Statistica Sinica **21** (2011), 659-678

MODEL SELECTION IN VALIDATION SAMPLING: AN ASYMPTOTIC LIKELIHOOD-BASED LASSO APPROACH

Chenlei Leng and Denis Heng-Yan Leung

National University of Singapore and Singapore Management University

Abstract: We propose an asymptotic likelihood-based LASSO approach for model selection in regression analysis when data are subject to validation sampling. The method makes use of an initial estimator of the regression coefficients and their asymptotic covariance matrix to form an asymptotic likelihood. This "working" objective function facilitates the formulation of the LASSO and the implementation of a fast algorithm. Our method circumvents the need to use a likelihood set-up that requires full distributional assumptions about the data. We show that the resulting estimator is consistent in model selection and that the method has lower prediction errors than a model that uses only the validation sample. Furthermore, we show that this formulation gives an optimal estimator in a certain sense. Extensive simulation studies are conducted for the linear regression model, the generalized linear regression model, and the Cox model. Our simulation results support our claims. The method is further applied to a dataset to illustrate its practical use.

Key words and phrases: Asymptotic likelihood-based LASSO, LASSO, least squares approximation, validation sampling.

1. Introduction

Non-standard sampling techniques are common in applied research. In studying the relationship between a response y and a set of covariates $\mathbf{x} \in \mathbb{R}^p$ via a parametric model $f(y; \mathbf{x}, \beta)$, the variables of interest can be so difficult or expensive to obtain that a cost-effective design must be implemented. Validation sampling designs, where the true response and covariate information is collected only on a subset and some proxy or surrogate measurements $(\tilde{y}, \tilde{\mathbf{x}})$ are available for all the subjects, arise naturally in this context. The subset with observations on $(y, \mathbf{x}, \tilde{y}, \tilde{\mathbf{x}})$ is usually referred to as the validation sample, whereas the subset with observations only on $(\tilde{y}, \tilde{\mathbf{x}})$ is referred to as the non-validation sample. The dimensionality of $\tilde{\mathbf{x}}$ is allowed to be different from that of $\mathbf{x}$. Surrogate response or covariates are commonly used in biomedical research. For example, Ellenberg and Hamilton (1989) and Wittes, Lakatos, and Probstfield (1989) discussed their application in clinical trials involving cancer and cardiovascular diseases; Alonzo, Pepe, and Lumley (2003) used them in mental health surveys.

Intuitively, crude proxy covariates that are correlated with the covariates of interest might help derive more efficient estimates for the parameters of interest, thus contributing to a better understanding of the relationship between the response and the covariates. The problem can be seen as a problem with missing data, a topic studied by Pepe (1992), Reilly and Pepe (1995), and Chen and Chen (2000), among others. Reviews of related methods can be found in Carroll, Ruppert, and Stefanski (1995) for auxiliary covariates, and in Leung (2001) for surrogate response. In this article, we only consider the situation where the validation sample is a simple random sample. The proposed method can be applied to other situations with some modifications; discussion of this is deferred to the conclusion of the paper.

If p is large, it would be useful to select a subset from $\mathbf{x}$ that explains most of the variation in the response without losing too much information. In other words, the interest is in determining the nonzero elements in β. The purpose of selecting the important covariates, or variable selection, is two-fold. First, identifying significant variables helps interpretation. Second, and maybe more importantly, the predictive performance of the fitted model can be improved by using only a subset of the covariates. Traditional variable selection tools usually suffer from instability and low accuracy (Breiman (1995)). But the recently developed penalized likelihood methods, noticeably the LASSO (Tibshirani (1996)) and the SCAD (Fan and Li (2001)), have been shown to produce competitive results. Their favorable performance can be attributed to the fact that variable selection and coefficient estimation are carried out simultaneously. Penalized likelihood methods have undergone rapid developments, for example in survival analysis (Tibshirani (1997); Fan and Li (2002)), semiparametric longitudinal data analysis (Fan and Li (2004)) and time series models (Wang, Li and Tsai (2007)). However, to our knowledge, variable selection under validation sampling has not been addressed.

The attractiveness of LASSO and other penalized likelihood methods in standard data situations naturally motivates their use in validation sampling situations. In order to use these methods, the likelihood or the loss function must be properly modeled. One method to model the likelihood is to use a fully parametric approach (Suh and Schafer (2002)), whereby the relationship between $(y, \mathbf{x}, \tilde{y}, \tilde{\mathbf{x}})$ is fully parametrized, but this approach is non-robust to model misspecification. Another approach is to model $(y, \mathbf{x})$ parametrically, while leaving the relationship between $(y, \mathbf{x}, \tilde{y}, \tilde{\mathbf{x}})$ unspecified and modelling it using a nonparametric method such as density estimation (*e.g.*, Pepe (1992) and Wang and Rao (2002)). The semi-parametric method is robust but is computationally intensive, especially when the dimension of $\mathbf{x}$ is high, which is precisely the situation of interest here.

Whether a parametric or semi-parametric approach is used to model the likelihood, the goal is to combine information about β from the validation and the non-validation samples. An alternative method (Chen and Chen (2000)) of combining information uses the regression estimation. The method assumes that information relating y and $\mathbf{x}$ is summarized in a set of estimating equations involving the unknown parameter β. There is a second set of estimating equations that summarizes the relationship between the surrogate response ($\tilde{y}$) and the surrogate covariates ($\tilde{\mathbf{x}}$) in terms of another set of parameters ϕ. The two sets of equations are individually solved using the data in the validation sample. If $\hat{\beta}$ and $\hat{\phi}$ represent the solutions to the equations, the estimate of β is the conditional mean of $\hat{\beta}$ given $\hat{\phi}$. We refer to this as the Chen-Chen method.

From a penalized model selection point of view, the Chen-Chen method has the estimator defined as the solution to a set of estimating equation, and, the usual loss function-based LASSO or SCAD are not applicable. We propose an asymptotic likelihood-based LASSO (ALL) method; it uses a preliminary estimate of β and its corresponding covariance matrix as input to a linear regression LASSO framework. The resulting estimate is consistent in model selection, and can be more efficient than the preliminary estimate; in particular, the proposed ALL estimator is shown to possess an asymptotic optimality property. In this paper, we discuss ALL estimators based on estimating equations from a linear regression model, a generalized linear model and a Cox model; the idea can be extended to other contexts.

The rest of this paper is organized as follows. Section 2 reviews the Chen-Chen estimator and the LASSO method. The asymptotic likelihood LASSO (ALL) is introduced in Section 3. Theoretical properties of the ALL estimator are given in Section 4. In Section 5, the finite sample behavior of the ALL estimator is compared to that of other estimators, using simulations. We also illustrate the practical application of the ALL estimator using a data set. Concluding remarks are given in Section 6. All proofs are relegated to the Appendix.

2. Background

2.1. The Chen-Chen estimator

Let V and $\bar{V}$ denote, respectively, the validation and the non-validation samples, both of which are simple random samples. Denote by $(y_i, \mathbf{x}_i, \tilde{y}_i, \tilde{\mathbf{x}}_i),\ i \in V$, the observations in V and $(\tilde{y}_j, \tilde{\mathbf{x}}_j),\ j \in \bar{V}$, the observations in $\bar{V}$. Let the sample size of V be n and the sample size of $\bar{V}$ be N, and $0 < \rho = \lim n/N < 1$ be the validation fraction. Suppose the conditional mean of y given $\mathbf{x}$ is $E(y|\mathbf{x}) = g(X;\beta)$. The true value of β is denoted by β^0. The goal is to estimate β using

the data in V and $\bar{V}$. Using only data in V, the estimating equation approach estimates β by solving

$$0 = \sum_{i \in V} \mathbf{s}_i(\beta) = \sum_{i \in V} \mathbf{w}_i \{y_i - g(\mathbf{x}_i; \beta), \}, \tag{2.1}$$

where $\mathbf{w}_i \in \mathbb{R}^p$ is a weight that may depend on $\mathbf{x}_i$. We denote the solution to (2.1) as $\hat{\beta}$. Since V is a random sample from the population, solving (2.1) gives an unbiased estimate of β. The estimating function reduces to the usual score function if the generalized linear model is assumed.

