Nonconcave Penalized Likelihood with A Diverging Number of Parameters

Jianqing Fan and Heng Peng
Presenter: Jiale Xu

March 12, 2010

Outlines

- Introduction
- Penalty function
- Properties of penalized likelihood estimation
- Proof of theorems
- Numerical examples

Background

- Traditional variable selection procedures (AIC, BIC and etc.) use a fixed penalty on the size of a model.
- Some new variable selection procedures suggest the use of a data adaptive penalty to replace fixed penalties.
- All the above procedures follow stepwise and subset selection procedures are computationally intensive, hard to derive sampling properties, and unstable.
- Most convex penalties produce shrinkage estimators of parameters that make trade-offs between bias and variance such as those in smoothing splines. However, they can create unnecessary biases when the true parameters are large and parsimonious models cannot be produced.

Background

- Fan and Li (2001) proposed a unified approach via nonconcave penalized least squares to automatically and simultaneously select variables and estimate the coefficients of variables.
- This method not only retains the good features of both subset selection and ridge regression, but also produces sparse solutions, ensures continuity of the selected models and has unbiased estimates for large coefficients.
- This is achieved by choosing suitable penalized nonconcave functions such as the smoothly clipped absolute deviation (SCAD) by Fan (1997). Other penalized least squares such as LASSO can also be studied under this unified work.
- The nonconcave penalized least-squares approach also corresponds to a Bayesian model selection with an improper prior and can be extended to likelihood-based models in various contexts.

Nonconcave penalized likelihood

Let $\log f(V, \beta)$ be the likelihood for a random vector V. This includes the likelihood of the form $\mathcal{L}(X^T\beta, Y)$ of the generalized linear model. Let $p_\lambda(|\beta_j|)$ be a nonconcave penalized function that is indexed by a regularization parameter λ. The penalized likelihood estimator then maximizes

$$(1.1) \quad \sum_{i=1}^{n} \log f(V_i, \beta) - \sum_{j=1}^{p} p_\lambda(|\beta_j|)$$

The parameter λ can be chosen by cross-validation. Various algorithms have been proposed to optimize such a high-dimensional nonconcave likelihood function, such as the modified Newton-Raphson algorithm by Fan and Li (2001).

Oracle estimator

If there were an oracle assisting us in selecting variables, then we would select variables only with nonzero coefficients and apply the MLE to this submodel and estimate the remaining coefficients as 0. This ideal estimator is called an oracle estimator.

A review on Fan and Li (2001)

- For the finite parameter case, Fan and Li (2001) established an "oracle property".
- It demonstrated that penalized likelihood estimators are asymptotically as efficient as this ideal oracle estimator for certain penalty functions, such as SCAD and the hard thresholding penalty.
- It also proposed a sandwich formula for estimating the standard error of the estimated nonzero coefficients and empirically verifying the consistency of the formula.
- It laid down important groundwork on variable selection problems, but their theoretical results are limited to the finite-parameter setting. While their results are encouraging, the fundamental problems with a growing number of parameters have not been addressed.

Objective in this paper

The objectives in this paper are to investigate the following asymptotic properties of a nonconcave penalized likelihood estimator.

- (Oracle property.) Under certain conditions of the likelihood function and penalty functions, if p_n does not grow too fast, then by the proper choice of λ_n there exists a penalized likelihood estimator such that $\hat{\beta}_{n2} = 0$ and $\hat{\beta}_{n1}$ behaves the same as the case in which $\beta_{n2} = 0$ is known in advance.
- (Asymptotic normality) As the length of $\hat{\beta}_{n1}$ depends on n, we will show that its arbitrary linear combination $A_n\hat{\beta}_{n1}$ is asymptotically normal, where A_n is a $q \times s_n$ matrix for any finite q.
- (Consistency of the sandwich formula.) Let $\hat{\Sigma}_n$ be an estimated covariance matrix for $\hat{\beta}_{n1}$, then it is consistent in the sense that $A_n^T\hat{\Sigma}_n A_n$ converges to the asymptotic covariance matrix of $A_n\hat{\beta}_{n1}$.
- (Likelihood ratio theory.) If one tests the linear hypothesis $H_0 : A_n\beta_{n1} = 0$ and uses the twice-penalized likelihood ratio statistic, then this statistic asymptotically follows a χ^2 distribution.

