial t_k)\right\} = \exp\big[(t^{\text{T}}t/2)\{1 + o(1)\}\big].$$

This is due to the fact that $\|t^*\| < \|t\| < (s_n^2 \log p_n)^{1/2}$ and that $|\partial^3 \bar{g}(t^*/n^{1/2})/(\partial t_i \partial t_j \partial t_k)| < C$ as $\|t\|/n^{1/2} \leqslant n^{-1/2}(s_n^2 \log p_n)^{1/2} \to 0$. Therefore $\log E\{\exp(t\eta)\} \leqslant a^2 t^{\text{T}}t/2$ for $\|t\| < (s_n^2 \log p_n)^{1/2}$, for some $a^2 > 1$ and n sufficiently large. □

LEMMA A1. *Under Assumptions 1–7, there exists a solution $\hat{\theta}_s$ to the score equation $U_n(\theta_s; Y) = \partial \ell_I(\theta_s, Y)/\partial \theta_s = 0$ that falls within a $(s_n^2 \log p_n/n)^{1/2}$-neighbourhood of θ_s^* for all $s \in \mathcal{S}$, with probability tending to 1 as $n \to \infty$.*

Proof of Lemma A1. For any unit vector v, let $\theta_s = \theta_s^* + C(s_n^2 \log p_n/n)^{1/2} v$ for some constant C. By Taylor expansion,

$$\ell_I(\theta_s) - \ell_I(\theta_s^*) = C(s_n^2 \log p_n/n)^{1/2} v^{\text{T}} U_n(\theta_s^*) + (1/2)C^2(s_n^2 \log p_n/n) v^{\text{T}} \ell_I^{(2)}(\tilde{\theta}_s) v,$$

where $\tilde{\theta}_s$ is within an η-neighbourhood of θ_s^* and $\ell_I^{(2)} = \partial^2 \ell_I(\theta_s)/\partial \theta_s^2$. By Assumption 6, which says that $E\{-\ell_I^{(2)}(\tilde{\theta}_s)\}$ has eigenvalues uniformly bounded away from zero and infinity, when θ_s is in the η-neighbourhood of θ_s^*, we have $v^{\text{T}} E\{-\ell_I^{(2)}(\tilde{\theta}_s)\} v = O_{\text{p}}(n)$. Using arguments similar to those in the proof of Lemma A6,

$$\max_{s\in\mathcal{S}} \big|\ell_{ij}^{(2)}(\tilde{\theta}_s) - E\{\ell_{ij}^{(2)}(\tilde{\theta}_s)\}\big| = O_{\text{p}}\{(ns_n \log p_n)^{1/2}\} = o_{\text{p}}[E\{\ell_{ij}^{(2)}(\tilde{\theta}_s)\}],$$

where $\ell_{ij}^{(2)}$ denotes the second derivative of ℓ_I with respect to indices i and j, for $i,j \in \{vw : v = 1, \ldots, K;\ w = 1, \ldots, d_s\}$. From Lemma A6, $\max_{s\in\mathcal{S}} \|U_n(\theta_{s,0})\| = (ns_n^2 \log p_n)^{1/2}$. By the Cauchy–Schwarz inequality, $v^{\text{T}} U_n(\theta_{s,0}) \leqslant \|v\|\, \|U_n(\theta_{s,0})\| = O_{\text{p}}\{(ns_n^2 \log p_n)^{1/2}\}$. Combining the results above yields that $\max_{s\in\mathcal{S}}\{\ell_I(\theta_s) - \ell_I(\theta_s^*)\} < 0$ in probability with the constant C chosen sufficiently large. This means that $\text{pr}[\max_{s\in\mathcal{S}}\{\ell_I(\theta_s) - \ell_I(\theta_s^*)\} < 0] \to 1$ as $n \to \infty$. Thus, with probability tending to 1, there exists a solution to the score equation which falls within a $(s_n^2 \log p_n/n)^{1/2}$-neighbourhood of θ_s^* for all $s \in \mathcal{S}$. □

LEMMA A2. *Under Assumptions 1–7, as* $n \to \infty$, $2\{\ell_I(\hat{\theta}_s) - \ell_I(\theta_s^*)\} = Q_s\{1 + o_p(1)\}$ *where* $Q_s = n^{-1}U_n(\theta_s^*)^T\{H_s(\theta_s^*)\}^{-1}U_n(\theta_s^*)$, *and* $o_p(1)$ *holds for all models* $s \in \mathcal{S}$.