To improve the efficiency in estimating β, Chen and Chen (2000) proposed incorporating the data in $\bar{V}$ in the following way. Suppose for the moment that a working model relating $\tilde{y}$ and $\tilde{\mathbf{x}}$ is postulated, $E(\tilde{y}|\tilde{\mathbf{x}}) = h(\tilde{\mathbf{x}}; \gamma)$. We may estimate γ by solving

$$0 = \sum_{i \in V} \tilde{\mathbf{s}}_i(\gamma) = \sum_{i \in V} \tilde{\mathbf{w}}_i \{\tilde{y}_i - h(\tilde{\mathbf{x}}_i; \gamma)\},$$

where $\tilde{\mathbf{w}}_i \in \mathbb{R}^q$ is a weight that may depend on $\tilde{\mathbf{x}}_i$. The solution $\hat{\gamma}$ is a consistent estimate of γ that satisfies the moment condition $E[\tilde{\mathbf{w}}\{\tilde{y} - h(\tilde{\mathbf{x}}; \gamma)\}]=0$, where the expectation is taken with respect to γ. The asymptotic distribution of $n^{1/2}\{(\hat{\beta} - \beta^0)', (\hat{\gamma} - \gamma^0)'\}'$ can easily be shown to be $N(0, D^{-1}CD^{-1})$, where $D = \mathrm{diag}(D_1, D_2)$ and

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}.$$

These matrices are consistently estimated by

$$\hat{D}_1 = n^{-1} \sum_{i \in V} \frac{\partial \mathbf{s}_i(\hat{\beta})}{\partial \beta}, \ \hat{D}_2 = n^{-1} \sum_{i \in V} \frac{\partial \tilde{\mathbf{s}}_i(\hat{\gamma})}{\partial \gamma},$$

$$\hat{C} = n^{-1} \sum_{i \in V} \{\mathbf{s}_i'(\hat{\beta}), \tilde{\mathbf{s}}_i'(\hat{\gamma})\}' \{\mathbf{s}_i'(\hat{\beta}), \tilde{\mathbf{s}}_i'(\hat{\gamma})\} = \begin{pmatrix} n^{-1} \sum_{i \in V} \mathbf{s}_i(\hat{\beta}) \mathbf{s}_i'(\hat{\beta}) & n^{-1} \sum_{i \in V} \mathbf{s}_i(\hat{\beta}) \tilde{\mathbf{s}}_i'(\hat{\gamma}) \\ n^{-1} \sum_{i \in V} \tilde{\mathbf{s}}_i(\hat{\gamma}) \mathbf{s}_i'(\hat{\beta}) & n^{-1} \sum_{i \in V} \tilde{\mathbf{s}}_i'(\hat{\gamma}) \tilde{\mathbf{s}}_i(\hat{\gamma}) \end{pmatrix}.$$

Therefore the asymptotic conditional distribution of $n^{1/2}(\hat{\beta} - \beta^0)$ given $n^{1/2}(\hat{\gamma} - \gamma^0)$ is multivariate normal with mean $n^{1/2} D_1^{-1} C_{12} C_{22}^{-1} D_2 (\hat{\gamma} - \gamma^0)$.

To exploit the information in the non-validation data set, another estimate $\bar{\gamma}$ of γ can be obtained by solving

$$0 = \sum_{j \in \bar{V}} \tilde{\mathbf{s}}_j(\gamma) = \sum_{j \in \bar{V}} \tilde{\mathbf{w}}_j \{\tilde{y}_j - h(\tilde{\mathbf{x}}_j; \gamma)\}.$$

By replacing γ^0 by $\bar{\gamma}$ and equating $n^{1/2}(\hat{\beta} - \beta^0)$ to its conditional mean, an improved estimate of β is given by $\bar{\beta} = \hat{\beta} - \hat{D}_1^{-1}\hat{C}_{12}\hat{C}_{22}^{-1}\hat{D}_2(\hat{\gamma} - \bar{\gamma})$, where $n^{1/2}(\bar{\beta} - \beta^0)$ is asymptotically normal with an asymptotic covariance matrix

$$\Sigma = D_1^{-1}C_{11}D_1^{-1} - (1-\rho)D_1^{-1}C_{12}C_{22}^{-1}C_{12}'D_1^{-1}; \tag{2.2}$$

this can be consistently estimated by

$$\Sigma_n = \hat{D}_1^{-1}\hat{C}_{11}\hat{D}_1^{-1} - (1 - \frac{n}{N})\hat{D}_1^{-1}\hat{C}_{12}\hat{C}_{22}^{-1}\hat{C}_{12}'\hat{D}_1^{-1}. \tag{2.3}$$

It can be shown that the variance of $n^{1/2}\hat{\beta}$ is $D_1^{-1}C_{11}D_1^{-1}$, and that $D_1^{-1}C_{12}C_{22}^{-1}C_{12}'D_1^{-1}$ is non-negative definite. Therefore, using $\bar{\beta}$ instead of $\hat{\beta}$ leads to a reduction of variance in the estimate, which is dependent on the validation fraction ρ.

Chen (2002) extended Chen and Chen's method to survival data. Let T_i, δ_i be the (possibly censored) failure time and the censoring indicator, respectively, for the i-th subject. Let $y_i(t) = \mathrm{I}(T_i \geq t)$ be the at-risk process. Associated with each subject is a set of possibly time dependent covariates $\mathbf{x}(t)$. The notations $\tilde{T}$, $\tilde{\mathbf{x}}(t)$, $\tilde{y}(t) = \mathrm{I}(\tilde{T} \geq t)$ are defined as in Section 2, but we allow them to depend on time t. To estimate the unknown parameter, the standard score function of the partial likelihood can be used,

$$Q(\beta) = \sum_{i \in V} \delta_i \left\{ \mathbf{x}_i(T_i) - \frac{S^{(1)}(T_i, \beta)}{S^{(0)}(T_i, \beta)} \right\}, \tag{2.4}$$

where $S^{(p)}(\beta, \cdot) = n^{-1}\sum_{i \in V} y_i(\cdot)\exp\{\beta'\mathbf{x}_i(\cdot)\}\mathbf{x}_i(\cdot)^p,\ p = 0, 1$. Take the solution to (2.4) as $\hat{\beta}$. Finally, $\hat{\beta}$ can be updated by $\bar{\beta} = \hat{\beta} - \hat{D}_1^{-1}\hat{C}_{12}\hat{C}_{22}\hat{D}_2(\hat{\gamma} - \bar{\gamma})$, where $\hat{\gamma}$ and $\bar{\gamma}$ are solutions to

$$\tilde{Q}(\gamma) = \sum_{i \in V} \delta_i \left\{ \tilde{\mathbf{x}}_i(\tilde{T}_i) - \frac{\tilde{S}^{(1)}(\tilde{T}_i, \gamma)}{\tilde{S}^{(0)}(\tilde{T}_i, \gamma)} \right\}, \text{ and } \bar{Q}(\gamma) = \sum_{i \in \bar{V}} \delta_i \left\{ \tilde{\mathbf{x}}_i(\tilde{T}_i) - \frac{\bar{S}^{(1)}(\tilde{T}_i, \gamma)}{\bar{S}^{(0)}(\tilde{T}_i, \gamma)} \right\},$$

respectively, and the matrices $\hat{D}_1^{-1}$, $\hat{C}_{12}$, $\hat{C}_{22}$ and $\hat{D}_2$ can be estimated as in Chen (2002). The details are omitted. Here

$$\tilde{S}^{(p)}(\cdot, \gamma) = n^{-1}\sum_{i \in V} \tilde{y}_i(\cdot)\exp\{\gamma'\tilde{\mathbf{x}}_i(\cdot)\}\tilde{\mathbf{x}}_i(\cdot)^p,\ p = 0, 1,$$

$$\bar{S}^{(p)}(\cdot, \gamma) = N^{-1}\sum_{i \in \bar{V}} \tilde{y}_i(\cdot)\exp\{\gamma'\tilde{\mathbf{x}}_i(\cdot)\}\tilde{\mathbf{x}}_i(\cdot)^p,\ p = 0, 1.$$

Furthermore, $\sqrt{n}(\bar{\beta} - \beta)$ is asymptotically $N(0, \Omega)$, where Ω can be consistently estimated by a formula similar to (2.3).

