Confidence Sets Based on Sparse Estimators Are Necessarily Large

Benedikt M. Pötscher

Department of Statistics, University of Vienna

August 2007

Confidence Sets Based on Sparse Estimators Are Necessarily Large

Benedikt M. Pötscher*
Department of Statistics, University of Vienna

Preliminary version: April 2007
This version: August 2007

Abstract

Confidence sets based on sparse estimators are shown to be large compared to more standard confidence sets, demonstrating that sparsity of an estimator comes at a substantial price in terms of the quality of the estimator. The results are set in a general parametric or semiparametric framework.

MSC Subject Classifications: Primary 62F25; secondary 62C25, 62J07

Keywords: sparse estimator, consistent model selection, post-model-selection estimator, penalized maximum likelihood, confidence set, coverage probability

1 Introduction

Sparse estimators have received increased attention in the statistics literature in recent years. An estimator for a parameter vector is called sparse if it estimates the zero components of the true parameter vector by zero with probability approaching one as sample size increases without bound. Examples of sparse estimator are (i) post-model-selection estimators following a consistent model selection procedure, (ii) thresholding estimators with a suitable choice of the thresholds, and (iii) many penalized maximum likelihood estimators (e.g., SCAD, LASSO, and variants thereof) when the regularization parameter is chosen in a suitable way. Many (but not all) of these sparse estimators also have the property that the asymptotic distribution of the estimator coincides with the asymptotic distribution of the (infeasible) estimator that uses the zero restrictions in the true parameter; see, e.g., Pötscher (1991, Lemma 1), Fan and Li (2001). This property – in the context of SCAD estimation – has been dubbed

*Department of Statistics, University of Vienna, Universitätsstrasse 5, A-1010 Vienna. Phone: +431 427738640. E-mail: benedikt.poetscher@univie.ac.at

the "oracle" property by Fan and Li (2001) and has received considerable attention in the literature, witnessed by a series of papers establishing the "oracle" property for a variety of estimators (e.g., Bunea (2004), Bunea and McKeague (2005), Fan and Li (2002, 2004), Zou (2006), Li and Liang (2007), Wang and Leng (2006), Wang, G. Li, and Tsai (2007), Wang, R. Li, and Tsai (2007), Zhang and Li (2007)).

The sparsity property and the closely related "oracle" property seem to intimate that an estimator enjoying these properties is superior to classical estimator like the maximum likelihood estimator (not possessing the "oracle" property). We show, however, that the sparsity property of an estimator does not translate into good properties of confidence sets based on this estimator. Rather we show in Section 2 that any confidence set based on a sparse estimator is necessarily large relative to more standard confidence sets, e.g., obtained from the maximum likelihood estimator, that have the same guaranteed coverage probability. Hence, there is a substantial price to be paid for sparsity, which is not revealed by the pointwise asymptotic analysis underlying the "oracle" property. Special cases of the general results provided in Section 2, have been observed in the literature: Kabaila (1995) notes that the "naive" confidence interval centered at Hodges estimator has infimal coverage probability that converges to zero as sample size goes to infinity; cf. also Beran (1992). [By the "naive" confidence interval we mean the interval one would construct in the usual way from the pointwise asymptotic distribution of Hodges estimator.] Similar results for "naive" confidence intervals centered at post-model-selection estimators that are derived from certain consistent model selection procedures can be found in Kabaila (1995) and Leeb and Pötscher (2005). [We note that these "naive" confidence intervals have coverage probabilities that converge to the nominal level *pointwise* in the parameter space, but these confidence intervals are – in view of the results just mentioned – not "honest" in the sense that the infimum over the parameter space of the coverage probabilities converges to a level that is *below* the nominal level.] Properties of confidence sets based on not necessarily sparsely tuned post-model-selection estimators are discussed in Kabaila (1995, 1998), Pötscher (1995), Leeb and Pötscher (2005), Kabaila and Leeb (2006).

The results discussed in the preceding paragraph show, in particular, that the "oracle" property is problematic as it gives a much too optimistic impression of the actual properties of an estimator. This problematic nature of the "oracle" property is also discussed in Leeb and Pötscher (2007) from a risk point of view; cf. also Yang (2005). The problematic nature of the "oracle" property is connected to the fact that the finite-dimensional distributions of these estimators converge to their limits pointwise in the parameter space but not uniformly. Hence, the limits often do not reveal the actual properties of the finite-sample distributions. An asymptotic analysis using a "moving parameter" asymptotics is possible and captures much of the actual behavior of the estimators, see Leeb and Pötscher (2005), Pötscher and Leeb (2007). These results lead to a view of these estimators that is less favorable then what is suggested by the "oracle" property.