Proof of Lemma A2. Consider a competing model s. Let $\ell_r^{(1)}$ denote $\partial\ell_I/\partial\theta_r$, let $\ell_{rt}^{(2)}$ denote $\partial^2\ell_I/(\partial\theta_r\partial\theta_t)$, and let $\ell_{rtu}^{(3)}$ denote $\partial^3\ell_I/(\partial\theta_r\partial\theta_t\partial\theta_u)$ for $r,t,u \in \{vw : v = 1,\ldots,K; w = 1,\ldots,d_s\}$. Let $H_{rt}(\theta_s^*)$ denote the (r,t)th entry of the Hessian matrix. Taylor expansion of $\ell_r^{(1)}(\hat{\theta}_s) = 0$ about θ_s^* gives the system of equations

$$0 = n^{-1}\ell_r^{(1)}(\hat{\theta}_s) = n^{-1}\ell_r^{(1)}(\theta_s^*) + \sum_t n^{-1}\ell_{rt}^{(2)}(\theta_s^*)(\hat{\theta}_s - \theta_s^*)_{[t]} + \sum_{tu}(2n)^{-1}\ell_{rtu}^{(3)}(\tilde{\theta}_s)(\hat{\theta}_s - \theta_s^*)_{[t]}(\hat{\theta}_s - \theta_s^*)_{[u]},$$

for some $\tilde{\theta}_s$ between θ_s^* and $\hat{\theta}_s$.

Here $n^{-1}\sum_t \ell_{rt}^{(2)}(\hat{\theta}_s - \theta_s^*)_{[t]} = \sum_t\{-H_{rt} + (n^{-1}\ell_{rt}^{(2)} + H_{rt})\}(\hat{\theta}_s - \theta_s^*)_{[t]}$, where $\ell_{rt}^{(2)}$ and H_{rt} are evaluated at θ_s^*. By Lemma A6, $\max_{s\in\mathcal{S}}(n^{-1}\ell_{rt}^{(2)} + H_{rt}) = (s_n \log p_n/n)^{1/2} = o_p(1)$, and we can write $\sum_t n^{-1}\ell_{rt}^{(2)}(\hat{\theta}_s - \theta_s^*)_{[t]} = \sum_t(-H_{rt})(\hat{\theta}_s - \theta_s^*)_{[t]}\{1 + o_p(1)\}$. By a similar argument, $n^{-1}\ell_{rtu}^{(3)}(\tilde{\theta}_s) = E\{n^{-1}\ell_{rtu}^{(3)}(\tilde{\theta}_s)\}\{1 + o_p(1)\}$. We rewrite $\sum_{tu}(2n)^{-1}\ell_{rtu}^{(3)}(\tilde{\theta}_s)(\hat{\theta}_s - \theta_s^*)_{[t]}(\hat{\theta}_s - \theta_s^*)_{[u]} = \sum_t[(1/2)(\hat{\theta}_s - \theta_s^*)_{[t]}\sum_u\{n^{-1}\ell_{rtu}^{(3)}(\tilde{\theta}_s)(\hat{\theta}_s - \theta_s^*)_{[u]}\}]$. By Lemma A1, $\sum_u n^{-1}\ell_{rtu}^{(3)}(\tilde{\theta}_s)(\hat{\theta}_s - \theta_s^*)_{[u]} = O_p\{(s_n^4 \log p_n/n)^{1/2}\} = o_p(1)$. This means that

$$0 = n^{-1}\ell_r^{(1)}(\hat{\theta}_s) = n^{-1}\ell_r^{(1)}(\theta_s^*) - \sum_t H_{rt}(\hat{\theta}_s - \theta_s^*)_{[t]}\{1 + o_p(1)\}.$$