2.2. The LASSO method

The least absolute shrinkage and selection operator (LASSO) introduced by Tibshirani (1996) has become an attractive approach to variable selection, as it permits simultaneous variable selection and parameter estimation. In a likelihood formulation, the LASSO estimator is the solution to the penalized likelihood objective function

$$l(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where $l(\beta)$ stands for a negative log-likelihood function or, more generally, a loss function, and λ is a tuning parameter that is usually determined by some information criterion, cross validation or one of its variants. In least squares regression, $l(\beta)$ is a quadratic function of β, and the whole solution path of β, denoted as $\hat{\beta}_\lambda$, can be effectively found with a computational complexity of a single ordinary least squares fit (Efron et al. (2004)). The fast computational algorithm referred to as the least angle regression (LARS) (Efron et al. (2004)) is largely responsible for the popularity of the LASSO. However, it is more difficult to develop fast implementation of the LASSO when $l(\beta)$ is a not a quadratic function, and various attempts have been made to identify the whole solution path (Rosset and Zhu (2007)). A related discussion can be found, for example, in Wang and Leng (2007), and references therein. From a computational point of view, it is attractive to develop a LASSO type of estimator in the context of validation sampling that effectively shares the computational advantage of the LARS algorithm.

A simple approach to implementing the LASSO method is to formulate a likelihood on the validation sample after discarding the surrogate data. However, such an analysis is guaranteed to give less efficient estimates. On the other hand, it is difficult or even impossible to write a complete likelihood using the validation and non-validation samples, as it requires the knowledge of several conditional distributions. This provides a strong motivation to develop alternative approaches to variable selection in this context.

3. The Asymptotic Likelihood-based LASSO

Based on the preliminary estimator $\bar{\beta}$ and a consistent estimator of its asymptotic covariance, we propose using the the asymptotic likelihood-based LASSO (Wang and Leng (2007)) that seeks to minimize

$$(\beta - \bar{\beta})' \Sigma_n^{-1} (\beta - \bar{\beta}) + \sum_{j=1}^{p} \lambda_j |\beta_j| \tag{3.1}$$

with respect to β. The λ_j are tuning parameters to be determined later. Following (Wang and Leng (2007)), λ_j can be replaced by $\lambda|\bar{\beta}_j|^{-\delta}$, where $\delta > 0$ (Zou (2006); Zhang and Lu (2007); Yuan and Lin (2007)). An advantage of this substitution is that we effectively reduce a p-dimensional optimization problem to one that is one-dimensional, which is important for easy implementation and fast computation. We fix $\delta = 1$, as suggested by Zou (2006). An important consideration in formulating the least squares optimization problem in (3.1) comes from the fact that the LARS algorithm developed by Efron et al. (2004) can be directly applied to obtain the whole solution path, which greatly facilitates the selection of a judicious tuning parameter. Indeed, Efron et al. (2004) showed that the solution $\hat{\beta}$ to (3.1) as a function of λ is piecewise linear with finite many connecting points. More remarkably, the connecting points can be identified via the LARS algorithm whose computational complexity is equivalent to that of an ordinary least squares fit. Thus, the implementation of this new approach is extremely easy. For convenience, we call formulation (3.1) the asymptotic likelihood-based LASSO, as $(\beta - \bar{\beta})'\Sigma_n^{-1}(\beta - \bar{\beta})$ comes from the asymptotic distribution of $\bar{\beta}$.

For choosing the tuning parameter λ, we propose the BIC criterion

$$\mathrm{BIC}_\lambda = (\bar{\beta}_\lambda - \bar{\beta})'\Sigma_n^{-1}(\bar{\beta}_\lambda - \bar{\beta}) + \log(N+n) \times \frac{df_\lambda}{(n+N)},$$

where $\bar{\beta}_\lambda$ is the estimate based on (3.1) when λ is used, and df_λ is the number of nonzero entries in $\bar{\beta}_\lambda$ (Zou, Hastie, and Tibshirani (2007)). The "effective" sample size in this approach is larger than n, since both the non-validation and the validation samples are used to determine $\bar{\beta}$. We use $N + n$ here because the validation sample is used twice.

4. Theoretical Properties

Without loss of generality, we assume that the nonzero index set is $\mathcal{A} = \{1, \ldots, d\}$, and the zero index set is $\mathcal{B} = \{d+1, \ldots, p\}$ such that $\beta^0 = (\beta_1^0, \ldots, \beta_d^0, 0, \ldots, 0)' = \big((\beta_{\mathcal{A}}^0)', (\beta_{\mathcal{B}}^0)'\big)'$. Furthermore, we partition $\bar{\beta}_\lambda$ as $(\bar{\beta}_{\lambda\mathcal{A}}, \bar{\beta}_{\lambda\mathcal{B}})$ and Σ as

$$\begin{pmatrix} \Sigma_{\mathcal{A}\mathcal{A}} & \Sigma_{\mathcal{A}\mathcal{B}} \\ \Sigma_{\mathcal{B}\mathcal{A}} & \Sigma_{\mathcal{B}\mathcal{B}} \end{pmatrix}.$$

Let $\sqrt{n}(\bar{\beta} - \beta^0) \to_p N(0, \Sigma)$, where Σ can be consistently estimated by Σ_n to be discussed later. Define $a_n = \max\{\lambda_j, j \le d\}$ and $b_n = \min\{\lambda_j, j > d\}$. Then we have the following.

Theorem 1. ($\sqrt{n}$-consistency) *If* $\sqrt{n}a_n \to_p 0$, *then* $\bar{\beta}_\lambda - \beta^0 = O_p(n^{-1/2})$.

Theorem 2. (Selection consistency) *If* $\sqrt{n}a_n \to_p 0$ *and* $\sqrt{n}b_n \to \infty$, *then* $Pr(\bar{\beta}_{\lambda\mathcal{B}} = 0) \to 1$.

Theorem 3. (Asymptotic normality) *If* $\sqrt{n}a_n \to_p 0$ *and* $\sqrt{n}b_n \to \infty$, *then* $\sqrt{n}(\bar{\beta}_{\lambda\mathcal{A}} - \beta^0_{\mathcal{A}}) \to N(0, \Gamma)$, *where* $\Gamma = \Sigma_{\mathcal{A}\mathcal{A}} - \Sigma_{\mathcal{A}\mathcal{B}}\Sigma^{-1}_{\mathcal{B}\mathcal{B}}\Sigma_{\mathcal{B}\mathcal{A}}$.

For $\lambda_j = \lambda|\bar{\beta}_j|^{-1}$, as long as λ satisfies $\sqrt{n}\lambda \to 0$ and $n\lambda \to \infty$, the conditions in Theorem1-3 are satisfied. In practice, to estimate the variance of $\bar{\beta}_{\lambda\mathcal{A}}$, we propose to use the sample version of Γ,

$$\Gamma_n = \Sigma_{n\mathcal{A}\mathcal{A}} - \Sigma_{n\mathcal{A}\mathcal{B}}\Sigma^{-1}_{n\mathcal{B}\mathcal{B}}\Sigma_{n\mathcal{B}\mathcal{A}}. \tag{4.1}$$

Theorems 1 and 2 still hold if we replace Σ_n by any stochastic positive definite matrix $S_n \in \mathbb{R}^{p\times p}$, as long as $S_n \to_p S$, where S is positive definite. The proofs follow trivially from those of Theorem 1 and 2 in Wang and Leng (2007). However, the choice Σ_n is asymptotically optimal in the following sense.

Theorem 4. (Asymptotic optimality) *Assume* $\sqrt{n}a_n \to_p 0$ *and* $\sqrt{n}b_n \to \infty$. *Let* $\check{\beta}_\lambda$ *be the solution to*

$$(\beta - \bar{\beta})' S_n (\beta - \bar{\beta}) + \sum_{j=1}^{p} \lambda_j |\beta_j|, \tag{4.2}$$

where $S_n \in \mathbb{R}^{p\times p}$ *is a positive definite matrix and* $S_n \to_p S$ *for a positive definite matrix* S. *Then,* $\sqrt{n}(\check{\beta}_{\lambda\mathcal{A}} - \beta^0_{\mathcal{A}}) \to N(0, \Lambda)$, *and* $\Lambda - \Gamma$ *is non-negative definite. Additionally,* $\Lambda = \Gamma$ *when* $S_n = \Sigma_n^{-1}$.

