arXiv:0704.1466v1 [math.ST] 11 Apr 2007

Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator

Hannes Leeb
Department of Statistics, Yale University
and
Benedikt M. Pötscher
Department of Statistics, University of Vienna

First version: November 2004
This version: March 2007

Abstract

We point out some pitfalls related to the concept of an oracle property as used in Fan and Li (2001, 2002, 2004) which are reminiscent of the well-known pitfalls related to Hodges' estimator. The oracle property is often a consequence of sparsity of an estimator. We show that any estimator satisfying a sparsity property has maximal risk that converges to the supremum of the loss function; in particular, the maximal risk diverges to infinity whenever the loss function is unbounded. For ease of presentation the result is set in the framework of a linear regression model, but generalizes far beyond that setting. In a Monte Carlo study we also assess the extent of the problem in finite samples for the smoothly clipped absolute deviation (SCAD) estimator introduced in Fan and Li (2001). We find that this estimator can perform rather poorly in finite samples and that its worst-case performance relative to maximum likelihood deteriorates with increasing sample size when the estimator is tuned to sparsity.

AMS 2000 Subject Classifications: Primary 62J07, 62C99; secondary 62E20, 62F10, 62F12

Key words and phrases: oracle property, sparsity, penalized maximum likelihood, penalized least squares, Hodges' estimator, SCAD, Lasso, Bridge estimator, hard-thresholding, maximal risk, maximal absolute bias, non-uniform limits

1 Introduction

Recent years have seen an increased interest in penalized least squares and penalized maximum likelihood estimation. Examples are the class of Bridge estimators introduced by Frank and Friedman (1993), which includes Lasso-type estimators as a special case (Knight and Fu (2000)), or the smoothly clipped absolute deviation (SCAD) estimator introduced in Fan and Li (2001) and further discussed in Fan and Li (2002, 2004), Fan and Peng (2004), and Cai, Fan, Li, and Zhou (2005). As shown in Fan and Li (2001), the SCAD estimator, with appropriate choice of the regularization (tuning) parameter, possesses a sparsity property, i.e., it estimates zero components of the true parameter vector exactly as zero with probability approaching one as sample size increases while still being consistent for the non-zero components. An immediate consequence of this sparsity property of the SCAD estimator is that the asymptotic distribution of this estimator remains the same whether or not the correct zero restrictions are imposed in the course of the SCAD estimation procedure. [This simple phenomenon is true more generally as pointed out, e.g., in Pötscher (1991, Lemma 1).] In other words, with appropriate choice of the regularization parameter, the asymptotic distribution of the SCAD estimator based on the overall model and that of the SCAD estimator derived from the most parsimonious correct model coincide. Fan and Li (2001) have dubbed this property the "oracle property" and have advertised this property of their estimator.[1] For appropriate choices of the regularization parameter, the sparsity and the oracle property are also possessed by several – but not all – members of the class of Bridge estimators (Knight and Fu (2000), p. 1361, Zou (2006)). Similarly, suitably tuned thresholding procedures give rise to sparse estimators.[2] Finally, we note that traditional post-model-selection estimators (e.g., maximum likelihood estimators following model selection) based on a consistent model selection procedure (for example, BIC or test procedures with suitably chosen critical values) are another class of estimators that exhibit the sparsity and oracle property; see Pötscher (1991) and Leeb and Pötscher (2005) for further discussion. In a recent paper, Bunea (2004) uses such procedures in a semiparametric framework and emphasizes the oracle property of the resulting estimator; see also Bunea and McKeague (2005).

At first sight, the oracle property appears to be a desirable property of an estimator as it seems to guarantee that, without knowing which components of the true parameter are zero, we can do (asymptotically) as well as if we knew the correct zero restrictions; that is, we can "adapt" to the unknown zero restrictions without paying a price. This is too good to be true, and

[1] The oracle property in the sense of Fan and Li should not be confused with the notion of an oracle inequality as frequently used elsewhere in the literature.

[2] These estimators do not satisfy the oracle property in case of non-orthogonal design.

it is reminiscent of the "superefficiency" property of the Hodges' estimator; and justly so, since Hodges' estimator in its simplest form is a hard-thresholding estimator exhibiting the sparsity and oracle property. [Recall that in its simplest form Hodges' estimator for the mean of an $N(\mu, 1)$-distribution is given by the arithmetic mean $\bar{y}$ of the random sample of size n if $|\bar{y}|$ exceeds the threshold $n^{-1/4}$, and is given by zero otherwise.] Now, as is well-known, e.g., from Hodges' example, the oracle property is an asymptotic feature that holds only *pointwise* in the parameter space and gives a misleading picture of the actual finite-sample performance of the estimator. In fact, the finite sample properties of an estimator enjoying the oracle property are often markedly different from what the pointwise asymptotic theory predicts; e.g., the finite sample distribution can be bimodal regardless of sample size, although the pointwise asymptotic distribution is normal. This is again well-known for Hodges' estimator. For a more general class of post-model-selection estimators possessing the sparsity and the oracle property this is discussed in detail in Leeb and Pötscher (2005), where it is, e.g., also shown that the finite sample distribution can "escape to infinity" along appropriate local alternatives although the pointwise asymptotic distribution is perfectly normal.[3] See also Knight and Fu (2000, Section 3) for related results for Bridge estimators. Furthermore, estimators possessing the oracle property are certainly not exempt from the Hajek-LeCam local asymptotic minimax theorem, further eroding support for the belief that these estimators are as good as the "oracle" itself (i.e., the infeasible "estimator" that uses the information which components of the parameter are zero).