Penalty function

3 principles for a good penalty function: unbiasedness, sparsity and continuity.

Consider a simple form of (1.1): $\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|)$.

- L_2-penalty $p_\lambda(|\theta|) = \lambda|\theta|^2$ (ridge regression) and L_q-penalty ($q > 1$) reduce variability via shrinking the solutions, but do not have the properties of sparsity.
- L_1-penalty $p_\lambda(|\theta|) = \lambda|\theta|$ (soft thresholding rule) and L_q-penalty ($q < 1$) functions result in sparse solutions, but cannot keep the estimators unbiased for large parameters.
- Hard thresholding penalty function $p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda)$ results in the hard thresholding rule $\hat{\theta} = zI(|z| > \lambda)$, but the estimator is not continuous in the data z.
- SCAD penalty satisfies all the three properties and is defined by

$$p'_\lambda = \lambda\{I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{a-1} I(\theta > \lambda)\}$$

Regularity condition on penalty

Let $a_n = max_{1\leq j\leq p_n}\{p'_\lambda(|\beta_{n0j}|)\}, \beta_{n0j} \neq 0$ and $b_n = max_{1\leq j\leq p_n}\{p''_\lambda(|\beta_{n0j}|)\}, \beta_{n0j} \neq 0$. Then we need to place the following conditions on the penalty functions:

(A) $\liminf_{n\to+\infty} \liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0$;
(B) $a_n = O(n^{-1/2})$;
(B′) $a_n = o(1/\sqrt{np_n})$;
(C) $b_n \to 0$ as $n \to +\infty$;
(C′) $b_n = o_p(1/\sqrt{p_n})$;
(D) there are constants C and D such that, when $\theta_1, \theta_2 > C\lambda_n$, $|p''_{\lambda n}(\theta_1) - p''_{\lambda n}(\theta_2)| \leq D|\theta_1 - \theta_2|$.

Regularity condition on penalty

Remark

- (A) makes the penalty function singular at the origin so that the PLE possess the sparsity property.
- (B) and (B′) ensure the unbiasedness property for large parameters and the existence of the root-n-consistent penalized likelihood estimator.
- (C) and (C′) guarantee that the penalty function does not have much more influence than the likelihood function on the penalized likelihood estimators.
- (D) is a smoothness condition that is imposed on the nonconcave penalty functions.
- Under (H) all the above are satisfied by SCAD and hard thresholding penalty, as $a_n = 0$ and $b_n = 0$ when n is large enough.

Regularity condition on likelihood functions

Due to the diverging number of parameters, some conditions have to be strengthened to keep uniform properties for the likelihood functions and sample series, compared to the conditions in the finite parameter case.

(E) For every n the observations $\{V_{ni}, i = 1, 2, \ldots, n\}$ are iid with the density $f_n(V_{n1}, \beta_n)$, which has a common support, and the model is identifiable. Furthermore, the following holds:

$$E_{\beta_n}\{\frac{\partial \log f_n(V_{n1}, \beta_n)}{\partial \beta_{nj}}\} = 0 \qquad for \quad j = 1, 2, \cdots, p_n \tag{1}$$

and

$$E_{\beta_n}\{\frac{\partial \log f_n(V_{n1}, \beta_n)}{\partial \beta_{nj}} \frac{\partial \log f_n(V_{n1}, \beta_n)}{\partial \beta_{nk}}\} = -E_{\beta_n}\{\frac{\partial^2 \log f_n(V_{n1}, \beta_n)}{\partial \beta_{nj} \partial \beta_{nk}}\} \tag{2}$$

(F) The Fisher information matrix

$$I_n(\beta_n) = E[\{\frac{\partial \log f_n(V_{n1}, \beta_n)}{\partial \beta_n}\}\{\frac{\partial \log f_n(V_{n1}, \beta_n)}{\partial \beta_n}\}^T] \tag{3}$$

