Electronic Journal of Statistics
Vol. 5 (2011) 309–328
ISSN: 1935-7524
DOI: 10.1214/11-EJS608

Automatic grouping using smooth-threshold estimating equations

Masao Ueki[*]

*Faculty of Medicine, Yamagata University, 2-2-2 Iida-Nishi, Yamagata,
Yamagata 990-9585, Japan*
e-mail: uekimrsd@nifty.com

and

Yoshinori Kawasaki[†]

*The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa,
Tokyo 190-8562, Japan*
e-mail: kawasaki@ism.ac.jp

Abstract: Use of redundant statistical model is often the case with practical data analysis. Redundancy widely investigated is inclusion of irrelevant predictors which is resolved by setting their coefficients to zero. On the other hand, it is also useful to consider overlapping parameters of which the values are similar. Grouping by regarding a set of parameters as a single parameter contributes to building intimate parameterization and increasing estimation accuracy by dimension reduction.

The paper proposes a data adaptive automatic grouping of parameters, which simultaneously enables variable selection that can yield sparse solution, by applying the smooth-thresholding. The new procedure is applicable to several estimation equation-based methods, and is shown to possess the oracle property. No convex optimization is needed for its implementation. Numerical examinations including large p small n situation are performed. Proposed automatic grouping applies to interaction modeling for Ohio wheeze data and for credit scoring data.

AMS 2000 subject classifications: Primary 62J07; secondary 62J10.
Keywords and phrases: Automatic grouping, lasso, smooth-thresholding, variable selection.

Received September 2010.

1. Introduction

It is typical in regression modeling such as for economic, social and clinical data that several kinds of predictors are sampled as many as possible to avoid oversight of important factors. Redundancy widely investigated is inclusion of irrelevant predictors which is resolved by setting the corresponding regression coefficient zero using model selection or hypothesis testing, i.e., variable selection

[*] Supported by Grant-in-Aid for Young Scientists (B) 21700318.

[†] Supported by Grant-in-Aid for Scientific Research (A) 21243019, (C) 21500287 and (C) 22530257.

problem. On the other hand, it can be useful to consider redundancy caused by inclusion of overlapping parameters where their regression coefficients are similar, which might come from functional reason. The similarity does not necessarily require that the predictors are highly correlated because it may arise that a latent factor common in the predictors influences the response although they behave independently. Thus, the concept of overlapping parameters is different from that of the grouping effect of [26] in terms of dealing with the association to the response.

Consider a d-dimensional multiple regression model $y_i = \sum_{j=1}^{d} X_{ij}\theta_j + \epsilon_i, i = 1, \ldots, n$, with regression coefficients θ_j, predictors X_{ij}, and random error ϵ_i. Suppose that there is a set of group indices $Q \subset \{1, \ldots, d\}$ such that $\theta_j = \theta_k$ for every pair of Q. Apparently, better estimation of θ_j for $j \in Q$ is on regarding them as a single parameter rather than as separate parameters provided that we know Q. This is equivalent to consider the revised regression model of

$$y_i = X_{jQ}\alpha_Q + \sum_{j \in \{1,\ldots,d\} \setminus Q} X_{ij}\alpha_j + \epsilon_i, i = 1, \ldots, n,$$

in which $X_{iQ} = \sum_{j \in Q} X_{ij}$, the grouped predictor, α_Q is the corresponding regression coefficient, and $\alpha_j = \theta_j$ for $j \in \{1, \ldots, d\} \setminus Q$. It in turn implies that a careful consideration of the validity of introducing X_{iQ} in the model is needed. For example, simply adding continuous and categorical predictors sounds rather odd, and scaling of predictors alters relationships between regression coefficients. The simplest case is that the predictors are all binary and independent each other, which turns out to be the problem of pooling categories in contingency table. Even without the independence between predictors, the regression model can represent additive effect by the number of falling under the category to the response, which seems a reasonable model in practical statistical modeling. Thus, we start with binary predictors in the following examples although the application is not limited to binary predictors. Binary predictor often is useful in handling categorical predictors as well as their interaction effect (See Section 5).

Since the overlapping parameters are unknown in advance, we need to infer it from the data. Exhaustive search of possible combinations of parameters is virtually impossible if the number of parameters is large. To address the issue this paper develops an automatic grouping method by using contemporary variable selection techniques [16, 26, 25, 18, 1, 4], enabling data adaptive grouping of parameters as in Bondell and Reich [1]. The proposed method is an extension of smooth-thresholding [18] that carries out simultaneous grouping and variable selection, i.e., zero parameter estimates are also produced. It is shown that the method possesses model selection consistency [6, 25, 18] in this context, termed the grouping consistency. Required optimization technique in smooth-thresholding of Ueki [18] is Newton–Raphson-type method which does not require convex programming, hence it is computationally efficient and stable. L_1-type penalization [1] is an alternative choice, but needs convex programming. Smooth-thresholding is therefore more advantageous, particularly, for estimation problem where iterative optimization is needed, such as the logistic regression and generalized estimating equation.

Our method requires an initial estimate as in the adaptive lasso [25]. The simplest choice is an unpenalized estimator when the number of parameters is smaller than the sample sizes. When the model has larger number of parameters than that of samples, we suggest to employ the elastic net as an initial estimate that can handle the grouping effect [26]. The tuning parameters included are selected by a BIC-type criterion [21, 20, 18]. Because the method is estimating equation-based, it has wide applicability, e.g., for the generalized estimation equation [14, 17] and the Buckley–James estimator [2, 13].

Our proposal is applicable to interaction modeling for categorical predictors. The aim is similar to that of Choi, Li and Zhu [4] who propose a variable selection method for the model with the strong heredity constraint [9]. The model obtained from their method is more interpretable, since it allows presence of interaction only when the corresponding main effects exist. The strong heredity constraint is appropriate in analyzing the designed experiments [3, 12]. However, other interaction models may be plausible in other fields, e.g., where interaction effect appears without main effect. In such situations, our method is useful owing to its flexibility for grouping of separate categories of interaction. Real data examples given in Section 5 demonstrate our procedure in detail.

2. Methodology

2.1. Grouping by smooth-thresholding

Suppose that we have a dataset $X = (X_1, \dots, X_n)$ of size n from a distribution having a parameter vector θ^* where $\theta^* = (\theta_1^*, \dots, \theta_d^*)^T$ is defined in Θ, a subset of R^d. This paper considers estimation of θ^* via estimating equations $u(\theta) = 0$, where $u(\theta) = \sum_{i=1}^n u(X_i; \theta)$, and $u(x; \theta) = \{u_j(x; \theta)\}_{j=1,\dots,d}$ are the d-dimensional vector-valued estimating function for estimating θ^*. For example, if u is the score function then it becomes the maximum likelihood estimation. Hereafter we assume that u satisfies $E\{u(X; \theta^*)\} = 0$, $E\{u(X; \theta^*)^2\} < \infty$, and some suitable regularity conditions for consistency and asymptotic normality of the full model estimator [e.g., 19, Ch. 5]. The additional assumption in this paper is that the full model involves overlapping parameters redundantly. Specific parameter structure underlying is given in Section 3.1. Turning to the example in Section 1, the saturated model has d parameters, $\theta = (\theta_1, \dots, \theta_d)^T$, but in reality consists of only $d - |Q| + 1$ intrinsic parameters such that $\theta_j = \alpha_Q$ for $j \in Q$ and $\theta_j = \alpha_j$ for otherwise. Notation $|\cdot|$ represents cardinality of the set. Our purpose is to identify genuine relationship, which is of course unknown, and simultaneously estimate it as accurately as possible. Smooth-thresholding [18] intends to connect the estimating equations of several submodels smoothly.

The full model estimator is a solution to the estimating equations $u_j(\theta) = 0\,(j = 1, \dots, d)$. According to [18], the smooth-thresholding modifies the above estimating equations to

$$(1 - \hat{\delta}_j)u_j(\theta) + \hat{\delta}_j\theta_j = 0,$$

for $j = 1, \dots, d$, where $\hat{\delta}_j = \hat{\delta}_j(\lambda, \delta) = \min(1, \lambda/|\hat{\theta}_j^{\text{ini}}|^{1+\gamma})$, $\hat{\theta}_j^{\text{ini}}$ is a suitable initial estimate for θ_j^*, and tuning parameters λ and γ. Note that the jth estimating equation reduces to $\theta_j = 0$ when $\hat{\delta}_j = 1$, creating sparse solution, and that $\hat{\delta}_j = 1$ is equivalent to $|\hat{\theta}_j^{\text{ini}}|^{1+\gamma} \leq \lambda$. The smooth-thresholding is closely related to the adaptive lasso [18].

