Published in final edited form as:
J Am Stat Assoc. 2012 January 1; 107(497): 223–232. doi:10.1080/01621459.2011.645783.

Likelihood-based selection and sharp parameter estimation*

Xiaotong Shen[1], **Wei Pan**[2], and **Yunzhang Zhu**[1]

[1]School of Statistics, University of Minnesota, Minneapolis, MN 55455

[2]Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455

Summary

In high-dimensional data analysis, feature selection becomes one means for dimension reduction, which proceeds with parameter estimation. Concerning accuracy of selection and estimation, we study nonconvex constrained and regularized likelihoods in the presence of nuisance parameters. Theoretically, we show that constrained L_0-likelihood and its computational surrogate are optimal in that they achieve feature selection consistency and sharp parameter estimation, under one necessary condition required for any method to be selection consistent and to achieve sharp parameter estimation. It permits up to exponentially many candidate features. Computationally, we develop difference convex methods to implement the computational surrogate through prime and dual subproblems. These results establish a central role of L_0-constrained and regularized likelihoods in feature selection and parameter estimation involving selection. As applications of the general method and theory, we perform feature selection in linear regression and logistic regression, and estimate a precision matrix in Gaussian graphical models. In these situations, we gain a new theoretical insight and obtain favorable numerical results. Finally, we discuss an application to predict the metastasis status of breast cancer patients with their gene expression profiles.

Keywords

Coordinate decent; continuous but non-smooth minimization; general likelihood; graphical models; nonconvex; (p, n)-asymptotics

1 Introduction

Feature selection is essential to battle the inherited “curse of dimensionality” in high-dimensional analysis. It removes non-informative features to derive simpler models for interpretability, prediction and inference. In cancer studies, for instance, a patient’s gene expression is linked to her metastasis status of breast cancer, for identifying cancer genes. In a situation as such, our ability of identifying cancer genes is as critical as a model’s predictive accuracy, where selection accuracy becomes extremely important to reproducible findings and generalizable conclusions. Towards accuracy of selection and parameter estimation, we address several core issues in high-dimensional likelihood-based selection.

Consider a selection problem with nuisance parameters, based on a random sample $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ with each Y_i following probability density $g(\boldsymbol{\theta}^0, y)$, where $\boldsymbol{\theta}^0 = (\boldsymbol{\beta}^0, \eta^0)$ is a true parameter vector, $\beta^0 \equiv (\beta_1^0, \cdots, \beta_p^0) = (\beta_{A_0}^0, \mathbf{0}_{A_0^c})$ and $\eta^0 \equiv (\eta_1^0, \cdots, \eta_q^2)$ are the parameters of

*Research supported in part by NSF grant DMS-0906616, NIH grants 1R01GM081535, 2R01GM081535, HL65462 and R01HL105397. The authors would like to thank the editor, the associate editor and anonymous referees for helpful comments and suggestions.

interest and nuisance parameters respectively, $A_0=\{j:\beta_j^0 \neq 0\}$ is a set of nonzero coefficients of $\boldsymbol{\beta}^0$ with size $|A_0| = p_0$, and $\mathbf{0}_{A_0^c}$ is a vector of 0's with ^{c} denoting the set complement. Here we estimate $(\boldsymbol{\beta}^0, A_0)$, where p may greatly exceed n, and $q = 0$ is permitted.

For estimation and selection, a likelihood is regularized with regard to $\boldsymbol{\beta}$, particularly when $p > n$. This leads to an information criterion:

$$-L(\theta)+\lambda\sum_{j=1}^{p} I(\beta_j \neq 0), \tag{1}$$

where $L(\theta)=\sum_{i=1}^{n}\log g(\theta, Y_i)$ is the log-likelihood based on $\boldsymbol{Y}$, $\lambda > 0$ is a regularization parameter, and $\sum_{j=1}^{p} I(\beta_j \neq 0)$ is the L_0-function penalizing an increase in a model's size. In (1), when $\boldsymbol{\theta} = \boldsymbol{\beta}$ without nuisance parameters, $\lambda = 1$ is Akaike's information criterion, $\lambda=\frac{\log n}{2}$ is Bayesian information criterion [21], among others. In fact, essentially all selection rules can be cast into the framework of (1).

Regularization (1) has been of considerable interest for its interpretability and computational merits. Yet its constrained counterpart (2) has not received much attention, which is

$$-L(\theta), \text{ subject to } \sum_{j=1}^{p} I(\beta_j \neq 0) \leq K, \tag{2}$$

where $K \geq 0$ is a tuning parameter corresponding to λ in (1). Minimizing (1) or (2) in $\boldsymbol{\theta}$ gives a global minimizer leading to an estimate $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{\hat{A}}, \mathbf{0}_{\hat{A}^c})^T$, with $\hat{A}$ the estimated A_0, where $\boldsymbol{\eta}$ is un-regularized and possibly profiled out. Note that (1) and (2) may not be equivalent in their global minimizers, which is unlike a convex problem.

This article systematically investigates constrained and regularized likelihoods involving nuisance parameters, for estimating zero components of $\boldsymbol{\beta}^0$ as well as nonzero ones of $\boldsymbol{\theta}^0$. This includes, but is not limited to, estimating nonzero entries of a precision matrix in graphical models.

There is a huge body on parameter estimation through L_1-regularization in linear regression; see, for instance, [16] for a comprehensive review. For feature selection, consistency of the Lasso [26] has been extensively studied under the irrepresentable assumption; c.f., [15] [34]. Other methods such as the SCAD [6] have been studied. Yet L_0-constrained or regularized likelihood remain largely unexplored. Despite progress, many open issues remain. First, what is the maximum number of candidate features allowed for a likelihood method to reconstruct informative features? Results, such as [13], seem to suggest that the capacity of handling exponentially many features may be attributed primarily to the exponential tail of a Gaussian distribution, which we show is not necessary. Second, can parameter estimation be enhanced through removal of zero components of $\boldsymbol{\beta}$? Third, can a selection method continue to perform well for parameters of interest in the presence of a large number of nuisance parameters, as in covariance selection for off-diagonal entries of a precision matrix?

This article intends to address the foregoing three issues. First, we establish finite-sample mis-selection error bounds for constrained L_0-likelihood as well as its computational surrogate, given (n, p_0, p), where the surrogate–a likelihood based on a truncated L_1-function (TLP) approximating the L_0-function, permits efficient computation; see Section

2.1 for a definition. On this basis, we establish feature selection consistency for them as $n, p \to \infty$, under one key condition that is necessary for any method to be selection consistent:

$$C_{\min}(\theta^0) \geq d_0 \frac{\log p}{n}, \tag{3}$$

where $C_{\min}(\theta^0) \equiv \inf_{\{\theta_A=((\beta_A,\mathbf{0}_{A^c}),\eta):A\neq A_0,|A|\leq p_0\}} \frac{h^2(\theta_A,\theta^0)}{\max(|A_0\setminus A|,1)}$, $d_0 > 0$ is a constant, $|\cdot|$ and $\setminus$ denote the size of a set and that of set difference, respectively, $h(\theta,\theta^0)=\frac{1}{2}\left(\int (g^{1/2}(\theta,y)-g^{1/2}(\theta^0,y))^2 d\mu(y)\right)^{1/2}$ is the Hellinger-distance for with respect to a dominating measure μ, and $g(\boldsymbol{\theta}, y)$ is a probability density for Y_1. As one consequence, exponentially many candidate features $p=\exp\left(n\frac{C_{\min}(\theta^0)}{d_0}\right)$ are permitted for selection consistency with a broad class of constrained likelihoods. This challenges the well established result that the maximum number of candidate features permitted for selection consistency depends highly on a likelihood's tail behavior, c.f., [4]. In fact, selection consistency continues to hold even if the error distribution does not have an exponential tail; see Proportion 1 for linear regression. Second, sharper parameter estimation results from accurate selection by L_0-likelihood and its surrogate as compared to that without such selection. For feature selection in linear regression and logistic regression, the optimal Hellinger risk of the oracle estimator, the maximum likelihood estimate (MLE) based on A_0 as if the true A_0 were known *a priori*, is recovered by these methods, which is of order of $\sqrt{\frac{p_0}{n}}$ and is uniform over a certain L_0-band of $\boldsymbol{\theta}^0$ excluding the origin. This is in contrast to the minimax rate $\sqrt{\frac{u\log(p/u)}{n}}$ with $u \geq p_0$ for estimation without feature selection in linear regression [17]. In other words, accurate selection by L_0-likelihood and its surrogate over the L_0-band improves accuracy of estimation after non-informative features are removed, without introducing additional bias to estimation. Moreover, in estimating a precision matrix in Gaussian graphical models, the foregoing conclusions extend but with a different rate at $\sqrt{\frac{p_0\log p}{n}}$, where a log p factor is due to estimation of $2p$ nuisance parameters as compared to logistic regression. Third, two difference of convex (DC) methods are employed for computation of (1) and (2), which relax nonconvex minimization through a sequence of convex problems.

Two disparate applications are considered, namely, feature selection in generalized linear models (GLMs), as well as estimation of a precision matrix in Gaussian graphical models. In GLMs, feature selection in nonlinear regression appears more challenging than linear regression for a high-dimensional problem. In statistical modeling of a precision matrix in Gaussian graphical models, two major approaches have emerged to exploit matrix sparsity by likelihood selection and neighborhood selection. Papers based on these two approaches include [15] [12] [28] [2] [8] [19] [20] [18], among others. As suggested by [19], existing methods may not perform well when the dimension of a matrix exceeds the sample size n, although they give estimates better than the sample covariance matrix. In addition, theoretical aspects for a likelihood approach remain to be under studied. In these situations, the proposed method compares favorably against its competitors in simulations, and novel theoretical results provide an insight into a selection process.

This article is organized as follows. Section 2 develops the proposed method for L_0-regularized and constrained likelihoods. Section 3 presents main theoretical results for selection consistency and parameter estimation involving selection, followed by a necessary condition for selection consistency. Section 4 applies the general method and theory to feature selection in GLMs. Section 5 is devoted to estimation of a precision matrix in Gaussian graphical models. Section 6 presents an application to predict the metastasis status

of breast cancer patients with their gene expression profiles. Section 7 contains technical proofs.

2 Method and computation

2.1 Method

In a high-dimensional situation, it is computationally infeasible to minimize a discontinuous cost function involving the L_0-function in (1) and (2). As a surrogate, we seek a good approximation of the L_0-function by the TLP, defined as $J(|z|)=\lambda \min\left(\frac{|z|}{\tau}, 1\right)$, with $\tau > 0$ a tuning parameter controlling the degree of approximation; see Figure 1 for a display. This τ decides which individual coefficients to be shrunk towards zero. The advantages of $J(|z|)$ are fourfold, although $J(z)$ has been considered in other contexts [9]:

1. (Surrogate) It performs the model selection task of the L_0-function, while providing a computationally efficient means. Note that the approximation error of the TLP function to the L_0-function becomes zero when τ is tuned such that $\tau < \min\{|\beta_k^0|: k \in A_0\}$, seeking the *sparsest* solution by minimizing the number of non-zero coefficients.
2. (Adaptive model selection through adaptive shrinkage) It performs adaptive model selection [23] through a computationally efficient means when λ is tuned. Moreover, it corrects the Lasso bias through adaptive shrinkage combining shrinkage with thresholding.
3. (Piecewise linearity) It is piecewise linear, gaining computational advantages.
4. (Low resolutions) It discriminates small from large coefficients through thresholding. Consequently, it is capable of handling many low-resolution coefficients, through tuning τ.