Remark 1. If we take S_n to be an identity matrix, the resulting estimator becomes a version of the soft-thresholding estimator. The jth element of the estimator in this case is easily seen as to be $\mathrm{sgn}(\bar{\beta}_j)(|\bar{\beta}_j| - \lambda_j/2)_+$, where $(s)_+ = s$ if $s > 0$ and 0 otherwise. The rate of convergence and the selection consistency is not affected, but the asymptotic variance becomes $\Sigma_{\mathcal{A}\mathcal{A}}$. According to Theorem 4, $\Sigma_{\mathcal{A}\mathcal{A}} - \Gamma$ is non-negative definite. Therefore, the ALL estimator gives smaller asymptotic variance than the soft-thresholding estimator.

The proposed BIC criterion can be shown to be consistent in terms of variable selection. The proof is similar to that of Theorem 4 in Wang and Leng (2007), and is omitted.

5. Simulations and Data Analysis

We conducted extensive numerical studies. The purpose was to compare the proposed approach with a few alternatives in terms of variable selection and estimation accuracy. The ALL method was further applied to a child psychopathology data set to illustrate its usefulness. To implement the ALL method, we made use of existing routines for fitting the linear regression model or any other mode to extract the estimating equations. After obtaining $\bar{\beta}$ and its asymptotic covariance Σ_n, we formed a LASSO objective function (3.1) and exploited the LARS

algorithm to compute the whole solution of $\hat{\beta}_\lambda$. The optimal λ was subsequently identified by BIC. The implementation of the ALL method was extremely easy and computationally fast.

5.1. Simulation studies

We compared the ALL estimator with the naive estimator (2.1) using the validation sample alone and without variable selection (VS), with LASSO based on (2.1) with only the validation sample (LASSO), and with Chen-Chen's estimator without variable selection (CC). For the first simulation study, we also included three estimator suggested by anonymous referees: the soft thresholding estimator by setting $S_n = I$ in (4.2) (ST); the covariance shrinkage estimator by setting $S_n = (aI + (1-a)\Sigma_n)^{-1}$ for $a \in \mathbb{R}$ in (4.2) (CS); a hybrid estimator by using $\hat{\beta}$ in (3.1) (HE). For simplicity, we took $a = 0.1$ for the estimator. Though we could have treated a as an additional tuning parameter in addition to λ. The oracle estimate is the estimate obtained using the Chen-Chen's estimator and excluding the zero covariates.

Two simulation studies were done. For each, 500 simulations were used for each scenario considered. For each estimator, β_{est}, we report the model error, $\mathrm{ME}(\beta_{est}) = (\beta_{est} - \beta^0)' E(\mathbf{x}\mathbf{x}')(\beta_{est} - \beta^0)$, the median relative model error (MRME) compared to the oracle, the average model size (MS), and the percentage of correct models identified (CM). Here the oracle estimator refers to the estimator one would use if the true sparsity pattern was known in advance (Fan and Li (2001)).

Simulation Study 1. For the first simulation study, we considered a linear model

$$y = \mathbf{x}'\beta + \sigma\epsilon,$$

where $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ and $\epsilon \sim N(0,1)$. We set $\sigma = 2,\ 4$, and $\mathbf{x}$ such that each covariate $\mathbf{x}_i$ was standard normal and the correlation between two covariates $\mathbf{x}_i$ and $\mathbf{x}_j$ was $(0.5)^{|i-j|}$. The non-validation sample was generated such that $\tilde{y} = y$ and $\tilde{\mathbf{x}}_i = \mathbf{x}_i + e_i$, where (a) $e \sim N(0, 0.5^2)$ or (b) $\log(e) \sim N(-0.5, 0.5)$. The non-validation sample size was taken to be $N = 50,\ 200,\ 500$ or 1,000, and the validation sample size was $n = 25,\ 50$ or 100. The working model for γ was the usual linear regression model. The MRME results are summarized in Table 1, and the MS and CM results for LASSO and ALL are presented in Table 2. The MS and CM results for ST, CS, and HE are omitted to save space.

A few observations can be made from Tables 1 and 2. First, ALL generally outperformed other methods in terms of MRME. Second, in terms of variable selection, ALL was generally better than LASSO when the error variance was high ($\sigma = 4$). When the error variance was small ($\sigma = 2$), the variable selection

Table 1. Median relative model errors for Simulation Study 1. Standard errors (in parentheses) are estimated via 500 bootstrap replications.

σ/e	N/n	ALL	LASSO	CC	VS	ST	CS	HE
4/(a)	50/25	$3.92_{(0.25)}$	$3.90_{(0.19)}$	$4.88_{(0.33)}$	$5.56_{(0.29)}$	$5.08_{(0.31)}$	$3.97_{(0.25)}$	$4.17_{(0.23)}$
	200/50	$2.47_{(0.10)}$	$4.46_{(0.23)}$	$3.83_{(0.19)}$	$6.24_{(0.28)}$	$3.73_{(0.22)}$	$2.92_{(0.18)}$	$4.66_{(0.26)}$
	200/100	$1.54_{(0.06)}$	$2.43_{(0.18)}$	$3.07_{(0.10)}$	$4.32_{(0.19)}$	$2.60_{(0.18)}$	$1.94_{(0.10)}$	$2.56_{(0.16)}$
	500/50	$2.40_{(0.16)}$	$5.09_{(0.31)}$	$3.88_{(0.25)}$	$6.80_{(0.33)}$	$3.87_{(0.28)}$	$2.99_{(0.18)}$	$4.95_{(0.25)}$
	500/100	$1.57_{(0.07)}$	$3.62_{(0.19)}$	$3.60_{(0.20)}$	$6.94_{(0.38)}$	$2.98_{(0.21)}$	$2.02_{(0.17)}$	$4.01_{(0.17)}$
	1000/50	$2.34_{(0.14)}$	$5.92_{(0.41)}$	$4.16_{(0.22)}$	$7.69_{(0.40)}$	$4.13_{(0.30)}$	$3.10_{(0.24)}$	$5.21_{(0.42)}$
	1000/100	$1.49_{(0.05)}$	$4.19_{(0.33)}$	$3.33_{(0.15)}$	$6.89_{(0.36)}$	$2.52_{(0.16)}$	$1.92_{(0.17)}$	$4.23_{(0.27)}$
/(b)	50/25	$3.40_{(0.15)}$	$3.98_{(0.28)}$	$4.74_{(0.25)}$	$6.12_{(0.31)}$	$4.91_{(0.26)}$	$3.57_{(0.19)}$	$4.44_{(0.23)}$
	200/50	$2.01_{(0.10)}$	$5.42_{(0.29)}$	$3.57_{(0.16)}$	$7.38_{(0.36)}$	$3.17_{(0.16)}$	$2.33_{(0.12)}$	$5.55_{(0.39)}$
	200/100	$1.62_{(0.08)}$	$3.13_{(0.17)}$	$3.15_{(0.16)}$	$5.31_{(0.25)}$	$2.32_{(0.14)}$	$1.88_{(0.11)}$	$3.12_{(0.15)}$
	500/50	$1.96_{(0.10)}$	$7.70_{(0.40)}$	$3.85_{(0.21)}$	$9.94_{(0.41)}$	$3.52_{(0.18)}$	$2.62_{(0.17)}$	$7.71_{(0.52)}$
	500/100	$1.48_{(0.06)}$	$4.81_{(0.37)}$	$3.50_{(0.18)}$	$8.57_{(0.44)}$	$2.73_{(0.21)}$	$1.92_{(0.13)}$	$5.31_{(0.31)}$
	1000/50	$1.93_{(0.10)}$	$8.11_{(0.63)}$	$4.12_{(0.32)}$	$12.13_{(0.62)}$	$3.78_{(0.26)}$	$2.61_{(0.16)}$	$8.22_{(0.73)}$
	1000/100	$1.34_{(0.07)}$	$6.49_{(0.41)}$	$3.57_{(0.13)}$	$10.73_{(0.57)}$	$2.48_{(0.11)}$	$1.91_{(0.12)}$	$7.02_{(0.39)}$
2/(a)	50/25	$2.46_{(0.20)}$	$2.49_{(0.11)}$	$4.88_{(0.27)}$	$4.38_{(0.24)}$	$4.44_{(0.27)}$	$3.49_{(0.15)}$	$2.95_{(0.16)}$
	200/50	$1.79_{(0.06)}$	$2.28_{(0.13)}$	$3.72_{(0.17)}$	$4.85_{(0.32)}$	$2.91_{(0.16)}$	$2.38_{(0.12)}$	$2.63_{(0.12)}$
	200/100	$1.36_{(0.04)}$	$1.68_{(0.09)}$	$3.21_{(0.14)}$	$4.05_{(0.23)}$	$1.88_{(0.10)}$	$1.50_{(0.07)}$	$1.83_{(0.10)}$
	500/50	$1.73_{(0.11)}$	$2.37_{(0.13)}$	$3.80_{(0.17)}$	$5.18_{(0.26)}$	$2.71_{(0.13)}$	$2.16_{(0.21)}$	$2.71_{(0.20)}$
	500/100	$1.40_{(0.05)}$	$1.87_{(0.09)}$	$3.41_{(0.12)}$	$4.76_{(0.25)}$	$1.97_{(0.10)}$	$1.63_{(0.08)}$	$2.13_{(0.12)}$
	1000/50	$1.91_{(0.09)}$	$2.23_{(0.17)}$	$3.91_{(0.22)}$	$5.30_{(0.35)}$	$2.68_{(0.18)}$	$2.31_{(0.16)}$	$2.96_{(0.16)}$
	1000/100	$1.26_{(0.04)}$	$1.73_{(0.09)}$	$3.09_{(0.19)}$	$4.71_{(0.22)}$	$1.78_{(0.09)}$	$1.52_{(0.05)}$	$2.10_{(0.13)}$
/(b)	50/25	$2.95_{(0.17)}$	$2.84_{(0.23)}$	$4.33_{(0.26)}$	$4.87_{(0.25)}$	$3.69_{(0.28)}$	$2.91_{(0.18)}$	$3.48_{(0.25)}$
	200/50	$1.66_{(0.07)}$	$2.42_{(0.13)}$	$3.52_{(0.18)}$	$5.43_{(0.33)}$	$2.23_{(0.13)}$	$1.66_{(0.07)}$	$3.11_{(0.14)}$
	200/100	$1.29_{(0.03)}$	$1.76_{(0.09)}$	$2.99_{(0.12)}$	$3.99_{(0.19)}$	$1.78_{(0.08)}$	$1.51_{(0.05)}$	$1.92_{(0.09)}$
	500/50	$1.61_{(0.08)}$	$2.88_{(0.18)}$	$3.72_{(0.15)}$	$5.75_{(0.35)}$	$2.38_{(0.16)}$	$1.86_{(0.09)}$	$3.73_{(0.20)}$
	500/100	$1.21_{(0.03)}$	$2.16_{(0.15)}$	$3.18_{(0.12)}$	$5.71_{(0.23)}$	$1.77_{(0.09)}$	$1.51_{(0.05)}$	$2.73_{(0.18)}$
	1000/50	$1.76_{(0.07)}$	$3.28_{(0.20)}$	$3.82_{(0.17)}$	$7.13_{(0.34)}$	$2.68_{(0.16)}$	$2.10_{(0.12)}$	$4.19_{(0.20)}$
	1000/100	$1.26_{(0.04)}$	$2.60_{(0.11)}$	$3.24_{(0.16)}$	$6.11_{(0.29)}$	$1.75_{(0.07)}$	$1.47_{(0.07)}$	$3.03_{(0.16)}$

