The Annals of Statistics
2012, Vol. 40, No. 3, 1403–1429
DOI: 10.1214/12-AOS1017

NONCONCAVE PENALIZED COMPOSITE CONDITIONAL LIKELIHOOD ESTIMATION OF SPARSE ISING MODELS[1]

By Lingzhou Xue, Hui Zou and Tianxi Cai

University of Minnesota, University of Minnesota and Harvard University

The Ising model is a useful tool for studying complex interactions within a system. The estimation of such a model, however, is rather challenging, especially in the presence of high-dimensional parameters. In this work, we propose efficient procedures for learning a sparse Ising model based on a penalized composite conditional likelihood with nonconcave penalties. Nonconcave penalized likelihood estimation has received a lot of attention in recent years. However, such an approach is computationally prohibitive under high-dimensional Ising models. To overcome such difficulties, we extend the methodology and theory of nonconcave penalized likelihood to penalized composite conditional likelihood estimation. The proposed method can be efficiently implemented by taking advantage of coordinate-ascent and minorization–maximization principles. Asymptotic oracle properties of the proposed method are established with NP-dimensionality. Optimality of the computed local solution is discussed. We demonstrate its finite sample performance via simulation studies and further illustrate our proposal by studying the Human Immunodeficiency Virus type 1 protease structure based on data from the Stanford HIV drug resistance database. Our statistical learning results match the known biological findings very well, although no prior biological information is used in the data analysis procedure.

1. Introduction. The Ising model was first introduced in statistical physics [Ising (1925)] as a mathematical model for describing magnetic interactions and the structures of ferromagnetic substances. Although rooted in physics, the Ising model has been successfully exploited to simplify complex interactions for network exploration in various research fields such as social-economics [Stauffer (2008)], protein modeling [Irback, Peterson and Potthast

Received September 2011; revised May 2012.
[1]Supported in part by NSF Grants DMS-08-46068 and DMS-08-54970.
AMS 2000 subject classifications. Primary 62G20, 62P10; secondary 90-08.
Key words and phrases. Composite likelihood, coordinatewise optimization, Ising model, minorization–maximization principle, NP-dimension asymptotic theory, HIV drug resistance database.

(1996)] and statistical genetics [Majewski, Li and Ott (2001)]. Following the terminology in physics, consider an Ising model with K magnetic dipoles denoted by X_j, $1 \le j \le K$. Each X_j equals $+1$ or -1, corresponding to the up or down spin state of the jth magnetic dipole. The energy function is defined as $E = -\sum_{i \neq j} \beta_{ij} \frac{X_i X_j}{4}$, where the coupling coefficient β_{ij} describes the physical interactions between dipoles i and j under the external magnetic field, $\beta_{ii} = 0$ and $\beta_{ij} = \beta_{ji}$ for any (i, j). According to Boltzmann's law, the joint distribution of $\mathbf{X} = (X_1, \ldots, X_K)$ should be

$$\Pr(X_1 = x_1, \ldots, X_K = x_K) = \frac{1}{Z(\boldsymbol{\beta})} \exp\left(\sum_{(i,j)} \frac{\beta_{ij} x_j x_i}{4}\right), \tag{1.1}$$

where $Z(\boldsymbol{\beta})$ is the partition function.

In this paper we focus on learning sparse Ising models; that is, many coupling coefficients are zero. Our research is motivated by the HIV drug resistance study where understanding the inter-residue couplings (interactions) could potentially shed light on the mechanisms of drug resistance. A suitable statistical learning method is to fit a sparse Ising model to the data, in order to discover the inter-residue couplings. More details are given in Section 5. In the recent statistical literature, penalized likelihood estimation has become a standard tool for sparse estimation. See a recent review paper by Fan and Lv (2010). In principle we can follow the penalized likelihood estimation paradigm to derive a sparse penalized estimator of the Ising model. Unfortunately, the penalized likelihood estimation method is very difficult to compute under the Ising model because the partition function $Z(\boldsymbol{\beta})$ is computationally intractable when the number of dipoles is relatively large. On the other hand, the composite likelihood idea [Lindsay (1988), Varin, Reid and Firth (2011)] offers a nice alternative. To elaborate, suppose we have N independent identically distributed (i.i.d.) realizations of $\mathbf{X}$ from the Ising model, denoted by $\{(x_{1n}, \ldots, x_{Kn}), n = 1, \ldots, N\}$. Let $\theta_j = P(X_i = x_j | \mathbf{X}_{(-j)})$, describing the conditional distribution of the jth dipole given the remaining dipoles, where $\mathbf{X}_{(-j)}$ denotes $\mathbf{X}$ with the jth element removed. By (1.1), it is easy see that for the nth observation,

$$\theta_{jn} = \frac{\exp(\sum_{k:\, k \neq j} \beta_{jk} x_{jn} x_{kn})}{\exp(\sum_{k:\, k \neq j} \beta_{jk} x_{jn} x_{kn}) + 1}.$$

Note that θ_{jn} does not involve the partition function. The conditional log-likelihood of the jth dipole, given the remaining dipoles, is given by

$$\ell^{(j)} = \frac{1}{N} \sum_{n=1}^{N} \log(\theta_{jn}).$$

As in Lindsay (1988) a composite log-likelihood function can be defined as

$$\ell_c = \sum_{j=1}^{K} \ell^{(j)}.$$

This kind of composite conditional likelihood was also called pseudo-likelihood in Besag (1974). Another popular type of composite likelihood is composite marginal likelihood [Varin (2008)]. Maximum composite likelihood is especially useful when the full likelihood is intractable. Such an approach has important applications in many areas including spatial statistics, clustered and longitudinal data and time series models. A nice review on the recent developments in composite likelihood can be found in Varin, Reid and Firth (2011).

To estimate a high-dimensional sparse Ising model, we consider the following penalized composite likelihood estimator:

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left\{ \ell_c(\boldsymbol{\beta}) - \sum_{j=1}^{K} \sum_{k=j+1}^{K} P_\lambda(|\beta_{jk}|) \right\}, \tag{1.2}$$

where $P_\lambda(t)$ is a positive penalty function defined on $[0, \infty)$. In this work we focus primarily on the LASSO penalty [Tibshirani (1996)] and smoothly clipped absolute deviation (SCAD) penalty [Fan and Li (2001)]. The LASSO penalty is $P_\lambda(t) = \lambda t$. The SCAD penalty is defined by

$$P'_\lambda(t) = \lambda \left\{ I(t \le \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right\}, \qquad t \ge 0; a > 2.$$

Following Fan and Li (2001) we set $a = 3.7$. We should make it clear that when $P_\lambda(t)$ is nonconcave, $\widehat{\boldsymbol{\beta}}$ should be understood as a good local maximizer of (1.2). See discussions in Section 2.

The optimization problem in (1.2) is very challenging because of two major issues: (1) the number of unknown parameters is $\frac{1}{2}K(K-1)$, and hence the optimization problem is high dimensional in nature; and (2) the penalty function is concave and nondifferentiable at zero, although ℓ_c is a smooth concave function. We propose to combine the strengths of coordinate-ascent and minorization–maximization, which results in two new algorithms, CMA and LLA–CMA, for computing a local solution of the nonconcave penalized composite likelihood. See Section 2 for details. With the aid of the new algorithms, the SCAD penalized estimators are able to enjoy computational efficiency comparable to that of the LASSO penalized estimator.

Fan and Li (2001) advocated the oracle properties of the nonconcave penalized likelihood estimator in the sense that it performs as well as the oracle estimator which is the hypothetical maximum likelihood estimator knowing the true submodel. Zhang (2010a) and Lv and Fan (2009) were among the first to study the concave penalized least-squares estimator with NP-dimensionality (p can grow faster than any polynomial function of n). Fan and Lv (2011) studied the asymptotic properties of nonconcave penalized likelihood for generalized linear models with NP-dimensionality. In this paper we show that the oracle model selection theory remains to hold nicely for nonconcave penalized composite likelihood with NP-dimensionality. Furthermore, we show that under certain regularity conditions the oracle estimator can be attained asymptotically via the LLA–CMA algorithm.

There is some related work in the literature. Ravikumar, Wainwright and Lafferty (2010) viewed the Ising model as a binary Markov graph and used a neighborhood LASSO-penalized logistic regression algorithm to select the edges. Their idea is an extension of neighborhood selection by LASSO regression proposed by Meinshausen and Bühlmann (2006) for estimating Gaussian graphical models. Höfling and Tibshirani (2009) suggested using the LASSO-penalized pseudo-likelihood to estimate binary Markov graphs. However, they did not provide any theoretical result nor application. In this paper we compare the LASSO and the SCAD penalized composite likelihood estimators and show the latter has substantial advantages with respect to both numerical and theoretical properties.

The rest of this paper is organized as follows. In Section 2, we introduce the CMA and LLA–CMA algorithms. The statistical theory is presented in Section 3. Monte Carlo simulation results are shown in Section 4. In Section 5 we present a real application of the proposed method to study the network structure of the amino-acid sequences of retroviral proteases using data from the Stanford HIV drug resistance database. Technical proofs are relegated to the Appendix.

2. Computing algorithms. In this section we discuss how to efficiently implement the penalized composite likelihood estimators. As mentioned before, the computational challenges come from (1) penalizing the concave composite likelihood with a nonconcave penalty which is not differentiable at zero; (2) the intrinsically high dimension of the unknown parameters. Zou and Li (2008) proposed the local linear approximation (LLA) algorithm to derive an iterative ℓ_1-optimization procedure for computing nonconcave penalized estimators. The basic idea behind LLA is the minorization–maximization principle [Lange, Hunter and Yang (2000), Hunter and Lange (2004), Hunter and Li (2005)]. Coordinate-ascent (or descent) algorithms [Tseng (1988)] have been successfully used for solving penalized estimators with LASSO-type penalties; see, for example, Fu (1998), Daubechies, Defrise and De Mol (2004), Genkin, Lewis and Madigan (2007), Yuan and Lin (2006), Meier, van de Geer and Bühlmann (2008), Wu and Lange (2008) and Friedman, Hastie and Tibshirani (2010). In this paper we combine the strengths of minorization–maximization and coordinatewise optimization to overcome the computational challenges.

2.1. *The CMA algorithm.* Let $\widetilde{\boldsymbol{\beta}}$ be the current estimate. The coordinate-ascent algorithm sequentially updates $\widetilde{\beta}_{ij}$ by solving the following univariate optimization problem:

$$\widetilde{\beta}_{jk} \Leftarrow \arg\max_{\beta_{jk}} \{\ell_c(\beta_{jk}; \beta_{j'k'} = \widetilde{\beta}_{j'k'}, (j', k') \neq (j,k)) - P_\lambda(|\beta_{jk}|)\}. \tag{2.1}$$

However, we do not have a closed-form solution for the maximizer of (2.1). The exact maximization has to be conducted by some numerical optimization routine, which may not be a good choice in the coordinate-ascent algorithm because the maximization routine needs to be repeated many times to reach convergence. On the other hand, one can find an update to increase, rather than maximize, the objective function in (2.1), maintaining the crucial ascent property of the coordinate-ascent algorithm. This idea is in line with the generalized EM algorithm [Dempster, Laird and Rubin (1977)] in which one seeks to increase the expected log likelihood in the M-step.