To treat nonconvex minimization, we replace the L_0-function by its surrogate $J(\cdot)$ to construct an approximation of (2) and that of (1):

$$-L(\theta), \text{ subject to } \sum_{j=1}^{p} J(|\beta_j|) \le K, \tag{4}$$

$$S(\theta) = -L(\theta) + \lambda \sum_{j=1}^{p} J(|\beta_j|), \tag{5}$$

where (5) is a dual problem of (4). To solve (5) and (4), we develop difference convex methods for the primal and dual problems, for efficient computation.

2.2 Unconstrained dual and constrained primal problems

Our DC method for the dual problem (5) begins with a DC decomposition of $S(\boldsymbol{\theta})$: $S(\boldsymbol{\theta}) = S_1(\boldsymbol{\theta}) - S_2(\boldsymbol{\beta})$, where $S_1(\theta) = -L(\theta) + \lambda \sum_{j=1}^{p} J_1(|\beta_j|)$, $S_2(\beta) = \lambda \sum_{j=1}^{p} J_2(|\beta_j|)$, $J_1(|\beta_j|) = \frac{|\beta_j|}{\tau}$, and $J_2(|\beta_j|) = \frac{|\beta_j|}{\tau} - \max\left(\frac{|\beta_j|}{\tau} - 1, 0\right)$. Without loss of generality, assume that $-L$ is convex in $\boldsymbol{\theta}$; otherwise, a DC decomposition of $-L$ is required and can be treated similarly. Given this DC decomposition, a sequence of upper approximations of $S(\boldsymbol{\theta})$ is constructed iteratively, say, at iteration m, with ∇S_2 a subgradient of S_2 in $|\boldsymbol{\beta}|$: $S^{(m)}(\boldsymbol{\theta}) = S_1(\boldsymbol{\theta}) - (S_2(\hat{\boldsymbol{\beta}}^{(m-1)}) + (|\boldsymbol{\beta}| - |\hat{\boldsymbol{\beta}}^{(m-1)}|)^T \nabla S_2(|\hat{\boldsymbol{\beta}}^{(m-1)}|))$, by successively replacing $S_2(\boldsymbol{\beta})$ by its minorization, where $|\cdot|$ for a

vector takes the absolute value in each component. After ignoring $S_2(\widehat{\beta}^{(m-1)}) - \frac{\lambda}{\tau}\sum_{j=1}^{p}|\widehat{\beta}_j^{(m-1)}|I(|\widehat{\beta}_j^{(m-1)}|>\tau)$ that is independent of $\boldsymbol{\theta}$, the problem reduces to

$$S^{(m)}(\theta) = -L(\theta) + \frac{\lambda}{\tau}\sum_{j=1}^{p}|\beta_j|I(|\widehat{\beta}_j^{(m-1)}| \le \tau). \tag{6}$$

Minimizing (6) in $\boldsymbol{\theta}$ yields its minimizer $\hat{\boldsymbol{\theta}}^{(m)}$. The process continues in m until termination occurs. Our unconstrained DC method is summarized as follows.

Algorithm 1

Step 1. (Initialization) Supply a good initial estimate $\hat{\boldsymbol{\theta}}^{(0)}$, such as the minimizer of $S_1(\boldsymbol{\theta})$.

Step 2. (Iteration) At iteration m, compute $\hat{\boldsymbol{\theta}}^{(m)}$ by solving (6).

Step 3. (Termination) Terminate when $S(\hat{\boldsymbol{\theta}}^{(m-1)}) - S(\hat{\boldsymbol{\theta}}^{(m)}) \le \varepsilon$, and no components of $\hat{\boldsymbol{\beta}}^{(m)}$ is at $\pm\tau$. Otherwise, add ε to that components whose absolute value is τ, and go to **Step 2**, where ε is the square root of the machine precision. Then the estimate $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{(m^*)}$, where m^* is the smallest index at the termination criterion.

In Algorithm 1, (6) reduces to a general weighted Lasso problem: $-L(\theta) + \sum_{j=1}^{p}\lambda_j|\beta_j|$, with $\lambda_j = \frac{\lambda}{\tau}I(|\widehat{\beta}_j^{(m-1)}| \le \tau)$. Therefore any efficient software is applicable.

For (4), we decompose the nonconvex constraint into a difference of two convex functions to construct a sequence of approximating convex constraints. This amounts to solving the mth subproblem in a parallel fashion as in (6):

$$\min_{\beta} -L(\theta), \quad \text{subject to } \frac{1}{\tau}\sum_{j=1}^{p}|\beta_j|I(|\widehat{\beta}_j^{(m-1)}| \le \tau) \le K - \sum_{j=1}^{p}I(|\widehat{\beta}_j^{(m-1)}|>\tau). \tag{7}$$

This leads to a constrained DC algorithm—**Algorithm 2** for solving (4) by replacing (5) in **Algorithm 1** by (4).

Algorithms 1 and 2 are a generalization of those in [24] for a general likelihood, where all the computational properties there extend to the present situation, including equivalence of the DC solutions of the two algorithms and their convergence. Next we shall work with (5) due to its computational advantage. For instance, a coordinate decent method that works well with (5) breaks down for (4), c.f., [24].

3 Theory

This section presents a general theory for accuracy of reconstruction of the oracle estimator $\hat{\boldsymbol{\theta}}^{ml} = (\hat{\boldsymbol{\beta}}^{ml}, \hat{\boldsymbol{\eta}}^{ml})$ with $\widehat{\beta}^{ml} = (\widehat{\beta}^{ml}_{A_0}, \mathbf{0}_{A_0})$ given A_0, which is the MLE provided that the knowledge about A_0 were known *a priori*. As direct consequences, feature selection consistency is studied as as well as optimal parameter estimation defined by the oracle estimator. In addition, a necessary condition for feature selection will be established as well. A parallel theory for regularized likelihood is similar and thus is omitted.

3.1 Constrained L_0-likelihood

In (2), assume that a global minimizer exists, denoted by $\hat{\boldsymbol{\theta}}^{L_0} = (\hat{\boldsymbol{\beta}}^{L_0}, \hat{\boldsymbol{\eta}}^{L_0})$ with $\widehat{\beta}^{L_0}=(\widehat{\beta}^{L_0}_{\widehat{A}^{L_0}}, \mathbf{0}_{(\widehat{A}^{L_0})^c})$. Write $\boldsymbol{\beta}$ as $(\boldsymbol{\beta}_A, \mathbf{0}|_{A^c})$, with $\boldsymbol{\beta}_A$ being $(\beta_1, \ldots, \beta_{|A|})^T$ for any subset $A \subset \{1, \cdots, p\}$ of nonzero coefficients.

Before proceeding, we define a complexity measure for the size of a space $\mathcal{F}$. The bracketing Hellinger metric entropy of $\mathcal{F}$, denoted by the function $H(\cdot, \mathcal{F})$, is defined by logarithm of the cardinality of the u-bracketing (of $\mathcal{F}$) of the smallest size. That is, for a bracket covering $S(\varepsilon, m)=\{f_1^l, f_1^u, \cdots, f_m^l, f_m^u\} \subset \mathcal{L}_2$ satisfying $\max_{1\le j\le m}\left\|f_j^u - f_j^l\right\|_2 \le \varepsilon$ and for any $f \in \mathcal{F}$, there exists a j such that $f_j^l \le f \le f_j^u$, *a.e. P*, then $H(u, \mathcal{F})$ is $\log(\min\{m : S(u, m)\})$, where $\|f\|_2 = \int f^2(z)d\mu$. For more discussions about metric entropy of this type, see [14].

Assumption A (Size of parameter space)—For some constant $c_0 > 0$ and any $\frac{\varepsilon}{2^4}<t<\varepsilon \le 1$, $H(t, \mathcal{B}_A) \le c_0(\log p)^2|A| \log(2\varepsilon/t)$, with $|A| \le p_0$, where $\mathcal{B}_A = \mathcal{F}_A \cap \{h(\boldsymbol{\theta}, \boldsymbol{\theta}^0) \le 2\varepsilon\}$ is a local parameter space, and $\mathcal{F}_A = \{g^{1/2}(\boldsymbol{\theta}, y) : \boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}), \boldsymbol{\beta} = (\boldsymbol{\beta}_A, 0)\}$ be a collection of square-root densities.

Theorem 1 (*Error bound and oracle properties*)—Under Assumption A, if $K = p_0$, then, there exists a constant $c_2 > 0$, say $c_2=\frac{2}{27}\frac{1}{963}$, such that for (n, p_0, p),

$$P(\widehat{\theta}^{L_0} \ne \widehat{\theta}^{ml}) \le \exp(-c_2 n C_{\min}(\theta^0)+2\log(p+1)+3)). \tag{8}$$

Moreover, under (3) with $d_0 > \max(\frac{2}{c_2}, (2c_0)^{1/2}c_4^{-1}\log(2^{1/2}/c_3))$, $\hat{\boldsymbol{\theta}}^{L_0}$ reconstructs the oracle estimator $\hat{\boldsymbol{\theta}}^{ml}$ with probability tending to one as $n, p \to \infty$. Three oracle properties hold as $n, p \to \infty$:

- **A.** (Selection consistency) Estimator $\hat{A}^{L_0}$ is selection consistent, that is, $P(\hat{A}^{L_0} \ne A_0) \to 0$.
- **B.** (Optimal parameter estimation) For $\boldsymbol{\theta}^0$, $Eh^2(\widehat{\theta}^{L_0}, \theta^0)=(1+o(1))Eh^2(\widehat{\theta}^{ml}, \theta^0)=O(\varepsilon_{n,p}^2)$ and $h^2(\widehat{\theta}^{L_0}, \theta^0)=O_p(\varepsilon_{n,p_0,p}^2)$, provided that $Eh^2(\hat{\boldsymbol{\theta}}^{ml}, \boldsymbol{\theta}^0)$ does not tend to zero too fast in that $\frac{c_2}{2}nC_{\min}(\theta^0)+\log Eh^2(\widehat{\theta}^{ml}, \theta^0) \to \infty$, where $\varepsilon_{n,p0p}$ is any solution for ε:
$$\int_{2^{-8}\varepsilon^2}^{2^{1/2}\varepsilon} H^{1/2}(t/c_3, \mathcal{B}_{A_0})dt \le c_4 n^{1/2}\varepsilon^2. \tag{9}$$
- **C.** (Uniformity over a L_0-band) The reconstruction holds uniformly over $B_0(u, l)$, namely, $\sup_{\boldsymbol{\theta}^0 \in B_0(u,l)} P(\hat{\boldsymbol{\theta}}^{L_0} \ne \hat{\boldsymbol{\theta}}^{ml}) \to 0$, where $B_0(u,l)$ is a L_0-band, defined as $\{(\beta, \eta_0): p_0=\sum_{j=1}^{p} I(\beta_j \ne 0) \le u, C_{\min}(\theta) \ge l\}$ with $0 < u \le \min(n, p)$, $l=d_0\sigma^2\frac{\log p}{n}$, and $u < \min(n, p)$. This implies feature selection consistency $\sup_{\boldsymbol{\theta}^0 \in B_0(u,l)} P(\hat{A}^{L_0} \ne A_0) \to 0$, and optimal parameter estimation $\frac{\sup_{\theta^0 \in B_0(u,l)} Eh^2(\widehat{\theta}^{L_0}, \theta^0)}{\sup_{\theta^0 \in B_0(u,l)} Eh^2(\widehat{\theta}^{ml}, \theta^0)} \to 1$, with $\sup_{\theta^0 \in B_0(u,l)} Eh^2(\widehat{\theta}^{ml}, \theta^0)=O(\varepsilon_{n,u,p}^2)$, provided that $Eh^2(\hat{\boldsymbol{\theta}}^{ml}, \boldsymbol{\theta}^0)$ does not tend to zero too fast in that $\frac{c_2}{2}nC_{\min}(\theta^0)+\log\sup_{\theta^0 \in B_0(u,l)} Eh^2(\widehat{\theta}^{ml}, \theta^0) \to \infty$.