Regularity condition on likelihood functions(Cont.)

satisfies conditions $0 < C_1 < \lambda_{min}\{I_n(\beta_n)\} \leq \lambda_{max}\{I_n(\beta_n)\} < C_2 < \infty$ for all n, and, for $j, k = 1, 2, \cdots, p_n$,

$$E_{\beta_n}\{\frac{\partial \log f_n(V_{n1}, \beta_n)}{\partial \beta_{nj}} \frac{\partial \log f_n(V_{n1}, \beta_n)}{\partial \beta_{nk}}\}^2 < C_3 < \infty \tag{4}$$

and

$$E_{\beta_n}\{\frac{\partial^2 \log f_n(V_{n1}, \beta_n)}{\partial \beta_{nj} \partial \beta_{nk}}\}^2 < C_4 < \infty \tag{5}$$

(G) There is a large enough open subset ω_n of $\Omega_n \in R^{p_n}$ which contains the true parameter point β_n, such that for almost all V_{ni} the density admits all third derivatives for all $\beta_n \in \omega_n$. Furthermore, there are functions M_{njkl} such that $|\frac{\log f_n(V_{n1}, \beta_n)}{\partial \beta_{nj} \beta_{nk} \beta_{nl}}| \leq M_{njkl}$

(H) Let the values of $\beta_{n01}, \beta_{n01}, \cdots, \beta_{n0s_n}$ be the nonzero and $\beta_{n0(s_n+1)}, \beta_{n02}, \cdots, \beta_{n0p_n}$ be zero. Then $\min_{1 \leq j \leq s_n} |\beta_{n0j}|/\lambda_n \to \infty$ as $n \to \infty$.

Regularity condition on likelihood functions

Remark

- Under (F) and (G), the second and fourth moments of the likelihood function are imposed.
- (H) is necessary for obtaining the oracle property. It shows the rate at which the penalized likelihood can distinguish nonvanishing parameters from 0. Its zero component can be relaxed as $\max_{s_{n+1} \leq j \leq p_n} |\beta_{n0j}| / \lambda_n \to 0$ as $n \to \infty$.

Oracle properties

Recall that $V_{ni}, i = 1, \cdots, n$,are iid r.v's with density $f_n(V_n, \beta_{n0})$. Let

$$L_n(\beta_n) = \sum_{i=1}^{n} \log f_n(V_{ni}, \beta_n)$$

be the log-likelihood function and let

$$Q_n(\beta_n) = L_n(\beta_n) - n \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_{nj}|)$$

be the penalized likelihood function.

Oracle properties

Theorem 1 Suppose that the density $f_n(V_n, \beta_{n0})$ satisfies conditions (E)-(G), and the penalty function $p_{\lambda_n}(\cdot)$ satisfies conditions (B)-(D). If $p_n^4/n \to 0$ as $n \to \infty$, then there is a local maximizer $\hat{\beta}_n$ of $Q(\beta_n)$ such that $\|\hat{\beta}_n - \beta_{n0}\| = O_p\{\sqrt{p_n}(n^{-1/2} + a_n)\}$.

Remark If a_n satisfies condition (B), that is, $a_n = O(n^{-1/2})$, then there is a root-(n/p_n)-consistent estimator. For a SCAD or hard thresholding penalty, and condition (H) is satisfied by the model, we have $a_n = 0$ when n is large enough. The root-(n/p_n)-consistent penalized likelihood estimator exists with probability tending to 1, and no requirements are imposed on the convergence rate of λ_n.