performances of ALL and LASSO were comparable. When $n = 25$ and $N = 50$, ALL outperformed LASSO for $\sigma = 4$ but was comparable when $\sigma = 2$. Third, the accuracy of ALL was higher when N or n was large. Fourth, LASSO performed better than CC when the error variance was small, but the reverse was true when the error variance was large. In addition, ST generally outperformed LASSO, but was inferior to ALL. CS, although inferior to ALL, was the second best in terms of performance. HE used a less efficient estimate of β, which was outperformed by ALL and CS in general. These observations confirm the theoretical results that higher efficiency can be achieved when the more efficient estimator of β and the optimal covariance matrix are used.

From Table 2, we note that the model selection results of ALL and LASSO were comparable when $\sigma = 2$. However, ALL was better than LASSO in model

Table 2. Average model sizes (MS) and proportions of the model which are correctly identified (CM) for Simulation Study 1.

				ALL			LASSO
σ	e	n		$N = 200$	$N = 500$	$N = 1{,}000$	
4	(a)	50	CM	44.8	46.8	50.6	37.0
			MS	3.46	3.40	3.34	3.03
		100	CM	71.0	78.0	82.2	65.2
			MS	3.23	3.22	3.16	3.18
	(b)	50	CM	54.0	57.2	60.8	34.4
			MS	3.39	3.50	3.43	3.01
		100	CM	70.4	79.0	82.2	57.8
			MS	3.27	3.22	3.19	3.22
2	(a)	50	CM	63.4	65.2	68.0	71.2
			MS	3.52	3.48	3.42	3.35
		100	CM	81.2	82.4	87.2	82.8
			MS	3.23	3.20	3.15	3.18
	(b)	50	CM	67.8	68.6	69.0	72.4
			MS	3.42	3.50	3.44	3.31
		100	CM	85.4	88.4	89.6	83.8
			MS	3.17	3.13	3.12	3.18

selection for $\sigma = 4$ because the non-validation sample was also used in ALL, an intuitive result.

Simulation Study 2. In this study, we considered two error-in-variables models. For the first, the true model was

$$g(\mathbf{x}) = \Pr(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{x}'\beta)},$$

where $\beta = (3, 0, 0, 1.5, 0, 0, 2, 0, 0)'$ and $\mathbf{x}$ was generated according to the first simulation study. The proxy variable $\tilde{\mathbf{x}}_i$ followed a normal distribution with mean $\mathbf{x}_i$ and variance $= 0.5^2$. Therefore, the non-validation sample had observations on $(y, \tilde{\mathbf{x}})$ and the validation sample had observations on $(y, \mathbf{x}, \tilde{\mathbf{x}})$. To estimate γ, we used a generalized linear model $h(\tilde{\mathbf{x}}; \gamma) = 1/\{1 + \exp(\tilde{\mathbf{x}}'\gamma)\}$ as the working model.

The second case was a surrogate response problem, where y was generated according to

$$y = \mathbf{x}'\beta + \epsilon, \tag{5.1}$$

with $\beta = (3, 0, 0, 1.5, 0, 0, 2, 0, 0)'$, $\epsilon \sim N(0, 3^2)$ and $\mathbf{x}$ was generated as in the first simulation study. The surrogate variable $\tilde{y}$ was binary with $\Pr(\tilde{y} = 1|y) = 1/\{1 + \exp(-3y)\}$. The non-validation data consisted of $(\tilde{y}, \mathbf{x})$ and the validation data consisted of $(y, \mathbf{x})$. We estimated β via ordinary least squares and γ using the

Table 3. Median relative model errors for Simulation Study 2.

Case	N	n	ALL	LASSO	CC	VS
(a)	1,000	100	2.90 (0.31)	10.50 (1.08)	5.41 (0.43)	6.88 (0.47)
		200	1.91 (0.18)	2.93 (0.33)	2.56 (0.12)	4.01 (0.24)
		400	1.81 (0.13)	2.13 (0.15)	2.28 (0.10)	3.10 (0.18)
	2,000	100	2.83 (0.21)	10.97 (1.00)	4.67 (0.30)	6.41 (0.47)
		200	2.25 (0.19)	3.23 (0.23)	2.68 (0.13)	4.01 (0.22)
		400	1.80 (0.20)	2.55 (0.20)	2.20 (0.08)	3.28 (0.19)
(b)	1,000	100	1.49 (0.06)	2.40 (0.15)	3.56 (0.14)	4.75 (0.22)
		200	1.23 (0.03)	1.61 (0.07)	3.35 (0.16)	4.64 (0.21)
		400	1.13 (0.03)	1.41 (0.05)	3.13 (0.17)	4.28 (0.22)
	2,000	100	1.48 (0.07)	2.08 (0.10)	3.70 (0.22)	5.12 (0.21)
		200	1.20 (0.03)	1.84 (0.07)	3.36 (0.13)	4.63 (0.16)
		400	1.08 (0.03)	1.47 (0.06)	2.94 (0.10)	4.26 (0.26)

Table 4. Standard deviations of estimators for Simulation Study 2 ($n = 400$).