First, we observe that for any β_{ij}

$$\frac{\partial^2 \ell_c(\boldsymbol{\beta})}{\partial \beta_{jk}^2} = -\frac{1}{N}\sum_{n=1}^{N}(\theta_{kn}(1-\theta_{kn}) + \theta_{jn}(1-\theta_{jn})) \geq -\frac{1}{2}. \tag{2.2}$$

Thus, by Taylor's expansion, we have

$$\ell_c(\beta_{jk}; \beta_{j'k'} = \widetilde{\beta}_{j'k'}, (j', k') \neq (j,k)) \geq Q(\beta_{jk}),$$

where

$$\begin{aligned} Q(\beta_{jk}) \equiv{}& \ell_c(\beta_{jk} = \widetilde{\beta}_{jk}; \beta_{j'k'} = \widetilde{\beta}_{j'k'}, (j', k') \neq (j,k)) \\ &+ \widetilde{z}_{jk}(\beta_{jk} - \widetilde{\beta}_{jk}) - \tfrac{1}{4}(\beta_{jk} - \widetilde{\beta}_{jk})^2, \end{aligned} \tag{2.3}$$

$$\widetilde{z}_{jk} = \frac{\partial \ell_c(\boldsymbol{\beta})}{\partial \beta_{jk}}\bigg|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} = \frac{1}{N}\sum_{n=1}^{N} x_{kn}x_{jn}(2 - \theta_{kn}(\widetilde{\beta}) - \theta_{jn}(\widetilde{\beta})). \tag{2.4}$$

Next, Zou and Li (2008) showed that

$$P_\lambda(|\beta_{jk}|) \leq P_\lambda(|\widetilde{\beta}_{jk}|) + P'_\lambda(|\widetilde{\beta}_{jk}|)\cdot(|\beta_{jk}| - |\widetilde{\beta}_{jk}|) \equiv L(|\beta_{jk}|). \tag{2.5}$$

Combining (2.3)–(2.5) we see that $Q(\beta_{jk}) - L(|\beta_{jk}|)$ is a minorization function of the objective function in (2.1). We update $\widetilde{\beta}_{jk}$ by

$$\widetilde{\beta}_{jk}^{\text{new}} = \arg\max_{\beta_{jk}} \{Q(\beta_{jk}) - L(|\beta_{jk}|)\}, \tag{2.6}$$

whose solution is given by $\widetilde{\beta}_{jk}^{\text{new}} = S(\widetilde{\beta}_{jk} + 2\widetilde{z}_{jk}, 2P'_\lambda(|\widetilde{\beta}_{jk}|))$ where $S(r,t) = \operatorname{sgn}(r)(|r| - t)_+$ denotes the soft-thresholding operator [Tibshirani (1996)]. The above arguments lead to Algorithm 1 below, which we call the coordinate-minorization-ascent (CMA) algorithm.

Algorithm 1 The CMA algorithm

(1) Initialization of $\widetilde{\boldsymbol{\beta}}$.

(2) Cyclic coordinate-minorization-ascent: sequentially update $\widetilde{\beta}_{ij}$ ($1 \leq j < k \leq K$) via soft-thresholding $\widetilde{\beta}_{jk} \Leftarrow S(\widetilde{\beta}_{jk} + 2\widetilde{z}_{jk}, 2P'_\lambda(|\widetilde{\beta}_{jk}|))$.

(3) Repeat the above cycle till convergence.

REMARK 1. It is easy to prove that Algorithm 1 has a nice ascent property which is a direct consequence of the minorization–maximizaton principle. Note that Algorithm 1 can be directly used to compute the LASSO-penalized composite likelihood estimator. We simply modify the coordinate-wise updating formula as $\widetilde{\beta}_{jk} \Leftarrow S(\widetilde{\beta}_{jk} + 2\widetilde{z}_{jk}, 2\lambda)$.

In practice we need to specify the λ value. BIC has been shown to perform very well for selecting the tuning parameter of the penalized likelihood estimator [Wang, Li and Tsai (2007)]. The BIC score is defined as

$$\widehat{\lambda} = \arg\max_{\lambda} \left\{ 2\ell_c(\widehat{\boldsymbol{\beta}}(\lambda)) - \log(n) \cdot \sum_{(j,k)} I(\widehat{\beta}_{jk}(\lambda) \neq 0) \right\}. \tag{2.7}$$

BIC is used to tune all methods considered in this work. We use SCAD1 to denote the SCAD solution computed by Algorithm 1 with the BIC tuned LASSO solution being the starting value.

For computational efficiency considerations, we implement Algorithm 1 by using the path-following idea and some other tricks, including warm-starts and active-set-cycling [Friedman, Hastie and Tibshirani (2010)]. We have implemented the algorithm in R language functions. The core cyclic coordinate-wise soft-thresholding operations were carried out in C.

REMARK 2. As suggested by a referee, the coordinate-gradient-ascent (CGA) algorithm is a natural alternative to Algorithm 1 for solving the LASSO-penalized composite likelihood estimator. The CGA algorithm has successfully used to solve other penalized models. See Genkin, Lewis and Madigan (2007), Meier, van de Geer and Bühlmann (2008), Städler, Bühlmann and van de Geer (2010) and Schelldorfer, Bühlmann and van de Geer (2011). In the CGA algorithm we need to find a good step size along the gradient direction to guarantee the ascent property after each coordinate-wise update. These extra computations are necessary for the CGA algorithm, but are not needed in the CMA algorithm. We have also implemented the CGA algorithm to solve the LASSO estimator and found that the CMA algorithm is about five times faster than the CGA algorithm. See Section 4 for the timing comparison details.

2.2. *Issues of local solution and the LLA–CMA algorithm.* The objective function in (1.2) is generally nonconcave if a nonconcave penalty function is used. Using Algorithm 1 we find a local solution to (1.2), but there is no guarantee that it is the global solution. A similar case is Schelldorfer, Bühlmann and van de Geer (2011) where the objective function is the LASSO-penalized maximum likelihood of a high-dimensional linear mixed-effects model, and the authors derived a coordinate-wise gradient descent algorithm to find a local solution.

It should not be considered as a special weakness of Algorithm 1 or other coordinate-wise descent algorithm as in Schelldorfer, Bühlmann and van de Geer (2011) that the algorithm can only find a local solution, because in the current literature there is no algorithm that can guarantee to find the global solution of nonconcave maximization (or nonconvex minimization) problems, especially when the dimension is huge. Consider, for example, the EM algorithm, which is perhaps the most famous algorithm in statistical literature. The EM algorithm often offers an elegant way to fit some statistical models that are formulated as nonconcave maximization problems. However, the EM algorithm provides a local solution in general. A recent application of the EM algorithm to high-dimensional modeling can be found in Städler, Bühlmann and van de Geer (2010) who considered a LASSO-penalized maximum likelihood estimator of a high-dimensional linear regression model with inhomogeneous errors that are modeled by a finite mixture of Gaussians. To handle the computational challenges in their problem, Städler, Bühlmann and van de Geer (2010) proposed a generalized EM algorithm in which a coordinate descent loop is used in the M-step and showed that the obtained solution is a local solution.

Our numerical results show that in the penalized composite likelihood estimation problem the SCAD performs much better than the LASSO. To offer theoretical understanding of their differences, it is important to show that the obtained local solution of the SCAD-penalized likelihood has better theoretical properties than the LASSO estimator. In Section 3 we establish the asymptotic properties of the LASSO estimator and a local solution of (1.2) with the SCAD penalty. However, a general technical difficulty in nonconcave maximization problems is to show that the computed local solution is the one local solution with proven theoretical properties. In Städler, Bühlmann and van de Geer (2010) and Schelldorfer, Bühlmann and van de Geer (2011), nice asymptotic properties are established for their proposed methods but it is not clear whether the computed local solutions could have those theoretical properties. The same issue exists in Fan and Lv (2011).

To circumvent the technical difficulty, we can consider combining the LLA idea [Zou and Li (2008)] and Algorithm 1 to solve (1.2) with a nonconcave penalty. The LLA algorithm turns a nonconcave penalization problem into a sequence of weighted LASSO penalization problems. Similar ideas of iterative LLA convex relaxation have been used in Candès, Wakin and Boyd (2008), Zhang (2010b) and Bradic, Fan and Wang (2011). Applying the LLA

Algorithm 2 The LLA–CMA algorithm

(1) Initialize $\widetilde{\boldsymbol{\beta}}^{(0)}$, and compute $w_{jk} = P'_\lambda(|\widetilde{\beta}^{(0)}_{jk}|)$.
(2) For $m = 0, 1, 2, 3, \ldots$, repeat the LLA iteration:
  (2.a) Use Algorithm 1 to solve $\widehat{\boldsymbol{\beta}}^{(m+1)}$ defined in (2.8);
  (2.b) Update the weights w_{jk} by $P'_\lambda(|\widetilde{\beta}^{(m+1)}_{jk}|)$.

algorithm to (1.2), we need to iteratively solve

$$\widehat{\boldsymbol{\beta}}^{(m+1)} = \arg\max_{\boldsymbol{\beta}} \left\{ \ell_c(\boldsymbol{\beta}) - \sum_{j=1}^{K} \sum_{k=j+1}^{K} w_{jk} \cdot |\beta_{jk}| \right\} \tag{2.8}$$

for $m = 0, 1, 2, \ldots$ where $w_{jk} = P'_\lambda(|\widetilde{\beta}^{(m)}_{jk}|)$. Note that Algorithm 1 can be used to solve (2.8) by simply modifying the coordinate-wise updating formula as $\widetilde{\beta}_{jk} \Leftarrow S(\widetilde{\beta}_{jk} + 2\widetilde{z}_{jk}, 2w_{jk})$. Therefore, we have the following LLA–CMA algorithm for computing a local solution of (1.2).

In Section 3 we show that if the LASSO estimator is $\widetilde{\boldsymbol{\beta}}^{(0)}$, then under certain regularity conditions the LLA–CMA algorithm finds the oracle estimator with high probability. These results suggest that we should take the following steps to compute the SCAD solution by the LLA–CMA algorithm.

The proposed LLA–CMA procedure for computing a SCAD estimator:

Step 1. Use Algorithm 1 to compute the LASSO solution path and find the LASSO estimator by BIC.

Step 2. Use the LASSO estimator as $\widetilde{\boldsymbol{\beta}}^{(0)}$ in the LLA–CMA algorithm to compute the solution path of the first iteration and use BIC to tune the first step solution. Then use the tuned first step solution as $\widetilde{\boldsymbol{\beta}}^{(0)}$ in the LLA–CMA algorithm to compute the solution path and use BIC to select λ. The resulting estimator is denoted by SCAD2.

Step 3. For the chosen λ of SCAD2, use Algorithm 2 to compute the fully converged SCAD solution with SCAD2 being the starting value. Denote this SCAD solution by SCAD2**.

The construction of SCAD2 follows an idea in Bühlmann and Meier (2008). Based on our experience, SCAD2** works slightly better than SCAD2, but the two are generally very close. Generally we recommend using SCAD2** in real applications.