We thus obtain $n^{-1}U_n(\theta_s^*) = H_s(\theta_s^*)(\hat{\theta}_s - \theta_s^*)\{1 + o_p(1)\}$ and $(\hat{\theta}_s - \theta_s^*) = n^{-1}H_s^{-1}(\theta_s^*)U_n(\theta_s^*)\{1 + o_p(1)\}$, while $o_p(1)$ holds for all models s. Next, Taylor expansion for the pseudo-loglikelihood leads to $\ell_I(\hat{\theta}_s) - \ell_I(\theta_s^*) = U_n(\theta_s^*)^T(\hat{\theta}_s - \theta_s^*) - (1/2)\sum_{rt} n(\hat{\theta}_s - \theta_s^*)_{[r]}(\hat{\theta}_s - \theta_s^*)_{[t]}H_{rt} + \tilde{R}_n$, where the error term is given by

$$\tilde{R}_n = \frac{1}{2}\sum_{rt}(\hat{\theta}_s - \theta_s^*)_{[r]}(\hat{\theta}_s - \theta_s)_{[t]}(\ell_{rt}^{(2)} + nH_{rt}) + \frac{1}{6}\sum_{rtu}(\hat{\theta}_s - \theta_s^*)_{[r]}(\hat{\theta}_s - \theta_s^*)_{[t]}(\hat{\theta}_s - \theta_s^*)_{[u]}\ell_{rtu}^{(3)}(\tilde{\theta}_s),$$

with $\tilde{\theta}_s$ between θ_s^* and $\hat{\theta}_s$. By arguments similar to those above, $(\ell_{rt}^{(2)} + nH_{rt})/(nH_{rt}) = o_p(1)$ and $\{\sum_u(\hat{\theta}_s - \theta_s^*)_{[u]}\ell_{rtu}^{(3)}(\tilde{\theta}_s)\}/nH_{rt} = o_p(1)$. Therefore

$$\ell_I(\hat{\theta}_s) - \ell_I(\theta_s^*) = U_n(\theta_s^*)^T(\hat{\theta}_s - \theta_s^*) - \left\{\frac{1}{2}\sum_{rt} n(\hat{\theta}_s - \theta_s^*)_{[r]}(\hat{\theta}_s - \theta_s^*)_{[t]}H_{rt}\right\}\{1 + o_p(1)\}.$$

This implies that $2\{\ell_I(\hat{\theta}_s) - \ell_I(\theta_s^*)\} = n^{-1}U_n(\theta_s^*)^T H_s(\theta_s^*)^{-1}U_n(\theta_s^*)\{1 + o_p(1)\}$, where $o_p(1)$ holds for all models $s \in \mathcal{S}$. □

LEMMA A3. *Under Assumptions 1–7,* $\max_{s\in\mathcal{S}}|2\{\ell_I(\hat{\theta}_s) - \ell_I(\theta_s^*)\}| = O_p(s_n \log p_n)$.

Proof of Lemma A3. By Lemma A2 we have $2\{\ell_I(\hat{\theta}_s) - \ell_I(\theta_s^*)\} = Q_s\{1 + o_p(1)\}$, where the quadratic approximation $Q_s = n^{-1}U_n(\theta_s^*)^T\{H_s(\theta_s^*)\}^{-1}U_n(\theta_s^*)$ and the term $o_p(1)$ both hold uniformly for every model s. Therefore, it suffices to show that $\max_{s\in\mathcal{S}}|Q_s| = O_p(s_n \log p_n)$. Based on the cumulant boundedness condition of $\ell^{(1)}(\theta_s^*; Y_{(i)})$ and the uniform boundedness of the eigenvalues of $V_s(\theta_s)$ in Assumption 6, $\eta = n^{-1/2}\{V_s(\theta_s^*)\}^{-1/2}U_n(\theta_s^*)$ satisfies the exponential moment condition

$$\log E\{\exp(\gamma^T\eta)\} \leqslant a^2\|\gamma\|^2/2$$

with $\gamma \in R^{d_s}$, $\|\gamma\| \leqslant (s_n^2 \log p_n)^{1/2}$ and some constant $a^2 > 1$. We scale the vector η by $\eta^* = \eta/a$, so that $\log E\{\exp(\gamma^{\mathrm{T}}\eta^*)\} \leqslant \|\gamma\|^2/2$ with $\|\gamma\| \leqslant (a^2 s_n^2 \log p_n)^{1/2} = g$. Define $B = V_s^{1/2}(\theta_s^*)\{H_s(\theta_s^*)\}^{-1}V_s^{1/2}(\theta_s^*)$ and $\tau = \lambda_{\max}(B)$. Because the eigenvalues of $H_s(\theta_s^*)$ and $V_s(\theta_s^*)$ are uniformly bounded away from zero and infinity, τ is bounded by a constant. We scale the matrix and let $B^* = B/\tau$. Then the maximum eigenvalue of B^* is 1. After the scaling, $Q_s = a^2\tau(\eta^*)^{\mathrm{T}}B^*\eta^* = a^2\tau Q_s^*$, where $Q_s^* = (\eta^*)^{\mathrm{T}}B^*(\eta^*)$.