		β_1		β_4		β_7	
Case	N	SD	SD_m	SD	SD_m	SD	SD_m
(a)	1,000	0.310	0.294	0.205	0.189	0.261	0.221
	2,000	0.269	0.273	0.224	0.175	0.224	0.204
(b)	1,000	0.145	0.134	0.139	0.131	0.140	0.131
	2,000	0.139	0.130	0.116	0.127	0.151	0.125

score function in the generalized linear model, that is, $h(\tilde{\mathbf{x}}; \gamma) = 1/\{1+\exp(\tilde{\mathbf{x}}'\gamma)\}$. The non-validation sample size was either $N = 2{,}000$ or 4,000, and the validation sample size was either $n = 200$ or $n = 400$. The two cases are denoted as (a) and (b) subsequently. From Table 3, we see that ALL outperformed other methods significantly in estimation accuracy. We also tested the accuracy of the asymptotic variance matrix in (4.1). The median absolute deviation divided by 0.6745, denoted by SD in Table 4, of 500 estimated coefficients in the 500 simulations can be regarded as the true standard error (Fan and Li (2001)). The median of the 500 estimated SD's, denoted by SD_m, measures the overall performance of the variance formula in (4.1). Although the formula (4.1) slightly underestimated the true standard deviation, it can be seen that the extent of underestimation was small. Thus, we conclude that the asymptotic covariance matrix estimation was satisfactory.

Simulation study 3. We considered an extension of the double sampling idea in the Cox model. Here independent observations were generated according to $h(t_i|\mathbf{x}_i) = \exp(\mathbf{x}_i'\beta)$, where t_i is the i-th subject's survival time and $\beta = (0.8, 0, 0, 1, 0, 0, 0.6, 0)'$. Again, $\mathbf{x}_i$ was generated in the same manner as in the previous simulation study. Furthermore, independent censoring times were

Table 5. Median relative model errors for Simulation Study 3.

Case	N	n	ALL	LASSO	CC	VS
(a)	200	100	1.95 (0.09)	2.17 (0.15)	2.83 (0.12)	3.74 (0.14)
		200	1.35 (0.06)	1.35 (0.06)	2.33 (0.08)	2.33 (0.08)
	500	100	1.95 (0.11)	2.85 (0.16)	2.91 (0.13)	4.65 (0.19)
		200	1.40 (0.08)	2.01 (0.11)	2.48 (0.09)	3.35 (0.15)
	1,000	100	2.21 (0.13)	2.79 (0.17)	3.22 (0.15)	4.79 (0.27)
		200	1.64 (0.08)	2.15 (0.14)	2.71 (0.11)	3.79 (0.22)
(b)	200	100	1.88 (0.09)	1.94 (0.14)	2.83 (0.19)	3.35 (0.17)
		200	1.33 (0.06)	1.36 (0.06)	2.51 (0.09)	2.51 (0.09)
	500	100	1.96 (0.11)	2.15 (0.14)	2.77 (0.12)	3.38 (0.19)
		200	1.45 (0.08)	1.56 (0.08)	2.45 (0.11)	2.82 (0.12)
	1,000	100	1.82 (0.08)	2.09 (0.10)	2.76 (0.11)	3.61 (0.17)
		200	1.44 (0.05)	1.75 (0.09)	2.54 (0.11)	3.34 (0.16)

generated from an exponential distribution with mean $u\exp\{\mathbf{x}_i'\beta\}$, where u was uniform on $[1,3]$. The censoring mechanism gave about 30% censored data. The non-validation sample was generated such that $\tilde{\mathbf{x}} = (\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_6, \mathbf{x}_7, \mathbf{x}_8)'$, a subset of the validation covariates. The value of $\tilde{T}$ was generated according to (a) $\tilde{T} = T$; or (b) $\tilde{T} = T + \varepsilon$, where $\varepsilon \sim N(0,1)$. For this simulation, some covariates were completely missing. The results are summarized in Table 5. We observe from Table 5 that the proposed ALL approach dominated other estimators in estimation accuracy. In terms of variable selection, the ALL method performed better than LASSO (results not shown).

5.2. Data analysis

We applied the proposed method to a dataset in child psychopathology. The data was from a study that was carried out between 1986-1989 in Eastern Connecticut, USA (Zhaner et al. (1993)). A goal of the study was to examine the geographical variation of psychopathology in rural-urban children. We focused on one of the main measures of psychopathology in children: behavioral ("externalizing") disturbances. In the study, the subjects (children) were first identified. One of the parents or the primary care provider was then contacted to provide a report on the child. With the consent of the parents and the school board, the child's teacher was also approached for rating. Information from each parent/primary care provider was gathered using the Child Behavioral Checklist (CBCL) (Achenbach and Edelbrock (1983)) while that from the teacher was obtained using the Teacher's Report Form (TRF) (Achenbach and Edelbrock (1986)). With over 3,500 published studies using the CBCL as of August 2001, the CBCL and TRF are arguably the most widely used measure in child psychopathology. Both CBCL and TRF are continuous scales ranging from 1 to 100,

with a higher score on either scale indicating more severe disturbance. CBCL and TRF scales were obtained on externalizing disturbance. The primary response was the rating of externalizing disturbance provided by the TRF. The parent's rating (CBCL rating) was used as the surrogate.

In total, 2,519 children, aged 6 - 11, were studied, with $2,501$ complete parent (CBCL) reports. The missing parent reports were due to un-scorable forms. There were $1,433$ children with complete teacher reports (TRFs). Most of the missing teacher reports were due to permission denied by the parents and/or the school board. The rest were due to non-response. As discussed in Horton and Lipsitz (2001), missingness of this magnitude is not uncommon in large surveys. The non-validation data therefore consist of the parents ratings and the covariates while the validation data consist of the parents and teachers ratings as well as the covariates. Based on the analysis in Leung and Qin (2006), we can assume that the two samples are random samples from the population.

There were eight covariates in this study: child's sex (CSEX: 1=boy, 0=girl); area (AREA: 1=cities, 0= rural-suburban); social economic status (SES: 1= high, 2=middle, 3=low); single mother (MOMSING: 1=yes, 0=no); mother distress (MOMSTRS: 1=yes, 0=no); family distress (FAMSTRS: 1=yes, 0=no); child with health problems (HLTHPRO: 1=yes, 0=no); child with academic problems (ACADPRO: 1=yes, 0=no). For AREA, "Large cities" and "Small cities" were combined as "Cities", and "Suburban" and "Rural" became "Rural-suburban". For SES, two binary dummy variables were created, with "High" as the baseline. Therefore, after re-coding, there were a total of nine binary covariates.

The mean externalizing rating in the teacher reports (TXEXT) was 50.9 (range 39-89), and that in the parent reports (PXEXT) was 49.3 (range 30-89). A summary of the raw data is given in Table 6. Linear regression was used to model the conditional means of TXEXT and PXEXT. Residuals (not shown) from the linear models based on either TXEXT or PXEXT showed no departure from normality.

We applied the four approaches to this data set. The solutions paths of ALL and LASSO, where the individual coefficients of $\bar{\beta}$ as functions of λ were plotted against the L_1 norm of $\bar{\beta}$, were presented in Figure 1, whereas the coefficient estimates were given in Table 7. The paths are obviously piecewise linear as the ALL estimator uses a quadratic objective function. LASSO and ALL gave similar estimates for AREA4, SES3 and ACADPRO. However, LASSO selected one more important variable, MOMSTRS, compared to ALL. Note that ALL is based on the Chen-Chen method, which also gave an insignificant MOMSTRS coefficient; while LASSO is based on the validation data alone, and the validation-only coefficient estimator for MOMSTRS was significant. In this example, the gain in using the surrogate is not that great, as evidenced by the similar values of

Table 6. Summary statistics for data in the psychopathology study.