The above discussion shows that the reasoning underlying the oracle property is misguided. Even worse, estimators possessing the sparsity property (which often entails the oracle property) necessarily have dismal finite sample performance: It is well-known for Hodges' estimator that the maximal (scaled) mean squared error grows without bound as sample size increases (e.g., Lehmann and Casella (1998), p.442), whereas the standard maximum likelihood estimator has constant finite quadratic risk. In this note we show that a similar unbounded risk result is in fact true for *any* estimator possessing the sparsity property. This means that there is a substantial price to be paid for sparsity even though the oracle property (misleadingly) seems to suggest otherwise. As discussed in more detail below, the bad risk behavior is a "local" phenomenon and furthermore occurs at points in the parameter space that are "sparse" in the sense that some of their coordinates are equal to zero. For simplicity of presentation and for reasons of comparability with the literature cited earlier, the result will be set in the framework of a linear regression model, but inspection of the proof shows that it easily extends far beyond that

[3] That pointwise asymptotics can be misleading in the context of model selection has been noted earlier in Hosoya (1984), Shibata (1986a), Pötscher (1991), and Kabaila (1995, 2002).

framework. For related results in the context of traditional post-model-selection estimators see Yang (2005) and Leeb and Pötscher (2005, Appendix C);[4] cf. also the discussion on "partially" sparse estimators towards the end of Section 2 below. The theoretical results in Section 2 are rounded out by a Monte Carlo study in Section 3 that demonstrates the extent of the problem in finite samples for the SCAD estimator of Fan and Li (2001). The reasons for concentrating on the SCAD estimator in the Monte Carlo study are (i) that the finite-sample risk behavior of traditional post-model-selection estimators is well-understood (Judge and Bock (1978), Leeb and Pötscher (2005)) and (ii) that the SCAD estimator – especially when tuned to sparsity – has been highly advertised as a superior procedure in Fan and Li (2001) and subsequent papers mentioned above.

2 Bad Risk Behavior of Sparse Estimators

Consider the linear regression model

$$y_t \quad = \quad x_t'\theta + \epsilon_t \qquad (1 \leq t \leq n) \tag{1}$$

where the $k \times 1$ nonstochastic regressors x_t satisfy $n^{-1}\sum_{t=1}^{n} x_t x_t' \to Q > 0$ as $n \to \infty$ and the prime denotes transposition. The errors ϵ_t are assumed to be independent identically distributed with mean zero and finite variance σ^2. Without loss of generality we freeze the variance at $\sigma^2 = 1$.[5] Furthermore, we assume that ϵ_t has a density f that possesses an absolutely continuous derivative df/dx satisfying

$$0 < \int_{-\infty}^{\infty} \left((df(x)/dx)/f(x)\right)^2 f(x)dx < \infty.$$

Note that the conditions on f guarantee that the information of f is finite and positive. These conditions are obviously satisfied in the special case of normally distributed errors. Let $P_{n,\theta}$ denote the distribution of the sample $(y_1, \ldots, y_n)'$ and let $E_{n,\theta}$ denote the corresponding expectation operator. For $\theta \in \mathbb{R}^k$, let $r(\theta)$ denote a $k \times 1$ vector with components $r_i(\theta)$ where $r_i(\theta) = 0$ if $\theta_i = 0$ and $r_i(\theta) = 1$ if $\theta_i \neq 0$. An estimator $\hat{\theta}$ for θ based on the sample $(y_1, \ldots, y_n)'$ is said to satisfy the sparsity-type condition if for every $\theta \in \mathbb{R}^k$

$$P_{n,\theta}\left(r(\hat{\theta}) \leq r(\theta)\right) \to 1 \tag{2}$$

[4] The unboundedness of the maximal (scaled) mean squared error of estimators following BIC-type model selection has also been noted in Hosoya (1984), Shibata (1986b), and Foster and George (1994).

[5] If the variance is not frozen at $\sigma^2 = 1$, the results below obviously continue to hold for each fixed value of σ^2, and hence hold a fortiori if the supremum in (3)–(4) below is also taken over σ^2.

holds for $n \to \infty$, where the inequality sign is to be interpreted componentwise. That is, the estimator is guaranteed to find the zero components of θ with probability approaching one as $n \to \infty$. Clearly, any sparse estimator satisfies (2). In particular, the SCAD estimator as well as certain members of the class of Bridge estimators satisfy (2) for suitable choices of the regularization parameter as mentioned earlier. Also, any post-model-selection estimator based on a consistent model selection procedure clearly satisfies (2). All these estimators are additionally also consistent for θ, and hence in fact satisfy the stronger condition $P_{n,\theta}(r(\hat{\theta}) = r(\theta)) \to 1$ for all $\theta \in \mathbb{R}^k$. [Condition (2) by itself is of course also satisfied by nonsensical estimators like $\hat{\theta} \equiv 0$, but is all that is needed to establish the subsequent result.] We now show that any estimator satisfying the sparsity-type condition (2) has quite bad finite sample risk properties. For purposes of comparison we note that the (scaled) mean squared error of the least squares estimator $\hat{\theta}_{LS}$ satisfies

$$E_{n,\theta}\left[n(\hat{\theta}_{LS} - \theta)'(\hat{\theta}_{LS} - \theta)\right] = \operatorname{trace}\left[\left(n^{-1}\sum_{t=1}^{n} x_t x_t'\right)^{-1}\right]$$

which converges to trace(Q^{-1}), and thus remains bounded as sample size increases.

Theorem 2.1 [6]*Let $\hat{\theta}$ be an arbitrary estimator for θ that satisfies the sparsity-type condition (2). Then the maximal (scaled) mean squared error of $\hat{\theta}$ diverges to infinity as $n \to \infty$, i.e.,*

$$\sup_{\theta \in \mathbb{R}^k} E_{n,\theta}\left[n(\hat{\theta} - \theta)'(\hat{\theta} - \theta)\right] \to \infty \tag{3}$$

for $n \to \infty$. More generally, let $l : \mathbb{R}^k \to \mathbb{R}$ be a nonnegative loss function. Then

$$\sup_{\theta \in \mathbb{R}^k} E_{n,\theta} l(n^{1/2}(\hat{\theta} - \theta)) \to \sup_{s \in \mathbb{R}^k} l(s) \tag{4}$$

for $n \to \infty$. In particular, if the loss function l is unbounded then the maximal risk associated with l diverges to infinity as $n \to \infty$.

The theorem says that, whatever the loss function, the maximal risk of a sparse estimator is – in large samples – as bad as it possibly can be.

[6] Theorem 2.1 and the ensuing discussion continue to apply if the regressors x_t as well as the errors ϵ_t are allowed to depend on sample size n, at least if the errors are normally distributed. The proof is analogous, except that one uses direct computation and LeCam's first lemma (cf., e.g., Lemma A.1 in Leeb and Pötscher (2006)) instead of Koul and Wang (1984) to verify contiguity. Also, the results continue to hold if the design matrix satisfies $\delta_n^{-1} \sum_{t=1}^{n} x_t x_t' \to Q > 0$ for some positive sequence δ_n other than n, provided that the scaling factor $n^{1/2}$ is replaced by $\delta_n^{1/2}$ throughout.