3. Theoretical results. In this section we establish the statistical theory for the penalized composite conditional likelihood estimator using the SCAD and the LASSO penalty, respectively. Such results allow us to compare the SCAD and the LASSO estimators theoretically.

In order to present the theory we need some necessary notation. For a matrix $\mathbf{A} = (a_{ij})$, we define the following matrix norms: the Frobenius norm $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$, the entry-wise ℓ_∞ norm $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$ and the matrix ℓ_∞ norm $\|\mathbf{A}\|_\infty = \max_i \sum_j |a_{ij}|$. Let $\boldsymbol{\beta}^* = \{\beta_{jk}^* : j < k\}$ denote the true coefficients, $\mathcal{A} = \{(j,k) : \beta_{jk}^* \neq 0, j < k\}$ and $s = |\mathcal{A}|$. Define $\rho(s,N) = \min_{(j,k)\in\mathcal{A}} |\beta_{jk}^*|$ which represents the weakness of the signal. Let H be the Hessian matrix of ℓ_c such that

$$H_{(j_1k_1),(j_2k_2)} = -\frac{\partial^2 \ell_c(\boldsymbol{\beta})}{\partial \beta_{j_1k_1}\,\partial \beta_{j_2k_2}},$$

$1 \le j_1 < k_1 \le K$ and $1 \le j_2 < k_2 \le K$. For simplicity we use $H^* = H(\boldsymbol{\beta}^*)$. We partition H and $\boldsymbol{\beta}$ according to $\mathcal{A}$ as $\begin{pmatrix} H_{\mathcal{A}\mathcal{A}} & H_{\mathcal{A}\mathcal{A}^c} \\ H_{\mathcal{A}^c\mathcal{A}} & H_{\mathcal{A}^c\mathcal{A}^c} \end{pmatrix}$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}}^T, \boldsymbol{\beta}_{\mathcal{A}^c}^T)^T$, respectively. We let

$$\mathbf{X}_{\mathcal{A}} = (X_j : (j,k) \text{ or } (k,j) \in \mathcal{A} \text{ for some } k)$$

and

$$\mathbf{x}_{\mathcal{A}n} = (x_{jn} : (j,k) \text{ or } (k,j) \in \mathcal{A} \text{ for some } k).$$

Finally, we define

$$\begin{aligned} b &= \lambda_{\min}(E[H^*_{\mathcal{A}\mathcal{A}}]), \\ B &= \lambda_{\max}(E[\mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^T]), \\ \phi &= \|E[H^*_{\mathcal{A}^c\mathcal{A}}](E[H^*_{\mathcal{A}\mathcal{A}}])^{-1}\|_\infty. \end{aligned}$$

Define the oracle estimator as $\widehat{\boldsymbol{\beta}}^{\text{oracle}} = (\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{hmle}}, 0)$ where

$$\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{hmle}} = \arg\max_{\boldsymbol{\beta}_{\mathcal{A}}} \ell_c((\boldsymbol{\beta}_{\mathcal{A}}, 0)).$$

If we knew the true submodel, then we would use the oracle estimator to estimate the Ising model.

THEOREM 3.1. *Consider the SCAD-penalized composite likelihood defined in (1.2). We have the following two conclusions:*

(1) *For any $R < \frac{b}{3B}\frac{\sqrt{N}}{s}$, we have*

$$\Pr\left(\|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{hmle}} - \boldsymbol{\beta}_{\mathcal{A}}^*\|_2 \le \sqrt{\frac{s}{N}}R\right) \ge 1 - \tau_1 \qquad (3.1)$$

with $\tau_1 = \exp(-R^2 \frac{b^2}{8^3}) + 2s^2 \exp(-\frac{N}{s^2}\frac{b^2}{2}) + 2s^2 \exp(-\frac{N}{s^2}\frac{B^2}{8})$.

(2) *Pick a λ satisfying $\lambda < \min(\frac{\rho(s,N)}{2a}, \frac{(2\phi+1)b^2}{3sB})$. With probability at least $1 - \tau_2$, $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$ is a local maximizer of the SCAD-penalized composite likeli-*

hood estimator where

$$\begin{aligned}\tau_2 &= \exp\left(-R_*^2 \frac{b^2}{8^3}\right) + K^2 \exp\left(-\frac{N\lambda^2}{32(2\phi+1)^2}\right)\\ &\quad + \exp\left(-\frac{N\lambda}{3B(2\phi+1)s}\frac{b^2}{8^3}\right) + K^2 s \exp\left(-\frac{Nb^2}{2s^3}\right) + 2s^2 \exp\left(-\frac{b^2 N}{8s^3}\right)\\ &\quad + 4s^2\left[\exp\left(-\frac{N}{s^2}\frac{b^2}{2}\right) + \exp\left(-\frac{N}{s^2}\frac{B^2}{8}\right)\right]\end{aligned} \tag{3.2}$$

and $R_* = \min(\frac{1}{2}\sqrt{\frac{N}{s}}\rho(s,N), \frac{b}{3B}\frac{\sqrt{N}}{s})$.

We also analyzed the theoretical properties of the LASSO estimator. If the LASSO can consistently select the true model, it must equal to the hypothetical LASSO estimator $(\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}, 0)$ where

$$\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} = \arg\max_{\boldsymbol{\beta}_{\mathcal{A}}}\left\{\ell_c((\boldsymbol{\beta}_{\mathcal{A}}, 0)) - \lambda \sum_{(j,k)\in\mathcal{A}} |\beta_{jk}|\right\}.$$

THEOREM 3.2. *Consider the LASSO-penalized composite likelihood estimator.*

(1) *Choose* λ *such that* $\lambda s < \frac{8b^2}{3B}$. $\Pr(\|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}\|_2 \le \frac{16\lambda\sqrt{s}}{b}) \ge 1 - \tau_1'$ *with*

$$\tau_1' = e^{-N\lambda^2/2} + 2s^2\left[\exp\left(\frac{-Nb^2}{2s^2}\right) + \exp\left(\frac{-NB^2}{8s^2}\right)\right].$$

(2) *Assume the ir-representable condition* $\phi \le 1 - \eta < 1$. *Choose* λ *such that* $\lambda s < \min(\frac{b^2}{16^2 B}\frac{\eta/3}{4-\eta}, \frac{8b^2}{3B})$. *Then* $(\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}, 0)$ *is the LASSO-penalized composite likelihood estimator with probability at least* $1 - \tau_2'$, *where*

$$\begin{aligned}\tau_2' &= e^{-N\lambda^2/2} + K^2 s \exp\left(-\frac{Nb^2\eta^2}{8s^3}\right) + K^2 \exp\left(-\frac{N\lambda^2\eta^2}{32(4-\eta)^2}\right)\\ &\quad + 2s^2\left[\exp\left(-\frac{Nb^2\eta^2}{2s^3(2-\eta)^2}\right) + \exp\left(\frac{-Nb^2}{2s^2}\right) + \exp\left(\frac{-NB^2}{8s^2}\right)\right].\end{aligned}$$

In Theorems 3.1 and 3.2 the three quantities b, B and ϕ do not need to be constants. We can obtain a more straightforward understanding of the properties of the penalized composite likelihood estimators by considering the asymptotic consequences of these probability bounds. To highlight the main point, we consider b, B and ϕ are fixed constants and derive the following asymptotic results.

COROLLARY 3.1. *Suppose that* b, B *and* ϕ *are fixed constants and further assume* $N \gg s^3 \log(K)$ *and* $\rho(s,N) \gg \sqrt{\frac{\log(K)}{N}}$.

(1) *Pick the SCAD penalty parameter* λ^{scad} *satisfying*

$$\lambda^{\mathrm{scad}} < \min\left(\frac{\rho(s,N)}{2a}, \frac{(2\phi+1)b^2}{3sB}\right), \qquad \lambda^{\mathrm{scad}} \gg \sqrt{\frac{\log(K)}{N}}.$$

With probability tending to 1, the oracle estimator is a local maximizer of the SCAD-penalized estimator and $\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{oracle}} - \boldsymbol{\beta}_{\mathcal{A}}^{*}\|_2 = O_P(\sqrt{\frac{s}{N}})$.

(2) *Assume the ir-representable condition in Theorem 3.2. Pick the LASSO penalty parameter* λ^{lasso} *satisfying*

$$\min\left(\frac{1}{\sqrt{s}}\rho(s,N), \frac{1}{s}\right) \gg \lambda^{\mathrm{lasso}} \gg \frac{1}{\sqrt{N}};$$

then the LASSO estimator consistently selects the true model and $\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{lasso}} - \boldsymbol{\beta}_{\mathcal{A}}^{*}\|_2 = O_P(\lambda^{\mathrm{lasso}}\sqrt{s})$.

REMARK 3. For the LASSO-penalized least squares, it is now known that the model selection consistency critically depends on the ir-representable condition [Zhao and Yu (2006), Meinshausen and Bühlmann (2006), Zou (2006)]. A similar condition is again needed in the LASSO-penalized composite likelihood. Furthermore, Corollary 3.1 shows that even when it is possible for the LASSO to achieve consistent selection, λ^{lasso} should be much greater than $\sqrt{\frac{1}{N}}$, which means that $\lambda^{\mathrm{lasso}}\sqrt{s} \gg \sqrt{\frac{s}{N}}$. So the LASSO yields larger bias than the SCAD.

REMARK 4. We have shown that asymptotically speaking the oracle estimator is in fact a local solution of the SCAD-penalized composite likelihood model. This property is stronger than the oracle properties defined in Fan and Li (2001). Our result is the first to show that the oracle model selection theory holds nicely for nonconcave penalized composite conditional likelihood models with NP-dimensionality. The usual composite likelihood theory in the literature is only applied to the fixed-dimension setting. Our result fills a long-standing gap in the composite likelihood literature.

What we have shown so far is the existence of a SCAD-penalized estimator that is superior to the LASSO-penalized estimator. Moreover, we would like to show that the computed SCAD estimator is equal to the oracle estimator. As discussed earlier in Section 2.2, such a result is very difficult to prove due to the nonconcavity of the penalized likelihood function. See also Fan and Lv (2011), Städler, Bühlmann and van de Geer (2010) and Schelldorfer, Bühlmann and van de Geer (2011).

If one can prove that the objective function has only one maximizer, then the computed solution and the theoretically proven solution must be the same. This idea has been used in Fan and Lv (2011) to study the nonconcave penalized generalized linear models and Bradic, Fan and Jiang (2011) to study the nonconcave penalized Cox proportional hazards models. Their arguments are based on the observation that the SCAD penalty function has a finite maximum concavity [Zhang (2010a), Lv and Fan (2009)]. Hence, if the smallest eigenvalue of the Hessian matrix of the negative log-likelihood is sufficiently large, the overall penalized likelihood function is concave and hence has a unique global maximizer. This argument requires that the sample size is greater than the dimension; otherwise, the Hessian matrix does not have full rank. To deal with the high-dimensional case, Fan and Lv (2011) further refined their arguments by considering a subspace denoted by $\mathbb{S}_s$, which is the union of all s-dimensional coordinate subspaces. Under some regularity conditions, Fan and Lv (2011) showed that the oracle estimator is the unique global maximizer in $\mathbb{S}_s$, which was referred to as restricted global optimality. Then by assuming that the computed solution has exactly s nonzero elements, it can be concluded that the computed solution is in $\mathbb{S}_s$ and hence equals the oracle estimator; see Proposition 3.b of Fan and Lv (2011). However, a fundamental problem with these arguments is that we have no idea whether the computed solution selects s nonzero coefficients, because s is unknown.