Parameter	Total	%	Validation sample	non-validation sample
AREA				
Cities	1,199	47.9	725	577
Rural-Suburbs	1,302	52.1	708	491
SES				
High	1,240	49.6	728	512
Middle	949	37.9	528	421
Low	312	12.5	177	135
MOMSING				
No	1,982	79.2	1,161	821
Yes	519	20.8	272	247
MOMSTRS				
No	2,110	84.4	1,199	911
Yes	391	15.6	234	157
HLTHPRO				
No	1,329	53.1	735	594
Yes	1,172	46.9	698	474
ACADPRO				
No	1,594	63.7	917	677
Yes	907	36.3	516	391
CSEX				
Girl	1,294	51.7	726	568
Boy	1,207	48.3	707	500
FAMSTRS				
No	905	36.2	515	390
Yes	1,596	63.8	918	678
PXEXT				
Mean(SD)			48.97 (10.13)	49.6 (10.49)

the parameter standard errors in the analysis with/without the surrogate data. This result is due to a relatively low correlation between the teacher's and the parents' ratings ($r = 0.37$). However, as pointed out by Kraemer et al. (2003), this type of correlation is common in studies of child psychopathology using multiple informants.

6. Conclusion

We have proposed a new asymptotic likelihood-based LASSO (ALL) for vari-

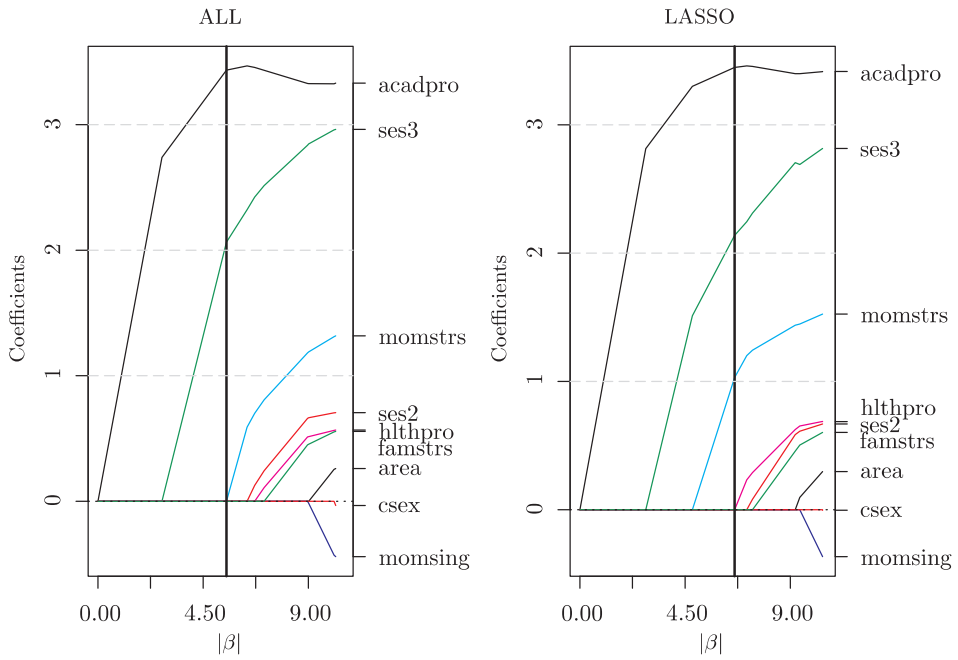

Figure 1. The solution paths for ALL and LASSO. The vertical lines indicate the fit via BIC.

Table 7. Analysis of the psychopathology data.

Variable	ALL	LASSO	CC	VS
INTERCEPT	48.14 (0.32)	48.10 (0.35)	47.01 (0.58)	46.91 (0.63)
AREA	-	-	0.27 (0.59)	0.30 (0.59)
SES2	-	-	0.71 (0.60)	0.67 (0.60)
SES3	2.06 (0.84)	2.14 (0.81)	2.96 (1.00)	2.82 (0.96)
MOMSING	-	-	-0.44 (0.84)	-0.36 (0.78)
MOMSTRS	-	1.03 (0.72)	1.32 (0.81)	1.52 (0.74)
HLTHPRO	-	-	0.57 (0.53)	0.69 (0.54)
ACADPRO	3.44 (0.56)	3.45 (0.56)	3.33 (0.59)	3.42 (0.57)
CSEX	-	-	-0.03 (0.53)	-0.00 (0.54)
FAMSTRS	-	-	0.56 (0.54)	0.60 (0.56)

able selection and coefficient estimation when data are collected via validation sampling. In contrast to other regularized model selection methods, such as the conventional LASSO and SCAD, the method requires neither a likelihood function nor a loss function. With simple preliminary estimates of the coefficients and their covariance matrix, the method is consistent in model selection. The ALL estimator has smaller asymptotic variance than Chen-Chen's estimator without variable selection, and that ALL is asymptotically optimal under the conditions

of our Theorem 4. In the finite sample situations we studied, ALL was superior to its competitors in all dimensions considered. The ALL estimator is very general and can be applied to other areas where a likelihood formulation is difficult.

In the paper, we took the validation and non-validation samples to be random subsets of the population. In practice, it is possible that the validation sample be selected using biased sampling techniques, where the probability of selection is dependent on some observable covariates. Thus if $u(\mathbf{x}_i) = u_i$ is the probability of selection of the i-th observation, and assuming that it can be estimated consistently by the data, then (2.1) is replaced by

$$0 = \sum_{i \in V} \mathbf{s}_i(\beta) = \sum_{i \in V} \frac{1}{u_i} \mathbf{w}_i \{y_i - g(\mathbf{x}_i; \beta)\}.$$

Furthermore, take

$$0 = \sum_{i \in V} \tilde{\mathbf{s}}_i(\gamma) = \sum_{i \in V} \frac{1}{u_i} \tilde{\mathbf{w}}_i \{\tilde{y}_i - h(\tilde{\mathbf{x}}_i; \gamma)\},$$
$$0 = \sum_{j \in \bar{V}} \tilde{\mathbf{s}}_j(\gamma) = \sum_{j \in \bar{V}} \frac{1}{1 - u_j} \tilde{\mathbf{w}}_j \{\tilde{y}_j - h(\tilde{\mathbf{x}}_j; \gamma)\},$$

where in this case $h(\tilde{\mathbf{x}}_j; \gamma)$ can be estimated by a method such as the inverse probability weighting method (Horvitz and Thompson (1952)). Then our method proceeds as before. Methods other than Chen and Chen's can also be used; see, for example, Chen, Leung, and Qin (2008), and the references therein. We will explore this direction in a future paper. In this paper, we have mainly focused on fixed dimensional problems with $p < n$. How to extend the current methodology to high dimensional data is another interesting topic.

Appendix: Proofs

The proofs of Theorem 1 and Theorem 2 follow straightforwardly from Wang and Leng (2007). Here we outline the proofs for Theorems 3 and 4.

Proof of Theorem 3. According to Theorem 2, with probability one, $\bar{\beta}_{\lambda\mathcal{B}} = 0$. Since $\bar{\beta}_\lambda$ is the minimizer of the objective function

$$(\beta - \bar{\beta})' \Sigma_n^{-1} (\beta - \bar{\beta}) + \sum \lambda_j |\beta_j|,$$

taking partial derivatives with respect to $\beta_\mathcal{A}$ shows that $\bar{\beta}_{\lambda\mathcal{A}}$ satisfies (Wang and Leng (2007))

$$\Omega_{\mathcal{A}\mathcal{A}}(\beta_\mathcal{A} - \bar{\beta}_\mathcal{A}) - \Omega_{\mathcal{A}\mathcal{B}} \bar{\beta}_\mathcal{B} + D(\bar{\beta}_\mathcal{A}) = 0, \tag{A.1}$$

where $\Omega = \Sigma_n^{-1}$ and $D(\bar{\beta}_{\mathcal{A}})$ is a d-dimensional vector with the jth component given by $(1/2)\lambda_j \mathrm{sgn}(\beta_j)$. Note that $\sqrt{n}\lambda_j \mathrm{sgn}(\beta_j) = o_p(1)$ as $\sqrt{n}a_n \to_p 0$. Therefore,

$$\sqrt{n}(\beta_{\lambda\mathcal{A}} - \beta^0_{\mathcal{A}}) = \sqrt{n}(\bar{\beta}_{\mathcal{A}} - \beta^0_{\mathcal{A}}) + \Omega^{-1}_{\mathcal{A}\mathcal{A}}\Omega_{\mathcal{A}\mathcal{B}}(\sqrt{n}\bar{\beta}_{\mathcal{B}}) + o_p(1).$$