Next we apply the large deviation result from Corollary 4.2 of Spokoiny & Zhilova (2013). Let $p_G = \mathrm{tr}(B^*)$ and $v_G^2 = 2\,\mathrm{tr}\{(B^*)^2\}$. Because $g^2 = a^2 s_n^2 \log p_n$, we have $g^2 > 2p_G$. Define w_c by $w_c(1+w_c)/(1+w_c^2)^{1/2} = gp_G^{-1/2}$. Define $\mu_c = \min\{w_c^2/(1+w_c^2),\ 2/3\}$. Further, define $y_c^2 = (1+w_c^2)p_G$ and $2x_c = \mu_c y_c^2 + \log[\det\{I_{d_s} - \mu_c(B^*)^2\}]$. Because $p_G = O(d_s)$, $v_G^2 = O(d_s)$, and the eigenvalues of B^* are all bounded away from zero uniformly, according to Spokoiny & Zhilova (2013), $x_c > g^2/4$ for n sufficiently large. For $v_G/18 \leqslant x \leqslant x_c$, $\mathrm{pr}\{Q_s^* \geqslant (p_G + 6x)\} \leqslant 2\exp(-x) + 8{\cdot}4\exp(-x_c)$. Choose $x = (7/6)s_n \log p_n$ so that $x < x_c$. Then

$$\mathrm{pr}\{Q_s^* \geqslant (p_G + 7s_n \log p_n)\} \leqslant 10{\cdot}4\exp\{-(7/6)s_n \log p_n\}.$$

By the Bonferroni inequality,

$$\max_{s\in\mathcal{S}} \mathrm{pr}(|Q_s^*| > 8s_n \log p_n) \leqslant \sum_s \mathrm{pr}(|Q_s^*| > p_G + 7s_n \log p_n) \to 0.$$

This means that Q_s is $O_{\mathrm{p}}(s_n \log p_n)$ uniformly for all s. □

LEMMA A4. *Under Assumptions 1–7, if $\gamma_n = 6\omega(1+\gamma)\log p_n$ for some $\gamma > 0$ or $\gamma_n = 6\omega(\log p_n + \log\log p_n)$, then* $\mathrm{pr}\{\max_{s\in S_+} Q_{s/T}/(d_s^* - d_T^*) \geqslant \gamma_n\} = o(1)$.

Proof of Lemma A4. Let $\eta_s = V_s(\theta_s^*)^{-1/2}U_n(\theta_s^*)$. Based on Assumption 7 and Lemma 1, $\log E\{\exp(\gamma^{\mathrm{T}}\eta_s)\} \leqslant a^2\|\gamma\|^2/2$ with $\gamma \in R^{d_s}$, $\|\gamma\|^2 \leqslant s_n^2 \log p_n$, and some constant $a^2 > 1$. If we scale the vector η_s and let $\eta_s^* = \eta_s/a$, then $\log E\{\exp(\gamma^{\mathrm{T}}\eta_s^*)\} \leqslant \|\gamma\|^2/2$ with $\|\gamma\| \leqslant (a^2 s_n^2 \log p_n)^{1/2} = g$. Given the matrix $B_s = V_s^{1/2}(\theta_s^*)M_{s/T}V_s^{1/2}(\theta_s^*)$, $\mathrm{tr}(B_s) = d_s^* - d_T^*$. Let $B_s^* = B_s/\tau$, where $\tau = \lambda_{\max}(B_s)$. Then the maximum eigenvalue of B_s^* is 1. After the scaling, $Q_{s/T} = a^2\tau Q_{s/T}^*$, where $Q_{sT}^* = (\eta_s^*)^{\mathrm{T}}B_s^*\eta_s^*$. Define $p_G = \mathrm{tr}(B_s^*)$ and $v_G = [2\,\mathrm{tr}\{(B_s^*)^2\}]^{1/2}$. Using the inequality for the trace of a matrix product (Fang et al., 1994), $v_G \leqslant (2p_G)^{1/2}$. Now we apply the large deviation result from Corollary 4.2 of Spokoiny & Zhilova (2013) and obtain that if $6x_c > K > v_G/3$,