Oracle properties

Proof Let $\alpha_n = \sqrt{p_n}(n^{-1/2} + a_n)$ and set $\|\mathbf{u}\| = C$, where C is a large constant. Our aim is to show that for any given ϵ there is a large C such that for large n we have

$$P\{\sup_{\|\mathbf{u}\|=C} Q_n(\beta_{n0} + \alpha_n \mathbf{u}) < Q_n(\beta_{n0})\} \geq 1 - \epsilon.$$

This implies that with probability tending to 1 there is a local maximum $\hat{\beta}_n$ in the ball $\{\beta_{n0} + \alpha_n \mathbf{u} : \|\mathbf{u}\| \leq C\}$ such that $\|\hat{\beta}_n - \beta_{n0}\| = O_p(\alpha_n)$.
Using $p_{\lambda_n}(0) = 0$, we have
$D_n(\mathbf{u}) = Q_n(\beta_{n0} + \alpha_n \mathbf{u}) - Q_n(\beta_{n0})$

$$\leq L_n(\beta_{n0} + \alpha_n \mathbf{u}) - L_n(\beta_{n0}) - n \sum_{j=1}^{s_n} \{p_{\lambda_n}(|\beta_{n0j} + \alpha_n u_j|) - p_{\lambda_n}(|\beta_{n0j})\}$$

$$= (I) + (II)$$

Oracle properties

Proof(cont.) Then by Taylor's expansion we obtain

$$(I) = \alpha_n \bigtriangledown^T L_n(\beta_{n0})\mathbf{u} + \frac{1}{2}\mathbf{u}^T \bigtriangledown^2 L_n(\beta_{n0})\mathbf{u}\alpha_n^2 + \frac{1}{6} \bigtriangledown^T \{\mathbf{u}^T \bigtriangledown^2 L_n(\beta_n^*)\mathbf{u}\}\mathbf{u}\alpha_n^3$$

$= I_1 + I_2 + I_3$, where the vector β_n^* lies between β_{n0} and $\beta_{n0} + \alpha_n \mathbf{u}$, and

$$(II) = -n\sum_{j=1}^{s_n}[n\alpha_n p'_{\lambda_n}(|\beta_{n0j}|)sgn(\beta_{n0j})u_j + n\alpha_n^2 p''_{\lambda_n}(\beta_{n0j})u_j^2\{1+o(1)\}] = I_4 + I_5$$

We could show that all terms I_1, I_3, I_4 and I_5 are dominated by $I_2 = -\frac{n\alpha_n^2}{2}\mathbf{u}^T I_n(\beta_{n0})\mathbf{u} + o_p(1)\cdot n\alpha_n^2\|\mathbf{u}\|^2$, which is negative. This completes the proof of Theorem 1.

Oracle properties

Lemma 1 Assume that conditions (A) and (E)-(H) are satisfied. If $\lambda_n \to 0$, $\sqrt{n/p_n}\lambda_n \to \infty$ and $p_n^5/n \to 0$ as $n \to \infty$, then with probability tending to 1, for any given β_{n1} satisfying $\|\beta_{n1} - \beta_{n01}\| = O_p(\sqrt{p_n/n})$ and any constant C,

$$Q\{(\beta_{n1}^T, 0)^T\} = \max_{\|\beta_{n2}\| \le C(p_n/n)^{1/2}} Q\{(\beta_{n1}^T, \beta_{n2}^T)^T\}$$

Oracle properties

Denote

$$\Sigma_{\lambda_n} = diag\{p''_{\lambda_n}(\beta_{n01}), \cdots, p''_{\lambda_n}(\beta_{n0s_n})\}$$

and

$$\mathbf{b}_n = \{p'_{\lambda_n}(|\beta_{n01}|)sgn(\beta_{n01}), \cdots, p'_{\lambda_n}(|\beta_{n0s_n}|)sgn(\beta_{n0s_n})\}^T.$$

Theorem 2 Under conditions (A)-(H) are satisfied, if $\lambda_n \to 0$, $\sqrt{n/p_n}\lambda_n \to \infty$ and $p_n^5/n \to 0$ as $n \to \infty$, then with probability tending to 1, the root-(n/p_n)-consistent local maximizer $\hat{\beta}_n = \binom{\hat{\beta}_{n1}}{\hat{\beta}_{n2}}$ in Theorem 1 must satisfy:(i) Sparsity: $\hat{\beta}_{n2} = 0$. (ii) Asymptotic normality: $\sqrt{n}A_n I_n^{-1/2}(\beta_{n01})\{I_n(\beta_{n01}) + \Sigma_{\lambda_n}\} \times [\hat{\beta}_{n1} - \beta_{n01} + \{I_n(\beta_{n01}) + \Sigma_{\lambda_n}\}^{-1}\mathbf{b}_n] \xrightarrow{\mathcal{D}} \mathcal{N}(0, G)$, where A_n is a $q \times s_n$ matrix such that $A_n A_n^T \to G$, and G is a $q \times q$ nonegative symmetric matrix.