The procedure of [18] can be viewed as a weighted L_2-penalization for $L(\theta)$ where L represents the criterion to be minimized, which corresponds to u, with a weighted L_2-penalty function $\sum_{j=1}^d \hat{w}_j \theta_j^2/2$ in which $\hat{w}_j = \hat{\delta}_j/(1-\hat{\delta}_j)$. Enlarging this perspective leads to a simple automatic grouping procedure as used in Bondell and Reich [1] who employ L_1-penalization. Instead of $\sum_{j=1}^d \hat{w}_j \theta_j^2/2$, we propose to use the penalty function

$$h(\theta) = \sum_{j=1}^{d} \sum_{k>j}^{d} \hat{w}_{jk} (\theta_j - \theta_k)^2/2,$$

where $\hat{w}_{jk} = \hat{\delta}_{jk}/(1-\hat{\delta}_{jk})$ and $\hat{\delta}_{jk} = \hat{\delta}_{jk}(\lambda, \delta) = \min(1, \lambda/|\hat{\theta}_j^{\text{ini}} - \hat{\theta}_k^{\text{ini}}|^{1+\gamma})$; Other choices of $\hat{\delta}_{jk}$ may be possible depending on the problem, such as the correlation between jth and kth predictors. Analogously to variable selection, restriction $\theta_j = \theta_k$ is produced from $\hat{\delta}_{jk} = 1$, which is equivalent to $|\hat{\theta}_j^{\text{ini}} - \hat{\theta}_k^{\text{ini}}|^{1+\gamma} \leq \lambda$, i.e., $L(\theta) + h(\theta)$ is minimized over the restricted space of

$$\bigcup_{j<k} \{\theta_j = \theta_k : |\hat{\theta}_j^{\text{ini}} - \hat{\theta}_k^{\text{ini}}|^{1+\gamma} \leq \lambda\},$$

for given tuning parameters λ and γ. Although it may arise that $\hat{\delta}_{12} = \hat{\delta}_{23} = 1$ but $\hat{\delta}_{13} < 1$, restriction $\hat{\delta}_{13} = 1$ is automatically imposed owing to the fact that there is only one free parameter under restrictions $\theta_1 = \theta_2$ and $\theta_2 = \theta_3$.

Furthermore, it is possible to engage variable selection as well as grouping. To this end, introduce the 0th parameter θ_0, which is always restricted to zero, and construct $h(\theta)$ in the same manner with $\hat{\theta}_0^{\text{ini}} = 0$. Then the penalty function is modified as

$$h(\theta) = \sum_{j=0}^{d} \sum_{k>j}^{d} \hat{w}_{jk} (\theta_j - \theta_k)^2/2 = \sum_{j=1}^{d} \sum_{k>j}^{d} \hat{w}_{jk} (\theta_j - \theta_k)^2/2 + \sum_{k>0}^{d} \hat{w}_{0k} \theta_k^2/2, \quad (2.1)$$

where $\hat{w}_{0k} = \hat{w}_k$ for each k, in which the second term is the penalty function of [18] for variable selection.

2.2. Implementation

A simple algebra rewrites the penalty function in (2.1) to a more convenient quadratic form

$$h(\theta) = \theta^T \hat{W} \theta/2,$$

where $\hat{W}$ is $d \times d$ matrix having the kth diagonal entry of $\sum_{j=0, j\neq k}^d \hat{w}_{jk}$ and off-diagonal (j,k)-entry of $-\hat{w}_{jk}$. Let $S_{\text{full}} = \{1, \dots, d\}$, and denote the jth unit vector in d-dimensional Euclidean space by e_j. We also represent the d-dimensional zero vector by e_0. Define the active set $\mathcal{A}$ by the complement of the inactive set,

$$\mathcal{A}^c = \{k \in S_{\text{full}} : \text{there exists } j \in S_{\text{full}} \cup \{0\} \text{ such that } j < k \text{ and } \hat{\delta}_{jk} = 1\},$$

i.e., the active set consists of indices smallest in each group. We also define d row vectors $R_1, \ldots, R_d$ where R_j's are e_j^T for $j \in \mathcal{A}$ and $e_{\xi(j)}^T$ for $j \in \mathcal{A}^c$. Here the function $\xi(j)$ is defined as follows. Let $f(j) = \min(i \in S_{\text{full}} \cup \{0\} : \hat{\delta}_{ji} = 1)$ and define $f^{(l)}(j) = f\{f^{(l-1)}(j)\}$ with $f^{(0)}(j) = f(j)$ for $l = 0, 1, 2, \ldots$. Then we define $\xi(j)$ to be $f^{(l)}(j)$ satisfying $f^{(l)}(j) = f\{f^{(l-1)}(j)\}$ for some non-negative integer l. The collection of $R_1, \ldots, R_d$ is denoted by the $d \times d$ matrix R, and the set $\{\theta \in \Theta : \theta = R\alpha\}$ with an intrinsic parameter $\alpha \in \Theta$ represents the resulting parameter space whose dimensionality is $|\mathcal{A}|$. Indeed, if jth column of R is zero vector, which corresponds to the event where index j is included in $\mathcal{A}^c$, α_j does not appear in $R\alpha$. Notably, $\theta_j = (R\alpha)_j$ for $j \in \mathcal{A}^c$ with $\xi(j) = 0$ is exactly zero, i.e., sparse solution.

Denote by $R_{\mathcal{A}}$ the $d \times |\mathcal{A}|$ matrix that consists of jth column vectors of R for $j \in \mathcal{A}$, and similarly, by $\alpha_{\mathcal{A}}$ the $|\mathcal{A}|$-dimensional active parameter vector and by $\hat{W}_{\mathcal{A}}$ the $|\mathcal{A}| \times |\mathcal{A}|$ sub-matrix of $\hat{W}$. Note that each component of $\hat{W}_{\mathcal{A}}$ is finite. Then, estimation for active parameters $\alpha_{\mathcal{A}}$ can be done by minimization of $L(R_{\mathcal{A}}\alpha_{\mathcal{A}}) + \alpha_{\mathcal{A}}^T \hat{W}_{\mathcal{A}} \alpha_{\mathcal{A}}/2$. By differentiating this with respect to $\alpha_{\mathcal{A}}$, we have the estimating equation

$$R_{\mathcal{A}}^T u(R_{\mathcal{A}}\alpha_{\mathcal{A}}) + \hat{W}_{\mathcal{A}}\alpha_{\mathcal{A}} = 0, \tag{2.2}$$

where Newton–Raphson-type procedures work effectively. The estimator in the full model space Θ is given by $\hat{\theta}_{\lambda,\gamma} = R_{\mathcal{A}}\hat{\alpha}_{\lambda,\gamma,\mathcal{A}}$ with a solution $\hat{\alpha}_{\lambda,\gamma,\mathcal{A}}$ to (2.2). We could introduce other types of penalization that carry out automatic grouping instead of the smooth-thresholding, e.g., L_1-penalty [1] and other penalties including elastic net, adaptive lasso, and smoothed clipped absolute deviation, following such a way of Johnson, Lin and Zeng [11]. However, we can suffer from finding the solution if we use those penalties that require convex programming. In contrast, the smooth-thresholding is free from this issue, which is an advantage over other penalization methods.

2.3. Illustrative example

It may be somewhat difficult to imagine the mechanism that underlies our grouping method. An application to simple estimation problem is thus helpful for illustrative purpose. Consider the loss function $L(\theta) = ||\theta - x||^2/2$ with a d-dimensional data vector x.

If we have $\hat{w}_{jk} \in [0, \infty)$ for all j and k, the matrix $\hat{W}$ defined in Section 2.2 is symmetric, thus so is $I + \hat{W}$. Since, by definition, $\theta^T \hat{W} \theta = 2h(\theta) \geq 0$ for each θ, we can see that $I + \hat{W}$ is invertible. The solution to (2.2) is thus written as $\hat{\theta} = (I + \hat{W})^{-1}x$. The following result is helpful in understanding our procedure.

Theorem 2.1. *All elements of* $(I + \hat{W})^{-1}$ *are non-negative.*

If we further assume that $\hat{w}_{0k} = 0$ for $k = 1, \ldots, d$, we have $\hat{W}1 = 0$ where 1 denotes the d-dimensional vector of ones. Using this fact, it can be readily seen that $(I + \hat{W})^{-1}1 = 1$, implying that sum to unity for each rows. Thus, the linear operator $(I + \hat{W})^{-1}$ permits re-allocation of each component of data vector x.

In more general situations where $\hat{w}_{0k} \geq 0$, it holds that $(I + \hat{W})1 = g$ where g is the d-vector whose jth element is $1 + \hat{w}_{0j}$. Hence, $(I + \hat{W})^{-1}g = 1$, which states that the re-allocation mentioned above applies to the data $x_j/(1 + \hat{w}_{0j})$ shrunken toward zero. The above argument justifies our grouping method.

2.4. Choice of initial estimate

Choice of initial estimate is in practice an important task. Theoretically, it should be root-n consistent (see Section 3). Although the simplest choice is the unpenalized full model estimator, when the number of parameters gets larger than that of samples, automatic variable selection method is desirable since the unpenalized estimators fail to perform. The lasso [16] however can not appropriately handle the grouped predictor variables, whereas the elastic net [26] can overcome this issue. Grouped predictor variables tend to appear in high-dimensional data. Consequently, we recommend using the elastic net as the initial estimate, in particular, for high-dimensional data. Numerical experiments given in Section 4 provide comparisons between the lasso and elastic net initial estimates.