Here we take a different route to tackle the local solution issue. Instead of trying to prove the uniqueness of maximizer, we directly analyze the local solution by the LLA–CMA algorithm and discuss under which regularity conditions the LLA–CMA algorithm can actually find the oracle estimator.

THEOREM 3.3. *Consider the SCAD-penalized composite likelihood estimator in (1.2). Let* $\widehat{\boldsymbol{\beta}}^{\mathrm{scad}}$ *be the local solution computed by Algorithm 2 (the LLA–CMA algorithm) with* $\widetilde{\boldsymbol{\beta}}^{(0)}$ *being the initial value. Pick a* λ *satisfying* $\lambda < \min(\frac{\rho(s,N)}{2a}, \frac{(2\phi+1)b^2}{3sB})$. *Write* $\tau_0 = \Pr(\|\widetilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_\infty > \lambda)$.

(1) *The LLA–CMA algorithm finds the oracle estimator after one LLA iteration with probability at least* $1 - \tau_0 - \tau_3$ *where*

$$\tau_3 = K^2 \exp\left(\frac{-N\lambda^2}{32(2\phi+1)^2}\right) + \exp\left(\frac{-N\lambda}{3B(2\phi+1)s}\frac{b^2}{8^3}\right) + K^2 s \exp\left(\frac{-Nb^2}{2s^3}\right) + 2s^2\left[\exp\left(-\frac{Nb^2}{8s^3}\right) + \exp\left(-\frac{N}{s^2}\frac{b^2}{2}\right) + \exp\left(-\frac{N}{s^2}\frac{B^2}{8}\right)\right].$$

(2) *The LLA–CMA algorithm converges after two LLA iterations and* $\widehat{\boldsymbol{\beta}}^{\mathrm{scad}}$ *equals the oracle estimator with probability at least* $1 - \tau_0 - \tau_2$, *where* τ_2 *is defined in (3.2).*

Theorem 3.3 can be used to drive the following asymptotic result.

COROLLARY 3.2. *Suppose that* b, B *and* ϕ *are fixed constants, and further assume* $N \gg s^3 \log(K)$ *and* $\rho(s, N) \gg \frac{\max(\sqrt{\log(K)}, 16\sqrt{s}/b)}{\sqrt{N}}$. *Consider the*

SCAD-penalized composite likelihood estimator with the SCAD penalty parameter λ^{scad} satisfying

$$\lambda^{\text{scad}} < \min\left(\frac{\rho(s,N)}{2a}, \frac{(2\phi+1)b^2}{3sB}\right), \qquad \lambda^{\text{scad}} \gg \sqrt{\frac{\log(K)}{N}}.$$

(1) *If $\tau_0 \to 0$, then with probability tending to one, the LLA–CMA algorithm converges after two LLA iterations and the LLA–CMA solution (or its one-step version) is equal to the oracle estimator.*

(2) *Consider using the LASSO estimator as $\widetilde{\boldsymbol{\beta}}^{(0)}$. Assume the ir-representable condition in Theorem 3.2, and pick the LASSO penalty parameter λ^{lasso} satisfying*

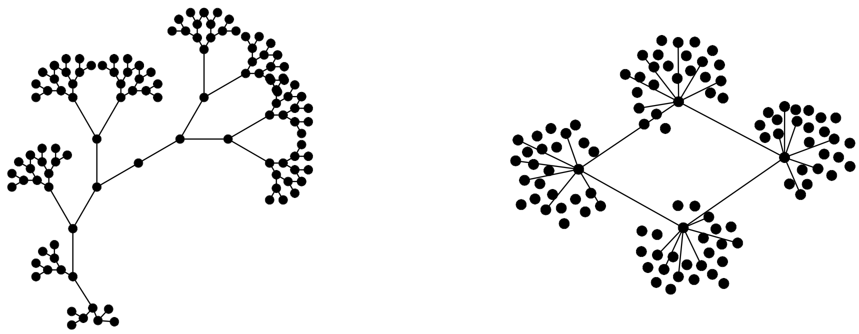
$$\frac{1}{\sqrt{N}} \ll \lambda^{\text{lasso}} \ll \min\left(\frac{1}{\sqrt{s}}\rho(s,N), \frac{1}{s}\right),$$
$$\lambda^{\text{lasso}} < \frac{\lambda^{\text{scad}}}{\sqrt{s}} \frac{b}{16}.$$

Then $\tau_0 \to 0$, and the conclusion in (1) holds.

REMARK 5. Part (1) of Corollary 3.2 basically says that any estimator that converges to $\boldsymbol{\beta}^*$ in probability at a rate faster than λ^{scad} can be used as the starting value in the LLA–CMA algorithm to find the oracle estimator with high probability. Note that such a condition is not very restrictive. Part (2) of Corollary 3.2 shows that the LASSO estimator satisfies that condition. We could also consider using other estimators as the starting value in the LLA–CMA algorithm. For example, we can use the neighborhood selection estimator as $\widetilde{\boldsymbol{\beta}}^{(0)}$. Following Ravikumar, Wainwright and Lafferty (2010) we assume an ir-representable condition for each of the K neighborhood LASSO-penalized logistic regression and some other regularity conditions. Then it is not hard to show that the neighborhood selection estimator is also a qualified starting value. In this work, we would like to faithfully follow the composite likelihood idea and hence prefer to use the LASSO-penalized composite likelihood estimator as the starting value in the LLA–CMA algorithm.

4. Simulation. In this section we use simulation to study the finite sample performance of the SCAD-penalized composite likelihood estimator. For comparison, we also include other two methods: neighborhood selection by LASSO-penalized logistic regression [Ravikumar, Wainwright and Lafferty (2010)] and the LASSO-penalized composite likelihood estimator.

For each coupling coefficient β_{jk}, the LASSO-penalized logistic method provides two estimates: $\widehat{\beta}_{j\mapsto k}$ based on the model for the jth dipole and $\widehat{\beta}_{k\mapsto j}$ based on the model for the kth dipole. Then we carry out two types of neighborhood selections: (i) aggregation by intersection (NSAI) based on $\widehat{\beta}_{jk}^{\text{NSAI}}$,



(A) Model 1: 127 dipoles and 126 non-zero coupling coefficients.

(B) Model 2: 104 dipoles and 24 non-zero coupling coefficients.

FIG. 1. *Plots of two simulated Ising models.*

and (ii) aggregation by union (NSAU) based on $\widehat{\beta}_{jk}^{\mathrm{NSAU}}$, where

$$\widehat{\beta}_{jk}^{\mathrm{NSAI}} = \begin{cases} 0, & \text{if } \widehat{\beta}_{j\mapsto k}\widehat{\beta}_{k\mapsto j} = 0, \\ \dfrac{\widehat{\beta}_{j\mapsto k} + \widehat{\beta}_{k\mapsto j}}{2}, & \text{otherwise,} \end{cases}$$

and

$$\widehat{\beta}_{jk}^{\mathrm{NSAU}} = \begin{cases} 0, & \text{if } \widehat{\beta}_{j\mapsto k} = 0 \text{ and } \widehat{\beta}_{k\mapsto j} = 0, \\ \widehat{\beta}_{j\mapsto k}, & \text{if } \widehat{\beta}_{j\mapsto k} \neq 0 \text{ and } \widehat{\beta}_{k\mapsto j} = 0, \\ \widehat{\beta}_{k\mapsto j}, & \text{if } \widehat{\beta}_{j\mapsto k} = 0 \text{ and } \widehat{\beta}_{k\mapsto j} \neq 0, \\ \dfrac{\widehat{\beta}_{j\mapsto k} + \widehat{\beta}_{k\mapsto j}}{2}, & \text{if } \widehat{\beta}_{j\mapsto k}\widehat{\beta}_{k\mapsto j} \neq 0. \end{cases}$$

As suggested by a referee, the relaxed LASSO [Meinshausen (2007)] was used in neighborhood selection to try to improve its estimation accuracy. In each neighborhood logistic regression model, we first found a subset model by using the LASSO-penalized logistic regression. We re-estimated the nonzero coefficients via the unpenalized logistic regression on the subset model.

BIC has been shown to perform very well for selecting the tuning parameter of the penalized likelihood estimator [Wang, Li and Tsai (2007), Städler, Bühlmann and van de Geer (2010), Schelldorfer, Bühlmann and van de Geer (2011)]. We used BIC to tune all competitors.

Two sparse Ising models were considered in our simulation. Their graphical structure is displayed in Figure 1 where solid dots represent the dipoles, and two dipoles are connected if and only if their coupling coefficient is nonzero. We generated the nonzero coupling coefficients as follows. If dipoles i and j are connected, we let β_{ij} be $t_{ij}s_{ij}$ where t_{ij} is a random variable following the uniform distribution on $[1, 2]$ and s_{ij} is a Bernoulli variable with

TABLE 1

Comparing different estimators using simulation models 1 and 2 with standard errors in the bracket. NSAI-relax and NSAU-relax mean that we use the relaxed LASSO to re-estimate the nonzero coefficients chosen by neighborhood selection method

	Model 1			**Model 2**		
	MSE	**NDE**	**FDR**	**MSE**	**NDE**	**FDR**
NSAI	22.96	138.9	0.09	8.16	26.8	0.16
	(0.18)	(0.4)	(0.01)	(0.12)	(0.2)	(0.01)
NSAU	17.34	197.3	0.36	6.38	39.7	0.39
	(0.14)	(1.0)	(0.01)	(0.16)	(0.5)	(0.01)
LASSO	21.33	332.5	0.62	12.19	117.1	0.79
	(0.13)	(3.8)	(0.04)	(0.12)	(3.0)	(0.05)
SCAD1	2.86	145.0	0.12	5.64	30.0	0.22
	(0.10)	(2.4)	(0.01)	(0.17)	(1.8)	(0.02)
SCAD2	2.43	129.2	0.07	4.41	26.1	0.17
	(0.05)	(0.5)	(0.01)	(0.13)	(0.7)	(0.02)
SCAD2**	2.42	128.6	0.06	4.39	25.7	0.16
	(0.05)	(0.5)	(0.01)	(0.13)	(0.6)	(0.02)
NSAI-relax	8.23	138.9	0.09	6.34	26.8	0.16
	(0.13)	(0.4)	(0.01)	(0.09)	(0.2)	(0.01)
NSAU-relax	4.44	197.3	0.36	5.67	39.7	0.39
	(0.10)	(0.4)	(0.01)	(0.10)	(0.5)	(0.01)

$\Pr(s_{ij} = 1) = \Pr(s_{ij} = -1) = 0.5$. For each model, we used Gibbs sampling to generate 100 independent datasets consisting 300 observations. For comparison, we use three measurements: the total number of discovered edges (NDE), the false discovery rate (FDR) and mean square errors (MSE).