Oracle properties

Proof As shown in Theorem 1, there is a root-n/p_n-consistent local maximizer $\hat{\beta}_n$ of $Q_n(\beta_n)$. By Lemma 1, part (i) holds that $\hat{\beta}_n$ has the form $(\hat{\beta}_{n1}, 0)^T$. We need only prove (ii), the asymptotic normality of the penalized nonconcave likelihood estimator $\hat{\beta}_{n1}$.
If we can show that

$$\{I_n(\beta_{n01}) + \Sigma_{\lambda_n}\}(\hat{\beta}_{n1} - \beta_{n01}) + \mathbf{b}_n = \frac{1}{n} \bigtriangledown L_n(\beta_{n01}) + o_p(n^{-1/2}).$$

then

$$\sqrt{n} A_n I_n^{-1/2}(\beta_{n01})\{I_n(\beta_{n01}) + \Sigma_{\lambda_n}\}[\hat{\beta}_{n1} - \beta_{n01} + \{I_n(\beta_{n01}) + \Sigma_{\lambda_n}\}^{-1}\mathbf{b}_n]$$

$$= \frac{1}{\sqrt{n}} A_n I_n^{-1/2}(\beta_{n01}) \bigtriangledown L_n(\beta_{n01}) + o_p\{A_n I_n^{-1/2}(\beta_{n01})\}.$$

By conditions of Theorem 2, we have the last term of $o_p(1)$.

Oracle properties

Proof (Cont.) Let $Y_{ni} = \frac{1}{\sqrt{n}} A_n I_n^{-1/2}(\beta_{n01}) \bigtriangledown L_{ni}(\beta_{n01}), \quad i = 1, 2, \cdots, n.$ It follows that, for any ϵ,

$$\sum_{i=1}^{n} E\|Y_{ni}\|^2 \mathbf{1}\{\|Y_{ni}\| > \epsilon\} \leq n\{E\|Y_{n1}\|^4\}^{1/2}\{P(\|Y_{n1}\| > \epsilon)\}^{1/2}.$$

Then we can show $P(\|Y_{n1}\| > \epsilon) = O(n^{-1})$ and $E\|Y_{n1}\|^4 = O(\frac{p_n^2}{n^2})$. Thus, we have $\sum_{i=1}^{n} E\|Y_{ni}\|^2 \mathbf{1}\{\|Y_{ni}\| > \epsilon\} = O(n\frac{p_n}{n}\frac{1}{\sqrt{n}}) = o(1)$. On the other hand we have $\sum_{i=1}^{n} cov(Y_{ni}) \to G$.

Thus, Y_{ni} satisfies the conditions of Lindeberg-Feller CLT. This also means that $\frac{1}{\sqrt{n}} A_n I_n^{-1/2}(\beta_{n01}) \bigtriangledown L_n(\beta_{n01})$ has an asymptotic multivariate normal distribution.

Oracle properties

Remark By Theorem 2 the sparsity and the asymptotic normality are still valid when the number of parameters diverges. The oracle property holds for the SCAD and the hard thresholding penalty function. When n is large enough, $\Sigma_{\lambda_n} = 0$ and $\mathbf{b}_n = 0$ for the SCAD and the hard thresholding penalty. Hence, Theorem 2(ii) becomes

$$\sqrt{n}A_n I_n^{1/2}(\beta_{n01})(\hat{\beta}_{01} - \beta_{n01}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, G)$$