3. Analyzing the method

3.1. Oracle property

In this section, we analyze the theoretical properties of the proposed method with respect to the oracle property [6, 25, 20, 21, 11, 18, 4]. In this section, we assume regularity conditions under which the full model estimator based on the estimating equations, $u(\theta) = 0$, is consistent and asymptotically normally distributed [e.g., 19, Ch.5]. We use the initial estimator $\hat{\theta}^{\text{ini}}$ possessing root-n consistency to the true parameter vector θ^*. Before mentioning the theoretical property we define $\mathcal{A}$ and R in Section 2.1 based on the true θ^* instead of $\hat{\theta}^{\text{ini}}$, and they are denoted by $\mathcal{A}^*$ and R^*. Using R^*, we define the intrinsic vector α^* such that $\theta^* = R^*\alpha^*_{\mathcal{A}^*}$. Full information on parameterization is condensed to a matrix R or R^*, and these are used to evaluate the model selection consistency in this context, termed the grouping consistency. The following theorem states existence of the tuning parameters that confers on the smooth-threshold estimator as the same good performance as the oracle estimator.

Theorem 3.1. *For any positive* λ *and* γ *such that* $n^{1/2}\lambda \to 0$ *and* $n^{(1+\gamma)/2}\lambda \to \infty$ *as* $n \to \infty$*, we have grouping consistency, i.e.* $P(R = R^*) \to 1$.

Provided that $R = R^*$ holds, the smooth-threshold estimating equation coincides with the oracle one (A.2) given in Appendix. Then, the solution to (A.2) denoted by $\hat{\alpha}_{\lambda,\gamma,\mathcal{A}^*}$ possesses an asymptotic normality:

Theorem 3.2. *Under the same assumptions in Theorem 3.1, we have asymptotic normality, i.e.* $n^{1/2}(\hat{\alpha}_{\lambda,\gamma,\mathcal{A}^*} - \alpha^*_{\mathcal{A}^*})$ *is asymptotically normally distributed with mean zero and the covariance matrix of the oracle estimator.*

The next section argues choice of tuning parameters to maintain the grouping consistency even after tuning parameter selection. Noting that the theorem holds for arbitrarily fixed $\gamma > 0$, pre-specification of γ is beneficial for saving computation [20]. We use $\gamma = 2$ throughout our numerical examples, because, in our numerical studies, other choices and data adaptive choice based on the BIC result in comparable performance. Consequently, λ is the only tuning parameter which we should determine.

3.2. Choice of tuning parameters

We propose the following BIC-type criterion to select the tuning parameter λ for the smooth-threshold estimator:

$$\mathrm{BIC}_\lambda = \ell(\hat{\theta}_\lambda) + \mathrm{df}_\lambda \log n, \tag{3.1}$$

where $\mathrm{df}_\lambda = |\mathcal{A}|$ and ℓ is a loss function such as the $-2\times$loglikelihood function. The selected λ minimizes the BIC. To consider properties of the BIC, we shall prepare some notations. Denote the estimator based on the estimating function u over a parameter space $\mathcal{H} \subset \Theta$ by $\hat{\theta}_{\mathcal{H}}$, which is a d-dimensional vector. Likewise, define $\tilde{\theta}_{\mathcal{H}}$ by $\ell(\tilde{\theta}_{\mathcal{H}}) = \inf_{\theta\in\mathcal{H}} \ell(\theta)$. Situations in which $\hat{\theta}_{\mathcal{H}} \neq \tilde{\theta}_{\mathcal{H}}$ occur when using the Wald-type loss $\ell_{\mathrm{w}}(\theta) = (\theta - \hat{\theta}_{\mathrm{full}})^T \hat{\mathrm{var}}(\hat{\theta}_{\mathrm{full}})(\theta - \hat{\theta}_{\mathrm{full}})$ [10, 20]. Denote $\Theta_R = \{\theta \in \Theta : \theta = R\alpha,\ \text{for each } \alpha \in \Theta\}$ with a $d \times d$ matrix R such that $R_{jk} \in \{0, 1\}$ and $\sum_{k=1}^d R_{jk} \in \{0, 1\}$. The true parameter space is also represented by Θ_{R^*}. We write Θ_{R^*} as Θ^* for brevity. In addition, we impose some assumptions on ℓ which are essentially the same as those given in [18].

Assumption 3.1. *For each* $\Theta \not\supset \Theta^*$, *we have* $\liminf_{n\to\infty} n^{-1}\{\ell(\tilde{\theta}_\Theta) - \ell(\theta^*)\} > 0$.

Assumption 3.2. *For each* $\Theta \supset \Theta^*$, *we have* $\ell(\tilde{\theta}_\Theta) - \ell(\theta^*) = O_p(1)$ *and* $\ell(\hat{\theta}_\Theta) - \ell(\theta^*) = O_p(1)$ *as* $n \to \infty$.

In the following, we give a justification for the use of the BIC. Define $\Lambda = \{\lambda : \lambda > 0\}$. Recalling that $\lambda \in \Lambda$ determines the matrix R, define the ideal tuning parameter set by $\Omega^* = \{\lambda \in \Lambda : \Theta^* = \Theta_R\}$. For sufficiently large n, Ω^* is not empty because Ω^* includes λ that satisfies $n^{1/2}\lambda \to 0$ and $n^{(1+\gamma)/2}\lambda \to \infty$ by Theorem 3.1. The overfitted tuning parameter set is defined by $\Omega_{\mathrm{O}} = \{\lambda \in \Lambda : \Theta^* \subset \Theta_R, R^* \neq R\}$ and the underfitted tuning parameter set by $\Omega_{\mathrm{U}} = \{\lambda \in \Lambda : \Theta^* \not\subset \Theta_R\}$.

Theorem 3.3. *Under Assumptions 3.1 and 3.2, for any $\lambda^* \in \Lambda$ such that $n^{1/2}\lambda^* \to 0$ and $n^{(1+\gamma)/2}\lambda^* \to \infty$ with a fixed $\gamma > 0$, as $n \to \infty$, it follows that*

$$P\left(\mathrm{BIC}_{\lambda^*} < \inf_{\lambda \in \Omega_{\mathrm{O}} \cup \Omega_{\mathrm{U}}} \mathrm{BIC}_{\lambda}\right) \to 1.$$

The proof goes in much the same line as [18] and thus is omitted. The above theorem states that the BIC selects tuning parameters such as in Theorems 3.1 and 3.2 rather than those in Ω_{O} or Ω_{U} for large n. It follows from the same argument as in Wang, Li and Tsai [21, Section 3.3] that the BIC selects a tuning parameter that yields the true model, i.e., the minimizer of the BIC enters in Ω^*. As a result, BIC enables consistent model selection, which is a justification for using the BIC as a tuning parameter selector. Even when a unique loss function is absent, typically when we resort to generalized estimating equations and Buckley–James estimator, [10] gives a justification for the Wald-type loss is based on a relationship to an approximate posterior probability conditional on the parameter estimates. Following [18], we prefer loss functions satisfying $\hat{\theta}_{\mathcal{H}} = \tilde{\theta}_{\mathcal{H}}$. Taking this into consideration, we can instead use an alternative loss function, the score-type loss $\ell_{\mathrm{s}}(\theta) = u(\theta)^T \hat{\mathrm{var}}\{u(\hat{\theta}_{\mathrm{full}})\}u(\theta)$, which is asymptotically equivalent to the Wald-type loss when $\theta = \hat{\theta}_{\mathcal{H}}$ because $u(\hat{\theta}_{\mathcal{H}}) \approx \nabla u(\hat{\theta}_{\mathrm{full}})(\hat{\theta}_{\mathcal{H}} - \hat{\theta}_{\mathrm{full}})$ for a sub-model $\mathcal{H}$. The score-type loss ℓ_s is the loss function such that $\hat{\theta}_{\mathcal{H}} = \tilde{\theta}_{\mathcal{H}}$ for a given $\mathcal{H}$.

3.3. Shortcut in model selection

The smooth-thresholding brings a further advantage in choosing tuning parameters if we use the BIC for which the loss function ℓ yielding $\hat{\theta}_{\mathcal{H}} = \tilde{\theta}_{\mathcal{H}}$ for a given $\mathcal{H}$. We then do not need to prepare candidate tuning parameters, typically, using arbitrary discretization. This is because we know the degrees of freedom of parameters, once tuning parameters and initial estimates are specified. Given γ, arrange $d(d-1)/2$ pairs of $|\hat{\theta}_j^{\mathrm{ini}} - \hat{\theta}_k^{\mathrm{ini}}|^{\gamma+1}$ in ascending order and denote them as $q_1 \le \cdots \le q_K$ where $K = d(d-1)/2$, and define $q_0 = 0$. Then, it is obvious that the change of active or inactive set occurs at $\lambda = q_i$ for each $i = 0, \ldots, K$, which in turn implies that the degree of freedom is constant for each $\lambda \in [q_i, q_{i+1})$. Therefore, minimizing BIC is equivalent to minimizing $\ell(\hat{\theta}_\lambda)$ for λ over $[q_i, q_{i+1})$. Since, by assumption, ℓ corresponds to the loss function in obtaining $\hat{\theta}_\lambda$, q_i is the minimizer of $\ell(\hat{\theta}_\lambda)$ for $\lambda \in [q_i, q_{i+1})$, because the smooth-threshold estimator $\hat{\theta}_\lambda$ is a solution to the weighted ridge penalized minimization with respect to the active parameters. Because $\hat{w}_{jk}$ is monotone increasing in λ, smaller λ alleviates the extent of penalization. Consequently, to seek the minimizer of BIC in λ, it suffices to evaluate only at each $\lambda = q_i$ for $i = 0, \ldots, K$, implying that we can find the exact minimizer of BIC with this shortcut procedure. It is also noteworthy that the procedure is similar to that of [24].