Based on Table 1, we make the following interesting observations:

- NSAU, while selecting larger models than NSAI, provides more accurate estimation. Neighborhood selection outperforms the LASSO-penalized composite likelihood estimator.
- Note that SCAD2** has the smallest MSE in both models. SCAD2** and SCAD2 gave almost identical results, and their improvement over SCAD1 is statistically significant. All three SCAD solutions perform much better than the LASSO for fitting penalized composite likelihood in terms of estimation and selection.
- The SCAD solutions and NSAI have similar model selection performance, but the SCAD is substantial better in estimation. Using the relaxed LASSO can improve the estimation accuracy of neighborhood selection methods, but their improved MSEs are still significantly higher than those of SCAD2 and SCAD2**.

In Table 2 we compare the run times of the three methods. LASSO-CGA denotes the coordinate gradient ascent algorithm for computing the LASSO

TABLE 2

*Total time (in seconds) for computing solutions at 100 penalization parameters, averaged over 3 replications. Timing was carried out on a laptop with an Intel Core 1.60 GHz processor. LASSO-CGA denotes a coordinate gradient ascent algorithm for computing the LASSO-penalized composite likelihood estimator. The timing of SCAD1, SCAD2 and SCAD2** includes the timing for computing the starting value*

(N, p)	**Neighborhood selection**	**LASSO**	**SCAD1**	**SCAD2**	**SCAD2****	**LASSO-CGA**
Model 1 (300, 7875)	51.1	32.7	67.9	84.7	95.1	179.8
Model 2 (300, 5356)	29.8	16.0	34.8	42.6	51.2	89.6

estimator. The computing time is about five times longer than that used by the CMA algorithm. Compared to the LASSO case, the run time for fitting the SCAD model is doubled or tripled, but it is still very manageable for the high-dimensional data.

5. Stanford HIV drug resistance data. We also illustrate our methods in a real example using a HIV antiretroviral therapy (ART) susceptibility dataset obtained from the Stanford HIV drug resistance database. Details of the database and related data sets can be found in Rhee et al. (2006). The data for analysis consists of virus mutation information at 99 protease residues (sites) for $N = 702$ isolates from the plasma of HIV-1-infected patients. This dataset has been previously used in Rhee et al. (2006) and Wu, Cai and Lin (2010) to study the association between protease mutations and susceptibility to ART drugs.

A well recognized problem with current ART treatment such as PIs for treating HIV is that individuals who initially respond to therapy may develop resistance to it due to viral mutations. HIV-1 protease plays a key role in the late stage of viral replication and its ability to rapidly acquire a variety of mutations in response to various PIs confers the enzyme with high resistance to ARTs. A high cooperativity has been observed among drug-resistant mutations in HIV-1 protease [Ohtaka, Schön and Freire (2003)]. The sequence data retrieved from treated patients is likely to include mutations that reflect cooperative effects originating from late functional constraints, rather than stochastic evolutionary noise [Atchley et al. (2000)]. However, the molecular mechanisms of drug resistance is yet to be elucidated. It is thus of great interest to study inter-residue couplings which might be relevant to protein structure or function and thus could potentially shed light on the mechanisms of drug resistance. We apply the proposed method to

TABLE 3
Application to HIVRT data. NSE is the number of "stable edges." $E[V]$ is the expected number of falsely selected edges. Its upper bounds were computed by Theorem 1 in Meinshausen and Bühlmann (2010)

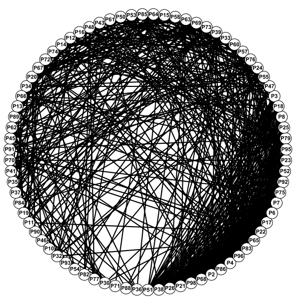
	NSAI	**NSAU**	**LASSO**	**SCAD1**	**SCAD2**	**SCAD2****
NDE	57	305	631	101	141	132
ME	26.38	36.34	18.35	18.30	16.76	16.74
			Stability selection			
NSE ($\pi_{\text{thr}} = 0.9$)	15	63	160	17	20	20
$E[V]$	≤ 3.2	≤ 48	$\leq$147.5	$\leq$4.3	$\leq$8.0	$\leq$7.2

the protease sequence data to investigate such inter-residue contacts. Our analysis only included $K = 79$ of the 99 residues that contain mutations.

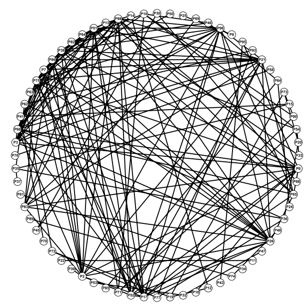
We split the data into a training set with 500 data and a test set with 202 data. Model fitting and selection were done on the training set and the test data were used to compare the model errors. For a given estimate $\widehat{\boldsymbol{\beta}}$ obtained from the training set, its model error is gauged by the value of composite likelihood evaluated on the test set, that is,

$$\mathrm{ME}(\widehat{\boldsymbol{\beta}}) = -\ell_c^{\text{test}}(\widehat{\boldsymbol{\beta}}) = -\frac{1}{202}\sum_{n=1}^{202}\sum_{j=1}^{79}\log(\theta_{jn}(\widehat{\boldsymbol{\beta}})).$$

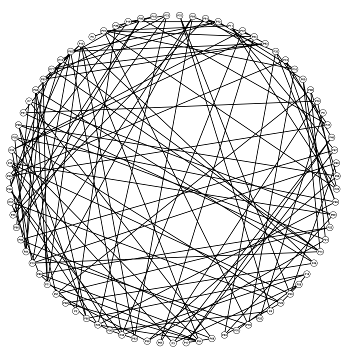
We report the analysis results in Table 3. There are total 3081 coupling coefficients to be estimated. Graphical presentations of the selected models are shown in Figure 2. Note that SCAD2 and SCAD2** again gave almost identical results and performed better SCAD1. We also performed stability selection [Meinshausen and Bühlmann (2010)] on each method to find "stable edges." A remarkable property of stability selection is that under some suitable conditions stability selection achieves finite sample control over the expected number of false discoveries in the set of "stable edges." We use the SCAD selector to explain the stability selection procedure. We took a random subsample of size 250 and fitted the SCAD model. The process was repeated 100 times. On average, SCAD1 selected 103.1 edges, SCAD2 selected 140.7 edges and SCAD2** chose 133.4 edges. For each coefficient β_{jk} we computed its frequency of being selected, denoted by $\widehat{\Pi}_{jk}$. The set of "stable edges" is defined as $\{(k,j) : \widehat{\Pi}_{kj} > \pi_{\text{thr}}\}$. In Table 3, we report the results using the threshold $\pi_{\text{thr}} = 0.9$, as suggested by Meinshausen and Bühlmann (2010). Stability selection found 17 edges in the SCAD1. SCAD2 and SCAD2** selected the same 20 stable edges. By Theorem 1 in Meinshausen and Bühlmann (2010), among these 17 stable edges selected by SCAD1, the expected number of false discoveries is no greater than 4.3, and



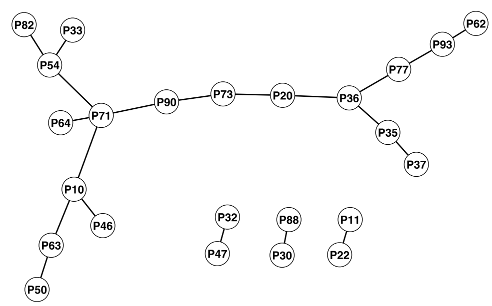
(A1) The LASSO model.



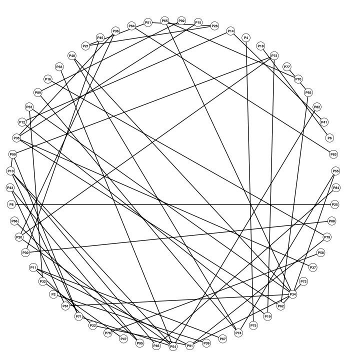
(A2) “Stable edges” in the LASSO model.



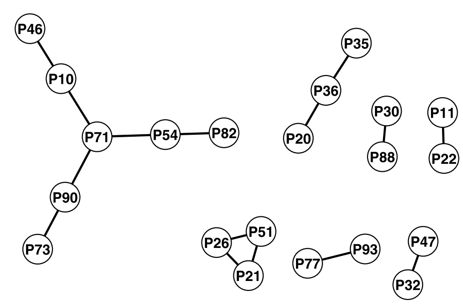
(B1) The SCAD2** model.



(B2) “Stable edges” in the SCAD2** model.



(C1) The model by neighborhood selection.



(C2) “Stable edges” in neighborhood selection.

FIG. 2. *Shown in the left three panels* (A1), (B1), (C1) *are the selected models by BIC. The right three panels* (A2), (B2), (C2) *show the stability selection results using* $\pi_{\mathrm{thr}} = 0.9$.

among the 20 stable edges selected by SCAD2 or SCAD2**, the expected number of false discoveries is at most 7.2. Likewise, we did stability selection with the LASSO selector and neighborhood selection, and the results are reported in Table 3 as well. Figure 2 shows the “stable edges” by stability selection. We see that the computed upper bounds are very useful for the

SCAD selector and NSAI and not so informative for the LASSO selector and NSAU. Interestingly, both NSAI and SCAD suggest there are about 12 true discoveries by stability selection. In fact, we found that NSAI and SCAD1 have 11 "stable edges" in common, and NSAI and SCAD2 (or SCAD2**) have 12 "stable edges" in common.

These results are consistent with some of the previous findings. For example, it has long been known that co-substitutions at residues 30 and 88 are most effective in reducing the susceptibility of nelfinavir [Liu, Eyal and Bahar (2008)]. Among the top 30 most common drug resistance mutations [Rhee et al. (2004)], 7 of those had a joint mutation at residues 54 and 82, the joint mutation at residues 88 and 30 was the second most common mutation among all. A co-mutation at residues 54, 82 and 90 was associated with high resistance to multiple drugs and an additional co-mutation at 46 was associated with an even higher level of resistance. It is interesting to note that using a larger set of isolates from treated HIV patients, Wu et al. (2003) reported (54, 82), (32, 47), (73, 90) as the three most highly correlated pairs. All these three pairs showed up as the stable edges in our analysis. Mutation at residue 71, often described as a compensatory or accessory mutation, has been reported as a critical mutation which appears to improve virus growth and contribute to resistance phenotype [Markowitz et al. (1995), Tisdale et al. (1995), Muzammil, Ross and Freire (2003)]. Accessory mutations contribute to resistance only when present with a mutation in the substrate cleft or flap or at residue 90 [Wu et al. (2003)]. The stable edges connect this accessory mutation with residues 90 and 54 (a flap residue), as well as with another flap residue at 46 through residue 10.