$$\mathrm{pr}\{Q_{s/T}^* > (p_G + K)\} \leqslant 2\exp(-K/6) + 8{\cdot}4\exp(-x_c),$$

where $x_c > g^2/4$ for large n. Choosing $L = \{(d_s^* - d_T^*)/\tau\}\{\gamma_n/(a^2-1)\}$, $\lim_{n\to\infty} L/(v_G/3) > 1$. Furthermore, since $\gamma_n(d_s^* - d_T^*) = O(s_n \log p_n)$, we have $L \leqslant 6x_c$. Using the relationship $d_s^* - d_T^* = (d_s - d_T)\bar{\tau}$, $p_G = (d_s^* - d_T^*)/\tau$, and the Bonferroni inequality, with $m' = d_s - d_T$ we have

$$\begin{aligned}
\mathrm{pr}\Big\{\max_{s\in S^+} Q_{s/T} > (d_s^* - d_T^*)\gamma_n\Big\} &\leqslant \sum_{s\in S+} \mathrm{pr}\{Q_{s/T}^* > (d_s^* - d_T^*)\gamma_n/(a^2\tau)\} \\
&= \sum_{s\in S+} \mathrm{pr}\{Q_{s/T}^* > p_G + p_G(\gamma_n/a^2 - 1)\} \\
&\leqslant \sum_{d_s=d_T+1}^{p_n} C(p_n - d_T, d_s - d_T)10{\cdot}4\exp\{-(\gamma_n/a^2 - 1)(d_s - d_T)\bar{\tau}/(6\tau)\} \\
&\leqslant \sum_{m'=1}^{p_n - d_T} C(p_n - d_T, m')10{\cdot}4\exp\{-m'(\gamma_n/a^2 - 1)/(6w)\} \\
&\leqslant \big[1 + 10{\cdot}4\exp\{-(\gamma_n/a^2 - 1)/(6w)\}\big]^{p_n - d_T} - 1.
\end{aligned}$$

Because a^2 can be chosen as close to 1 as desired by increasing the sample size n, it can be seen that the choices of $\gamma_n = 6w(1+\gamma)\log p_n$ and $\gamma_n = 6w(\log p_n + \log\log p_n)$ lead to $\lim_{n\to\infty}[1+10{\cdot}4\exp\{-(\gamma_n/a^2-1)/(6w)\}]^{p_n-d_T} = 1$. □

LEMMA A5. *Let $Z_1,\ldots,Z_n$ be independent random variables. If each Z_i has zero mean and unit variance and satisfies the cumulant boundedness condition in Definition 1, then*

$$\mathrm{pr}\left\{\sum_{i=1}^{n} Z_n > (2ns_n\log p_n)^{1/2}\right\} = o(p_n^{-s_n}).$$

Proof of Lemma A5. By Taylor expansion, for $|t| \leqslant \delta$, the cumulant generating function for Z_i is

$$g_i(t) = t^2/2 + g_i^{(3)}(t^*)t^3/6$$

for some $0 \leqslant |t^*| \leqslant |t| \leqslant \delta$. Let $\bar{g}^{(3)}(t) = n^{-1}\sum_i g_i^{(3)}(t)$. Because each $g_i^{(3)}$ is uniformly bounded, the average $\bar{g}^{(3)}$ is also bounded. For any $|t|/n^{1/2} \leqslant \delta$, the moment generating function of $n^{-1/2}\sum_{i=1}^n Z_i$ is

$$\phi_n(t) = \exp\{t^2/2 + \bar{g}^{(3)}(t^*/n^{1/2})t^3/(6n^{1/2})\}.$$

For convenience, let $b_n = (2{\cdot}1s_n\log p_n)^{1/2}$. It can be shown that

$$I\left(n^{-1/2}\sum_{i=1}^{n} Z_i > b_n\right) \leqslant \exp\left\{t\left(n^{-1/2}\sum_{i=1}^{n} Z_i - b_n\right)\right\}$$

for any $t > 0$. Then

$$\begin{aligned}\mathrm{pr}\left(n^{-1/2}\sum_{i=1}^{n} Z_i > b_n\right) &\leqslant E\left[\exp\left\{t\left(n^{-1/2}\sum_{i=1}^{n} Z_i - b_n\right)\right\}\right] \\ &= \exp\{t^2/2 + \bar{g}^{(3)}(t^*/n^{1/2})t^3/(6n^{1/2}) - b_n t\} = \exp[(t^2/2)\{1+o(1)\} - b_n t].\end{aligned}$$