4. Numerical experiments

For our simulation studies, we evaluate the accuracy of grouping using the following measures. Let $\mathcal{B}$ be the index set of the group which the jth parameter θ_j^* belongs to, and define, for each j, proportion of correctly fused (PCF) with the other members in its group, and that of incorrectly fused (PICF) with the members in other groups across N simulations by

$$\begin{aligned}\mathrm{PCF}_j &= \frac{1}{N}\frac{1}{|\mathcal{B}|-1}\sum_{r=1}^{N}\sum_{i\in\mathcal{B}\setminus\{j\}} I(\hat{\theta}_{j,r}=\hat{\theta}_{i,r}) \quad \text{and} \\ \mathrm{PICF}_j &= \frac{1}{N}\frac{1}{|\mathcal{B}^c|}\sum_{r=1}^{N}\sum_{i\in\mathcal{B}^c} I(\hat{\theta}_{j,r}=\hat{\theta}_{i,r}),\end{aligned}$$

where $\hat{\theta}_{i,r}$ denotes the ith parameter estimate for rth simulation and $\mathcal{B}^c$ is the complement of $\mathcal{B}$. Here $I(\cdot)$ represents the indicator function. If the obtained grouping is complete, the PCFs and PICFs are one and zero, respectively. For zero parameters, we also define the proportions of correctly and incorrectly setting zero by

$$\mathrm{PC}_0 = \frac{1}{N}\frac{1}{|\mathcal{M}|}\sum_{r=1}^{N}\sum_{i\in\mathcal{M}} I(\hat{\theta}_{i,r}=0) \quad \text{and} \quad \mathrm{PIC}_0 = \frac{1}{N}\frac{1}{|\mathcal{M}^c|}\sum_{r=1}^{N}\sum_{i\in\mathcal{M}^c} I(\hat{\theta}_{i,r}=0),$$

where $\mathcal{M}=\{j\in S_{\text{full}}:\theta_j^*=0\}$ and $\mathcal{M}^c$ is its complement.

On the other hand, the estimation performance is evaluated by the average relative absolute error (ARAE) compared with the oracle estimator, $\mathrm{ARAE}(\hat{\theta}) = N^{-1}\sum_{r=1}^{N}\sum_{j=1}^{d}|\hat{\theta}_{j,r}-\theta_j^*|/\sum_{j=1}^{d}|\hat{\theta}_{j,r}^{o}-\theta_j^*|$ in N simulations. Here $\hat{\theta}_{j,r}^{o}$ denotes the oracle estimator of the jth parameter for rth simulation obtained under which the genuine parameterization is in hand. The estimator is better if ARAE is closer to unity.

4.1. Logistic regression model

We consider an example with binary response with 10 predictor variables in which first two variables are quantitative and others are dichotomous $X = (X_1,\dots,X_{10})$. Response variable Y_i is sampled independently from Bernoulli trial with success rate of $1/\{1+\exp(-X_i\beta^*)\}$. We consider two models having coefficients vector of

$$\begin{aligned}&\text{Model 1:} \quad \beta^* = (-1.3, 0.5, 0, 0, 0, 4, 4, 0, -4, -4)^T, \\ &\text{Model 2:} \quad \beta^* = (-1.3, 0.5, 1, -1, -3, 0, 4, -5, -2, 2)^T, \qquad (4.1)\end{aligned}$$

where first two of β^* are assumed to be out of interest for grouping. Regarding eight parameters remained, model 1 has two intrinsic parameters $(-4, 0, 4)$, while model 2 has eight different parameters with a zero parameter. Since

Table 1

Results in logistic regression example. $ARAE(\hat{\theta})$, average relative absolute error of the proposed automatic grouping method; $ARAE_{\mathrm{ini}}$, average relative absolute error of the initial estimator; AMD, average model dimension; $p_0^u/p_0^c/p_0^o$, Three proportions regarding zero parameter identification (%); PCM, proportion of identifying correct model (%); Values in parenthesis denote the standard error

Model	n/ρ	ARAE($\hat{\theta}$)	ARAE$_{\mathrm{ini}}$	AMD	$p_0^u/p_0^c/p_0^o$	PCM
1	100/0	$1.9_{(0.1)}$	$6.5_{(0.45)}$	$2.9_{(0.04)}$	5/67/28	37
	100/0.5	$2.2_{(0.14)}$	$9.3_{(0.68)}$	$3.1_{(0.04)}$	14/50/35	23
	200/0	$1.2_{(0.05)}$	$3.1_{(0.14)}$	$2.3_{(0.03)}$	0/91/ 9	75
	200/0.5	$1.7_{(0.1)}$	$5.1_{(0.4)}$	$2.6_{(0.04)}$	1/85/14	61
2	100/0	$0.7_{(0.03)}$	$1.1_{(0.01)}$	$3.5_{(0.05)}$	84/11/5	0
	100/0.5	$0.7_{(0.03)}$	$1.2_{(0.04)}$	$3.4_{(0.05)}$	86/ 9/5	0
	200/0	$1.2_{(0.03)}$	$1.1_{(0.01)}$	$4.1_{(0.04)}$	74/21/5	0
	200/0.5	$1.2_{(0.03)}$	$1.1_{(0.01)}$	$3.8_{(0.05)}$	76/16/8	0

Table 2

Results in logistic regression example. $PCF_j/PICF_j$, proportion of correctly fused with the other members in its group/proportion of incorrectly fused with the members in other groups (%), for indices js to be grouped, where the corresponding coefficients under models 1 and 2 are presented; PC_0/PIC_0, proportion of correctly setting zero/proportion of incorrectly setting zero (%)

n/ρ	PCF$_j$/PCIF$_j$, $j=3,\ldots,10$								PC$_0$/PIC$_0$
	$\beta_3^*=0$	$\beta_4^*=0$	$\beta_5^*=0$	$\beta_6^*=4$	$\beta_7^*=4$	$\beta_8^*=0$	$\beta_9^*=-4$	$\beta_{10}^*=-4$	
100/0	83/4	84/3	83/3	60/2	60/2	84/3	60/2	60/3	90/1
100/0.5	78/6	77/6	78/6	43/4	43/5	75/7	48/5	48/4	85/4
200/0	95/0	95/0	95/0	88/0	88/0	95/0	88/0	88/0	96/0
200/0.5	92/1	92/1	92/1	74/1	74/1	92/1	73/1	73/0	95/0
	$\beta_3^*=1$	$\beta_4^*=-1$	$\beta_5^*=-3$	$\beta_6^*=0$	$\beta_7^*=4$	$\beta_8^*=-5$	$\beta_9^*=-2$	$\beta_{10}^*=2$	
100/0	–/26	–/28	–/16	–/27	–/6	–/8	–/24	–/20	82/25
100/0.5	–/28	–/30	–/17	–/29	–/6	–/9	–/27	–/24	78/29
200/0	–/19	–/20	–/14	–/17	–/3	–/6	–/17	–/12	89/16
200/0.5	–/23	–/25	–/14	–/23	–/5	–/7	–/22	–/15	88/20

model 2 has no group, the grouping procedure is redundant and initial estimate will be more efficient. On the other hand, the jth predictor variable X_{ij} for ith sample is generated in the following way. First, define latent variable Z_i which is 10-dimensional multivariate normal random variable with mean zero and covariance matrix whose (j,k)-entry is $\rho^{|j-k|}$. Then define $X_{1i}=Z_{1i}, X_{2i}=Z_{2i}$ and $X_{ij}=I(Z_{ij}>1)$, the latter corresponds to about 16% occurrence being exposed. The experiments validate all combinations of $n\in\{100,200\}$ and $\rho\in\{0,0.5\}$, where unpenalized logistic regression estimates are used as the initial estimates.