The L_0-method consistently reconstructs the oracle estimator when the degree of separation exceeds the minimal level, precisely under (3). As a result, selection consistency is established for the L_0-method. This, combined with that in Theorem 3, suggests that the L_0-method is optimal in feature selection against any method, matching up with the lower bound requirement under the degree of separation with respect to (p, p_0, n) except a constant factor $d_0 > 0$ in Theorem 3. Moreover, the optimality extends further to parameter estimation, where sharper parameter estimation is obtained from accurate L_0 selection, achieving the optimal Hellinger risk of the oracle estimator asymptotically. By comparison, such a result is not expected for L_1-regularization. As suggested in [17], selection consistency of Lasso does not give sharper parameter estimation, where the rate of convergence of a L_1-method in the L_2 risk remains to be $\sqrt{\frac{p_0 \log(p/p_0)}{n}}$ in linear regression. This is because a L_1-method is nonadaptive and overpenalizes large coefficients as a result of shrinking small coefficients towards zero. Similarly, in feature selection in logistic regression, the L_0-method is expected to give rise to better estimation precision than a L_1-method, although a parallel result for a L_1-method has not been available. Finally, the uniform result in (C) is over a L_0-band $B_0(u, l)$, which is not expected over a L_0-ball $B_0(u, 0)$ in view of the result of Theorem 3.

3.2 Constrained truncated L_1-likelihood

For constrained truncated L_1-likelihood, one additional regularity condition–Assumption B is assumed, which is generally met with a smooth likelihood; see Section 4 for an example. It requires the Hellinger-distance to be smooth so that the TLP approximation to the L_0-function becomes adequate through tuning τ.

Assumption B—For some constants $d_1 - d_3 > 0$,

$$h^2(\theta, \theta^0) \geq d_1 h^2(\theta_{\tau^+}, \theta^0) - d_3 p \tau^{d_2}, \quad A^{\tau^+} \equiv \{j : |\beta_j| \geq \tau\}, \tag{10}$$

where $\boldsymbol{\theta}_{\tau}{}^{+} = (\beta_1 I(|\beta_1| \geq \tau), \cdots, \beta_p I(|\beta_p| \geq \tau), \eta_1, \cdots, \eta_q)$.

Theorem 2 (*Error bound and oracle properties*)—Under Assumption A with $\mathcal{F}_A$ replaced by $\{g^{1/2}(\boldsymbol{\theta}, y) : \boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}) : \boldsymbol{\beta} = (\boldsymbol{\beta}_A, \boldsymbol{\beta}_{A^c}), \|\boldsymbol{\beta}_{A^c}\|_{\ell_\infty} \leq \tau\}$, say $0 \leq \tau \leq c' \varepsilon$ for some constant c', and Assumption B, if $K = p_0$ and $\tau \leq \max(c', (d_1 C_{min}(\boldsymbol{\theta}^0)/2pd_3)^{1/d_2})$, then there exists a constant $c_2 > 0$, such that for any (n, p_0, p),

$$P(\widehat{\theta}^T \neq \widehat{\theta}^{ml}) \leq \exp\left(-c_2 n C_{\min}(\theta^0) + 2\log(p+1) + 3\right). \tag{11}$$

Moreover, under (3) with sufficiently large constant $d_0 > 0$, $\hat{\boldsymbol{\theta}}^T$ has the three oracle properties (A)–(C) of $\hat{\boldsymbol{\theta}}^{L_0}$, provided that $\frac{c_2 d_1}{4} n C_{\min}(\theta^0) + \log E h^2(\widehat{\theta}^{ml}, \theta^0) \to \infty$. For (C), $\tau \leq (\frac{d_1 l}{2pd_3})^{1/d_2}$ is required as well as $\frac{c_2 d_1}{4} n C_{\min}(\theta^0) + \log \sup_{\theta^0 \in B_0(u,l)} E h^2(\widehat{\theta}^{ml}, \theta^0) \to \infty$.

Remark—Constants in Theorem 1 can be made precise. For instance, $c_2 = \frac{4}{27} \frac{1}{1926}$ and $d_0 > \max(\frac{4}{c_2 d_1}, (2c_0)^{1/2} c_4^{-1} \log(2^{1/2}/c_3))$.

Theorem 2 says that the oracle properties of the L_0-function are attained by its computational surrogate when τ is sufficiently small.

3.3 Necessary condition for selection consistency

This section establishes the necessary condition (3) by estimating the minimal value d_0 in (3), required for feature selection consistency.

Let $K(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = E \log g(\boldsymbol{\theta}_1, Y)/g(\boldsymbol{\theta}_2, Y)$ be the Kullback-Leibler loss for $\boldsymbol{\theta}_1$ versus $\boldsymbol{\theta}_2$, where E is taken with regard to $g(\boldsymbol{\theta}_1, Y)$. Let $\gamma_{\min}(\theta^0) \equiv \min\{|\beta_k^0|: k \in A_0\} > 0$.

Assumption C—For a constant $r > 0$, $K(\theta_j, \theta_k) \le r\gamma_{\min}^2(\beta)$. Here $\{\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j, \boldsymbol{\eta}^0), j = 1, \cdots, p\}$ is a set of parameters, where $\beta_j = \sum_{k=1}^{p_0} \gamma_{min} \boldsymbol{e}_k - \gamma_{min} \boldsymbol{e}_j$; $j = 1, \cdots, p_0\}$, and $\beta_j = \sum_{k=1}^{p_0} \gamma_{min} \boldsymbol{e}_k + \gamma_{min} \boldsymbol{e}_j$; $j = p_0 + 1, \cdots, p$, and $\boldsymbol{e}_j = (\underbrace{0, \ldots, 0}_{j-1}, 1, \underbrace{0, \ldots, 0}_{p-j-1})^T$. Assume that $s \equiv \inf_{\theta^0} \frac{C_{\min}(\theta^0)}{\gamma_{\min}^2(\theta^0)} > 0$.

Theorem 3 (*Necessary condition for feature selection consistency*)—Under Assumption *C*, for any constant $c_* \in (0, 1)$, any (n, p_0, p) with $p_0 \le p/2$, and any $\boldsymbol{\eta}^0$, we have

$$\inf_{\widehat{A}} \sup_{\{\beta^0: C_{\min}(\theta^0) = R^*\}} P\left(\widehat{A} \ne A_0\right) \ge c_*, \tag{12}$$

with $R^* = \frac{s(1-c_*)\log p}{4rn}$. Moreover,

$$\inf_{\widehat{A}} \sup_{\theta^0 \in B_0(u,l)} P\left(\widehat{A} \ne A_0\right) \ge c_*, \quad as \quad n, p \to \infty, \tag{13}$$

where $u \le \min(p/2, n)$, $l = d_0 \frac{\log p}{n}$, and $d_0 = \frac{(1-c^*)s}{4r}$.

Theorem 3 says that feature selection inconsistency occurs when $d_0 < \frac{s}{4r}$ in (3). There the minimal value $d_0 = \frac{s}{4r}$ yields a requirement for feature selection consistency in (3).

4 Generalized linear models

For GLMs, observations $Y_i = (Z_i, \boldsymbol{X}_i)$ are paired, response Z_i is assumed to follow an exponential family with density function $g(z_i; \theta_i, \varphi) = \exp\{[z_i\theta_i - b(\theta_i)]/a(\varphi) + c(z_i, \varphi)\}$, where θ_i is the natural parameter that is related to the mean $\mu_i = E(z_i) = b'(\theta_i)$, and φ is a dispersion parameter. With a link function g, a regression model becomes $\eta_i = g(\mu_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i$. The penalized likelihood for estimating regression coefficient vectors $\boldsymbol{\beta}$ is $-L(\beta) + \sum_{j=1}^{p} J(|\beta_j|; \lambda, \tau)$, where $L(\beta) = \sum_{i=1}^{n} [z_i\mu_i - b(\mu_i)]/a(\varphi) + c(z_i, \varphi)$ is the log-likelihood, and $J(|\beta_j|; \lambda, \tau) = \frac{\lambda}{\tau} \min(|\beta_j|, \tau)$ is the TLP penalty.

For parameter estimation and feature selection, we apply **Algorithm 1**, where (6) becomes a series of weighted lasso for GLMs, for which some existing routines are applicable, for simplicity. In implementation, we use the function wtlassoglm() in R package SIS.

Next we examine effectiveness of the proposed method through simulated examples in feature selection. In linear regression and logistic regression, the Lasso, SCAD [6], SCAD-OS, TLP and TLP-OS are compared in terms of predictive accuracy and identification of the true model, where SCAD-OS and TLP-OS are SCAD and TLP with only one iteration step

in the DC iterative process, and SCAD-OS is proposed in [33]. The latter four methods use the Lasso as an initial estimate.

4.1 Simulations

For simulations predictors $\boldsymbol{X}_i$'s are iid from $N(0, \boldsymbol{V})$, where $\boldsymbol{V}$ is a $p \times p$ matrix whose ijth element is $0.5^{|i-j|}$. In linear regression, $Z_i = \boldsymbol{\beta}^T\boldsymbol{X}_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$; $i = 1, \cdots, n$, and random error ε_i is independent of $\boldsymbol{X}_i$; in logistic regression, a binary response is generated from logit $\Pr(Z_i = 1) = \boldsymbol{\beta}^T\boldsymbol{X}_i$. In both cases, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ with $\beta_1 = 1$, $\beta_2 = 0.5$ and $\beta_5 = 0.75$; $\beta_j = 0$ for $j \neq 1, 2, 5$. This set-up was similar to that considered in [33]; here we examine various situations with respect to p, n. Each simulation is based on 1000 independent replications.

For any given tuning parameter λ, all other methods use the Lasso estimate as an initial estimate. For each method, we choose its tuning parameter values by maximizing the log-likelihood based on a common tuning dataset with an equal sample size of the training data and independent of the training data. This is achieved through a grid search over 21 λ values returned by glmnet() for all the methods, and additionally over a grid of 10 τ values that are the 9th-, 19th-, 29th-, ⋯, 99th-percentiles of the final Lasso estimate for the TLP.

The model error (ME) is used to evaluate predictive performance of $\hat{\boldsymbol{\beta}}$, defined as $ME(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T\boldsymbol{V}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$, which is the prediction error minus σ^2 in linear regression, corresponding to the test error over an independent test sample of size $T = \infty$. In our context, the median ME's are reported over 1000 simulation replications, due to possible skewness of the distribution of ME. In addition, the mean parameter estimates of the nonzero elements of $\boldsymbol{\beta}$ will be reported, together with the mean true positive (TP) and mean false positive (FP) numbers: $\#TP = \sum_{j=1}^{p} I(\beta_j \neq 0, \widehat{\beta}_j \neq 0)$ and $\#FP = \sum_{j=1}^{p} I(\beta_j = 0, \widehat{\beta}_j \neq 0)$.

For linear regression, simulation results are reported for the cases of $p = 12$, 500, 1000, $n = 50$, 100, and $\sigma^2 = 1$ in Table 1. As suggested by Table 1, the TLP performs best: it gives the smallest estimation and prediction error as measured by the ME, the smallest mean false positive number (FP) while maintaining a comparable mean number of true positives (TP) around 3. Most critically, as p increases, the TLP's performance remains much more stable than its competitors. On a relative basis, the TLP outperforms its competitors more in more difficult situations.

For logistic regression, simulation results are summarized for the cases of $p = 12$, 200, 500 and $n = 100$, 200 in Table 2. As expected, the TLP continues to outperform other methods with the smallest median ME's. It gives less biased estimates than the Lasso estimates. The TLP's superior performance remains strong over other methods, as p increases.