Upon choosing $l(s) = |s_i|$, where s_i denotes the i-th coordinate of s, relation (4) shows that also the maximal (scaled) absolute bias of each component $\hat{\theta}_i$ diverges to infinity.

Applying relation (4) to the loss function $l^*(s) = l(c's)$ shows that (4) holds mutatis mutandis also for estimators $c'\hat{\theta}$ of arbitrary linear contrasts $c'\theta$. In particular, using quadratic loss $l^*(s) = (c's)^2$, it follows that also the maximal (scaled) mean squared error of the linear contrast $c'\hat{\theta}$ goes to infinity as sample size increases, provided $c \neq 0$.

Proof of Theorem 2.1: It suffices to prove (4).[7] Now, with $\theta_n = -n^{-1/2}s$, $s \in \mathbb{R}^k$ arbitrary, we have

$$\begin{aligned}\sup_{u\in\mathbb{R}^k} l(u) &\geq \sup_{\theta\in\mathbb{R}^k} E_{n,\theta} l(n^{1/2}(\hat{\theta}-\theta)) \geq E_{n,\theta_n} l(n^{1/2}(\hat{\theta}-\theta_n)) \\ &\geq E_{n,\theta_n}[l(n^{1/2}(\hat{\theta}-\theta_n))\mathbf{1}(\hat{\theta}=0)] = l(-n^{1/2}\theta_n)P_{n,\theta_n}(r(\hat{\theta})=0) \\ &= l(s)P_{n,\theta_n}(r(\hat{\theta})=0). \qquad (5)\end{aligned}$$

By the sparsity-type condition we have that $P_{n,0}(r(\hat{\theta}) = 0) \to 1$ as $n \to \infty$. Since the model is locally asymptotically normal under our assumptions (Koul and Wang (1984), Theorem 2.1 and Remark 1; Hajek and Sidak (1967), p.213), the sequence of probability measures P_{n,θ_n} is contiguous w.r.t. the sequence $P_{n,0}$. Consequently, the far r.h.s. of (5) converges to $l(s)$. Since $s \in \mathbb{R}^k$ was arbitrary, the proof is complete. ■

Inspection of the proof shows that Theorem 2.1 remains true if the supremum of the risk in (4) is taken only over open balls of radius ρ_n centered at the origin as long as $n^{1/2}\rho_n \to \infty$. Hence, the bad risk behavior is a local phenomenon that occurs in a part of the parameter space where one perhaps would have expected the largest gain over the least squares estimator due to the sparsity property. [If the supremum of the risk in (4) is taken over the open balls of radius $n^{-1/2}\rho$ centered at the origin where $\rho > 0$ is now fixed, then the proof still shows that the limit inferior of this supremum is not less than $\sup_{\|s\|<\rho} l(s)$.] Furthermore, for quadratic loss $l(s) = s's$, a small variation of the proof shows that these "local" results continue to hold if the open balls over which the supremum is taken are not centered at the origin, but at an arbitrary θ, as long as θ possesses at least one zero component. [It is easy to see that this is more generally true for any nonnegative loss function l satisfying, e.g., $l(s) \geq l(\pi_i(s))$ for every $s \in \mathbb{R}^k$ and an index i with $\theta_i = 0$, where π_i represents the projection on the i-th coordinate axis.]

Inspection of the proof also shows that – at least in the case of quadratic loss – the element s can be chosen to point in the direction of a standard basis vector. This then shows that the bad risk behavior occurs at parameter values that themselves are "sparse" in the sense of having

[7] Note that the expectations in (3) and (4) are always well-defined.

many zero coordinates.

If the quadratic loss $n(\hat{\theta}-\theta)'(\hat{\theta}-\theta)$ in (3) is replaced by the weighted quadratic loss $(\hat{\theta}-\theta)'\sum_{t=1}^{n} x_t x_t'(\hat{\theta}-\theta)$, then the corresponding maximal risk again diverges to infinity. More generally, let l_n be a nonnegative loss function that may depend on sample size. Inspection of the proof of Theorem 2.1 shows that

$$\limsup_{n\to\infty}\sup_{u\in\mathbb{R}^k} l_n(u) \geq \limsup_{n\to\infty}\sup_{\|\theta\|<n^{-1/2}\rho} E_{n,\theta} l_n(n^{1/2}(\hat{\theta}-\theta)) \geq \sup_{\|u\|<\rho}\limsup_{n\to\infty} l_n(u), \tag{6}$$

$$\liminf_{n\to\infty}\sup_{u\in\mathbb{R}^k} l_n(u) \geq \liminf_{n\to\infty}\sup_{\|\theta\|<n^{-1/2}\rho} E_{n,\theta} l_n(n^{1/2}(\hat{\theta}-\theta)) \geq \sup_{\|u\|<\rho}\liminf_{n\to\infty} l_n(u) \tag{7}$$

hold for any $0<\rho\leq\infty$. [In case $0<\rho<\infty$, the lower bounds in (6)-(7) can even be improved to $\limsup_{n\to\infty}\sup_{\|u\|<\rho} l_n(u)$ and $\liminf_{n\to\infty}\sup_{\|u\|<\rho} l_n(u)$, respectively.[8] It then follows that in case $\rho=\infty$ the lower bounds in (6)-(7) can be improved to $\sup_{0<\tau<\infty}\limsup_{n\to\infty}\sup_{\|u\|<\tau} l_n(u)$ and $\sup_{0<\tau<\infty}\liminf_{n\to\infty}\sup_{\|u\|<\tau} l_n(u)$, respectively.]