Tables 1 and 2 summarize the result of numerical experiments repeated 500 times. The oracle estimators for models 1 and 2 are the logistic regression estimator through the parameterization of $\beta=\beta(\alpha)=(\alpha_1,\alpha_2,0,0,0,\alpha_3,\alpha_3,0,\alpha_4,\alpha_4)^T$ and $\beta(\alpha)=(\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5,0,\alpha_6,\alpha_7,\alpha_8,\alpha_9)^T$. The third and fourth columns in Table 1 give ARAE($\hat{\theta}$) and ARAE$_{\mathrm{ini}}$ = ARAE($\hat{\theta}_{\mathrm{ini}}$), respectively. The fifth column provides the average model dimension (AMD) where the parameters out of interest are excluded, hence, the ideal model dimensions are 2 and 7 for models 1 and 2, respectively. The sixth column presents three quantities, p_0^c, p_0^o, and p_0^u: p_0^c is the proportions whether all zero parameters are correctly set zero, and set nonzero for all nonzero parameters; p_0^o is proportion whether all nonzero parameters are correctly set nonzero but not all zero parameters are correctly set zero; p_0^u equals to $1 - p_0^c - p_0^o$, i.e., other cases. The seventh column (PCM) presents the proportion of identifying correct model which accounts for whether true zero parameters are set zero exactly. Table 2 presents the PCFs and PICFs defined above for each coefficients, which inform performance of grouping in detail. In addition, PC_0 and PIC_0 are shown.

For simulations under model 1, the fact that the ARAEs shown in Table 1 for the proposed grouping method are all less than those of initial estimator emphasizes improvement to the unpenalized estimator. Table 2 illustrates success both in zero parameter identification and in grouping, particularly, for larger sample size setting, which coincides with the theoretical results given in Section 3.

On the other hand, PCF_js are not defined since there is no group for model 2, and are denoted by "$-$" in Table 2. The initial estimator works well in model 2, which is shown in Table 1 through $\mathrm{ARAE}_{\mathrm{ini}}$s whose values are close to unity. ARAEs for the proposed grouping method are comparable to $\mathrm{ARAE}_{\mathrm{ini}}$s, insisting the advantage of our method. Table 2 also implies that the missclassification rate can be reduced by increasing the number of samples.

4.2. Generalized estimating equation

The second example applies to the generalized estimating equation, where each sample of size n has four sets of observation together with 10 predictor variables same as those in the previous example. Response vector Y_i is independently generated from 4-dimensional multivariate normal of $N(X_i\beta^*, V)$ where V is the covariance matrix whose (j,k)-element is $0.667^{|j-k|}$, i.e., AR(1) working correlation is specified. We use two models given in equation (4.1).

The experiments validate all combinations of $n \in \{100, 200\}$, and $\rho \in \{0, 0.5\}$ where unpenalized estimates of generalized estimating equation, where working correlation is correctly specified, are used for the initial estimates. Tables 3 and 4 summarize the numerical experiments repeated 500 times. Simulations in the current example result in more accurate grouping and estimation performance compared with those in the logistic regression example. Although the score-type loss is preferable to the Wald-type loss as stated in Section 3.2, the results of model selection were almost the same. Therefore, we show the result for score-type loss only. Moreover, Tables 3 and 4 also report the improvement of performance as the increase of sample size.

4.3. Large p *small* n *situation*

In this section we consider a large p and small n situation in normal linear model, where n and p represent the number of samples and parameters, respectively. We use $p = 10n$ dichotomous predictor variables $X = (X_1, \ldots, X_{10n})$. As in the

Table 3

Results in generalized estimating equation example. ARAE($\hat{\theta}$), average relative absolute error of the proposed automatic grouping method; ARAE$_{\mathrm{ini}}$, average relative absolute error of the initial estimator; AMD, average model dimension; $p_0^u/p_0^c/p_0^o$, Three proportions regarding zero parameter identification (%); PCM, proportion of identifying correct model (%); Values in parenthesis denote the standard error

Model	n/ρ	ARAE($\hat{\theta}$)	ARAE$_{\mathrm{ini}}$	AMD	$p_0^u/p_0^c/p_0^o$	PCM
1	100/0	$1.7_{(0.07)}$	$3.4_{(0.1)}$	$2.6_{(0.04)}$	0/73/27	62
	100/0.5	$1.7_{(0.06)}$	$3.5_{(0.09)}$	$2.6_{(0.04)}$	0/72/28	60
	200/0	$1.4_{(0.05)}$	$3.2_{(0.08)}$	$2.3_{(0.03)}$	0/84/16	77
	200/0.5	$1.4_{(0.04)}$	$3.6_{(0.1)}$	$2.3_{(0.03)}$	0/86/14	80
2	100/0	$1.1_{(0.01)}$	$1.1_{(0.01)}$	$6.7_{(0.03)}$	2/ 93/5	67
	100/0.5	$1.2_{(0.02)}$	$1.2_{(0.01)}$	$6.6_{(0.03)}$	3/ 88/9	59
	200/0	$1.0_{(0.01)}$	$1.2_{(0.01)}$	$7.0_{(0.01)}$	0/ 99/1	96
	200/0.5	$1.0_{(0.00)}$	$1.1_{(0.01)}$	$7.0_{(0.01)}$	0/100/0	95

Table 4

Results in generalized estimating equation example. PCF$_j$/PICF$_j$, proportion of correctly fused with the other members in its group/proportion of incorrectly fused with the members in other groups (%), for indices js to be grouped, where the corresponding coefficients under models 1 and 2 are presented; PC$_0$/PIC$_0$, proportion of correctly setting zero/proportion of incorrectly setting zero (%)

n/ρ	PCF$_j$/PCIF$_j$, $j=3,\ldots,10$								PC$_0$/PIC$_0$
	$\beta_3^*=0$	$\beta_4^*=0$	$\beta_5^*=0$	$\beta_6^*=4$	$\beta_7^*=4$	$\beta_8^*=0$	$\beta_9^*=-4$	$\beta_{10}^*=-4$	
100/0	85/0	85/0	84/0	87/0	87/0	84/0	88/0	88/0	89/0
100/0.5	83/0	84/0	83/0	86/0	86/0	84/0	88/0	88/0	88/0
200/0	90/0	92/0	91/0	94/0	94/0	92/0	94/0	94/0	94/0
200/0.5	92/0	92/0	92/0	95/0	95/0	92/0	96/0	96/0	95/0
	$\beta_3^*=1$	$\beta_4^*=-1$	$\beta_5^*=-3$	$\beta_6^*=0$	$\beta_7^*=4$	$\beta_8^*=-5$	$\beta_9^*=-2$	$\beta_{10}^*=2$	
100/0	−/2	−/2	−/2	−/1	−/0	−/0	−/3	−/2	95/0
100/0.5	−/3	−/3	−/2	−/2	−/0	−/0	−/4	−/2	91/1
200/0	−/0	−/0	−/0	−/0	−/0	−/0	−/0	−/0	99/0
200/0.5	−/0	−/0	−/0	−/0	−/0	−/0	−/0	−/0	100/0

preceding examples, Z_{ij} is defined as $I(X_{ij} > 1)$ for $j = 1, \ldots, 10n$ with latent variable X_i of $10n$-dimensional multivariate normal distribution. Continuous response variables Y are sampled from the multivariate linear regression model of $y_i = X_i\beta^* + \epsilon_i$ with independent and identical random error e_i from $N(0, \sigma_0)$, where β^* is $10n$-dimensional coefficient vector. The following two models are considered:

$$\text{Model 3:} \quad \beta^* = (4, 4, -4, -4, 0_{10n-4}^T)^T,$$
$$\text{Model 4:} \quad \beta^* = (4, -5, -2, 2, 0_{10n-4}^T)^T,$$

where 0_{10n-4} is the zero vector of $(10n-4)$-dimension. Model 3 has three intrinsic parameters $(-4, 0, 4)$, while model 4 has five intrinsic parameters. Excluding zero parameters, the ideal model dimensions are 2 and 4, respectively. In both models, we expect that $10n-4$ zero components are exactly set zero. Since the ordinary least-squares break down in large p small n situation, we use penalized regression methods, the lasso and elastic net, for the initial estimates. As we mentioned in Section 2.3, the elastic net is expected to work better in high-

Table 5

Results in large p small n example. $\hat{\theta}_{\text{ini}}$, initial estimator used; ARAE($\hat{\theta}$), average relative absolute error of the proposed automatic grouping method; ARAE$_{\text{ini}}$, average relative absolute error of the initial estimator; AMD, average model dimension; $p_0^u/p_0^c/p_0^o$, Three proportions regarding zero parameter identification (%); PCM, proportion of identifying correct model (%); Values in parenthesis denote the standard error