4.2 Theory for feature selection

This section establishes some theoretical results to gain an insight into performance of the proposed method in feature selection. Let $Y = (Z, \boldsymbol{X})$, and $g(\beta, Z) = \frac{1}{2\sqrt{\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(Z - \beta^T\boldsymbol{X})^2)$ and $g(\boldsymbol{\beta}, Z) = p^Z(1-p)^{1-Z}$ in linear and logistic regression. Assume that $\beta^T\boldsymbol{x} = \beta_A^T\boldsymbol{x}_A$ belongs to a compact parameter space for any model size $|A| \leq p_0$. In this case, selection does not involve nuisance parameters, where $\boldsymbol{\theta} = \boldsymbol{\beta}$. Under (14), we establish feature selection consistency as well as optimal parameter estimation for the TLP:

$$C_{\min}(\beta^0) = \min_{A:|A|\leq p_0, A\neq A_0} \frac{1}{\max(|A_0\setminus A|,1)}\left(\beta^0_{A_0\setminus A}\right)^T\left(\sum_{A_0\setminus A} - \sum_{A_0\setminus A,A}\sum_A^{-1}\sum_{A,A_0\setminus A}\right)\beta^0_{A_0\setminus A} \geq d_0\frac{\log p}{n}, \tag{14}$$

where $d_0 > 0$ is a constant independent of (n, p, p_0), and Σ_B is a sub-matrix given a subset B of predictors, of covariance matrix Σ with the jkth element $Cov(X_j, X_k)$, independent of $\boldsymbol{\beta}^0$. A simpler but stronger condition can be used for verification of (14):

$$\gamma^2_{\min} \min_{A:|A|\le p_0, A\ne A_0} c_{\min}\left(\sum\nolimits_{A_0\setminus A} - \sum\nolimits_{A_0\setminus A,A}\sum\nolimits_{A}^{-1}\sum\nolimits_{A,A_0\setminus A}\right) \ge d_0\frac{\log p}{n}, \tag{15}$$

where $\gamma_{\min}=\gamma_{\min}(\beta^0) \equiv \min\{|\beta^0_k| : \beta^0_k \ne 0\}$ is the resolution level of the true regression coefficients, $\min_{A:|A|\le p_0, A\ne A_0} c_{\min}\left(\sum_{A_0\setminus A} - \sum_{A_0\setminus A,A}\sum_{A}^{-1}\sum_{A,A_0\setminus A}\right) \ge \min_{B\supset A_0:|B|\le 2p_0} c_{\min}(\sum_B)$, and $c_{\min}$ denotes the smallest eigenvalue. Note that (14) is necessary for any method to be selection consistent except constant d_0 if $\min_{A:|A|\le p_0, A\ne A_0} c_{\min}\left(\sum_{A_0\setminus A} - \sum_{A_0\setminus A,A}\sum_{A}^{-1}\sum_{A,A_0\setminus A}\right) > 0$.

Proposition 1—Under (14), the constrained MLE $\hat{\boldsymbol{\beta}}^T$ of (4) consistently reconstructs the oracle estimate $\hat{\boldsymbol{\beta}}^{ml}$. As $n, p \to \infty$, feature selection consistency is established for the TLP as well as optimal parameter estimation $Eh^2(\widehat{\beta}^T, \beta^0) = Eh^2(\widehat{\beta}^{ml}, \beta^0) = O(\frac{p_0}{n})$ under the Hellinger distance $h(\cdot, \cdot)$. Moreover, the results hold uniformly over a L_0-band $B_0(u, l) = \{\beta_0 : \sum_{j=1}^{p} I(\beta^0_j \ne 0) \le u, \gamma^2_{\min}(\beta^0) \min_{B\supset A_0:|B|\le 2p_0} c_{\min}(\sum_B) \ge l\}$, with $0 < u \le \min(n, p)$, $l = d_0\sigma^2\frac{\log p}{n}$, that is, as $n, p \to \infty$,

$$\sup_{\beta^0\in B_0(u,l)} P\left(\widehat{\beta}^T \ne \widehat{\beta}^{ml}\right) \to 0, \quad \frac{\sup_{\beta^0\in B_0(u,l)} Eh^2(\widehat{\beta}^T, \beta^0)}{\sup_{\beta^0\in B_0(u,l)} Eh^2(\widehat{\beta}^{ml}, \beta^0)} \to 1,$$

with $\sup_{\beta^0\in B_0(u,l)} Eh^2(\widehat{\beta}^{ml}, \beta^0) = d^*\frac{u}{n}$ for some d^*.

Various conditions have been proposed for studying feature selection consistency in linear regression. In particular, a condition on $\gamma_{\min}$ is usually imposed, in addition to assumptions on the design matrix $\boldsymbol{X}$ such as the sparse Riesz condition in [32]. To compare (14) with existing assumptions for consistent selection, note that these assumptions imply a fixed design version of (14) by necessity of consistent feature selection. For instance, as showed in [32], the sparse Riesz condition with dimension restriction and $\gamma^2_{\min} \ge c'\frac{\log(p-u)}{n}$, required for the minimum concavity penalty to be consistent, imply (15) with p replaced by $p - u$ thus (14) when p/u bounded away from 1, where $u \ge p_0$. Moreover, the number of over-selected variables is proved to be bounded but may not tend to zero for thresholding Lasso in Theorem 1.1 of [35], under a restrictive eigenvalue condition [1] and a requirement on $\gamma_{\min}$. Finally, in linear regression, only finite variance σ^2 is required for the proposed method, which is in contrast to a commonly used assumption on sub-Gaussian distribution of ε_i.

In conclusion, the computational surrogate–the TLP method indeed shares desirable oracle properties of the L_0-method, which is optimal against any selection method, for feature selection and parameter estimation.

5 Estimation of a precision matrix

Given n random samples from a p-dimensional normal distribution $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we estimate the inverse covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ that is $p \times p$ positive definite, denoted by $\boldsymbol{\Omega} \succ 0$. For estimation of $(\boldsymbol{\mu}, \boldsymbol{\Omega})$, the log-likelihood is proportional to

$$\frac{n}{2}\log det(\boldsymbol{\Omega}) - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \mu)^T\boldsymbol{\Omega}(\boldsymbol{Y}_i - \mu). \tag{16}$$

The profile log-likelihood for $\boldsymbol{\Omega}$, after $\boldsymbol{\mu}$ is maximized out, is proportional to $\frac{n}{2}\log det(\boldsymbol{\Omega}) - \frac{1}{2}tr(\boldsymbol{S}\boldsymbol{\Omega})$, where $\overline{\boldsymbol{Y}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{Y}_i$ and $\boldsymbol{S} = n^{-1}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}})(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}})^T$ are the corresponding sample mean and covariance matrix, *det* and *tr* denote the determinant and trace. In (16), the number of unknown parameters p^2 in $\boldsymbol{\Omega}$ can greatly exceed the sample size n in the presence of $2p$ nuisance parameters $(\boldsymbol{\mu}, \{\Omega_{jj} : j = 1, \cdots, p\})$, where Ω_{jk} denotes the jkth elements of $\boldsymbol{\Omega}$. To avoid non-identifiability in estimation, we regularize off-diagonal elements of $\boldsymbol{\Omega}$ in (16) through a nonnegative penalty function $J(\cdot)$ for the $\frac{p(p-1)}{2}$ parameters of interest:

$$S(\boldsymbol{\Omega}) = \log det(\boldsymbol{\Omega}) - \frac{1}{2}tr(\boldsymbol{S}\boldsymbol{\Omega}) - \sum_{j,k=1, j\neq k}^{p} J(\{\Omega_{jk}, j \neq k\}). \tag{17}$$

In estimation, the TLP function $J(\{\Omega_{jk}, j \neq k\}) = \frac{\lambda}{\tau}\min(|\Omega_{jk}|, \tau)$ is employed for both parameter estimation and covariance selection in (17). Towards this end, we apply **Algorithm 1** to solve (6) sequentially, which reduces to a series of weighted graphical lasso problems, and is solved by taking advantage of existing software. In implementation, we use R package glasso [8] for (6).

5.1 Simulations

Simulations are performed, where a tridiagonal precision matrix is used as in [7]. In particular, $\boldsymbol{\Sigma}$ is AR(1)-structured with its ij-element being $\sigma_{ij} = \exp(-a|s_i - s_j|)$, and $s_1 < s_2 < \cdots < s_p$ are randomly chosen: $s_i - s_{i-1} \sim \text{Unif}(0.5, 1)$, for some $a > 0$; $i = 2, \cdots, p$. The following situations are considered: $(n, p) = (120, 30)$ or $(n, p) = (120, 200)$, and $a = 0.9$ or $a = 0.6$, based on 100 replications.

Five competing methods are compared, including Lasso, adaptive Lasso (ALasso), SCAD-OS and SCAD, and TLP-OS and TLP. ALasso uses weight $\lambda/|\widehat{\beta}_j^{(0)}|^{\gamma}$, where $\hat{\boldsymbol{\beta}}^{(0)}$ is an initial estimate and $\gamma = 1/2$ as in [7].

To measure performance of estimator $\hat{\boldsymbol{\Omega}}$, we use the entropy loss and quadratic loss: $\text{loss}_1(\boldsymbol{\Omega}, \hat{\boldsymbol{\Omega}}) = tr(\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\Omega}}) - \log|\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\Omega}}| - p$, and $\text{loss}_2(\boldsymbol{\Omega}, \hat{\boldsymbol{\Omega}}) = tr(\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\Omega}} - I)^2$, as well as the true positive (TP) and false positive (FP) numbers: $\#TP = \Sigma_{i,j}I(\Omega_{ij} \neq 0, \hat{\Omega}_{ij} \neq 0)$; $\#FP = \Sigma_{i,j}I(\omega_{ij} = 0, \hat{\Omega}_{ij} \neq 0)$.

For small $p = 30$, TLP and TLP-OS are always among the winners. It is also confirmed that the one-step approximation to SCAD or TLP gives similar performance to that of the fully iterated SCAD or TLP, respectively. For large p, to save computing time, as advocated in [7], we only run SCAD-OS and TLP-OS. In such a situation, an improvement of TLP-OS over other methods is more substantial for large $p = 200$ than for small = 30. Overall, the

proposed method delivers higher performance in low-dimensional and high-dimensional situations, respectively.

5.2 Theory for precision matrix

To perform theoretical analysis, we specify a parameter space Θ in which $\mathbf{\Omega} \in 0$ with $0 < \max_{1\le j\le p}|\Omega_{jj}| \le M_2$, $c_{\min}(\mathbf{\Omega}) \ge M_1 > 0$, for some constants M_1, $M_2 > 0$, independent of (n, p, p_0). Let $A = \{(j, k) : j \ne k, \Omega_{jk} \ne 0\}$ be the set of nonzero off diagonal elements of $\mathbf{\Omega}$, where $|A| = p_0$ is an even number by symmetry of $\mathbf{\Omega}$, and $\mathbf{\Omega}$ depends on A. Results in Theorem 1 imply that the constrained MLE yields covariance selection consistency under one assumption:

$$C_{\min}(\mathbf{\Omega}^0) \ge d_0 \frac{\log p}{n}, \tag{18}$$

which is necessary for covariance selection consistency indeed for any method, up to constant d_0 when $c_{\min}(H) > 0$, where $d_0 > 0$ is a constant independent of (n, p, p_0), and $H = \left(\frac{\partial^2(-\log\det(\mathbf{\Omega}))}{\partial^2\mathbf{\Omega}}\right)|_{\Omega=\Omega_0}$ is the $p^2 \times p^2$ Hessian matrix of $-\log det(\mathbf{\Omega})$, whose $(\Omega_{jk}, \Omega_{j'k'})$ element is $tr(\Sigma^0\Delta_{jk}\Sigma^0\Delta_{j'k'})$, c.f., [3], Δ_{jK} is a $p \times p$ with the jk-element being 1 and 0 otherwise. Sufficiently, (18) can be verified using

$$C_{\min}(\mathbf{\Omega}^0) \ge \gamma^2_{min} c_{\min}(H), \tag{19}$$

with $\gamma_{\min}(\mathbf{\Omega}^0) \equiv \gamma_{\min} = \min\{|\Omega^0_{jk}| : \Omega^0_{jk} \ne 0, j \ne k\}$.