Next we briefly discuss the case where an estimator $\hat{\theta}$ only has a "partial" sparsity property (and consequently a commensurable oracle property) in the following sense: Suppose the parameter vector θ is partitioned as $\theta=(\alpha',\beta')'$ and the estimator $\hat{\theta}=(\hat{\alpha}',\hat{\beta}')'$ only finds the true zero components in the subvector β with probability converging to one. E.g., $\hat{\theta}$ is a traditional post-model-selection estimator based on a consistent model selection procedure that is designed to only identify the zero components in β. A minor variation of the proof of Theorem 2.1 immediately shows again that the maximal (scaled) mean squared error of $\hat{\beta}$, and hence also of $\hat{\theta}$, diverges to infinity, and the same is true for linear combinations $d'\hat{\beta}$ as long as $d\neq 0$. [This immediately extends to linear combinations $c'\hat{\theta}$, as long as c charges at least one coordinate of $\hat{\beta}$ with a nonzero coefficient.][9] However, if the parameter of interest is α rather than β, Theorem 2.1 and its proof (or simple variations thereof) do not apply to the mean squared error of $\hat{\alpha}$ (or its linear contrasts). Nevertheless, the maximal (scaled) mean squared error of $\hat{\alpha}$ can again be shown to diverge to infinity, at least for traditional post-model-selection estimators $\hat{\theta}$ based on a consistent model selection procedure; see Leeb and Pötscher (2005, Appendix C).

While the above results are set in the framework of a linear regression model with nonstochastic regressors, it is obvious from the proof that they extend to much more general models such as regression models with stochastic regressors, semiparametric models, nonlinear models,

[8] Note that the local asymptotic normality condition in Koul and Wang (1984) as well as the result in Lemma A.1 in Leeb and Pötscher (2006) imply contiguity of P_{n,θ_n} and $P_{n,0}$ not only for $\theta_n=\gamma/n^{1/2}$ but more generally for $\theta_n=\gamma_n/n^{1/2}$ with γ_n a bounded sequence.

[9] In fact, this variation of the proof of Theorem 2.1 shows that the supremum of $E_{n,\theta} l(n^{1/2}(\hat{\beta}-\beta))$, where l is an arbitrary nonegative loss function, again converges to the supremum of the loss function.

time series models, etc., as long as the contiguity property used in the proof is satisfied. This is in particular the case whenever the model is locally asymptotically normal, which in turn is typically the case under standard regularity conditions for maximum likelihood estimation.

3 Numerical Results on the Finite Sample Performance of the SCAD Estimator

We replicate and extend Monte Carlo simulations of the performance of the SCAD estimator given in Example 4.1 of Fan and Li (2001); we demonstrate that this estimator, when tuned to enjoy a sparsity property and an oracle property, can perform quite unfavorably in finite samples. Even when not tuned to sparsity, we show that the SCAD estimator can perform worse than the least squares estimator in parts of the parameter space, something that is not brought out in the simulation study in Fan and Li (2001) as they conducted their simulation only at a single point in the parameter space (which happens to be favorable to their estimator).

Consider n independent observations from the linear model (1) with $k = 8$ regressors, where the errors ϵ_t are standard normal and are distributed independently of the regressors. The regressors x_t are assumed to be multivariate normal with mean zero. The variance of each component of x_t is equal to 1 and the correlation between the i-th and the j-th component of x_t, i.e., $x_{t,i}$ and $x_{t,j}$, is $\rho^{|i-j|}$ with $\rho = 0.5$. Fan and Li (2001) consider this model with $n = 40$, $n = 60$, and with the true parameter equal to $\theta_0 = (3, 1.5, 0, 0, 2, 0, 0, 0)'$; cf. also Tibshirani (1996, Section 7.2). We consider a whole range of true values for θ at various sample sizes, namely $\theta_n = \theta_0 + (\gamma/\sqrt{n}) \times \eta$ for some vector η and for a range of γ's as described below. We do this because (i) considering only one choice for the true parameter in a simulation may give a wrong impression of the actual performance of the estimators considered, and (ii) because our results in Section 2 suggest that the risk of sparse estimators can be large for parameter vectors which have some of its components close to, but different from, zero.

The SCAD estimator is defined as a solution to the problem of minimizing the penalized least squares objective function

$$\frac{1}{2}\sum_{t=1}^{n}(y_t - x_t'\theta)^2 + n\sum_{i=1}^{k} p_\lambda(|\theta_i|)$$

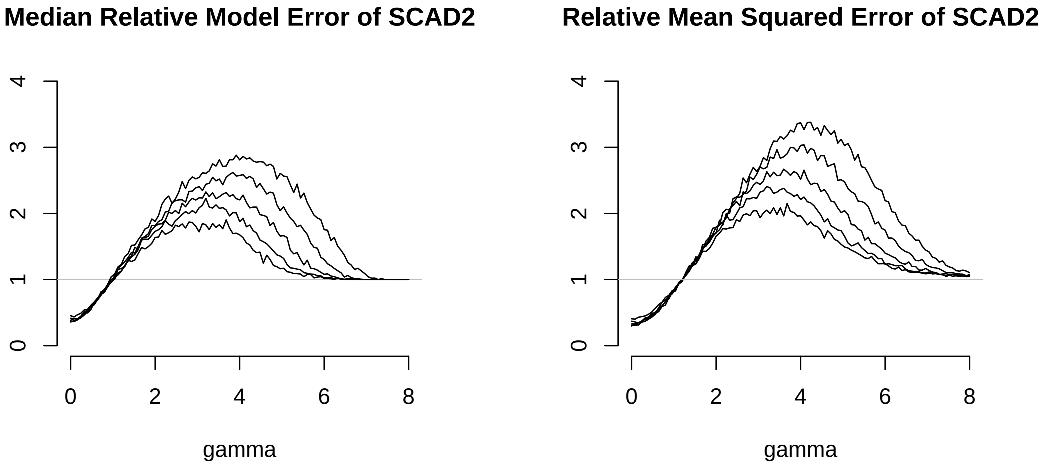
where the penalty function p_λ is defined in Fan and Li (2001) and $\lambda \geq 0$ is a tuning parameter. The penalty function p_λ contains also another tuning parameter a, which is set equal to 3.7 here, resulting in a particular instance of the SCAD estimator which is denoted by SCAD2 in Example

4.1 of Fan and Li (2001). According to Theorem 2 in Fan and Li (2001) the SCAD estimator is guaranteed to satisfy the sparsity property if $\lambda \to 0$ and $\sqrt{n}\lambda \to \infty$ as samples size n goes to infinity.