Estimation of covariance matrix

As in Fan and Li (2001), by the sandwich formula let

$$\hat{\Sigma}_n = n\{\bigtriangledown^2 L_n(\hat{\beta}_{n1}) - n\Sigma_{\lambda_n}(\hat{\beta}_{n1})\}^{-1}$$

$$\times \widehat{cov}\{\bigtriangledown L_n(\hat{\beta}_{n1})\}\{\bigtriangledown^2 L_n(\hat{\beta}_{n1}) - n\Sigma_{\lambda_n}(\hat{\beta}_{n1})\}^{-1}$$

be the estimated covariance matrix of $\hat{\beta}_{n1}$, where

$$\widehat{cov}\{\bigtriangledown L_n(\hat{\beta}_{n1})\} = \{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial L_{ni}(\hat{\beta}_{n1})}{\partial \beta_j}\frac{\partial L_{ni}(\hat{\beta}_{n1})}{\partial \beta_k}\}$$

$$-\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial L_{ni}(\hat{\beta}_{n1})}{\partial \beta_j}\}\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial L_{ni}(\hat{\beta}_{n1})}{\partial \beta_k}\}$$

Denote by $\Sigma_n = \{I_n(\beta_{n01} + \Sigma_{\lambda_n}(\beta_{n01})\}^{-1} I_n(\beta_{n01})\{I_n(\beta_{n01} + \Sigma_{\lambda_n}(\beta_{n01})\}^{-1}$ the asymptotic variance of $\hat{\beta}_{n1}$ in Theorem 2(ii).

Consistency of the sandwich formula

Theorem 3 If conditions (A)-(H) are satisfied and $p_n^5/n \to 0$ as $n \to \infty$, then we have

$$A_n \hat{\Sigma}_n A_n^T - A_n \Sigma_n A_n^T \xrightarrow{p} 0 \qquad n \to \infty \tag{6}$$

for any $q \times s_n$ matrix A_n such that $A_n A_n^T = G$, where q is any fixed integer.

Theorem 3 not only proves a conjecture of Fan and Li (2001) about the consistency of the sandwich formula for the standard error matrix, but also extends the result to the situation with a growing number of parameters. It offers a way to construct a confidence interval for the esimtates of parameters.

Likelihood ratio test

Consider the problem of testing linear hypotheses:

$$H_0 : A_n\beta_{n01} = 0 \quad vs. \quad H_1 : A_n\beta_{n01} \neq 0$$

where A_n is a $q \times s_n$ matrix and $A_n A_n^T = I_q$ for a fixed q.
A natural ratio test for the problem is

$$T_n = 2\{\sup_{\Omega_n} Q(\beta_n|\mathbf{V}) - \sup_{\Omega_n, A_n\beta_{n1}=0} Q(\beta_n|\mathbf{V})\}.$$

Theorem 4 When conditions (A)-(H), (B′) and (C′)are satisfied, under H_0 we have

$$T_n \xrightarrow{\mathcal{D}} \chi^2_q \tag{7}$$

provided that $p_n^5/n \to 0$ as $n \to \infty$.

Simulation study

Consider the autoregressive model:

$$X_i = \beta_1 X_{i-1} + \beta_2 X_{i-2} + \cdots + \beta_p X_{i-p_n} + \epsilon, \qquad i = 1, 2, \cdots, n,$$

where $\beta = (11/4, -23/6, 37/12, -13/9, 1/3, 0, \cdots, 0)^T$ and ϵ is white noise with variance σ^2. In the simulation experiments 400 samples of sizes 100, 200, 400 and 800 with $p_n = [4n^{1/4}] - 5$ are from this model. The SCAD penalty is employed. The results are summarized in the following tables.

Simulation Results

Simulation results for the time series model

		MRME (%)			Average number of zero coeficients	
n	p_n	Oracle/LS	PLS/LS	Oracle/PLS	Correct	Incorrect
100	7	75.33	89.21	80.17	1.34 [67%]	0.49
200	10	50.61	69.64	73.27	3.91 [78%]	0.39
400	12	40.03	59.57	73.06	5.78 [83%]	0.22
800	16	31.75	49.05	70.08	9.49 [86%]	0.10

Median of estimators for coefficients of time series model

n	p_n	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
100	7	2.678	−3.616	2.739	−1.096	0
200	10	2.711	−3.696	2.856	−1.240	0.242
400	12	2.729	−3.769	2.959	−1.333	0.293
800	16	2.737	−3.792	3.023	−1.383	0.306
True	—	2.750	−3.833	3.083	−1.444	0.333