APPENDIX: TECHNICAL PROOFS

Before presenting the proof, we first define some useful quantities. The score functions of the negative composite likelihood ($-\ell^{(j)}$) and the Hessian matrices are defined as follows:

$$\psi_k^{(j)} = -\frac{\partial \ell^{(j)}(\boldsymbol{\beta}^{(j)})}{\partial \beta_{jk}} = \frac{1}{N}\sum_{n=1}^{N} x_{jn}x_{kn}(\theta_{jn}-1), \qquad k \neq j,$$

$$H_{k_1,k_2}^{(j)} = -\frac{\partial^2 \ell^{(j)}(\boldsymbol{\beta}^{(j)})}{\partial \beta_{jk_1}\,\partial \beta_{jk_2}} = \frac{1}{N}\sum_{n=1}^{N} x_{k_1 n}x_{k_2 n}(1-\theta_{jn})\theta_{jn}, \qquad k_1, k_2 \neq j.$$

Similarly, let ψ be the score function of $-\ell_c$ such that $\psi_{(jk)} = \frac{\partial -\ell_c(\boldsymbol{\beta})}{\partial \beta_{jk}}$ for $1 \leq j < k \leq K$. By definition we have the following identities: $\psi_{(jk)} = \psi_k^{(j)} + \psi_j^{(k)}$. In what follows we write $\psi^* = \psi(\boldsymbol{\beta}^*)$.

PROOF OF THEOREM 3.1. We first prove part (1).

Consider $V(\boldsymbol{\alpha}_{\mathcal{A}}) = -\ell_c(\boldsymbol{\beta}^*_{\mathcal{A}} + d_N\boldsymbol{\alpha}_{\mathcal{A}}) + \ell_c(\boldsymbol{\beta}^*_{\mathcal{A}})$ and its minimizer is $\widetilde{\boldsymbol{\alpha}}^{\mathrm{hmle}}_{\mathcal{A}} = \frac{1}{d_N}(\widetilde{\boldsymbol{\beta}}^{\mathrm{hmle}}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}})$. By definition, $V(\widetilde{\boldsymbol{\alpha}}^{\mathrm{hmle}}_{\mathcal{A}}) \le V(\mathbf{0}) = 0$. Fix any $R > 0$ and consider any $\boldsymbol{\alpha}_{\mathcal{A}}$ satisfying $\|\boldsymbol{\alpha}_{\mathcal{A}}\|_2 = R$. Using Taylor's expansion, we know that, for some $t \in [0,1]$ and $\boldsymbol{\beta}(t) = \boldsymbol{\beta}^*_{\mathcal{A}} + t d_N \boldsymbol{\alpha}_{\mathcal{A}}$,

$$\begin{aligned} V(\boldsymbol{\alpha}_{\mathcal{A}}) &= d_N \boldsymbol{\alpha}^T_{\mathcal{A}} \psi^*_{\mathcal{A}} + \tfrac{1}{2} d_N^2 \boldsymbol{\alpha}^T_{\mathcal{A}} H^*_{\mathcal{A}\mathcal{A}} \boldsymbol{\alpha}_{\mathcal{A}} \\ &\quad + \tfrac{1}{2} d_N^2 \boldsymbol{\alpha}^T_{\mathcal{A}} [H_{\mathcal{A}\mathcal{A}}(\boldsymbol{\beta}(t)) - H^*_{\mathcal{A}\mathcal{A}}] \boldsymbol{\alpha}_{\mathcal{A}} \\ &\equiv T_1 + T_2 + T_3. \end{aligned} \tag{A.1}$$

Note that $E[\psi^*_{\mathcal{A}}] = 0$ and $\|\psi^*_{\mathcal{A}}\|_\infty \le 2$. By the Cauchy–Schwarz inequality, $|\boldsymbol{\alpha}^T_{\mathcal{A}} \psi^*_{\mathcal{A}}| \le 2\sqrt{s}R$. Using Hoeffding's inequality, we have

$$\Pr(T_1 \ge -d_N \varepsilon) \le \exp\left(-\frac{N\varepsilon^2}{8sR^2}\right). \tag{A.2}$$

For the second term, we first have $T_2 \ge \frac{d_N^2}{2} \lambda_{\min}(H^*_{\mathcal{A}\mathcal{A}}) R^2$. Each entry of H^* is between $-\frac{1}{2}$ and $\frac{1}{2}$. Thus Hoeffding's inequality and the union bound yield

$$\Pr\left(\|H_j^{(N)} - H_j\|_F^2 \ge \frac{b^2}{4}\right) \le 2s^2 \exp\left(-N\frac{b^2}{2s^2}\right).$$

So by the inequality $\lambda_{\min}(H^*_{\mathcal{A}\mathcal{A}}) \ge b - \|H^*_{\mathcal{A}\mathcal{A}} - E[H^*_{\mathcal{A}\mathcal{A}}]\|_F$, we have

$$\Pr(T_2 \ge d_N^2 b R^2/4) \ge 1 - 2s^2 \exp\left(-\frac{Nb^2}{2s^2}\right). \tag{A.3}$$

For $|T_3|$, let $\lambda_{\max}(\frac{1}{N}\sum_{n=1}^N \mathbf{x}_{\mathcal{A}n}\mathbf{x}^T_{\mathcal{A}n}) = B_N$. Define $\bar{\eta}_{jn}(\boldsymbol{\beta}) = \theta_{jn}(1-\theta_{jn})(2\theta_{jn} - 1)$. Using the mean value theorem, we have that, for some $t' \in [0,t]$ and $\boldsymbol{\beta}(t') = \boldsymbol{\beta}^*_{\mathcal{A}} + t' d_N \boldsymbol{\alpha}_{\mathcal{A}}$,

$$\begin{aligned} |T_3| &= \frac{d_N^3}{2}\left|\frac{1}{N}\sum_n \sum_{j=1}^K \sum_{\substack{k_1 \ne j \\ k_2 \ne j}} \alpha_{jk_1}\alpha_{jk_2} x_{k_1 n} x_{k_2 n} t' \bar{\eta}_{jn}(\boldsymbol{\beta}(t'))\left(\sum_{k' \ne j} \alpha_{jk'} x_{jn} x_{k'n}\right)\right| \\ &\le \frac{d_N^3}{2}\left(\frac{\sqrt{s R^2}}{4}\right) \cdot \left(2B_N \sum_{(j,k)\in\mathcal{A}} \alpha_{jk}^2\right) = \frac{d_N^3 B_N}{4}\sqrt{s}R^3. \end{aligned} \tag{A.4}$$

In the last step we have used $|\bar{\eta}_{jn}(\boldsymbol{\beta}(t'))| \le \frac{1}{4}$ for any j and $\boldsymbol{\alpha}_{\mathcal{A}^c} = 0$. Moreover, $B_N \le B + \|\frac{1}{N}\sum_{n=1}^N \mathbf{x}_{\mathcal{A}n}\mathbf{x}^T_{\mathcal{A}n} - E[\mathbf{x}_{\mathcal{A}}\mathbf{x}^T_{\mathcal{A}}]\|_F$. Since $x_{jn} = \pm 1$, we apply Hoeffding's inequality and the union bound to obtain the following probability bound:

$$\Pr\left(\left\|\frac{1}{N}\sum_{n=1}^N \mathbf{x}_{\mathcal{A}n}\mathbf{x}^T_{\mathcal{A}n} - E[\mathbf{x}_{\mathcal{A}}\mathbf{x}^T_{\mathcal{A}}]\right\|_F \ge B/2\right) \le 2s^2 \exp\left(-\frac{NB^2}{8s^2}\right),$$

which leads to

$$\Pr\left(|T_3| \le \frac{3d_N^3 B}{8}\sqrt{s}R^3\right) \ge 1 - 2s^2 \exp\left(-\frac{NB^2}{8s^2}\right). \tag{A.5}$$

Taking $R < \frac{b}{3B}\frac{\sqrt{N}}{s}$ and combining (A.2) (A.3) and (A.5), we have

$$T_1 + T_2 + T_3 \ge \frac{bR^2}{8}d_N^2 - \frac{3B}{8}R^3 d_N^3\sqrt{s} > 0$$

with probability at least $1-\tau_1$. Thus, the convexity of V implies that

$$\Pr\left(\|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{hmle}} - \boldsymbol{\beta}_{\mathcal{A}}^*\|_2 \le \sqrt{\frac{s}{N}}R\right) \ge 1 - \tau_1.$$

We now prove part (2). First, we show that if $\min_{(j,k)\in\mathcal{A}} |\widetilde{\beta}_{jk}^{\mathrm{hmle}}| > a\lambda$ and $\|\psi_{\mathcal{A}^c}(\widehat{\boldsymbol{\beta}}^{\mathrm{oracle}})\|_\infty \le \lambda$, then $\widehat{\boldsymbol{\beta}}^{\mathrm{oracle}}$ is a local maximizer of $\ell_c(\boldsymbol{\beta}) - \sum_{(j,k)} P_\lambda(|\beta_{jk}|)$. To see that, consider a small ball of radius t with $\widehat{\boldsymbol{\beta}}^{\mathrm{oracle}}$ being the center. Let $\boldsymbol{\beta}$ be any point in the ball. So $\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{\mathrm{oracle}}\|_2 \le t$. Clearly, for a sufficiently small t we have $\min_{(j,k)\in\mathcal{A}} |\beta_{jk}| > a\lambda$ and $\max_{(j,k)\in\mathcal{A}^c} |\beta_{jk}| < \lambda$. By Taylor's expansion we have

$$\begin{aligned}
&\left\{-\ell_c(\boldsymbol{\beta}) + \sum_{(j,k)} P_\lambda(|\beta_{jk}|)\right\} - \left\{-\ell_c(\widehat{\boldsymbol{\beta}}^{\mathrm{oracle}}) + \sum_{(j,k)} P_\lambda(|\widehat{\beta}_{jk}^{\mathrm{oracle}}|)\right\} \\
&\quad = (\boldsymbol{\beta}_{\mathcal{A}} - \widetilde{\boldsymbol{\beta}}^{\mathrm{hmle}})^T \psi_{\mathcal{A}^c}(\widehat{\boldsymbol{\beta}}^{\mathrm{oracle}}) + \frac{1}{2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{\mathrm{oracle}})^T H(\boldsymbol{\beta}')(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{\mathrm{oracle}}) \\
&\qquad + \sum_{(j,k)\in\mathcal{A}^c} \lambda|\beta_{jk}| \\
&\quad \ge \sum_{(j,k)\in\mathcal{A}^c} (\lambda - |\psi_{(jk)}(\widehat{\boldsymbol{\beta}}^{\mathrm{oracle}})|)|\beta_{jk}| \ge 0.
\end{aligned}$$

A probability bound for the event of $\min_{(j,k)\in\mathcal{A}} |\widetilde{\beta}_{jk}^{\mathrm{hmle}}| > a\lambda$ is given by

$$\begin{aligned}
&\Pr\left(\min_{(j,k)\in\mathcal{A}} |\widetilde{\beta}_{jk}^{\mathrm{hmle}}| > a\lambda\right) \\
&\quad \ge \Pr\left(\|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{hmle}} - \boldsymbol{\beta}_{\mathcal{A}}^*\|_2 \le \sqrt{\frac{s}{N}}R_*\right) \qquad \text{(A.6)} \\
&\quad \ge 1 - \exp\left(-R_*^2\frac{b^2}{8^3}\right) - 2s^2\exp\left(-\frac{N}{s^2}\frac{b^2}{2}\right) - 2s^2\exp\left(-\frac{N}{s^2}\frac{B^2}{8}\right).
\end{aligned}$$