Proposition 2—Under (18), the constrained MLE $\hat{\mathbf{\Omega}}^T$ of (4) consistently reconstructs the oracle estimator $\hat{\mathbf{\Omega}}^{ml}$. As $n, p \to \infty$, covariance selection consistency is established for the TLP as well as optimal parameter estimation $Eh^2(\widehat{\mathbf{\Omega}}^T, \mathbf{\Omega}^0) = (1+o(1))Eh^2(\widehat{\mathbf{\Omega}}^{ml}, \mathbf{\Omega}^0) = O(\frac{p_0 \log p}{n})$, where $h^2(\mathbf{\Omega}, \mathbf{\Omega}^0) = 1 - \sqrt{\frac{(det(\mathbf{\Omega})det(\mathbf{\Omega}^0))^{1/2}}{det(\frac{\Omega+\Omega^0}{2})}}$ is the squared Hellinger distance for $\mathbf{\Omega}$ versus $\mathbf{\Omega}^0$. Moreover, the above results hold uniformly over a L_0-band $B_0(u, l) = \{\mathbf{\Omega}^0 : \sum_{j,k=1, j\ne k}^{p} I(\Omega^0_{jk} \ne 0) \le u, \gamma^2_{min}(\mathbf{\Omega}^0) c_{\min}(H) \le l\}$, with $0 < u \le \min(n, p)$ and $l = d_0\sigma^2 \frac{\log p}{n}$, that is, as $n, p \to \infty$,

$$\sup_{\mathbf{\Omega}^0 \in B_0(u,l)} P\left(\widehat{\mathbf{\Omega}}^T \ne \widehat{\mathbf{\Omega}}^{ml}\right) \to 0, \quad \frac{\sup_{\mathbf{\Omega}^0 \in B_0(u,l)} Eh^2(\widehat{\Omega}^T, \mathbf{\Omega}^0)}{\sup_{\mathbf{\Omega}^0 \in B_0(u,l)} Eh^2(\widehat{\Omega}^{ml}, \mathbf{\Omega}^0)} \to 1,$$

with $\sup_{\mathbf{\Omega}^0 \in B_0(u,l)} Eh^2(\widehat{\mathbf{\Omega}}^{ml}, \mathbf{\Omega}^0) = d^* \frac{u \log p}{n}$ for some $d^* > 0$.

In short, the TLP method is optimal against any method in covariance selection, permitting p up to exponentially large in the sample size, or $p^2 \le p_0 \exp\left(n \frac{\gamma^2_{\min} c_{\min}(H)}{d_0}\right)$. Moreover, as a result of accurate selection of this method, parameter estimation can be sharply enhanced at an order of $\sqrt{\frac{p_0 \log p}{n}}$, as measured by the Hellinger distance, after zero off-diagonal elements are removed. Note that the log p factor is due to estimation of $2p$ nuisance parameters as

compared to the rate of $\sqrt{\frac{p_0}{n}}$ in logistic regression. In view of the result in Lemma 1, this result seems to be consistent with the minimax rate $\sqrt{\frac{\log p}{n}}$ under the L_∞ matrix norm [20].

6 Metastasis status of breast cancer patients

We apply the penalized logistic regression methods to analyze a microarray gene expression dataset of [29], where our objectives are (1) to develop a model predicting the metastasis status, and (2) to identify cancer genes, for breast cancer patients. Among the 286 patients, metastasis was detected in 106 patients during follow-ups within 5 years after surgery. Their expression profiles were obtained from primary breast tumors with Affymetrix HG-133a GeneChips.

In [29], a 76-gene signature was developed based on a training set of 115 patients, which yielded a misclassification error rate of 64/171=37.4% when applied to the remaining samples. [30] compared the performance of a variety of classifiers using a subset of 245 genes drawn from 33 cancer-related pathways: based on a 10-fold cross-validation (CV). Their non-parametric pathway-based regression method yielded the smallest error rate at 29%, while random forest, bagging and Support Vector Machine (SVM) had error rates of 33%, 35% and 42%.

In our analysis, we first performed a preliminary screening of the genes using a marginal t-test to select the top p genes with most significant p-values, based on the training data for each fold of a 10-fold-CV. Then the training data were split into two parts to fit penalized logistic models and to select tuning parameters, respectively. The results were summarized in Table 3, including the total misclassification errors and average model sizes (i.e. non-zero estimates) based on 10-fold CV. A final model is obtained by fitting the best model selected from a 10-fold CV to the entire data set.

With regard to prediction, no large difference is seen among various methods, with the error rates ranging from 102/286=35.7% (of TLP and TLP-OS with $p = 200$) to 118/286=41.3% (of ALasso with $p = 200$). The TLP performed similar to the TLP-OS, both were among the winners. In addition, the Lasso gave the least sparse models while the SCAD gave the most sparse models.

With regard to identifying cancer genes, the Lasso, TLP-OS and TLP yield the same model, identifying the largest number of cancer genes, whereas the SCAD and SCAD-OS give the most sparse models with only at most 2 cancer genes, and ALasso only yields 10 cancer genes. Here cancer genes are defined according to the Cancer Gene Database [10].

In summary, the TLP and TLP-OS identify a good proportion of cancer genes and lead to a model giving a reasonably good predictive accuracy of the metastasis status. In this sense, they perform well with regard to the foregoing two objectives.

References

1. Bickel P, Ritov Y, Tsybakov A. Simultaneous analysis of lasso and dantzig selector. Ann Statist. 2008; 37:1705–1732.
2. Benerjee O, Ghmoui LE, dAspremont A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. J Mach Learn Res. 2008; 9:485–516.
3. Boyd, S.; Vandenberghe, L. Convex optimization. Cambridge Univ. Press; 2004.
4. Chen J, Chen Z. Extended Bayesian information criterion for model selection with large model space. Biometrika. 2008; 95:759–771.
5. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Statist. 2004; 32:407–499.

6. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Amer Statist Assoc. 2001; 96:1348–1360.
7. Fan J, Feng Y, Wu Y. Network exploration via the adaptive Lasso and SCAD penalties. Ann Appl Statist. 2009; 3:521–541.
8. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9:432–441. [PubMed: 18079126]
9. Gasso, G.; Raotomamonjy, A.; Canu, S. Recovering sparse signals with nonconvex penalties and DC programming. 2009. Submitted
10. Higgins ME, Claremont M, Major JE, Sander C, Lash AE. Cancer-Genes: a gene selection resource for cancer genome projects. Nucleic Acids Research. 2007; 35 (suppl 1):D721–D726. [PubMed: 17088289]
11. Ibragimov, IA.; Has'minskii, RZ. Statistical estimation. Springer; New York: 1981.
12. Li H, Gui J. Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. Biostat. 2006; 7:302–317.
13. Kim Y, Choi H, Oh HS. Smoothly clipped absolute deviation of high dimensions. J Amer Statist Assoc. 2008; 103:1665–1673.
14. Kolmogorov AN, Tihomirov VM. ε-entropy and ε-capacity of sets in function spaces. Uspekhi Mat Nauk. 1959; 14:3–86. [In Russian. English translation, *Ameri.* Math. Soc. transl. 2, 17, 277–364. (1961)].
15. Meinshausen N, Buhlmann P. High dimensional graphs and variable selection with the lasso. Ann Statist. 2006; 34:1436–1462.
16. Negahban S, Wainwright M, Ravikumar P, Yu B. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. 2010
17. Raskutti, G.; Wainwright, M.; Yu, B. Tech Report. UC Berkeley: 2009. Minimax rates of estimation for high-dimensional linear regression over l_q balls.
18. Rocha, G.; Zhao, P.; Yu, B. Technical Report. UC Berkeley: 2008. A path following algorithm for sparse pseudo-likelihood inverse covariance estimation; p. 759
19. Rothman A, Bickel P, Levina E, Zhu J. Sparse permutation invariant covariance estimation. Electronic J Statist. 2008; 2:494–515.
20. Rothman A, Bickel P, Levina E, Zhu J. A new approach to Cholesky-based covariance regularization in high dimensions. Biometrika. 2009 To appear.
21. Schwarz G. Estimating the dimension of a model. Ann Statist. 1978; 6:461–64.
22. Shen X, Wong WH. Convergence rate of sieve estimates. Ann Statist. 1994; 22:580–615.
23. Shen X, Ye J. Adaptive model selection. J Amer Statist Assoc. 2002; 97:210–221.
24. Shen X, Zhu Y, Pan W. Necessary and sufficient conditions towards feature selection consistency. 2010 Unpublished manuscript.
25. Stanica P, Montgomery AP. Good lower and upper bounds on binomial coefficients. J Ineq in Pure Appl Math. 2001; 2:art 30.
26. Tibshirani R. Regression shrinkage and selection via the LASSO. JRSS-B. 1996; 58:267–288.
27. Wong WH, Shen X. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. The Annals of Statistics. 1995; 23:339–362.
28. Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. Biometrika, bf. 2007; 94:19–35.
29. Wang Y, Klijn JG, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005; 365:671–679. [PubMed: 15721472]
30. Wei Z, Li H. Nonparametric pathway-based regression models for analysis of genomic data. Biostatistics. 2007; 8:265–284. [PubMed: 16772399]
31. Yuan M. High dimensional inverse covariance matrix estimation via linear programming. J Mach Learning Res. 2010; 11:2261–2286.
32. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. Ann Statist. 2010; 38:894–942.

33. Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models (with discussion). Ann Statist. 2008; 36:1509–1566.
34. Zhao P, Yu B. On model selection consistency of Lasso. JMLR. 2006; 7:2541–2563.
35. Zhou S. Thresholded Lasso for high dimensional variable selection and statistical estimation. Technical report. 2010

7 Appendix

Proof of Theorem 1

The proof uses a large deviation probability inequality of [27] to treat one-sided log-likelihood ratios with constraints. This enables us to obtain sharp results without a moment condition on both tails of the log-likelihood ratios.

When $K = p_0$, $|\hat{A}^{L0}| \le p_0$. If $\hat{A}^{L0} = A_0$, then $\hat{\boldsymbol{\beta}}^{L0} = \hat{\boldsymbol{\beta}}^{ml}$. Let a class of candidate subsets be $\{A : A \neq A_0, |A| \le p_0\}$ for feature selection. Note that $A \subset \{1, \cdots, p\}$ can be partitioned into $(A \setminus A_0) \cup (A_0 \cap A)$. Let $B_{kj} = \{\boldsymbol{\theta} = (\boldsymbol{\beta}_A, \mathbf{0}, \boldsymbol{\eta}) : A \neq A_0, |A_0 \cap A| = k, |A \setminus A_0| = j, (p_0 - k)C_{\min}(\boldsymbol{\theta}^0) \le h^2(\boldsymbol{\theta}, \boldsymbol{\theta}^0)\} \subset \mathcal{F}_A$; $k = 0, \cdots, p_0 - 1$, $j = 1, \cdots, p_0 - k$. Note that B_{kj} consists of $\binom{p_0}{k}\binom{p - p_0}{j}$ different elements A's of sizes $|A_0 \cap A| = k$ and $|A \setminus A_0| = j$. By definition, $\{\theta=(\beta_A, \mathbf{0}, \eta):A \neq A_0:C_{\min}(\theta^0) \le h^2(\theta, \theta^0), |A| \le p_0\} \subset \cup_{k=0}^{p_0-1}\cup_{j=1}^{p_0} B_{kj}$. Hence

$$
\begin{aligned}
P(\widehat{\theta}^{L_0} \neq \widehat{\theta}^{ml}) &\le P^*\left(\sup_{\theta=(\beta,\eta);\beta=(\beta_A,\mathbf{0}),A\neq A_0,|A|\le p_0} (L(\theta) - L(\widehat{\theta}^{ml}))>0\right) \\
&\le P^*\left(\sup_{\theta=(\beta,\eta):\beta=(\beta_A,\mathbf{0}),A\neq A_0,|A|\le p_0} (L(\theta) - L(\theta^0))>0\right) \\
&\le \sum_{A\subset\{1,\cdots,p\}:A\neq A_0,|A|\le p_0} P^*\left(\sup_{\theta=(\beta,\eta);\beta=(\beta_A,\mathbf{0})} (L(\theta) - L(\theta^0)) \ge 0\right) \equiv I
\end{aligned}
$$

where P^* is the outer measure and $L(\hat{\boldsymbol{\theta}}^{ml}) \ge L(\boldsymbol{\theta}^0)$ by definition.