Using the MATLAB code provided to us by Runze Li, we have implemented the SCAD2 estimator in R. [The code is available from the first author on request.] Two types of performance measures are considered: The 'median relative model error' studied by Fan and Li (2001), and the relative mean squared error. The median relative model error is defined as follows: For an estimator $\hat{\theta}$ for θ, define the model error $ME(\hat{\theta})$ by $ME(\hat{\theta}) = (\hat{\theta}-\theta)'\Sigma(\hat{\theta}-\theta)$, where Σ denotes the variance/covariance matrix of the regressors. Now define the relative model error of $\hat{\theta}$ (relative to least squares) by $ME(\hat{\theta})/ME(\hat{\theta}_{LS})$, with $\hat{\theta}_{LS}$ denoting the least squares estimator based on the overall model. The median relative model error is then given by the median of the relative model error. The relative mean squared error of $\hat{\theta}$ is given by $E[(\hat{\theta}-\theta)'(\hat{\theta}-\theta)]/E[(\hat{\theta}_{LS}-\theta)'(\hat{\theta}_{LS}-\theta)]$.[10] Note that we have scaled the performance measures such that both of them are identical to unity for $\hat{\theta} = \hat{\theta}_{LS}$.

Setup I: For SCAD2 the tuning parameter λ is chosen by generalized cross-validation (cf. Section 4.2 of Fan and Li (2001)). In the original study in Fan and Li (2001), the range of λ's considered for generalized cross-validation at sample sizes $n = 40$ and $n = 60$ is $\{\delta(\hat{\sigma}/\sqrt{n}) : \delta = 0.9, 1.1, 1.3, \ldots, 2\}$; here, $\hat{\sigma}^2$ denotes the usual unbiased variance estimator obtained from a least-squares fit of the overall model. For the simulations under Setup I, we re-scale this range of λ's by $\log n/\log 60$. With this, our results for $\gamma = 0$ replicate those in Fan and Li (2001) for $n = 60$; for the other (larger) sample sizes that we consider, the re-scaling guarantees that $\lambda \to 0$ and $\sqrt{n}\lambda \to \infty$ and hence, in view of Theorem 2 in Fan and Li (2001), guarantees that the resulting estimator enjoys the sparsity condition. [For another choice of λ see Setup VI.] We compute Monte Carlo estimates for both the median relative model error and the relative mean squared error of the SCAD2 estimator for a range of true parameter values, namely $\theta_n = \theta_0 + (\gamma/\sqrt{n}) \times (0, 0, 1, 1, 0, 1, 1, 1)'$ for 101 equidistant values of γ between 0 and 8, and for sample sizes $n = 60$, 120, 240, 480, and 960, each based on 500 Monte Carlo replications (for comparison, Fan and Li (2001) use 100 replications). Note that the performance measures are symmetric about $\gamma = 0$, and hence are only reported for nonnegative values of γ. The results are summarized in Figure 1 below. [For better readability, points in Figure 1 are joined by lines.]

[10] The mean squared error of $\hat{\theta}_{LS}$ is given by $E\,\mathrm{trace}((X'X)^{-1})$ which equals $\mathrm{trace}(\Sigma^{-1})/(n-9) = 38/(3n-27)$ by von Rosen (1988, Theorem 3.1).


Median Relative Model Error of SCAD2

Relative Mean Squared Error of SCAD2

Figure 1: Monte Carlo performance estimates under the true parameter $\theta_n = \theta_0 + (\gamma/\sqrt{n}) \times (0,0,1,1,0,1,1,1)'$, as a function of γ. The left panel gives the estimated median relative model error of SCAD2 for sample sizes $n = 60, 120, 240, 480, 960$. The right panel gives the corresponding results for the estimated relative mean squared error of SCAD2. Larger sample sizes correspond to larger maximal errors. For comparison, the gray line at one indicates the performance of the ordinary least squares estimator.

In the Monte Carlo study of Fan and Li (2001), only the parameter value θ_0 is considered. This corresponds to the point $\gamma = 0$ in the panels of Figure 1. At that particular point in the parameter space, SCAD2 compares quite favorably with least squares. However, Figure 1 shows that there is a large range of parameters where the situation is reversed. In particular, we see that SCAD2 can perform quite unfavorably when compared to least squares if the true parameter, i.e., θ_n, is such that some of its components are close to, but different from, zero. In line with Theorem 2.1, we also see that the worst-case performance of SCAD2 deteriorates with increasing sample size: For $n = 60$, ordinary least squares beats SCAD2 in terms of worst-case performance by a factor of about 2 in both panels of Figure 1; for $n = 960$, this factor has increased to about 3; and increasing n further makes this phenomenon even more pronounced. We also see that, for increasing n, the location of the peak moves to the right in Figure 1, suggesting that the worst-case performance of SCAD2 (among parameters of the form $\theta_n = (\gamma/\sqrt{n}) \times (0,0,1,1,0,1,1,1)'$) is attained at a value γ_n, which is such that $\gamma_n \to \infty$ with n. In view of the proof of Theorem 2.1, this is no surprise.[11] [Of course, there may be other parameters at any given sample size for

[11] See Section 2.1 and Footnote 14 in Leeb and Pötscher (2005) for related discussion.

which SCAD2 performs even worse.] Our simulations thus demonstrate: If each component of the true parameter is either very close to zero or quite large (where the components' size has to be measured relative to sample size), then the SCAD estimator performs well. However, if some component is in-between these two extremes, then the SCAD estimator performs poorly. In particular, the estimator can perform poorly precisely in the important situation where it is statistically difficult to decide whether some component of the true parameter is zero or not. Poor performance is obtained in the worst case over a neighborhood of one of the lower-dimensional models, where the 'diameter' of the neighborhood goes to zero slower than $1/\sqrt{n}$.