Model	$n/\sigma/\rho$	$\hat{\theta}_{\text{ini}}$	ARAE($\hat{\theta}$)	ARAE$_{\text{ini}}$	AMD	$p_0^u/p_0^c/p_0^o$	PCM
3	100/1/0	lasso	$11.6_{(1.75)}$	$17.2_{(3.24)}$	$5.1_{(0.17)}$	0/24/76	22
		enet	$5.1_{(0.61)}$	$35.8_{(6.34)}$	$3.3_{(0.13)}$	2/62/37	49
	100/1/0.5	lasso	$7.8_{(0.70)}$	$12.8_{(0.87)}$	$5.0_{(0.17)}$	0/30/70	29
		enet	$5.8_{(0.78)}$	$27.3_{(1.92)}$	$3.6_{(0.13)}$	10/61/29	36
	100/2/0	lasso	$8.4_{(0.69)}$	$11.8_{(0.76)}$	$4.7_{(0.15)}$	7/21/72	16
		enet	$7.6_{(0.62)}$	$15.7_{(1.07)}$	$3.1_{(0.13)}$	35/30/36	20
	100/2/0.5	lasso	$10.2_{(1.42)}$	$13.5_{(1.65)}$	$4.7_{(0.16)}$	20/17/62	12
		enet	$11.1_{(1.89)}$	$17.1_{(1.86)}$	$3.3_{(0.15)}$	46/18/36	12
	200/1/0	lasso	$5.2_{(0.47)}$	$10.4_{(0.60)}$	$3.9_{(0.14)}$	0/45/55	43
		enet	$2.7_{(0.32)}$	$25.0_{(1.51)}$	$2.5_{(0.09)}$	0/79/21	76
	200/1/0.5	lasso	$6.7_{(0.64)}$	$12.4_{(0.80)}$	$4.5_{(0.17)}$	0/40/60	40
		enet	$2.7_{(0.30)}$	$29.8_{(2.04)}$	$2.8_{(0.11)}$	0/82/18	70
	200/2/0	lasso	$7.4_{(1.10)}$	$13.3_{(1.06)}$	$4.0_{(0.14)}$	0/42/58	40
		enet	$6.0_{(1.31)}$	$19.7_{(1.70)}$	$3.6_{(0.13)}$	0/50/50	45
	200/2/0.5	lasso	$6.7_{(0.79)}$	$12.5_{(0.98)}$	$4.3_{(0.14)}$	0/38/62	32
		enet	$5.7_{(0.59)}$	$17.9_{(1.41)}$	$3.9_{(0.13)}$	7/44/48	30
4	100/1/0	lasso	$5.2_{(0.30)}$	$6.6_{(0.25)}$	$5.9_{(0.13)}$	7/14/78	14
		enet	$6.4_{(0.35)}$	$11.9_{(0.43)}$	$4.0_{(0.12)}$	64/ 8/28	8
	100/1/0.5	lasso	$6.2_{(0.40)}$	$8.3_{(0.47)}$	$5.4_{(0.14)}$	29/ 7/64	6
		enet	$8.9_{(0.75)}$	$13.1_{(0.75)}$	$3.0_{(0.13)}$	92/ 0/8	0
	100/2/0	lasso	$5.1_{(0.22)}$	$5.7_{(0.22)}$	$3.5_{(0.15)}$	81/ 2/17	2
		enet	$5.4_{(0.24)}$	$6.6_{(0.25)}$	$2.4_{(0.11)}$	98/ 0/2	0
	100/2/0.5	lasso	$5.9_{(0.24)}$	$6.6_{(0.26)}$	$2.6_{(0.13)}$	94/ 0/6	0
		enet	$6.5_{(0.28)}$	$7.4_{(0.29)}$	$1.7_{(0.11)}$	98/ 1/2	0
	200/1/0	lasso	$4.3_{(0.25)}$	$6.2_{(0.23)}$	$5.0_{(0.10)}$	18/22/60	22
		enet	$3.5_{(0.26)}$	$12.9_{(0.48)}$	$4.8_{(0.10)}$	14/36/50	36
	200/1/0.5	lasso	$4.3_{(0.29)}$	$7.3_{(0.35)}$	$5.7_{(0.11)}$	7/14/79	14
		enet	$5.7_{(0.37)}$	$14.7_{(0.68)}$	$4.3_{(0.15)}$	56/22/22	22
	200/2/0	lasso	$4.1_{(0.18)}$	$6.0_{(0.22)}$	$4.2_{(0.11)}$	45/ 9/46	8
		enet	$4.8_{(0.23)}$	$7.8_{(0.30)}$	$3.5_{(0.11)}$	67/ 4/29	4
	200/2/0.5	lasso	$5.1_{(0.23)}$	$6.7_{(0.23)}$	$3.6_{(0.12)}$	74/ 4/22	4
		enet	$5.6_{(0.26)}$	$8.2_{(0.28)}$	$2.7_{(0.12)}$	94/ 0/6	0

dimensional data, particularly, when grouped parameters exist, i.e., model 3. This is validated by the simulation studies through comparisons with the lasso and elastic net initial estimates, where ridge tuning parameter of 1 for elastic net is used throughout, whereas the tuning parameter for L_1 is chosen by the BIC both in the lasso and elastic net. Fixed choice for ridge parameter is unacceptable in practical data analysis, but it is only for illustrative purposes. We mention how to choose the ridge parameter, in practice, through credit scoring data example in Section 5.2. Our simulation studies validate all combinations of $n \in \{100, 200\}$, $\sigma \in \{1, 2\}$, and $\rho \in \{0, 0.5\}$ for models 3 and 4. Tables 5 and 6 show the results of numerical experiments repeated 200 times. Notably, grouping with the elastic net initial estimate is better than that with the lasso initial

Table 6

Results in large p small n example. $\hat{\theta}_{\text{ini}}$, initial estimator used; $PCF_j/PICF_j$, proportion of correctly fused with the other members in its group/proportion of incorrectly fused with the members in other groups (%), for indices js to be grouped, where the corresponding coefficients (nonzero only) under models 1 and 2 are presented; PC_0/PIC_0, proportion of correctly setting zero/proportion of incorrectly setting zero (%)

$n/\sigma/\rho$	$\hat{\theta}_{\text{ini}}$	$\text{PCF}_j/\text{PCIF}_j$, $j = 1, \ldots, 4$				PC_0/PIC_0
		$\beta_1^* = 4$	$\beta_2^* = 4$	$\beta_3^* = -4$	$\beta_4^* = -4$	
100/1/0	lasso	39/0	39/0	43/0	43/0	100/0
	enet	69/0	69/0	71/1	71/0	100/0
100/1/0.5	lasso	40/0	40/0	40/0	40/0	100/0
	enet	52/0	52/5	54/4	54/0	100/3
100/2/0	lasso	39/2	39/3	40/2	40/3	100/3
	enet	52/14	52/17	56/18	56/19	100/17
100/2/0.5	lasso	30/4	30/13	27/11	27/4	100/8
	enet	42/14	42/28	40/31	40/11	100/21
200/1/0	lasso	60/0	60/0	61/0	61/0	100/0
	enet	91/0	91/0	90/0	90/0	100/0
200/1/0.5	lasso	50/0	50/0	48/0	48/0	100/0
	enet	78/0	78/0	80/0	80/0	100/0
200/2/0	lasso	57/0	57/0	61/0	61/0	100/0
	enet	61/0	61/0	69/0	69/0	100/0
200/2/0.5	lasso	50/0	50/0	46/0	46/0	100/0
	enet	48/0	48/2	48/4	48/0	100/2
		$\beta_1^* = 4$	$\beta_2^* = -5$	$\beta_3^* = -2$	$\beta_4^* = 2$	
100/1/0	lasso	−/0	−/0	−/4	−/4	100/2
	enet	−/5	−/0	−/41	−/47	100/23
100/1/0.5	lasso	−/4	−/0	−/19	−/22	100/11
	enet	−/28	−/5	−/57	−/82	100/43
100/2/0	lasso	−/16	−/4	−/62	−/65	100/37
	enet	−/31	−/10	−/81	−/83	100/51
100/2/0.5	lasso	−/35	−/14	−/78	−/88	100/54
	enet	−/62	−/30	−/75	−/95	100/66
200/1/0	lasso	−/0	−/0	−/9	−/9	100/5
	enet	−/0	−/0	−/4	−/10	100/4
200/1/0.5	lasso	−/0	−/0	−/3	−/3	100/2
	enet	−/3	−/0	−/15	−/54	100/18
200/2/0	lasso	−/2	−/0	−/27	−/32	100/15
	enet	−/9	−/0	−/47	−/53	100/27
200/2/0.5	lasso	−/13	−/1	−/46	−/67	100/32
	enet	−/28	−/6	−/50	−/90	100/44

estimate in model 3, while the result reverses in most cases of model 4. The grouping effect of elastic net may explain this observation. Remarkably, from the fact that the ARAEs are less than those of initial estimate, the automatic grouping is likely to improve the estimation accuracy.