For I, we apply Theorem 1 of [27] to bound each term. Towards this end, we verify the entropy condition (3.1) there for the local entropy over $\mathcal{B}_A$. Note that under Assumption A $\varepsilon=\varepsilon_{n,p_0,p}=(2c_0)^{1/2}c_4^{-1}\log(2^{1/2}/c_3)\log p(\frac{p_0}{n})^{1/2}$ satisfies there with respect to $\varepsilon > 0$, that is,

$$
\sup_{\{0\le|A|\le p_0\}} \int_{2^{-8}\varepsilon^2}^{2^{1/2}\varepsilon} H^{1/2}(t/c_3, \mathcal{B}_A)dt \le p_0^{1/2}2^{1/2}\varepsilon\log(2/2^{1/2}c_3) \le c_4 n^{1/2}\varepsilon^2. \tag{20}
$$

for some constant $c_3 > 0$ and c_4, say $c_3 = 10$ and $c_4=\frac{(2/3)^{5/2}}{512}$. Moreover, by Theorem 2.6 of [25], $\binom{b}{a} \le \frac{b^{b+1/2}}{\sqrt{2\pi}a^{a+1/2}(b-a)^{b-a+1/2}} \le \exp((a+1/2)\log(b/a)+a)$ for any integers $a < b$. By (3), $C_{\min}(\theta^0) \ge \varepsilon^2_{n,p_0,p}$ implies (20), provided that $d_0>(2c_0)^{1/2}c_4^{-1}\log(2^{1/2}/c_3)$. Using the facts about binomial coefficients: $\sum_{j=0}^{p_0-k}\binom{p - p_0}{j} \le (p - p_0+1)^{p_0-k}$ and $\binom{p_0}{i} \le p_0^i$, we obtain, by

Theorem 1 of [27], that for a constant $c_2 > 0$, say $c_2=\frac{4}{27}\frac{1}{1926}$, I is upper bounded by

$$\sum_{k=0}^{p_0-1}\sum_{j=0}^{p_0-k} P^*(\sup_{\theta\in B_{kj}}(L(\theta)-L(\theta^0))\geq 0) \leq 4\sum_{k=0}^{p_0-1}\binom{p_0}{k}\exp(-c_2n(p_0-k)C_{\min}(\theta^0))\sum_{j=0}^{p_0-k}\binom{p-p_0}{j}$$
$$\leq 4\sum_{i=1}^{p_0}\exp\left(-i(c_2nC_{\min}(\theta^0)-\log(p-p_0+1)-\log p_0)\right)$$
$$\leq R\left(\exp(-(c_2nC_{\min}(\theta^0)-\log(p-p_0+1)-\log p_0))\right),$$

where $R(x) = x/(1-x)$ is the exponentiated logistic function. Note, moreover, that $I \leq 1$ and $\log(p-p_0+1)+\log p_0 \leq 2\log(p+1)/2 \leq 2\log\frac{p+1}{2}$. Then

$$I \leq 5\exp(-c_2nC_{\min}(\theta^0)+2\log\frac{p+1}{2}) \leq \exp(-c_2nC_{\min}(\theta^0)+2\log(p+1)+3).$$

Finally, (A) follows from $P(\hat{A}^{L0}\neq A_0)\leq P(\hat{\boldsymbol{\theta}}^{L0}\neq\hat{\boldsymbol{\theta}}^{ml})$, (8) and (3) with $d_0>\frac{2}{c_2}$, as $n, p\to\infty$. For (B), let $G=\{\hat{\boldsymbol{\theta}}^{L0}\neq\hat{\boldsymbol{\theta}}^{ml}\}$ and $P(G)\leq 8\exp(-c_2nC_{\min}/4)$ by (8) and (3). For the risk property, $Eh^2(\hat{\boldsymbol{\theta}}^{L0},\boldsymbol{\theta}^0)\leq Eh^2(\hat{\boldsymbol{\theta}}^{ml},\hat{\boldsymbol{\theta}}^0)+Eh^2(\hat{\boldsymbol{\theta}}^{L0},\boldsymbol{\theta}^0)I(G)$ is upper bounded by

$$Eh^2(\widehat{\theta}^{ml},\theta^0)+4\exp(-c_2nC_{\min}/2)=(1+o(1))Eh^2(\widehat{\theta}^{ml},\theta^0),$$

using the fact that $h(\hat{\boldsymbol{\theta}}^{L0},\boldsymbol{\theta}^0)\leq 1$. Then (B) is established. Similarly (C) follows. This completes the proof.

Proof of Theorem 2

The proof is basically the same as that in Theorem 1 with a modification that A is replaced by A^{τ^+}. Now $B_{kj}=\{\boldsymbol{\theta}_{\tau^+}: A^{\tau^+}\neq A_0, |A_0\cap A^{\tau^+}|=k, |A^{\tau^+}\setminus A_0|=j, (d_1(p_0-k)C_{\min}(\boldsymbol{\theta}^0)-d_3p\tau^{d_2})\leq h^2(\boldsymbol{\theta}_{\tau^+},\boldsymbol{\theta}^0)\}$; $j=1,\cdots,p_0$. Then

$$\{\theta=(\beta_A,\mathbf{0},\eta): A\neq A_0: \sum_{j=1}^{p}J(|\beta_j|)\leq p_0, C_{\min}(\theta^0)\leq h^2(\theta,\theta^0)\}\subset\cup_{k=0}^{p_0-1}\cup_{j=1}^{p_0-k}B_{kj}.$$

When $K=p_0$, $\sum_{j=1}^{p}J(|\beta_j|)\leq p_0$, implying that $|\hat{A}^+|\leq p_0$. If $|\hat{A}^+|=p_0$, then $\sum_{j=1}^{p}|\beta_j|I(|\beta_j|\leq\tau)=0$, implying that $\hat{\boldsymbol{\theta}}^T=\hat{\boldsymbol{\theta}}^{ml}$. Then we focus our attention to the case of $A^+\neq A_0$. Note that, with $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\eta})$ and $\boldsymbol{\beta}=(\boldsymbol{\beta}_A,\mathbf{0})$,

$$P^*\left(\sup_{\theta:A\neq A_0,\sum_{j=1}^{p}J(|\beta_j|)\leq p_0}(L(\theta)-L(\theta^0))\geq 0\right)\leq\sum_{j=1}^{p_0}P^*\left(\sup_{\theta\in B_j}(L(\theta))-L(\theta^0))\geq 0\right)$$
$$\leq 4\sum_{k=0}^{p_0-1}\sum_{j=0}^{p_0-k}\binom{p-p_0}{j}\binom{p_0}{k}\exp(-c_2n(d_1C_{\min}(\theta^0)-d_3p\tau^{d_2}))$$
$$\leq 5\exp(-(c_2d_1/2)nC_{\min}(\theta^0)+2\log\tfrac{p+1}{2})\leq\exp(-c_2nC_{\min}(\theta^0)+2\log(p+1)+3),$$

provided that $\tau\leq(d_1C_{\min}(\boldsymbol{\theta}^0)/2pd_3)^{1/d_2}$. The rest of the proof proceeds as in the proof of Theorem 1. This completes the proof.

Proof of Theorem 3

The main idea of the proof is the same as that for Theorem 1 of [24], which constructs an approximated least favorable situation for feature selection and uses Fano's Lemma. According to Fano's Lemma [11], for any mapping $T = T(Y_1, \cdots, Y_n)$ taking values in $S = \{1, \cdots, |S|\}$, $|S|^{-1}\sum_{j=1}^{|S|} P_j(T(Y_1, \cdots, Y_n) = j) \le \sum_{1 \le j,k \le |S|} \frac{nK(q_j,q_k)+\log 2}{|S|^2 \log(|S|-1)}$, where $K(q_j, q_k) = \int q_j \log(q_j/q_k)$ is the Kullback-Leibler information for densities q_j versus q_k corresponding P_j and P_k.

To construct an approximated least favorable set of parameters S for A_0 versus A_0^c, define $\boldsymbol{\beta}$ to be $(\gamma_{\min}\mathbf{1}_{|A_0|}, \mathbf{0}_{|A_0^c|})$. Let $S = \{\theta_j = (\beta_j, \eta^0)\}_{j=0}^{p}$ be a collection of parameters with components equal to γ_{min} or 0 satisfying that for any $1 \le j, j' = p$, $\left\|\beta_{j'} - \beta_j\right\|^2 \le 4\gamma_{min}^2$, as defined in Assumption C. Then for any $\boldsymbol{\theta}_j, \boldsymbol{\theta}_k \in S$, $K(\theta_j, \theta_k) \le r\gamma_{\min}^2 \le n\frac{r}{s} C_{\min}(\theta^0)$ by Assumption C.

By Fano's lemma, $|S|^{-1}\sum_{j \in S} P_j(T = j) \le \frac{n(r/s)C_{\min}(\theta^0)+\log 2}{\log p}$, implying that

$$\sup_{\{\theta : C_{\min}(\theta^0) = R^*\}} P(\widehat{A} \ne A_0) \ge 1 - \frac{nrC_{\min}(\theta^0) + s\log 2}{s\log p},$$

bounded below by c_* with $R^* = \frac{s(1-c_*)\log p}{4rn}$. This yields (12). For (13), it follows that $R^* \ge l$ with $l = d_1 \frac{\log p}{n}$ and $d_1 = \frac{(1-c^*)s}{4r}$, for any $\boldsymbol{\theta}^0 \in B_0(u, l)$. This completes the proof.

Proof of Proposition 1

We now verify Assumptions A–C. Note that

$$h^2(\beta, \beta^0) = 2E\left(1 - \exp(-\frac{1}{8}(\beta^T \boldsymbol{X} - (\beta^0)^T \boldsymbol{X})^2)\right)$$

for linear regression, and $h^2(\boldsymbol{\beta}, \boldsymbol{\beta}^0)$ is

$$\frac{1}{2}\left(E(\mu^{1/2}((\beta^0)^T \boldsymbol{X}) - \mu^{1/2}(\beta^T \boldsymbol{X}))^2 + (1 - \mu((\beta^0)^T \boldsymbol{X}))^{1/2} - (1 - \mu(\beta^T \boldsymbol{X}))^{1/2}\right),$$

for logistic regression, where $\mu(s) = (1 + \exp(s))^{-1}$.