We have also re-run our simulations for other experimental setups; the details are given below. Since our findings for these other setups are essentially similar to those summarized in Figure 1, we first give a brief overview of the other setups and summarize the results before proceeding to the details. In Setups II and III we consider slices of the 8-dimensional performance measure surfaces corresponding to directions other than the one used in Setup I: In Setup II the true parameter is of the form $\theta_0 + (\gamma/\sqrt{n}) \times (0,0,1,1,0,0,0,0)'$, i.e., we consider the case where some components are exactly zero, some are large, and others are in-between. In Setup III, we consider a scenario in-between Setup I and Setup II, namely the case where the true parameter is of the form $\theta_0 + (\gamma/\sqrt{n}) \times (0,0,1,1,0,1/10,1/10,1/10)'$. The method for choosing λ in these two setups is the same as in Setup I. The results in these additional setups are qualitatively similar to those shown in Figure 1 but slightly less pronounced. In further setups we also consider various other rates for the SCAD tuning parameter λ. By Theorem 2 of Fan and Li (2001), the SCAD estimator is sparse if $\lambda \to 0$ and $\sqrt{n}\lambda \to \infty$; as noted before, for Figure 1, λ is chosen by generalized cross-validation from the set $\Lambda_n = \{\delta(\hat{\sigma}/\sqrt{n})(\log(n)/\log(60)) : \ \delta = 0.9, 1.1, 1.3, \ldots, 2\}$; i.e., we have $\sqrt{n}\lambda = O_p(\log(n))$. The magnitude of λ has a strong impact on the performance of the estimator. Smaller values result in 'less sparse' estimates, leading to less favorable performance relative to least squares at $\gamma = 0$, but at the same time leading to less unfavorable worst-case performance; the resulting performance curves are 'flatter' than those in Figure 1. Larger values of λ result in 'more sparse' estimates, improved performance at $\gamma = 0$, and more unfavorable worst-case performance; this leads to performance curves that are 'more spiked' than those in Figure 1. In Setups IV and V we have re-run our simulations with γ chosen from a set Λ_n as above, but with $\log(n)/\log(60)$ replaced by $(n/60)^{1/10}$ as well as by $(n/60)^{1/4}$, resulting in $\sqrt{n}\lambda = O_p(n^{1/10})$ and $\sqrt{n}\lambda = O_p(n^{1/4})$, respectively. In Setup IV, where $\sqrt{n}\lambda = O_p(n^{1/10})$, we get results similar to, but less pronounced than, Figure 1; this is because Setup IV leads to λ's smaller than in Setup I. In Setup V, where $\sqrt{n}\lambda = O_p(n^{1/4})$, we get similar but more pronounced results when compared to Figure 1; again, this is so because Setup V leads to larger λ's than Setup I. A final setup

(Setup VI) in which we do not enforce the conditions for sparsity is discussed below after the details for Setups II-V are presented.

Setups II and III: In Setup II, we perform the same Monte Carlo study as in Setup I, the only difference being that the range of θ's is now $\theta_n = \theta_0 + (\gamma/\sqrt{n}) \times (0,0,1,1,0,0,0,0)'$ for 101 equidistant values of γ between 0 and 8. The worst-case behavior in this setup is qualitatively similar to the one in Setup I but slightly less pronounced; we do not report the results here for brevity. In Setup III, we again perform the same Monte Carlo study as in Setup I, but now with $\theta_n = \theta_0 + (\gamma/\sqrt{n}) \times (0,0,1,1,0,1/10,1/10,1/10)'$ for 101 equidistant values of γ between 0 and 80. Note that here we consider a range for γ wider than that in Scenario I and II, where we had $0 \leq \gamma \leq 8$. Figure 2 gives the results for Setup III.

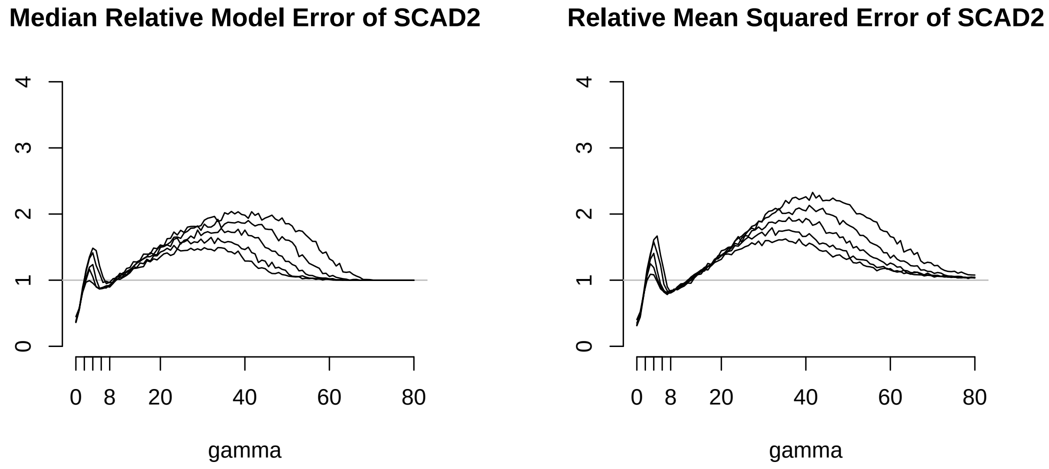


Figure 2: Monte Carlo performance estimates under the true parameter $\theta_n = \theta_0 + (\gamma/\sqrt{n}) \times (0,0,1,1,0,1/10,1/10,1/10)'$, as a function of γ. See the legend of Figure 1 for a description of the graphics.

The same considerations as given for Figure 1 also apply to Figure 2. The new feature in Figure 2 is that the curves are bimodal. Apparently, this is because now there are two regions, in the range of γ's under consideration, for which some components of the underlying regression parameter θ_n are neither very close to zero nor quite large (relative to sample size): Components 3 and 4 for γ around 5 (first peak), and components 6, 7, and 8 for γ around 40 (second peak).

Setups IV and V: Here we perform the same simulations as in Setup I, but now with the range of λ's considered for generalized cross-validation given by $\{\delta(\hat{\sigma}/\sqrt{n})(n/60)^{1/10} : \delta =$

$0.9, 1.1, 1.3, \ldots, 2\}$ for Setup IV, and by $\{\delta(\hat{\sigma}/\sqrt{n})(n/60)^{1/4} : \ \delta = 0.9, 1.1, 1.3, \ldots, 2\}$ for Setup V. Setup IV gives 'less sparse' estimates while Setup V gives 'more sparse' estimates relative to Setup I. The results are summarized in Figures 3 and 4 below. Choosing the SCAD tuning-parameter λ so that the estimator is 'more sparse' clearly has a detrimental effect on the estimator's worst-case performance.