Now consider $\Pr(\|\psi_{\mathcal{A}^c}(\widehat{\boldsymbol{\beta}}^{\mathrm{oracle}})\|_\infty < \lambda)$. There exists some $t \in [0,1]$ such that

$$\psi(\widehat{\boldsymbol{\beta}}^{\mathrm{oracle}}) = \psi(\boldsymbol{\beta}^*) + H^*(\widehat{\boldsymbol{\beta}}^{\mathrm{oracle}} - \boldsymbol{\beta}^*) + r, \tag{A.7}$$

where $r = (H(\boldsymbol{\beta}^* + t(\widehat{\boldsymbol{\beta}}^{\text{oracle}} - \boldsymbol{\beta}^*)) - H^*)(\widehat{\boldsymbol{\beta}}^{\text{oracle}} - \boldsymbol{\beta}^*)$. Note $\psi_{\mathcal{A}}(\widehat{\boldsymbol{\beta}}^{\text{oracle}}) = 0$, so

$$\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}} = (H^*_{\mathcal{A}A})^{-1}(-\psi_{\mathcal{A}} - r_{\mathcal{A}}).$$

Then $\|\psi_{\mathcal{A}^c}(\widehat{\boldsymbol{\beta}}^{\text{oracle}})\|_\infty \le \lambda$ becomes

$$\|H^*_{\mathcal{A}^cA}(H^*_{\mathcal{A}A})^{-1}(-\psi_{\mathcal{A}} - r_{\mathcal{A}}) + \psi_{\mathcal{A}^c} + r_{\mathcal{A}^c}\|_\infty \le \lambda,$$

which is guaranteed if

$$(\|H^*_{\mathcal{A}^cA}(H^*_{\mathcal{A}A})^{-1}\|_\infty + 1)(\|\psi\|_\infty + \|r\|_\infty) \le \lambda.$$

Therefore we have a simple lower bound for $\Pr(\|\psi_{\mathcal{A}^c}(\widehat{\boldsymbol{\beta}}^{\text{oracle}})\|_\infty \le \lambda)$.

$$\begin{aligned}
&\Pr(\|\psi_{\mathcal{A}^c}(\widehat{\boldsymbol{\beta}}^{\text{oracle}})\|_\infty \le \lambda)\\
&\quad > 1 - \Pr(\|H^*_{\mathcal{A}^cA}(H^*_{\mathcal{A}A})^{-1}\|_\infty > 2\phi) - \Pr\left(\|\psi\|_\infty > \frac{\lambda}{4\phi+2}\right)\\
&\qquad - \Pr\left(\|r\|_\infty > \frac{\lambda}{4\phi+2}\right).
\end{aligned}$$

Using Hoeffding's inequality and the union bound, we have

$$\Pr\left(\|\psi\|_\infty \le \frac{\lambda}{4\phi+2}\right) \ge 1 - K^2 \exp\left(-\frac{N\lambda^2}{128(\phi+1/2)^2}\right). \tag{A.8}$$

Write $\boldsymbol{\alpha} = \widetilde{\boldsymbol{\beta}}^{\text{hmle}} - \boldsymbol{\beta}^*$, and thus $\boldsymbol{\alpha}_{\mathcal{A}^c} = 0$. By the mean value theorem, we have a bound for $r_{(jk)}$:

$$\begin{aligned}
|r_{(jk)}| &= \Bigg|\frac{1}{N}\sum_{n=1}^N \sum_{k_2\ne j}\sum_{k'\ne j} x_{kn}x_{jn}x_{k_2n}x_{k'n}\alpha_{jk_2}\alpha_{jk'}t'\bar{\eta}_{jn}(\boldsymbol{\beta}(t'))\\
&\qquad + \frac{1}{N}\sum_{n=1}^N \sum_{j_2\ne k}\sum_{j'\ne k} x_{jn}x_{kn}x_{j_2n}x_{j'n}\alpha_{kj_2}\alpha_{kj'}t'\bar{\eta}_{kn}(\boldsymbol{\beta}(t'))\Bigg|\\
&\le B_N \cdot \|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}\|_2^2.
\end{aligned}$$

In the last step we have used $|\bar{\eta}_{jn}(\boldsymbol{\beta}(t'))| \le \frac{1}{4}$ for any j and $\boldsymbol{\alpha}_{\mathcal{A}^c} = 0$. Moreover, recall that

$$B_N \le B + \left\|\frac{1}{N}\sum_{n=1}^N \mathbf{x}_{\mathcal{A}n}\mathbf{x}_{\mathcal{A}n}^T - E[\mathbf{x}_{\mathcal{A}}\mathbf{x}_{\mathcal{A}}^T]\right\|_F.$$

Thus

$$\begin{aligned}
\Pr\left(\|r\|_\infty < \frac{\lambda}{4\phi+2}\right) &\ge 1 - \exp\left(\frac{-N\lambda}{3B(2\phi+1)s}\frac{b^2}{8^3}\right) - 2s^2\exp\left(\frac{-Nb^2}{2s^2}\right)\\
&\qquad - 2s^2\exp\left(\frac{-NB^2}{8s^2}\right).
\end{aligned} \tag{A.9}$$

For notation convenience define $c = \|(E[H^*_{\mathcal{A}A}])^{-1}\|_\infty \le \sqrt{s}\|(E[H^*_{\mathcal{A}A}])^{-1}\|_2$ and

$$\begin{aligned}
\delta &= \|H^*_{\mathcal{A}^cA}(H^*_{\mathcal{A}A})^{-1} - E[H^*_{\mathcal{A}^cA}](E[H^*_{\mathcal{A}A}])^{-1}\|_\infty,\\
\delta_1 &= \|(H^*_{\mathcal{A}A})^{-1} - (E[H^*_{\mathcal{A}A}])^{-1}\|_\infty,\\
\delta_2 &= \|H^*_{\mathcal{A}A} - E[H^*_{\mathcal{A}A}]\|_\infty,\\
\delta_3 &= \|H^*_{\mathcal{A}^cA} - E[H^*_{\mathcal{A}^cA}]\|_\infty.
\end{aligned}$$

Then by definition

$$\begin{aligned}
\delta &= \|(H^*_{\mathcal{A}^cA} - E[H^*_{\mathcal{A}^cA}])((H^*_{\mathcal{A}A})^{-1} - (E[H^*_{\mathcal{A}A}])^{-1})\\
&\qquad + E[H^*_{\mathcal{A}^cA}](E[H^*_{\mathcal{A}A}])^{-1}(-H^*_{\mathcal{A}A} + E[H^*_{\mathcal{A}A}])(H^*_{\mathcal{A}A})^{-1}\\
&\qquad\qquad + (H^*_{\mathcal{A}^cA} - E[H^*_{\mathcal{A}^cA}])(E[H^*_{\mathcal{A}A}])^{-1}\|_\infty\\
&\le \delta_3\delta_1 + \phi\delta_2\|(H^*_{\mathcal{A}A})^{-1}\|_\infty + \delta_3 c\\
&\le \delta_3\delta_1 + \phi(c+\delta_1)\delta_2 + \delta_3 c.
\end{aligned}$$

Note that

$$\begin{aligned}
\delta_1 &= \|(H^*_{\mathcal{A}A})^{-1}(E[H^*_{\mathcal{A}A}] - H^*_{\mathcal{A}A})(E[H^*_{\mathcal{A}A}])^{-1}\|_\infty\\
&\le \|(H^*_{\mathcal{A}A})^{-1}\|_\infty \cdot \|E[H^*_{\mathcal{A}A}] - H^*_{\mathcal{A}A}\|_\infty \cdot \|(E[H^*_{\mathcal{A}A}])^{-1}\|_\infty\\
&\le (\delta_1 + c)\delta_2 c.
\end{aligned}$$

Hence as long as $\delta_2 c < 1$ we have $\delta_1 \le \frac{\delta_2 c^2}{1-\delta_2 c}$ and $\delta \le (\delta_3 + \phi\delta_2)\frac{c}{1-\delta_2 c}$.

$$\begin{aligned}
\Pr\left(\delta_2 < \frac{1}{4c}\right) &\ge 1 - \Pr\left(\|H^*_{\mathcal{A}^cA} - E[H^*_{\mathcal{A}^cA}]\|_{\max} > \frac{1}{4cs}\right)\\
&\ge 1 - 2s^2\exp\left(-\frac{N}{8c^2s^2}\right),
\end{aligned}\tag{A.10}$$

$$\begin{aligned}
\Pr\left(\delta_3 < \frac{\phi}{2c}\right) &\ge 1 - \Pr\left(\|H^*_{\mathcal{A}^cA} - E[H^*_{\mathcal{A}^cA}]\|_{\max} > \frac{\phi}{4cs}\right)\\
&\ge 1 - K^2 s\exp\left(-\frac{N\phi^2}{2c^2s^2}\right).
\end{aligned}\tag{A.11}$$

Finally we have $c \le \sqrt{s}/b$. Therefore, part (2) is proven by combining (A.6), (A.8) (A.9) and (A.10), (A.11). This completes the proof. □

PROOF OF THEOREM 3.2. The proof is relegated to a supplementary file [Xue, Zou and Cai (2010)] for the sake of space. □

PROOF OF COROLLARY 3.1. It follows directly from Theorems 3.1 and 3.2; thus we omit its proof here. □

PROOF OF THEOREM 3.3. Under the event $\|\widetilde{\boldsymbol\beta}^{(0)} - \boldsymbol\beta^*\|_\infty \le \lambda$, we have $|\widetilde{\boldsymbol\beta}^{(0)}_{jk}| \le \lambda$ for $(j,k) \in \mathcal{A}^c$ and $|\widetilde{\boldsymbol\beta}^{(0)}_{jk}| \ge a\lambda$ for $(j,k) \in \mathcal{A}$. Therefore, $\widetilde{\boldsymbol\beta}^{(1)}$ is the solution of the following penalized composite likelihood:

$$\widehat{\boldsymbol\beta}^{(1)} = \arg\max_{\boldsymbol\beta}\left\{\ell_c(\boldsymbol\beta) - \lambda \sum_{(j,k)\in\mathcal{A}^c} |\beta_{jk}|\right\}. \tag{A.12}$$

It turns out that $\widehat{\boldsymbol\beta}^{\text{oracle}}$ is the global solution of (A.12) under the additional probability event that $\{\|\psi_{\mathcal{A}^c}(\widehat{\boldsymbol\beta}^{\text{oracle}})\|_\infty \le \lambda\}$. To see this, we observe that for any $\boldsymbol\beta$,

$$\begin{aligned}
&\left(-\ell_c(\boldsymbol\beta) + \lambda \sum_{(j,k)\in\mathcal{A}^c} |\beta_{jk}|\right) - \left(-\ell_c(\widehat{\boldsymbol\beta}^{\text{oracle}}) + \lambda \sum_{(j,k)\in\mathcal{A}^c} |\widehat{\beta}^{\text{oracle}}_{jk}|\right)\\
&\quad \ge \sum_{(j,k)\in\mathcal{A}^c} (\lambda - |\psi_{(jk)}(\widehat{\boldsymbol\beta}^{\text{oracle}})|) \cdot |\beta_{jk}|\\
&\quad \ge 0,
\end{aligned}$$

where we used the convexity of $-\ell_c$. In the proof of Theorem 3.1 we have shown that

$$\begin{aligned}
&\Pr(\|\psi_{\mathcal{A}^c}(\widehat{\boldsymbol\beta}^{\text{oracle}})\|_\infty > \lambda)\\
&\quad < K^2 \exp\left(-\frac{N\lambda^2}{32(2\phi+1)^2}\right) + \exp\left(-\frac{N\lambda}{3B(2\phi+1)s}\frac{b^2}{8^3}\right)\\
&\qquad + K^2 s \exp\left(-\frac{Nb^2}{2s^3}\right)\\
&\qquad + 2s^2\left[\exp\left(-\frac{b^2 N}{8s^3}\right) + \exp\left(-\frac{N}{s^2}\frac{b^2}{2}\right) + \exp\left(-\frac{N}{s^2}\frac{B^2}{8}\right)\right]\\
&\quad \equiv \tau_3.
\end{aligned}$$

Therefore, the LLA–CMA algorithm finds the oracle estimator with probability at least $1 - \tau_3 - \Pr(\|\widetilde{\boldsymbol\beta}^{(0)} - \boldsymbol\beta^*\|_\infty > \lambda)$. This proves part (1).