Assumption A follows from [14]. Note that $A = A^{\tau^+} \cup A_2^{\tau^-}$ and $\left|\frac{\partial h^2(\beta,\beta^0)}{\partial \beta_j}\right| \le \frac{1}{2}E(|X_j|)$, for $1 \le j \le p$ and $\boldsymbol{\beta} \in \mathcal{R}^p$. Thus

$$|h^2(\beta, \beta^0) - h^2(\beta_{\tau^+}, \beta^0)| = \tau\left|\sum_{j \in A^{\tau^-}} \frac{\partial h^2(\beta, \beta^0)}{\partial \beta_j}\right|_{\beta = \beta^\star} \le 2\tau \sum_{j \in A^{\tau^+}} E(|X_j|) \le 2\tau p \max_j \sum_{jj}.$$

Then Assumption B is fulfilled with $d_1 = d_2 = 1$ and $d_3 = 2 \max_j \Sigma_{jj}$.

To simplify (3), we derive an inequality through some straightforward calculations: with $\tilde{\boldsymbol{\beta}} = ((\boldsymbol{\beta}_A, \mathbf{0}) - (\mathbf{0}, \boldsymbol{\beta}_{A_0}))$

$$C_{\min}(\beta^0) \geq c_1^* \min_{\beta_A : A \neq A_0, |A| \leq p_0} |A_0 \backslash A|^{-1} E(\beta_A \boldsymbol{X}_A - \beta_{A_0} \boldsymbol{X}_{A_0})^2$$
$$\geq c_1^* \min_{\beta_A : A \neq A_0, |A| \leq p_0} |A_0 \backslash A|^{-1} \tilde{\beta}^T \Sigma_{A \cup A_0} \tilde{\beta} \geq \gamma_{\min}^2 \min_{B : |B| \leq 2p_0, A_0 \subset B} c_{\min}(\Sigma_B).$$

for some constant $c_1^* > 0$, because the derivative of $1 - \exp(-\frac{1}{8}x^2)$ and $(1 + \exp(x))^{-1/2}$ are bounded away from zero under the compactness assumption. This leads to (14). By Theorem 2, the TLP has the properties (A)–(C) there, through tuning.

Finally, $K(\beta_j, \beta_k) \leq cE(\beta_{A_j} \boldsymbol{X}_{A_j} - \beta_{A_k} \boldsymbol{X}_{A_k})^2 \leq r\gamma_{\min}^2$ by the compactness assumption, where $r = c \max_{(A_j, A_k)} E(\boldsymbol{\beta}_{A_j} \boldsymbol{X}_{A_j} - \boldsymbol{\beta}_{A_k} \boldsymbol{X}_{A_k})^2$. By Theorem 3, (14) except a constant $d_0 > 0$ is necessary for any method to be feature selection consistent. This completes the proof.

Proof of Proposition 2

To obtain the desired results, Theorems 1–3 are applied. First a lower bound of $C_{\min}(\boldsymbol{\theta}^0)$ is derived to simplify (3). Given the squared Hellinger distance

$h^2(\theta, \theta^0) = 1 - \sqrt{\frac{(det(\boldsymbol{\Omega}) det(\boldsymbol{\Omega}^0))^{1/2}}{det(\frac{\boldsymbol{\Omega}+\boldsymbol{\Omega}^0}{2})}} e^{-\frac{1}{4}(\mu-\mu^0)^T(\boldsymbol{\Omega}+\boldsymbol{\Omega}^0)(\mu-\mu^0)}$, by strong convexity of $-\log det(\boldsymbol{\Omega})$, c.f., [3], for any $\boldsymbol{\theta} \in \Theta$ and a constant $c^* > 0$ depending on M_1,

$$-2\log(1 - h^2(\theta, \theta^0)) \geq -\frac{1}{2}(\log\det(\boldsymbol{\Omega}) + \log\det(\boldsymbol{\Omega}^0)) + \log\det\left(\frac{\boldsymbol{\Omega}+\boldsymbol{\Omega}^0}{2}\right)$$
$$\geq \frac{1}{8} tr((\boldsymbol{\Omega}^*)^{-1}(\boldsymbol{\Omega} - \boldsymbol{\Omega}^0)(\boldsymbol{\Omega}^*)^{-1}(\boldsymbol{\Omega} - \boldsymbol{\Omega}^0)) \geq c^* |A^0 \backslash A| c_{\min}(H) \gamma_{\min}^2,$$

where A^0 and A are as defined in Section 4.2, and $\boldsymbol{\Omega}^*$ is an intermediate value between $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}^0$; see A.4.3 of [3] of such an expansion. Moreover, $C_{\min}(\boldsymbol{\theta}^0) \geq \inf_{\boldsymbol{\Omega}_A A \neq A_0, |A| \leq p_0} \log(1 - h^2(\boldsymbol{\Omega}, \boldsymbol{\Omega}^0))$, yielding (18).

For Assumption A, note that $|\Omega_{jk}| \leq (\Omega_{jj}\Omega_{kk}) \leq M_2$; $j \neq k$, because $\boldsymbol{\Omega} \succ 0$ and $det(\boldsymbol{\Omega})$ is bounded away from zero. To calculate the bracketing Hellinger metric entropy, we apply Proposition 1 of [22]. Let $\boldsymbol{\Omega}_A$ be a submatrix, consisting of p_0 nonzero off-diagonal elements of $\boldsymbol{\Omega}$. Note that $g(\boldsymbol{\theta}, \boldsymbol{y})$ of $\boldsymbol{Y}_1$ is proportional to $h_0(\boldsymbol{\theta}_A, \boldsymbol{y}) \Pi_{j \in A^c} h_j(\theta_j, y_1)$, where

$h_0(\theta_A, \boldsymbol{y}) = (det(\boldsymbol{\Omega}_A))^{n/2} \exp(-\frac{1}{2}(\boldsymbol{y}_A - \mu_A)^T \boldsymbol{\Omega}_A (\boldsymbol{y}_A - \mu_A))$, $h_j(\theta_j, \boldsymbol{y}) = det(\Omega_{jj}) \exp(-\frac{1}{2}(y_j - \mu_j)^2 \Omega_{jj})$, and $\boldsymbol{y}_A$ and $\boldsymbol{\theta}_A$ are the sub-vectors of $\boldsymbol{y}$ and $\boldsymbol{\theta}$ corresponding to $\boldsymbol{\Omega}_A$. Then for some constants $k_j > 0$; $j = 1, 2$ and any $\bar{\Omega}, \boldsymbol{\Omega} \in \Theta$,

$$|g^{1/2}(\bar{\theta}, \boldsymbol{y}) - g^{1/2}(\theta, \boldsymbol{y})| \leq k |g(\bar{\theta}, \boldsymbol{y}) - g(\theta, \boldsymbol{y})|$$
$$\leq k_1 k_2^{p-p_0} \left(|h_0(\bar{\theta}_A, \boldsymbol{y}) - h_0(\theta_A, \boldsymbol{y})| + \sum_{j \in A^c} |h_j(\bar{\theta}_j, \boldsymbol{y}) - h_j(\theta_j, \boldsymbol{y})| \right).$$

This implies that $H(t, \mathcal{R}^p) = c_0(|A| \log(2\varepsilon p/t) + \log((p - |A|)p/t))$ by [14] for some constant c_0, which in turn yields Assumption A. For Assumption B, note that, for $j \neq k = 1, \cdots, p$, for any $\boldsymbol{\theta} \in \Theta$,

$$\left|\frac{\partial h^2(\theta, \theta^0)}{\partial \Omega_{jk}}\right| = \frac{1}{4}\left|(1 - h^2(\theta, \theta^0))tr\left((2\left(\frac{\boldsymbol{\Omega} + \boldsymbol{\Omega}^0}{2}\right)^{-1} - \boldsymbol{\Omega}^{-1})\boldsymbol{\Delta}_{jk}\right)\right|,$$

which is upper bounded by $\left|\frac{1}{c_{\min}(\boldsymbol{\Omega}) + c_{\min}(\boldsymbol{\Omega}^0)} + \frac{1}{4c_{\min}(\boldsymbol{\Omega})}\right| \leq \frac{2}{M_1}$; $j \neq k = 1, \cdots, p$. With $A = A^{\tau^+} \cup A_2^{\tau^-}$, $|h^2(\theta, \theta^0) - h^2(\theta_{\tau^+}, \theta^0)| = \tau\left|\sum_{j \in A^{\tau^-}} \frac{\partial h^2(\theta, \theta^0)}{\partial \Omega_{jk}}\right|_{\boldsymbol{\Omega} = \boldsymbol{\Omega}^\star}\Big|$. This implies Assumption B with $d_1 = d_2 = 1$ and $d_3 = \frac{2}{M_1}$. For Assumption C, note that the Kullback-Leibler for $\boldsymbol{\theta}^0$ versus $\boldsymbol{\theta}$ is $\frac{1}{2n}(\log \frac{det(\boldsymbol{\Omega}^0)}{det(\boldsymbol{\Omega})} + tr(\boldsymbol{\Omega}\sum^0) + (\mu - \mu^0)^T \boldsymbol{\Omega}(\mu - \mu^0) - n)$, which is upper bounded by $h^2(\boldsymbol{\theta}, \boldsymbol{\theta}^0)$, because the likelihood ratios are uniformly bounded. An application of Taylor's expansion as in verification of Assumption A yields that $-2\log(1 - h^2(\theta, \theta^0)) \leq r\gamma_{\min}^2$, where $r = c^* c_{\max}(H)$, leading to Assumption C.

The results in (18) follow from Theorems 1–3 with $\varepsilon_{n,p_0,p} = \max((p_0 \log p)^{1/2}, (\log(p - p_0)p_0)^{1/2})n^{-1/2} = \sqrt{\frac{p_0 \log p}{n}}$ by solving (9). This completes the proof.

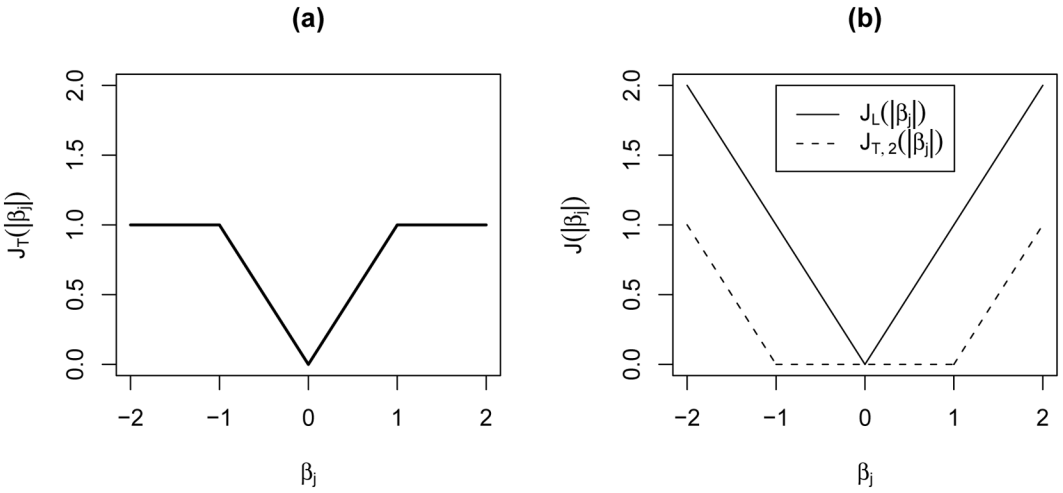

Figure 1.
Truncated L_1 function $J_t(|\beta_j|)$ with $\tau = 1$ in (a), and its DC decomposition into a dfference of two convex functions J_L and $J_{T,2}$ in (b).