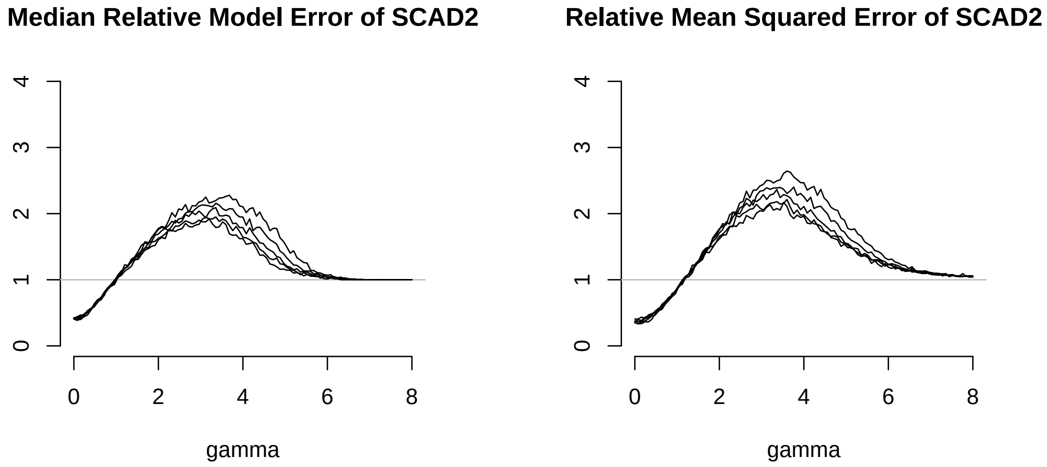


Figure 3: Monte Carlo performance estimates under the true parameter $\theta_n = \theta_0 + (\gamma/\sqrt{n}) \times (0, 0, 1, 1, 0, 1, 1, 1)'$ as a function of γ; the SCAD tuning parameter λ is chosen as described in Setup IV.

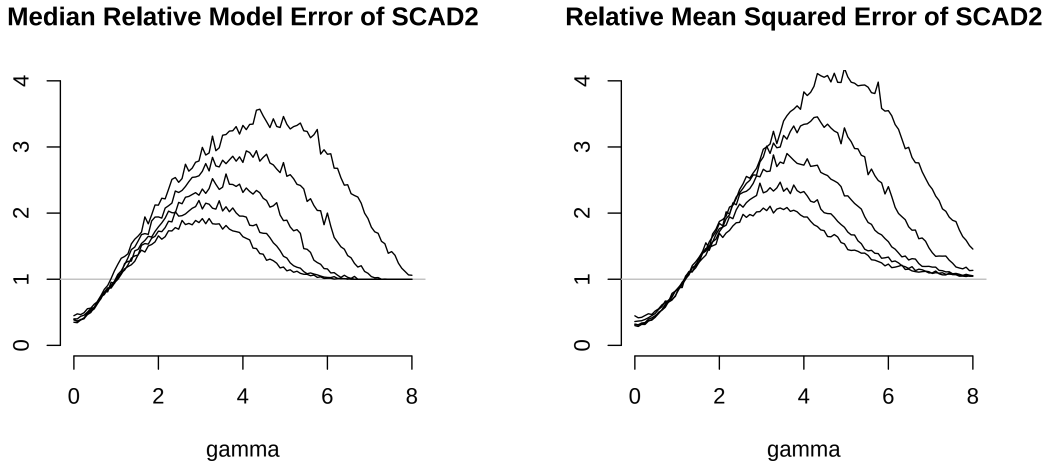


Figure 4: Monte Carlo performance estimates under the true parameter $\theta_n = \theta_0 + (\gamma/\sqrt{n}) \times (0,0,1,1,0,1,1,1)'$, as a function of γ; the SCAD tuning parameter λ is chosen as described in Setup V.

In all setups considered so far we have enforced the conditions $\lambda \to 0$ and $\sqrt{n}\lambda \to \infty$ to guarantee sparsity of the resulting SCAD estimator as risk properties of sparse estimators are the topic of the paper. In response to a referee we further consider Setup VI which is identical to Setup I, except that the range of λ's over which generalized cross-validation is effected is given by $\{\delta(\hat{\sigma}/\sqrt{n}) : \ \delta = 0.9, 1.1, 1.3, \ldots, 2\}$, which is precisely the range considered in Fan and Li (2001). Note that the resulting λ does now *not* satisfy the conditions for sparsity given in Theorem 2 of Fan and Li (2001). The results are shown in Figure 5 below. The findings are similar to the results from Setup I, in that SCAD2 gains over the least squares estimator in a neighborhood of θ_0, but is worse by approximately a factor of 2 over considerable portions of the range of γ, showing once more that the simulation study in Fan and Li (2001) does not tell the entire truth. What is, however, different here from the results obtained under Setup I is that – not surprisingly at all – the worst case behavior now does not get worse with increasing sample size. [This is akin to the boundedness of the worst case risk of a post-model-selection estimator based on a conservative model selection procedure like AIC or pre-testing with a sample-size independent critical value.]

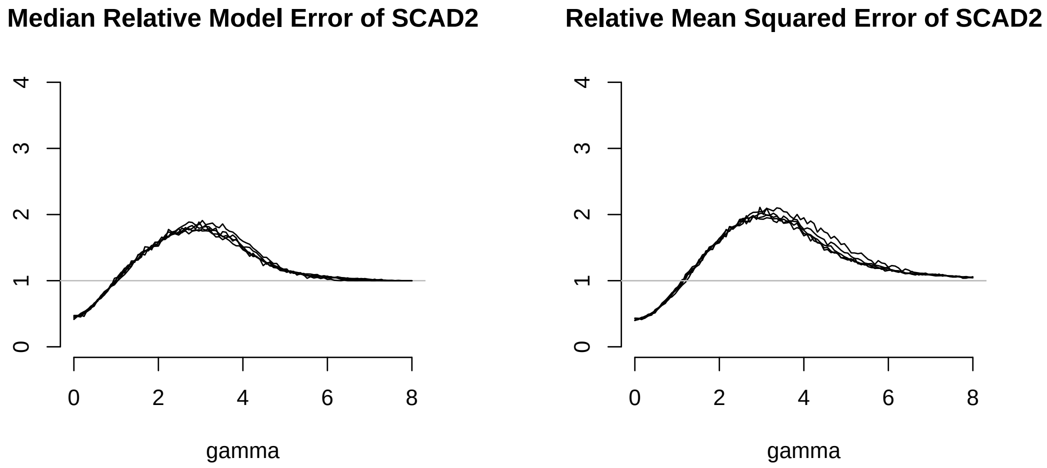


Figure 5: Monte Carlo performance estimates under the true parameter $\theta_n = \theta_0 + (\gamma/\sqrt{n}) \times (0,0,1,1,0,1,1,1)'$, as a function of γ; the SCAD tuning parameter λ is chosen as described

in Setup VI.