Simulation Results

Standard deviations (multiplied by 1000) of estimators for time series model

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_3$		$\hat{\beta}_4$		$\hat{\beta}_5$	
n	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})
100	120	91 (5.1)	337	230 (29.8)	525	285 (66.6)	451	177 (87.2)	249	79 (66.7)
200	76	66 (2.8)	221	174 (15.2)	340	231 (58)	348	170 (87.2)	243	64 (49.5)
400	50	47 (1.2)	149	126 (4.5)	222	169 (8.8)	204	125 (9.0)	129	47 (3.90)
800	35	34 (0.7)	99	90 (3.1)	145	121 (8.5)	132	90 (14.1)	63	34 (12.5)

Simulation Results

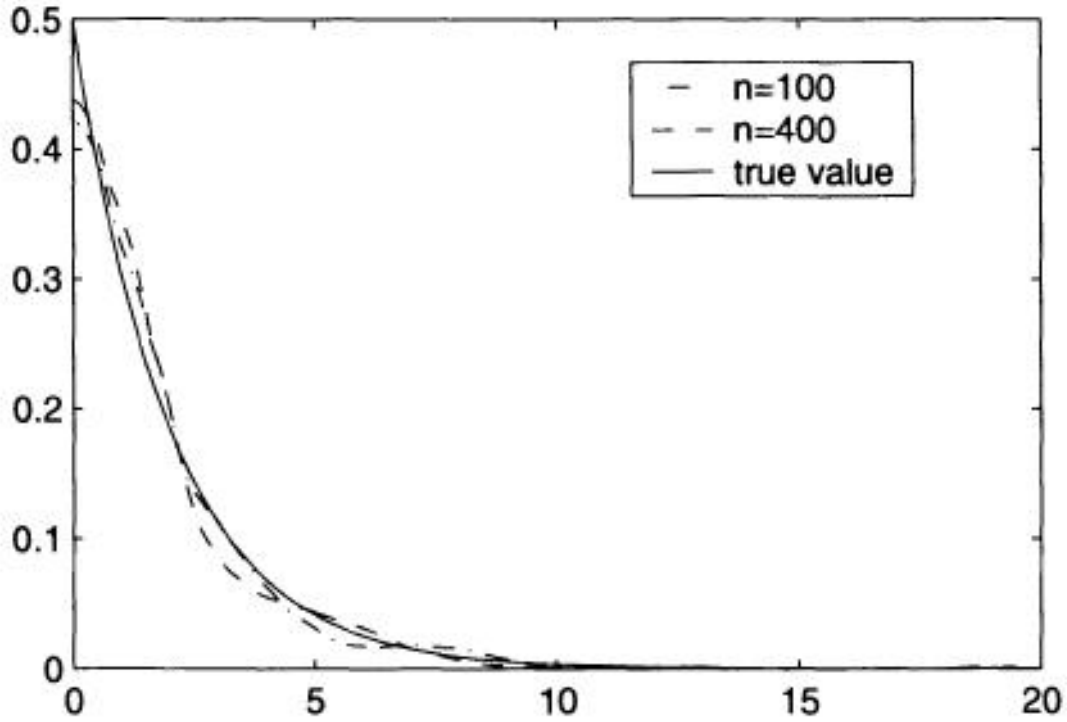

FIG. 4. *Estimated densities of the likelihood ratio statistics for* $n = 100$ *(dot-dash) and* $n = 400$ *(long-dash) along with the density of the* χ^2_2 *distribution (solid).*

Summary

- In most model selection problems the number of parameters should be large and grow with the sample size.
- Some asymptotic properties of the nonconcave penalized likelihood are established for situations in which the number of parameters tends to ∞ as the sample size increases.
- Under regularity conditions an oracle property and asymptotic normality of the PLE are established.
- Consistency of the sandwich formula of the covariance matrix is demonstrated. And nonconcave penalized likelihood ratio statistics are discussed.

Thank You!