Table 1

Median ME's, means (SD in parentheses) nonzero coefficients (β_1, β_2, β_5), and true positive (TP) and false positive (FP) numbers of nonzero estimates, for linear regression, based on 1000 simulation replications.

n	*p*	Method	ME	$\beta_1 = 1$	$\beta_2 = .5$	$\beta_5 = .75$	#TP	#FP
50	12	Lasso	.129	.91(.17)	.41(.18)	.60(.16)	2.98(0.14)	3.82(2.39)
		SCAD-OS	.109	1.02(.19)	.40(.22)	.68(.18)	2.92(0.27)	2.50(1.97)
		SCAD	.118	1.04(.20)	.39(.24)	.71(.18)	2.88(0.32)	2.30(1.90)
		TLP-OS	.088	1.01(.18)	.41(.20)	.68(.17)	2.94(0.25)	1.65(2.04)
		TLP	.090	1.01(.19)	.41(.21)	.69(.17)	2.92(0.27)	1.57(1.98)
50	500	Lasso	.431	.76(.19)	.29(.18)	.41(.17)	2.90(0.30)	14.7(10.48)
		SCAD-OS	.327	1.01(.24)	.25(.25)	.52(.24)	2.70(0.47)	14.81(8.69)
		SCAD	.301	1.09(.26)	.21(.27)	.59(.26)	2.53(0.53)	12.25(7.63)
		TLP-OS	.150	1.02(.21)	.39(.26)	.67(.22)	2.75(0.45)	4.27(6.86)
		TLP	.143	1.02(.21)	.39(.26)	.68(.22)	2.75(0.45)	4.10(6.89)
50	1000	Lasso	.501	.72(.19)	.28(.18)	.37(.18)	2.88(0.33)	17.20(11.49)
		SCAD-OS	.370	.99(.25)	.26(.25)	.51(.26)	2.67(0.49)	18.76(9.60)
		SCAD	.327	1.08(.26)	.20(.28)	.57(.29)	2.49(0.54)	15.19(8.41)
		TLP-OS	.182	1.01(.20)	.40(.27)	.66(.25)	2.72(0.47)	5.43(8.69)
		TLP	.175	1.02(.20)	.40(.27)	.66(.25)	2.72(0.47)	5.06(8.30)
100	12	Lasso	.063	.94(.12)	.44(.13)	.65(.11)	3.00(0.00)	3.94(2.42)
		SCAD-OS	.042	1.01(.12)	.45(.14)	.72(.11)	2.99(0.08)	2.17(2.04)
		SCAD	.042	1.02(.13)	.45(.15)	.74(.11)	2.99(0.09)	2.06(2.01)
		TLP-OS	.037	1.01(.12)	.45(.13)	.71(.11)	3.00(0.06)	1.51(1.99)
		TLP	.036	1.01(.12)	.45(.13)	.72(.11)	3.00(0.07)	1.46(1.95)
100	500	Lasso	.186	.84(.12)	.36(.12)	.52(.11)	3.00(0.04)	15.61(11.04)
		SCAD-OS	.118	1.06(.14)	.32(.18)	.66(.14)	2.94(0.24)	14.92(10.45)
		SCAD	.121	1.10(.15)	.30(.21)	.71(.12)	2.89(0.31)	14.20(10.00)
		TLP-OS	.036	1.01(.12)	.47(.14)	.72(.11)	2.99(0.12)	3.64(6.57)
		TLP	.035	1.01(.12)	.46(.14)	.72(.11)	2.99(0.12)	3.49(6.69)

J Am Stat Assoc. Author manuscript; available in PMC 2013 January 01.

n	p	Method	ME	$\beta_1 = 1$	$\beta_2 = .5$	$\beta_3 = .75$	#TP	#FP
100	1000	Lasso	.211	.83(.13)	.34(.13)	.51(.12)	3.00(0.06)	18.10(12.50)
		SCAD-OS	.142	1.06(.15)	.30(.19)	.66(.15)	2.91(0.29)	19.70(13.35)
		SCAD	.147	1.10(.15)	.27(.22)	.72(.14)	2.83(0.38)	18.80(12.74)
		TLP-OS	.037	1.01(.13)	.46(.15)	.74(.12)	2.97(0.18)	3.93(7.04)
		TLP	.037	1.01(.13)	.46(.15)	.74(.12)	2.97(0.18)	3.80(6.94)

Table 2

Median ME's, means (SD in parentheses) nonzero coefficients (β_1, β_2, β_5), and true positive (TP) and false positive (FP) numbers of nonzero estimates, for logistic regression, based on 1000 simulation replications.

n	*p*	Method	ME	$\beta_1 = 1$	$\beta_2 = .5$	$\beta_5 = .75$	#TP	#FP
100	12	Lasso	.388	.80(.27)	.35(.25)	.49(.27)	2.8(0.4)	3.8(2.2)
		SCAD-OS	.416	1.03(.37)	.39(.36)	.61(.40)	2.5(0.7)	1.6(1.9)
		SCAD	.472	1.10(.41)	.38(.39)	.67(.44)	2.4(0.7)	1.1(1.9)
		TLP-OS	.350	.98(.35)	.36(.32)	.59(.35)	2.7(0.5)	1.8(2.0)
		TLP	.355	.98(.35)	.35(.32)	.58(.35)	2.6(0.6)	1.8(2.0)
100	200	Lasso	.947	.57(.25)	.20(.19)	.26(.21)	2.6(0.6)	11.7(7.1)
		SCAD-OS	.733	.96(.45)	.23(.36)	.40(.41)	2.0(0.7)	3.1(2.9)
		SCAD	.827	1.08(.53)	.23(.42)	.46(.53)	1.7(0.6)	1.1(1.4)
		TLP-OS	.649	.99(.42)	.31(.36)	.49(.46)	2.2(0.7)	3.8(5.2)
		TLP	.664	.99(.43)	.30(.37)	.48(.47)	2.2(0.7)	3.6(5.2)
100	500	Lasso	1.166	.48(.24)	.18(.19)	.19(.19)	2.4(0.7)	13.6(9.1)
		SCAD-OS	.867	.84(.48)	.23(.35)	.29(.37)	1.8(0.7)	3.9(3.6)
		SCAD	.847	1.00(.57)	.25(.46)	.34(.50)	1.6(0.6)	1.3(1.5)
		TLP-OS	.791	.93(.45)	.30(.39)	.38(.45)	2.0(0.7)	4.4(6.5)
		TLP	.811	.94(.46)	.29(.40)	.38(.45)	2.0(0.7)	4.1(6.4)
200	12	Lasso	.203	.87(.20)	.39(.20)	.57(.20)	3.0(0.2)	4.3(2.4)
		SCAD-OS	.173	1.06(.25)	.44(.28)	.72(.26)	2.8(0.4)	1.6(2.2)
		SCAD	.202	1.08(.25)	.45(.30)	.77(.25)	2.8(0.5)	1.2(2.1)
		TLP-OS	.155	1.00(.24)	.40(.24)	.67(.24)	2.9(0.3)	1.8(2.1)
		TLP	.157	1.00(.24)	.40(.24)	.67(.24)	2.9(0.3)	1.8(2.1)
200	200	Lasso	.540	.68(.18)	.27(.17)	.38(.17)	2.9(0.3)	14.1(8.9)
		SCAD-OS	.271	1.07(.26)	.34(.32)	.64(.32)	2.6(0.6)	3.2(3.3)
		SCAD	.262	1.12(.29)	.30(.37)	.68(.36)	2.3(0.6)	0.8(1.4)
		TLP-OS	.204	1.04(.25)	.40(.29)	.68(.27)	2.7(0.5)	3.3(5.5)
		TLP	.204	1.04(.26)	.39(.30)	.68(.28)	2.7(0.5)	3.2(5.8)

n	p	Method	ME	$\beta_1 = 1$	$\beta_2 = .5$	$\beta_3 = .75$	#TP	#FP
200	500	Lasso	.651	.64(.17)	.24(.16)	.33(.15)	2.9(0.3)	18.0(10.5)
		SCAD-OS	.289	1.07(.27)	.31(.32)	.57(.31)	2.5(0.5)	4.1(4.0)
		SCAD	.262	1.13(.28)	.29(.37)	.66(.36)	2.3(0.6)	1.4(1.7)
		TLP-OS	.231	1.04(.27)	.39(.30)	.65(.29)	2.7(0.5)	4.1(6.8)
		TLP	.231	1.04(.27)	.38(.30)	.65(.30)	2.7(0.5)	3.8(6.8)

Table 3

Averaged (with SD in parentheses) entropy loss ($loss_1$), quadratic loss ($loss_2$), true positive (TP) and false positive (FP) numbers of nonzero parameters based on 100 simulations, for estimating a precision matrix in Gaussian graphical models in Section 4.

Set-up	Method	$loss_1$	$loss_2$	#TP	#FP
$p = 30$, $a = 0.9$	Lasso	1.55(.15)	2.96(.42)	88.0(.0)	314.0(41.6)
	ALasso	1.02(.15)	1.99(.37)	88.0(.0)	95.5(30.2)
	SCAD-OS	0.93(.16)	1.99(.44)	88.0(.0)	126.0(39.4)
	SCAD	0.74(.16)	1.60(.42)	87.9(.5)	85.5(18.0)
	TLP-OS	0.66(.18)	1.47(.47)	87.9(.5)	28.1(21.4)
	TLP	0.63(.18)	1.39(.48)	87.8(.7)	22.4(17.0)
$p = 30$, $a = 0.6$	Lasso	1.69(.16)	3.28(.46)	88.0(.0)	342.5(35.5)
	ALasso	1.01(.15)	1.97(.37)	88.0(.0)	103.9(17.4)
	SCAD-OS	0.75(.14)	1.61(.36)	88.0(.0)	83.8(29.0)
	SCAD	0.56(.12)	1.20(.30)	88.0(.2)	26.1(15.0)
	TLP-OS	0.57(.14)	1.26(.37)	88.0(.0)	14.7(13.0)
	TLP	0.54(.14)	1.18(.36)	88.0(.0)	7.3(10.6)
$p = 200$, $a = 0.9$	Lasso	20.16(.50)	34.50(1.85)	597.9(.4)	4847.8(614.7)
	ALasso	10.62(.53)	19.64(1.20)	597.3(1.2)	936.8(37.9)
	SCAD-OS	11.46(.60)	24.03(1.67)	597.7(.8)	2453.6(251.2)
	TLP-OS	6.16(.77)	13.99(2.10)	593.6(3.0)	284.8(158.0)
$p = 200$, $a = 0.6$	Lasso	24.86(.54)	46.18(3.72)	598.0(.0)	6161.7(863.0)
	ALasso	11.06(.48)	21.53(1.18)	598.0(.0)	1526.1(118.6)
	SCAD-OS	9.43(.49)	20.87_1.77)	598.0(.0)	2754.8(523.6)
	TLP-OS	4.45(.48)	9.89(1.29)	597.7(.7)	185.5(76.3)

Table 4

Analysis results with various numbers (p) of predictors for the breast cancer data. The numbers of total classification errors (#Err), including false positives (#FP), and mean numbers of nonzero estimates (#Nonzero) from 10-fold CV, and the total numbers of nonzero estimates and cancer genes in the final models are shown.

		10-fold CV			Final model	
p	**Method**	**#Err**	**#FP**	**#Nonzero**	**#Nonzero**	**#Cancer genes**
200	Lasso	107	17	40.1	62	13
	ALasso	118	27	18.8	39	9
	SCAD-OS	107	4	9.5	15	2
	SCAD	107	1	4.7	2	0
	TLP-OS	102	8	33.5	62	13
	TLP	102	8	33.2	62	13
400	Lasso	107	19	46.9	95	26
	ALasso	112	19	14.4	32	10
	SCAD-OS	108	8	11.1	15	2
	SCAD	106	0	4.1	2	0
	TLP-OS	106	15	40.1	95	26
	TLP	106	14	38.2	95	26

J Am Stat Assoc. Author manuscript; available in PMC 2013 January 01.