If we further consider the event $\{\min_{(j,k)\in\mathcal{A}} |\widehat{\beta}^{\text{oracle}}_{jk}| > a\lambda\}$. Then $\widetilde{\boldsymbol\beta}^{(2)}$ is the solution of the following penalized composite likelihood $\max_{\boldsymbol\beta}\{\ell_c(\boldsymbol\beta) - \lambda\sum_{(j,k)\in\mathcal{A}^c} |\beta_{jk}|\}$, which implies that $\widetilde{\boldsymbol\beta}^{(2)} = \widetilde{\boldsymbol\beta}^{(1)}$, and hence the LLA loop will stop. From (A.6) we have obtained a probability bound for the event of $\{\min_{(j,k)\in\mathcal{A}} |\widehat{\beta}^{\text{oracle}}_{jk}| \le a\lambda\}$ as follows:

$$\begin{aligned}
&\Pr\left(\min_{(j,k)\in\mathcal{A}} |\widetilde{\beta}^{\text{hmle}}_{jk}| \le a\lambda\right)\\
&\quad \le \exp\left(-R_*^2 \frac{b^2}{8^3}\right) + 2s^2 \exp\left(-\frac{N}{s^2}\frac{b^2}{2}\right) + 2s^2 \exp\left(-\frac{N}{s^2}\frac{B^2}{8}\right)\\
&\quad \equiv \tau_4.
\end{aligned}$$

Then we have $\widetilde{\boldsymbol{\beta}}^{(m)} = \widetilde{\boldsymbol{\beta}}^{(1)} = \widehat{\boldsymbol{\beta}}^{\text{oracle}}$ for $m = 2, 3, \ldots$ which means the LLA–CMA algorithm converges after two LLA iteration and finds the oracle estimator with probability at least $1 - \tau_3 - \Pr(\|\widetilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_\infty > \lambda) - \tau_4$. Note that $\tau_3 + \tau_4 = \tau_2$. This proves part (2). $\square$

Proof of Corollary 3.2. Part (1) follows directly from Theorem 3.3. We only prove part (2). With the chosen λ^{lasso}, Theorem 3.2 shows that with probability tending to one, $\widehat{\boldsymbol{\beta}}^{\text{lasso}}_{\mathcal{A}} = \widetilde{\boldsymbol{\beta}}_{\mathcal{A}}$, $\widehat{\boldsymbol{\beta}}^{\text{lasso}}_{\mathcal{A}^c} = 0$ and $\Pr(\|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}\|_2 \leq 16\lambda^{\text{lasso}}\sqrt{s}/b) \to 0$. Note that $16\lambda^{\text{lasso}}\sqrt{s}/b < \lambda^{\text{scad}}$ and $\|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}\|_\infty \leq \|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}\|_2$, we then conclude $\tau_0 = \Pr(\|\widehat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^*\|_\infty \leq \lambda^{\text{scad}}) \to 0$. $\square$

Acknowledgments. We thank the Editor, Associate Editor and referees for their helpful comments.

SUPPLEMENTARY MATERIAL

Supplementary materials for "Non-concave penalized composite likelihood estimation of sparse Ising models" (DOI: 10.1214/12-AOS1017SUPP; .pdf). In this supplementary file, we provide a complete theoretical analysis of the LASSO-penalized composite likelihood estimator for sparse Ising models.

REFERENCES

Atchley, W. R., Wollenberg, K. R., Fitch, W. M., Terhalle, W. and Dress, A. W. (2000). Correlations among amino acid sites in bHLH protein domains: An information theoretic analysis. *Mol. Biol. Evol.* **17** 164–178.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **36** 192–236. MR0373208

Bradic, J., Fan, J. and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73** 325–349. MR2815779

Bradic, J., Fan, J. and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.* **39** 3092–3120.

Bühlmann, P. and Meier, L. (2008). Discussion: "One-step sparse estimates in nonconcave penalized likelihood models," by H. Zou and R. Li. *Ann. Statist.* **36** 1534–1541. MR2435444

Candès, E. J., Wakin, M. B. and Boyd, S. P. (2008). Enhancing sparsity by reweighted l_1 minimization. *J. Fourier Anal. Appl.* **14** 877–905. MR2461611

Daubechies, I., Defrise, M. and De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.* **57** 1413–1457. MR2077704

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **39** 1–38. MR0501537

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. MR2640659

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

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33** 1–22.

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *J. Comput. Graph. Statist.* **7** 397–416. MR1646710

Genkin, A., Lewis, D. D. and Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics* **49** 291–304. MR2408634

Höfling, H. and Tibshirani, R. (2009). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *J. Mach. Learn. Res.* **10** 883–906. MR2505138

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *Amer. Statist.* **58** 30–37. MR2055509

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33** 1617–1642. MR2166557

Irback, A., Peterson, C. and Potthast, F. (1996). Evidence for nonrandom hydrophobicity structures in protein chains. *Proc. Natl. Acad. Sci. USA* **93** 533–538.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Z. Physik* **31** 53–258.

Lange, K., Hunter, D. R. and Yang, I. (2000). Optimization transfer using surrogate objective functions (with discussion). *J. Comput. Graph. Statist.* **9** 1–59. MR1819865

Lindsay, B. G. (1988). Composite likelihood methods. In *Statistical Inference from Stochastic Processes (Ithaca, NY, 1987). Contemporary Mathematics* **80** 221–239. Amer. Math. Soc., Providence, RI. MR0999014

Liu, Y., Eyal, E. and Bahar, I. (2008). Analysis of correlated mutations in HIV-1 protease using spectral clustering. *Bioinformatics* **24** 1243–1250.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37** 3498–3528. MR2549567

Majewski, J., Li, H. and Ott, J. (2001). The Ising model in physics and statistical genetics. *Am. J. Hum. Genet.* **69** 853–862.

Markowitz, M., Mo, H., Kempf, D. J., Norbeck, D. W., Bhat, T. N., Erickson, J. W. and Ho, D. D. (1995). Selection and analysis of human immunodeficiency virus type 1 variants with increased resistance to ABT-538, a novel protease inhibitor. *Journal of Virology* **69** 701–706.

Meier, L., van de Geer, S. and Bühlmann, P. (2008). The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 53–71. MR2412631

Meinshausen, N. (2007). Relaxed Lasso. *Comput. Statist. Data Anal.* **52** 374–393. MR2409990

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 417–473. MR2758523

Muzammil, S., Ross, P. and Freire, E. (2003). A major role for a set of non-Active site mutations in the development of HIV-1 protease drug resistance. *Biochemistry* **42** 631–638.

Ohtaka, H., Schön, A. and Freire, E. (2003). Multidrug resistance to HIV-1 protease inhibition requires cooperative coupling between distal mutations. *Biochemistry* **42** 13659–13666.

Ravikumar, P., Wainwright, M. J. and Lafferty, J. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319.

Rhee, S.-Y., Liu, T., Ravela, J., Gonzales, M. J. and Shafer, R. W. (2004). Distribution of human immunodeficiency virus type 1 protease and reverse transcriptase mutation patterns in 4,183 persons undergoing genotypic resistance testing. *Antimicrob. Agents Chemother.* **48** 3122–3126.

Rhee, S. Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L. and Shafer, R. W. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proc. Natl. Acad. Sci. USA* **103** 17355–17360.

Schelldorfer, J., Bühlmann, P. and van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scand. J. Stat.* **38** 197–214. MR2829596

Städler, N., Bühlmann, P. and van de Geer, S. (2010). ℓ_1-penalization for mixture regression models. *TEST* **19** 209–256. MR2677722

Stauffer, D. (2008). Social applications of two-dimensional Ising models. *American Journal of Physics* **76** 470–473.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. MR1379242

Tisdale, M., Myers, R. E., Maschera, B., Parry, N. R., Oliver, N. M. and Blair, E. D. (1995). Cross-resistance analysis of human immunodeficiency virus type 1 variants individually selected for resistance to five different protease inhibitors. *Antimicrob. Agents Chemother.* **39** 1704–1710.

Tseng, P. (1988). Coordinate ascent for maximizing nondifferentiable concave functions. Technical Report LIDS-P, 1840, Massachusetts Institute of Technology, Laboratory for Information and Decision Systems.

Varin, C. (2008). On composite marginal likelihoods. *AStA Adv. Stat. Anal.* **92** 1–28. MR2414624

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21** 5–42. MR2796852

Wang, H., Li, R. and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94** 553–568. MR2410008

Wu, M., Cai, T. and Lin, X. (2010). Testing for regression coefficients in lasso regularized regression. Technical report, Harvard Univ.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.* **2** 224–244. MR2415601

Wu, T. D., Schiffer, C. A., Gonzales, M. J., Taylor, J., Kantor, R., Chou, S., Israelski, D., Zolopa, A. R., Fessel, W. J. and Shafer, R. W. (2003). Mutation patterns and structural correlates in human immunodeficiency virus type 1 protease following different protease inhibitor treatments. *J. Virol.* **77** 4836–4847.

Xue, L., Zou, H. and Cai, T. (2010). Supplement to "Nonconcave penalized composite conditional likelihood estimation of sparse Ising models." Technical report, School of Statistics, Univ. Minnesota. Available at http://users.stat.umn.edu/~zouxx019/ftpdir/supplement/supplement-NPCL.pdf.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

Zhang, C.-H. (2010a). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

Zhang, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.* **11** 1081–1107. MR2629825

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443

L. Xue
H. Zou
School of Statistics
University of Minnesota
Minneapolis, Minnesota 55455
USA
E-mail: lzxue@stat.umn.edu
zouxx019@umn.edu

T. Cai
Department of Biostatistics
Harvard University
Boston, Massachusetts 02115
USA
E-mail: tcai@hsph.harvard.edu