Letting $t = b_n$,

$$\mathrm{pr}\left\{\sum_{i=1}^{n} Z_i > (2{\cdot}1ns_n\log p_n)^{1/2}\right\} \leqslant \exp\left[-(1/2)b_n^2\{1+o(1)\}\right] = o(p_n^{-s_n}),$$

completing the proof. □

LEMMA A6. *Under Assumptions 1–7,*

$$\max_{s\in\mathcal{S}}\left|\sum_{i=1}^{n}\ell_I(\theta_s^*; Y_{(i)}) - E\{\ell_I(\theta_s^*; Y_{(i)})\}\right| = O_{\mathrm{p}}\{(ns_n\log p_n)^{1/2}\},$$

$$\max_{s\in\mathcal{S}}\left|\sum_{i=1}^{n}\partial\ell_I(\theta_s^*; Y_{(i)})/\partial\theta_j\right| = O_{\mathrm{p}}\{(ns_n\log p_n)^{1/2}\},$$

$$\max_{s\in\mathcal{S}}\left|\sum_{i=1}^{n}\partial^2\ell_I(\theta_s^*;Y_{(i)})/(\partial\theta_j\partial\theta_k)-E\{\partial^2\ell_I(\theta_s^*;Y_{(i)})/(\partial\theta_j\partial\theta_k)\}\right|=O_{\mathrm{p}}\{(ns_n\log p_n)^{1/2}\},$$

$$\max_{s\in\mathcal{S}}\left|\sum_{i=1}^{n}\partial^3\ell_I(\theta_s;Y_{(i)})/(\partial\theta_j\partial\theta_k\partial\theta_l)-E\{\partial^3\ell_I(\theta_s;Y_{(i)})/(\partial\theta_j\partial\theta_k\partial\theta_l)\}\right|=O_{\mathrm{p}}\{(ns_n\log p_n)^{1/2}\},$$

with $j,k,l\in\{vw:v=1,\ldots,K;\ w=1,\ldots,d_s\}$ *and* $\|\theta_s-\theta_s^*\|\leqslant\delta$.

Proof of Lemma A6. Because $\ell_I(\theta_s^*;Y_{(i)})$ satisfies the cumulant boundedness condition in Definition 1, its first and second moments are bounded uniformly. Given a model s, by Lemma A5,

$$\mathrm{pr}\left(\sum_{i=1}^{n}\big[\ell_I(\theta_s^*;Y_{(i)})-E\{\ell_I(\theta_s^*;Y_{(i)})\}\big]\big/\mathrm{var}\{\ell_I(\theta_s^*;Y_{(i)})\}>(2{\cdot}1ns_n\log p_n)^{1/2}\right)=o(p_n^{-s_n}).$$

Because there are $p_n^{s_n}$ models in the model space, by the Bonferroni inequality,

$$\mathrm{pr}\left(\max_{s\in\mathcal{S}}\sum_{i=1}^{n}\big[\ell_I(\theta_s^*;Y_{(i)})-E\{\ell_I(\theta_s^*;Y_{(i)})\}\big]>C(2{\cdot}1ns_n\log p_n)^{1/2}\right)\leqslant o(p_n^{-s_n})p_n^{s_n}\to 0,$$

where C is the upper bound for $\mathrm{var}\{\ell_I(\theta_s^*;Y_{(i)})\}$. Similar arguments apply to the results for the first, second and third derivatives of the pseudo-loglikelihood. □

References

Bach, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.* **9**, 1179–225.

Breheny, P. & Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statist. Comp.* **25**, 173–87.

Chen, J. H. & Chen, Z. H. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–71.

Claeskens, G. & Consentino, F. (2008). Variable selection with incomplete covariate data. *Biometrics* **64**, 1062–9.

Cox, D. R. & Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729–37.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

Fan, J. & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20**, 101–48.