4 Conclusion

We have shown that sparsity of an estimator leads to undesirable risk properties of that estimator. The result is set in a linear model framework, but easily extends to much more general parametric and semiparametric models, including time series models. Sparsity is often connected to a so-called "oracle property". We point out that this latter property is highly misleading and should not be relied on when judging performance of an estimator. Both observations are not really new, but worth recalling: Hodges' construction of an estimator exhibiting a deceiving pointwise asymptotic behavior (i.e., the oracle property in today's parlance) has led mathematical statisticians to realize the importance uniformity has to play in asymptotic statistical results. It is thus remarkable that today – more than 50 years later – we observe a return of Hodges' estimator in the guise of newly proposed estimators (i.e., sparse estimators). What is even more surprising is that the deceiving pointwise asymptotic properties of these estimators (i.e., the oracle property) are now advertised as virtues of these methods. It is therefore perhaps fitting to repeat Hajek's (1971, p.153) warning:

> "Especially misinformative can be those limit results that are not uniform. Then the limit may exhibit some features that are not even approximately true for any finite n."

The discussion in the present paper as well as in Leeb and Pötscher (2005) shows in particular that distributional or risk behavior of consistent post-model-selection estimators is not as sometimes believed, but is much worse.

The results of this paper should not be construed as a criticism of shrinkage-type estimators including penalized least squares (maximum likelihood) estimators per se. Especially if the dimension of the model is large relative to sample size, some sort of shrinkage will typically be beneficial. However, achieving this shrinkage through sparsity is perhaps not such a good idea (at least when estimator risk is of concern). It certainly cannot simply be justified through an appeal to the oracle property.[12]

[12] In this context we note that "superefficiency" per se is not necessarily detrimental in higher dimensions as witnessed by the Stein phenomenon. However, not all forms of "superefficiency" are created equal, and "superefficiency" generated through sparsity of an estimator typically belongs to the undesirable variety as shown in the paper.

Acknowledgements

A version of this paper was previously circulated in 2004. We are grateful to the editor Ron Gallant and the referees as well as to Hemant Ishwaran, Paul Kabaila, Richard Nickl, and Yuhong Yang for helpful comments.

5 References

Bunea, F. (2004): Consistent covariate selection and post model selection inference in semiparametric regression. *Annals of Statistics* 32, 898-927.

Bunea, F. & I. W. McKeague (2005): Covariate selection for semiparametric hazard function regression models. *Journal of Multivariate Analysis* 92, 186-204.

Cai, J., Fan, J., Li, R., & H. Zhou (2005): Variable selection for multivariate failure time data, *Biometrika* 92, 303-316.

Fan, J. & R. Li (2001): Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96, 1348-1360.

Fan, J. & R. Li (2002): Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics* 30, 74-99.

Fan, J. & R. Li (2004): New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association* 99, 710-723.

Fan, J. & H. Peng (2004): Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics* 32, 928-961.

Foster D. P. & E. I. George (1994): The risk inflation criterion for multiple regression. *Annals of Statistics* 22, 1947-1975.

Frank, I. E. & J. H. Friedman (1993): A statistical view of some chemometrics regression tools (with discussion). *Technometrics* 35, 109-148.

Hajek, J. (1971): Limiting properties of likelihoods and inference. In: V. P. Godambe and D. A. Sprott (eds.), *Foundations of Statistical Inference: Proceedings of the Symposium on the Foundations of Statistical Inference, University of Waterloo, Ontario, March 31 – April 9, 1970*, 142-159. Toronto: Holt, Rinehart & Winston.

Hajek, J. & Z. Sidak (1967): *Theory of Rank Tests.* New York: Academic Press.

Hosoya, Y. (1984): Information criteria and tests for time series models. In: O. D. Anderson (ed.), *Time Series Analysis: Theory and Practice* 5, 39-52. Amsterdam: North-Holland.

Judge, G. G. & M. E. Bock (1978): *The Statistical Implications of Pre-test and Stein-rule Estimators in Econometrics*. Amsterdam: North-Holland.

Kabaila, P. (1995): The effect of model selection on confidence regions and prediction regions. *Econometric Theory* 11, 537-549.

Kabaila, P. (2002): On variable selection in linear regression. *Econometric Theory* 18, 913-915.

Knight, K. & W. Fu (2000): Asymptotics of lasso-type estimators. *Annals of Statistics* 28, 1356-1378.

Koul, H. L. & W. Wang (1984): Local asymptotic normality of randomly censored linear regression model. *Statistics & Decisions*, Supplement Issue No. 1, 17-30.

Lehmann, E. L. & G. Casella (1998): *Theory of Point Estimation*. Springer Texts in Statistics. New York: Springer-Verlag.

Leeb, H. & B. M. Pötscher (2005): Model selection and inference: facts and fiction. *Econometric Theory* 21, 21-59.

Leeb, H. & B. M. Pötscher (2006): Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory* 22, 69-97. (Correction, ibid., forthcoming.)

Pötscher, B. M. (1991): Effects of model selection on inference. *Econometric Theory* 7, 163-185.

Shibata R. (1986a): Consistency of model selection and parameter estimation. *Journal of Applied Probability, Special Volume* 23A, 127-141.

Shibata R. (1986b): Selection of the number of regression variables; a minimax choice of generalized FPE. *Annals of the Institute of Statistical Mathematics* 38, 459-474.

Tibshirani, R. J. (1996): Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Ser. B* 58, 267-288.

Von Rosen, D. (1988): Moments for the inverted Wishart distribution. *Scandinavian Journal of Statistics* 15, 97-109.

Yang, Y. (2005): Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* 92, 937-950.

Zou, H. (2006): The adaptive lasso and its orcale properties. *Journal of the American Statistical Association* 101, 1418-1429.