5. Application to real data

5.1. *Ohio wheeze data*

We first analyze the Ohio data which is a subset of the six cities study, a longitudinal study of the health effects of air pollution [22]. The dataset contains

Table 7

Results for Ohio wheeze data. Coef$_{\text{full}}$*, coefficients estimated under full model;* P_{full}*,* P*-value for the hypothesis that the coefficient is zero; Coef*$_{\text{AVS}}$*, coefficients estimated by Ueki's (2009) automatic variable selection; Group code, number of groups created by the proposed method ("Not grouped" indicates predictors out of interest of grouping); Coef*$_{\text{AG}}$*, coefficients estimated by the proposed automatic grouping*

Predictors	Coef$_{\text{full}}$	P_{full}	Coef$_{\text{AVS}}$	Group code	Coef$_{\text{AG}}$
$\text{age} = 7, \text{smoke} = 0$	0.13	0.45	0	0	0
$\text{age} = 8, \text{smoke} = 0$	0.05	0.78	0	0	0
$\text{age} = 10, \text{smoke} = 0$	−0.34	0.06	−0.4	1	−0.38
$\text{age} = 7, \text{smoke} = 1$	0.18	0.48	0	0	0
$\text{age} = 8, \text{smoke} = 1$	0.46	0.05	0.27	2	0.32
$\text{age} = 9, \text{smoke} = 1$	0.32	0.18	0	2	0.32
$\text{age} = 10, \text{smoke} = 1$	−0.03	0.9	0	0	0
Intercept	−1.8	<0.001	−1.69	Not grouped	−1.72

complete records on 537 children from Steubenville, Ohio, each of whom was examined annually at ages 7 through 10. This dataset was previously analyzed by Zeger, Liang and Albert [23] and Fitzmaurice and Laird [7]. The repeated binary response is the wheezing status (1 = yes, 0 = no) of a child at each occasion. Maternal smoking was categorized as 1 if the mother smoked regularly and 0 otherwise. Previous studies that treat the age as a continuous variable found weak effect of maternal smoking. We re-analyzed the data by treating the age as a qualitative variable for four categories of ages $7, 8, 9$, and 10 as in Fitzmaurice and Laird [7]. This strategy creates eight interactions between age and smoking status. Applying the generalized estimating equations with binomial model and exchangeable working correlation generated the result given in Table 7, where baseline is set to $I(\text{age} = 9, \text{smoke} = 0)$. P-values obtained imply weak difference from the baseline for $I(\text{age} = 10, \text{smoke} = 0)$ and $I(\text{age} = 8, \text{smoke} = 1)$. Variable selection of [18] selects the model having only these two variables. However, our automatic grouping additionally identifies an interaction $I(\text{age} = 9, \text{smoke} = 1)$, which has not ever been specified. The model obtained implies presence of interactions between maternal smoking and age different effect from baseline. Both Wald- and score-type losses in BIC leaded to the identical conclusion. This example points out that the grouping can detect variables that are missed by variable selection alone.

5.2. Credit scoring data

We apply the developed grouping method to the credit-scoring data analyzed in [5]. The dataset consists of 1000 consumers' credits from a southern German bank, and the aim is to model the probability that a client will not pay back the credit. The response variable is "creditability" which is given in dichotomous ($y = 0$ for creditworthy, $y = 1$ for not creditworthy), and 20 factors are available. We fit logistic regression model. According to [5], we analyzed seven risk factors, H_1: running account (trichotomous, no/good/bad), H_3: duration of credit in months (metrical), H_4: amount of credit in DM (metrical), H_5:

payment of previous credits (dichotomous, good/bad), H_6: intended use (dichotomous, private/professional), H_7 and H_8: dummies for gender and marital status (dichotomous, man/woman, and live alone/not live alone). We consider second order interactions only for categorical factors, H_1, H_5, H_6, H_7, and H_8 in the following way. Define dummy predictors that represent interaction between H_1 and H_5 by $X_1 = I(H_1 = \text{no}, H_5 = \text{bad}), X_2 = I(H_1 = \text{good}, H_5 = \text{good}), X_3 = I(H_1 = \text{good}, H_5 = \text{bad}), X_4 = I(H_1 = \text{bad}, H_5 = \text{good})$, and $X_5 = I(H_1 = \text{bad}, H_5 = \text{bad})$, thereby we have $(3 \times 2 - 1) = 5$ dummy predictors, in which $I(H_1 = \text{no}, H_5 = \text{good})$ is set to baseline. This strategy leads to $4 \times (3 \times 2 - 1) = 20$ dummy predictors of interactions (H_1, H_5), (H_1, H_6), and (H_1, H_7); Similarly we have $3 \times (2 \times 2 - 1) = 9$ dummy predictors of interactions $(H_5, H_6), (H_5, H_7)$, and (H_5, H_8); Consequently $20 + 9 + 6 + 3 = 38$ dummy predictors are created in total. The logistic regression model at the start, including H_3, H_4, and intercept, which are out of interest of grouping, is

$$\text{logit}\{P(y = 1|X)\} = \beta_0 + \sum_{j=1}^{38} X_j \beta_j + H_3 \beta_{39} + H_4 \beta_{40}.$$

Although an initial estimate is needed to apply our method, singularity incurred failure of standard logistic regression estimation, instead, the ridge-penalized logistic regression is utilized. Subsequently, smooth-thresholding applies to the ridge-penalized loglikelihood function of $L(\beta) = \ell(\beta) + \lambda_2 ||\beta||^2$, where λ_2 is the ridge penalty and ℓ is the loglikelihood function of the logistic regression model. Since different λ_2s lead to different models, we chose λ_2 that minimizes the BIC value of the final model resulted from grouping for given λ_2. This strategy resulted in the model given in Table 8, showing that three groups coded by 1, 2, and 3, were appeared; each group is interpreted as effective factors measuring creditworthy, not creditworthy, and less creditworthy, respectively, by considering magnitude of estimated coefficients. Predictors not shown in Table 8 are concluded to be irrelevant by the assignment of zero coefficient. Table 8 further provides the outputs of standard logistic regression based on the grouped predictors for form's sake, implying that three predictors are all significant although the P-value is not reliable because they are computed after model selection.

Appendix

Proof of Theorem 2.1

For the proof, Farkas' lemma is useful, which is familiar in convex analysis [see, e.g., 15].

Lemma A.1 (Farkas). *For given $n \times m$ matrix A and $x \in R^n$, the following statements (i) and (ii) are equivalent: (i). For any $\theta \in R^m$ such that $A\theta = x$, we have $\theta \geq 0$. (ii). For any $y \in R^n$ such that $A^T y \leq 0$, we have $x^T y \leq 0$. Here $\leq$ and $\geq$ for vectors means component-wise inequalities, i.e., $x \leq 0$ means that $x_j \leq 0$ holds for every component.*

Table 8

Results for credit scoring data. Survived predictors, interactions that are survived by the proposed automatic grouping; Group code, number of groups created by the proposed method ("Not grouped" indicates predictors out of interest of grouping; group 0 means the parameters set to zero); Coef$_{\mathrm{AG}}$*, coefficients estimated by the proposed method, where those in identical group are omitted; Coef*$_{\mathrm{SLR}}$*, recalculated parameter estimates by the standard logistic regression based on grouped predictors obtained from our method;* P_{SLR}*, output of* P*-value corresponding to Coef*$_{\mathrm{SLR}}$

Survived predictors	Group code	Coef$_{\mathrm{AG}}$	Coef$_{\mathrm{SLR}}$	P_{SLR}
$H_1 =$ good, $H_5 =$ good	1	−0.39	−0.37	<0.001
$H_1 =$ good, $H_6 =$ private	1			
$H_1 =$ good, $H_7 =$ man	1			
$H_1 =$ good, $H_7 =$ woman	1			
$H_1 =$ good, $H_8 =$ not live alone	1			
$H_1 =$ no, $H_5 =$ bad	2	0.7	0.76	0.003
$H_5 =$ bad, $H_6 =$ professional	2			
$H_1 =$ no, $H_6 =$ private	3	0.3	0.33	<0.001
$H_1 =$ no, $H_6 =$ professional	3			
$H_1 =$ bad, $H_7 =$ woman	3			
$H_1 =$ no, $H_7 =$ woman	3			
$H_1 =$ no, $H_8 =$ live alone	3			
$H_5 =$ bad, $H_7 =$ woman	3			
$H_5 =$ bad, $H_8 =$ live alone	3			
$H_6 =$ professional, $H_7 =$ man	3			
$H_6 =$ professional, $H_8 =$ live alone	3			
Intercept	Not grouped	−1.71	−1.82	<0.001
H_3	Not grouped	0.032	0.034	<0.001
H_4	Not grouped	0.000028	0.000029	0.38

To prove the theorem statement, it suffices to show that $(I+\hat{W})^{-1}e_j \geq 0$ for each $j = 1, \ldots, d$, where e_j is d-vector whose jth element is unity, zero otherwise. Farkas' lemma is applied for $\theta = A^{-1}x$ in which $A = I+\hat{W}$ and $x = e_j$. For any y such that $A^T y \leq 0$, we may only have to show that $x^T y \leq 0$, or equivalently, $y_j = e_j^T y \leq 0$.

Inequalities $A^T y = (I+\hat{W})y \leq 0$ are represented component-wisely as,

$$\left(1 + \hat{w}_{i0} + \sum_{k\geq 1, k\neq i}^{d} \hat{w}_{ik}\right) y_i - \sum_{k=1,k\neq i}^{d} \hat{w}_{ik} y_k \leq 0, \tag{A.1}$$

for $i = 1, \ldots, d$. Assume without loss of generality that y_1 is the maximum among $y_1, \ldots, y_d$ and let $a = \sum_{k>1}^{d} \hat{w}_{1k}$. It follows from (A.1) that

$$0 \geq (1 + \hat{w}_{10} + a)y_1 - \sum_{k>1}^{d} \hat{w}_{1k} y_k \geq (1+a)y_1 - y_1 a = y_1,$$

implying $y \leq 0$. The proof is completed.