Fan, J. & Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Info. Theory* **57**, 5467–84.

Fan, J. & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928–61.

Fang, Y., Loparo, K. A. & Feng, X. (1994) Inequalities for the trace of matrix product. *IEEE Trans. Auto. Contr.* **39**, 2489–90.

Foster, D. P. & George, E. I. (1994) The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947–75.

Gao, X. & Song, P. X.-K. (2010). Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *J. Am. Statist. Assoc.* **105**, 1531–40.

Garcia, R., Ibrahim, J. G. & Zhu, H. (2010). Variable selection for regression models with missing data. *Statist. Sinica* **20**, 149–65.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31**, 1208–11.

Guo, X., Zhang, H., Wang, Y. & Wu, J. (2015). Model selection and estimation in high dimensional regression models with group SCAD. *Statist. Prob. Lett.* **103**, 86–92.

Huang, J., Breheny, P. & Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statist. Sci.* **27**, 481–99.

Iwamoto, T., Bianchini, G., Booser, D., Qi, Y., Coutant, C., Shiang, C. Y., Santarpia, L., Matsuoka, J., Hortobagyi, G. N., Symmans, W. F. et al. (2011). Gene pathways associated with prognosis and chemotherapy sensitivity in molecular subtypes of breast cancer *J. Nat. Cancer Inst.* **103**, 264–72.

JOE, H. & LEE, Y. (2009). On weighting of bivariate margins in pairwise likelihood. *J. Mult. Anal.* **100**, 670–85.
KIM, Y., KWON, S. & CHOI, H. (2012). Consistent model selection criteria on high dimensions. *J. Mach. Learn. Res.* **13**, 1037–57.
KWON, S. & KIM, Y. (2012). Large sample properties of the SCAD-penalized maximum likelihood estimation on high dimensions. *Statist. Sinica* **22**, 629–53.
LINDSAY, B. G. (1988). Composite likelihood methods. In *Statistical Inference from Stochastic Processes*, N. U. Prabhu, ed. Providence, Rhode Island: American Mathematical Society, pp. 221–39.
LINDSAY, B. G., YI, G. Y. & SUN, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statist. Sinica* **21**, 71–105.
MEIER, L., VAN DE GEER, S. & BÜHLMANN, P. (2008). The group lasso for logistic regression. *J. R. Statist. Soc.* B **70**, 53–71.
MEINSHAUSEN, N. & BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436–62.
NARDI, Y. & RINALDO, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Statist.* **2**, 605–33.
RIBATET, M., COOLEY, D. & DAVISON, A. C. (2012). Bayesian inference from composite likelihood, with an application to spatial extremes. *Statist. Sinica* **22**, 813–45.
SPOKOINY, V. & ZHILOVA, M. (2013). Sharp deviation bounds for quadratic forms. *Math. Meth. Statist.* **22**, 100–13.
TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.
VARIN, C. (2008). On composite marginal likelihoods. *Adv. Statist. Anal.* **92**, 1–28.
VARIN, C. & VIDONI, P. (2005). A note on composite likelihood inference and model selection. *Biometrika* **92**, 519–28.
VARIN, C. & VIDONI, P. (2006). Pairwise likelihood inference for ordinal categorical time series. *Comp. Statist. Data Anal.* **51**, 2365–73.
WANG, Y., KLIJN, J. G., ZHANG, Y., SIEUWERTS, A. M., LOOK, M. P., YANG, F., TALANTOV, D., TIMMERMANS, M., MEIJER-VAN GELDER, M. E., YU, J. ET AL. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671–9.
WANG, L., LI, H. & HUANG, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Am. Statist. Assoc.* **103**, 1556–69.
WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.
XU, X. & REID, N. (2011). On the robustness of maximum composite likelihood estimate. *J. Statist. Plan. Infer.* **141**, 3047–54.
YUAN, M. & LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B **68**, 49–67.
ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894–942.
ZHANG, Y. & SHEN, X. (2010). Model selection procedure for high-dimensional data. *Statist. Anal. Data Mining* **3**, 350–8.
ZHAO, P., ROCHA, G. & YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* **37**, 3468–97.
ZHAO, P. & YU, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–63.
ZOU, H. (2006) The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–29.

[*Received on* 29 *January* 2015. *Editorial decision on* 27 *January* 2017]