Since $\Omega = \Sigma^{-1}$, it is easy to see that $\Omega^{-1}_{\mathcal{A}\mathcal{A}}\Omega_{\mathcal{A}\mathcal{B}} = -\Sigma_{n\mathcal{A}\mathcal{B}}\Sigma^{-1}_{n\mathcal{B}\mathcal{B}}$. Thus, we can write

$$\sqrt{n}(\beta_{\lambda\mathcal{A}} - \beta^0_{\mathcal{A}}) = \sqrt{n}(\bar{\beta}_{\mathcal{A}} - \beta^0_{\mathcal{A}}) - \Sigma_{n\mathcal{A}\mathcal{B}}\Sigma^{-1}_{n\mathcal{B}\mathcal{B}}(\sqrt{n}\bar{\beta}_{\mathcal{B}}) + o_p(1).$$

The asymptotic distribution of $\bar{\beta}_{\lambda\mathcal{A}}$ follows by noting that the asymptotic distribution of $\bar{\beta}$ is $N(0, \Sigma)$ and that $\Sigma_n \to_p \Sigma$.

Proof of Theorem 4. We partition S_n as

$$\begin{pmatrix} S_{n\mathcal{A}\mathcal{A}} & S_{n\mathcal{A}\mathcal{B}} \\ S_{n\mathcal{B}\mathcal{A}} & S_{n\mathcal{B}\mathcal{B}} \end{pmatrix}.$$

Denote the estimator as $\check{\beta}_{\lambda}$. It follows from the proof of Theorem 3 that

$$\sqrt{n}(\check{\beta}_{\lambda\mathcal{A}} - \beta^0_{\mathcal{A}}) = \sqrt{n}(\bar{\beta}_{\mathcal{A}} - \beta^0_{\mathcal{A}}) + S^{-1}_{n\mathcal{A}\mathcal{A}}S_{n\mathcal{A}\mathcal{B}}(\sqrt{n}\bar{\beta}_{\mathcal{B}}) + o_p(1).$$

Since $\sqrt{n}(\bar{\beta} - \beta^0)$ follows $N(0, \Sigma)$, therefore, $\sqrt{n}(\check{\beta}_{\lambda\mathcal{A}} - \beta^0_{\mathcal{A}}) \to_d N(0, \Lambda)$, where $\Lambda = \Sigma_{\mathcal{A}\mathcal{A}} + S^{-1}_{\mathcal{A}\mathcal{A}}S_{\mathcal{A}\mathcal{B}}\Sigma_{\mathcal{B}\mathcal{B}}S_{\mathcal{B}\mathcal{A}}S^{-1}_{\mathcal{A}\mathcal{A}} + 2S^{-1}_{\mathcal{A}\mathcal{A}}S_{\mathcal{A}\mathcal{B}}\Sigma_{\mathcal{B}\mathcal{A}}$. We can write

$$\Lambda = \Gamma + (S^{-1}_{\mathcal{A}\mathcal{A}}S_{\mathcal{A}\mathcal{B}} + \Sigma_{\mathcal{A}\mathcal{B}}\Sigma^{-1}_{\mathcal{B}\mathcal{B}})\Sigma_{\mathcal{B}\mathcal{B}}(S^{-1}_{\mathcal{A}\mathcal{A}}S_{\mathcal{A}\mathcal{B}} + \Sigma_{\mathcal{A}\mathcal{B}}\Sigma^{-1}_{\mathcal{B}\mathcal{B}})'.$$

Therefore $\Lambda - \Gamma$ is a non-negative definite matrix for any S, and $\Lambda = \Gamma$ if $S^{-1}_{\mathcal{A}\mathcal{A}}S_{\mathcal{A}\mathcal{B}} = -\Sigma_{\mathcal{A}\mathcal{B}}\Sigma^{-1}_{\mathcal{B}\mathcal{B}}$, which is satisfied by taking $\Sigma = S^{-1}$ and, correspondingly, $S_n = \Sigma_n^{-1}$.

References

Achenbach, T. M. and Edelbrock, C. S. (1983). *Manual for the Child Behavior Checklist and the Revised Child Behavior Profile.* University of Vermont Department of Psychiatry, Burlington, VT.

Achenbach, T. M. and Edelbrock, C. S. (1986). *Manual for the Teacher Report Form and Teacher Version of the Child Profile.* University of Vermont Department of Psychiatry, Burlington, VT.

Alonzo, T. and Pepe, M. S. and Lumley, T. (2003). Estimating disease prevalence in two-phase studies. *Biostatistics* **4**, 313-326.

Breiman, L. (1995). Better subset regression using the non-negative garrote. *Technometrics* **37**, 373-384.

Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models.* Chapman and Hall, London.

Chen, Y. (2002). Cox regression in cohort studies with validation sampling. *J. Roy. Statist. Soc. Ser. B* **64**, 51-62.

Chen, Y. and Chen, H. (2000). A unified approach to regression analysis under double-sampling designs. *J. Roy. Statist. Soc. Ser. B* **62**, 449-460.

Chen, S. X., Leung, D. and Qin, J. (2008). Improving semiparametric estimation using surrogate data. *J. Roy. Statist. Soc. Ser. B* **70**, 803-823.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32**, 407-451.

Ellenberg, S. and Hamilton, J. (1989). Surrogate endpoints in clinical trials: Cancer. *Statist. Medicine* **8**, 405-413.

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

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30**, 74-99.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Amer. Statist. Assoc.* 99, 710-723.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47**, 663-685.

Horton, N. and Lipsitz, S. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *Amer. Statist.* **55**, 244-254.

Kraemer, H. C., Jeffrey R. M., Jennifer C. A., Marilyn J. E., Boyce, W. T. and Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *Amer. J. Psychiatry* **160**, 1566-1577.

Leung, D. H. Y. (2001). Statistical methods for clinical studies in the presence of surrogate end points. *J. Roy. Statist. Soc. Ser A* **164**, 485-503.

Leung, D. H. Y. and Qin, J. (2006). Empirical likelihood based estimation for surrogate covariate data. *Appl. Statist.* **55**, 379-376.

Pepe, M. S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika* **79**, 355-365.

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* 82, 299-314.

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *Ann. Statist.* **35**, 1012-1030.

Suh, E. Y. and Schafer, D. W. (2002). Semiparametric maximum likelihood for nonlinear regression with measurement errors. *Biometrics* 58, 448-453.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statist. Medicine* **16**, 385-395.

Wang, H. and Leng, C. (2007). Unified LASSO estimation via least squares approximation. *J. Amer. Statist. Assoc.* **102**, 1039-1048.

Wang, H., Li, G. and Tsai, C. L. (2007). Regression coefficients and autoregressive order shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B* **69**, 63-78.

Wang, Q. and Rao, J. N. K. (2002). Empirical likelihood-based inference in linear errors-in-covariables models with validation data. *Biometrika* **89**, 345-358.

Wittes, J., Lakatos, E. and Probstfield, J. (1989). Surrogate endpoints in clinical trials: Cardiovascular disease. *Statist. Medicine* **8**, 415-425.

Yuan, M. and Lin, Y. (2007). On the nonnegative garrote estimator. *J. Roy. Statist. Soc. Ser. B* **69**, 143-161.

Zahner, G., Jacobs, J., Freeman, D. H. and Trainor, K. F. (1993). Rural-urban children psychopathology in a Northeastern US state: 1986-1989. *J. Amer. Academy of Child and Adolescent Psychiatry* **32**, 378-387.

Zhang, H. H. and Lu, W. (2007). Adaptive-LASSO for Cox's proportional hazard model. *Biometrika* **94**, 691-703

Zou, H. (2006). The adaptive LASSO and its oracle properties. *J. Amer. Statist. Assoc.* 101, 1418–1429.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the degrees of freedom of the LASSO. *Ann. Statist.* **35**, 2173-2192.

Department of Statistics and Applied Probability, National University of Singapore, Singapore.

E-mail: stalc@nus.edu.sg

School of Economics, Singapore Management University, Singapore.

E-mail: denisleung@smu.edu.sg

(Received March 2009; accepted October 2009)