Proof of Theorem 3.1

First we consider pairs that have a common parameter value. We abbreviate S_{full} simply as S. Let $\mathcal{B} = \{(j,k) \in S_2 : \theta_j^* = \theta_k^*, j < k\}$ where $S_2 = \{(j,k) \in$

$S \times S : j < k\}$, and let $\mathcal{M} = \{j \in S : \theta_j^* = 0\}$. First, by the triangle inequality we have $\{\lambda^{1/(1+\gamma)} < \max(\max_{(j,k)\in\mathcal{B}} |\hat\theta_j^{\text{ini}} - \hat\theta_k^{\text{ini}}|, \max_{j\in\mathcal{M}} |\hat\theta_j^{\text{ini}}|)\} \subset \{\lambda^{1/(1+\gamma)} < 2\max_{j\in S} |\hat\theta_j^{\text{ini}} - \theta_j^*|\}$. This argument and root-n consistency of $\hat\theta^{\text{ini}}$ yield that $P\{\min(\min_{(j,k)\in\mathcal{B}} \hat\delta_{jk}, \min_{j\in\mathcal{M}} \hat\delta_{0j}) < 1\} \leq P(\lambda^{1/(1+\gamma)} < 2\max_{j\in S} |\hat\theta_j^{\text{ini}} - \theta_j^*|) \leq \sum_{j\in S} P(|\hat\theta_j^{\text{ini}} - \theta_j^*| > 0.5\lambda^{1/(1+\gamma)}) \leq 0.5d\lambda^{1/(1+\gamma)}O(n^{-1/2})$. The right-hand side tends to zero as $\lambda n^{(1+\gamma)/2} \to \infty$.

Second we consider pairs that have different parameter values. Let $\mathcal{C} = \{(j,k) \in S_2 : \theta_j^* \neq \theta_k^*, j < k\}$, and let $\mathcal{N} = \{j \in S : \theta_j^* \neq 0\}$. We have for any $\epsilon > 0$ and, $P(\max_{(j,k)\in\mathcal{C}} \hat\delta_{jk} > n^{-1/2}\epsilon) = P(\lambda n^{1/2}/\epsilon > \min_{(j,k)\in\mathcal{C}} |\hat\theta_j^{\text{ini}} - \hat\theta_k^{\text{ini}}|^{1+\gamma}) \leq P\{(\lambda n^{1/2}/\epsilon)^{1/(1+\gamma)} > \min_{(j,k)\in\mathcal{C}} |\theta_j^* - \theta_k^*| - \max_{(j,k)\in\mathcal{C}} |\hat\theta_j^{\text{ini}} - \hat\theta_k^{\text{ini}} - \theta_j^* + \theta_k^*|\} \leq \sum_{j\in\mathcal{A}^*} \mathrm{P}\{2|\hat\theta_j^{\text{ini}} - \theta_j^*| > \min_{(j,k)\in\mathcal{C}} |\theta_j^* - \theta_k^*| - (\lambda n^{1/2}/\epsilon)^{1/(1+\gamma)}\}$. The right-hand side tends to zero as $\lambda n^{1/2} \to 0$ using root-n consistency of $\hat\theta^{\text{ini}}$. Thus $\max_{(j,k)\in\mathcal{C}} \hat\delta_{jk} = o_p(n^{-1/2})$. Similar argument concludes that $\max_{j\in\mathcal{N}} \hat\delta_{0j} = o_p(n^{-1/2})$. These in turn imply that $P\{\max(\max_{(j,k)\in\mathcal{C}} \hat\delta_{jk}, \max_{j\in\mathcal{N}} \hat\delta_{0j}) < 1\} \to 1$. Consequently, the statement is proved.

Proof of Theorem 3.2

Let $R^*_{\mathcal{A}^*}$ be $R_{\mathcal{A}}$ evaluated at $\mathcal{A}^*$ and R^* instead of $\mathcal{A}$ and R. From Theorem 3.1, the smooth-threshold estimating equations for $j \in \mathcal{A}^*$ coincide with (2.2) in which $R_{\mathcal{A}}$ is replaced by $R^*_{\mathcal{A}^*}$. Let $d_0 = |\mathcal{A}^*|$. By rearranging,

$$R^{*T}_{\mathcal{A}^*} u(R^*_{\mathcal{A}^*}\alpha_{\mathcal{A}^*}) + \hat W^* \alpha_{\mathcal{A}^*} = 0, \tag{A.2}$$

where $\hat W^*$ is the $d_0 \times d_0$ sub-matrix of $\hat W$ corresponding to the index set $\mathcal{A}^*$, and each components of $\hat W^*$ are finite. The first term corresponds to the oracle estimating equation, while the second term is the penalty term. Note that $\sum_{j,k\in\mathcal{A}^*} |(\hat W^*)_{jk}| \leq \sum_{j\in\mathcal{A}^*} |(\hat W^*)_{jj}| + \sum_{j,k\in\mathcal{A}^*, j\neq k} |(\hat W^*)_{jk}|$. Since it holds that $|(\hat W^*)_{jj}| = \hat w_{0j} + \sum_{k\in\mathcal{A}^*, k\neq j} \hat w_{jk} \leq w_{\max} d_0$, and that $|(\hat W^*)_{jk}| = \hat w_{jk} \leq w_{\max}$, in which $w_{\max} = \max_{j,k\in\mathcal{A}^*\cup\{0\}, j\neq k} \hat w_{jk}$. Therefore we have $\sum_{j,k\in\mathcal{A}^*} |(\hat W^*)_{jk}| \leq d_0^2 w_{\max} + d_0(d_0-1)w_{\max} = 2d_0^2 w_{\max}$. By $\hat w_{jk} = \hat\delta_{jk}/(1-\hat\delta_{jk})$ and the argument in the proof of theorem 3.1, we have $w_{\max} = o_p(n^{-1/2})$. Hence the $\hat W^*$ is asymptotically negligible; similar arguments can be found in [8]. Consequently we have the statement.

Acknowledgements

We would like to thank the Editor, the Associate Editor, and a referee for their helpful comments and suggestions which improve our presentation greatly.

References

[1] Bondell, H. D. and Reich, B. J. (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics* **65** 169–177. MR2665858

[2] Buckley, J. J. and James, I. R. (1979). Linear regression with censored data. *Biometrika* **66** 429–36.

[3] Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics* **24** 407–499. MR1394738

[4] Choi, N. H., Li, W. and Zhu, J. (2010). Variable seletion with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association* **105** 354–364. MR2656056

[5] Fahrmeier, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Model, 2nd Edition.* New York: Springer.

[6] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360. MR1946581

[7] Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80** 141–151.

[8] Fu, W. J. (2003). Penalized estimating equations. *Biometrics* **59** 126–32. MR1978479

[9] Hamada, M. and Wu, C. (1992). Analisis of designed experiments with complex aliasing. *Journal of Quality Technology* **24** 130–137.

[10] Jiang, W. and Liu, X. (2004). Consistent model selection based on parameter estimates. *Journal of Statistical Planning and Inference* **121** 265–283. MR2038821

[11] Johnson, B. A., Lin, D. Y. and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association* **103** 672–680. MR2435469

[12] Joseph, V. (2006). A Bayesian approach to the design and analysis of fractionated experiments. *Technometrics* **48** 219–229. MR2277676

[13] Lai, T. L. and Ying, Z. (1991). Large sample theory of a modified Buckley–James estimator for regression analysis with censored data. *Annals of Statistics* **19** 1370–1402. MR1126329

[14] Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430

[15] Rockafellar, R. T. (1979). *Convex Analysis.* Princeton University Press. MR1451876

[16] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58** 267–288. MR1379242

[17] Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer. MR2233926

[18] Ueki, M. (2009). A note on automatic variable selection using smooth-threshold estimationg equations. *Biometirka* **96** 1005–1011.

[19] van der Vaart, A. W. (1998). *Asymptotic Statistics*. New York: Cambridge University Press. MR1652247

[20] Wang, H. and Leng, C. (2007). Unified lasso estimation via least squares approximation. *Journal of the American Statistical Association* **102** 1039–48. MR2411663

[21] Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94** 553–68. MR2410008

[22] Ware, J. H., Dockery, D. W., Spiro, A. III, Speizer, F. E. and Fenis, B. G. Jr (1984). Passive smoking, gas cooking and respiratory health in children living in six cities. *American Review of Respiratory Disease* **129** 366–374.

[23] Zeger, S. L., Liang, K. Y. and Albert, P. A. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44** 1049–1060. MR0980999

[24] Zheng, X. and Loh, W. Y. (1995). Consistent variable selection in linear models. *Journal of the American Statistical Association* **90** 151–156. MR1325122

[25] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429. MR2279469

[26] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* **67** 301–320. MR2137327