The remainder of the paper is organized as follows: In Section 2 we provide the main results showing that confidence sets based on sparse estimators are necessarily large. These results are extended to "partially" sparse estimators in Section 2.1. In Section 3 we consider a thresholding estimator as a simple example of a sparse estimator, construct a confidence set based on this estimator, and discuss its properties.

2 On the size of confidence sets based on sparse estimators

Suppose we are given a sequence of statistical experiments

$$\left\{P_{n,\theta} : \theta \in \mathbb{R}^k\right\} \qquad n = 1, 2, \ldots \tag{1}$$

where the probability measures $P_{n,\theta}$ live on suitable measure spaces $(\mathcal{X}_n, \mathfrak{X}_n)$. [Often $P_{n,\theta}$ will arise as the distribution of a random vector $(y_1', \ldots, y_n')'$ where y_i takes values in a Euclidean space. In this case $\mathcal{X}_n$ will be an n-fold product of that Euclidean space and $\mathfrak{X}_n$ will be the associated Borel σ-field; also n will then denote sample size.] We assume further that for every $\gamma \in \mathbb{R}^k$ the sequence of probability measures

$$\left\{P_{n,\gamma/\sqrt{n}} : n = 1, 2, \ldots\right\}$$

is contiguous w.r.t. the sequence

$$\left\{P_{n,0} : n = 1, 2, \ldots\right\}.$$

This is a quite weak assumption satisfied by many statistical experiments; for example, it is certainly satisfied whenever the experiment is locally asymptotically normal. The above assumption that the parameter space is $\mathbb{R}^k$ is made only for simplicity of presentation and is by no means essential, see Remark 6.

Let $\hat{\theta}_n$ denote a sequence of estimators, i.e., $\hat{\theta}_n$ is a measurable function on $\mathcal{X}_n$ taking values in $\mathbb{R}^k$. We say that the estimator $\hat{\theta}_n$ (more precisely, the sequence of estimators) is *sparse* if for every $\theta \in \mathbb{R}^k$ and $i = 1, \ldots, k$

$$\lim_{n\to\infty} P_{n,\theta}\left(\hat{\theta}_{n,i} = 0\right) = 1 \quad \text{holds whenever} \quad \theta_i = 0. \tag{2}$$

Here $\hat{\theta}_{n,i}$ and θ_i denote the i-th component of $\hat{\theta}_n$ and of θ, respectively. That is, the estimator is guaranteed to find the zero components of θ with probability approaching one as $n \to \infty$. [The focus on zero-values in the coordinates of θ is of course arbitrary. Furthermore, note that Condition (2) is of course satisfied for nonsensical estimators like $\hat{\theta}_n \equiv 0$. The sparse estimators mentioned in Section 1, however, are more sensible as they are typically also consistent for θ.]

We are interested in confidence sets for θ based on $\hat{\theta}_n$. Let C_n be a random set in $\mathbb{R}^k$ in the sense that $C_n = C_n(\omega)$ is a subset of $\mathbb{R}^k$ for every $\omega \in \mathcal{X}_n$ with the property that for every $\theta \in \mathbb{R}^k$

$$\{\omega \in \mathcal{X}_n : \theta \in C_n(\omega)\}$$

is measurable, i.e., belongs to $\mathfrak{X}_n$. We say that the random set C_n is *based on* the estimator $\hat{\theta}_n$ if C_n satisfies

$$P_{n,\theta}\left(\hat{\theta}_n \in C_n\right) = 1 \tag{3}$$

for every $\theta \in \mathbb{R}^k$. For example, if C_n is a k-dimensional interval (box) of the form

$$\left[\hat{\theta}_n - a_n, \hat{\theta}_n + b_n\right] \tag{4}$$

where a_n and b_n are random vectors in $\mathbb{R}^k$ with only nonnegative coordinates, then condition (3) is trivially satisfied. Here we use the notation $[c,d] = [c_1,d_1] \times \cdots \times [c_k,d_k]$ for vectors $c = (c_1,\ldots,c_k)'$ and $d = (d_1,\ldots,d_k)'$. We also use the following notation: For a subset A of $\mathbb{R}^k$, let

$$\operatorname{diam}(A) = \sup\{\|x-y\| : x \in A, y \in A\}$$

denote the diameter of A (measured w.r.t. the usual Euclidean norm $\|\cdot\|$); furthermore, if e is an arbitrary element of $\mathbb{R}^k$ of length 1, and $a \in A$ let

$$\operatorname{ext}(A,a,e) = \sup\{\lambda \geq 0 : \lambda e + a \in A\}.$$

That is, $\operatorname{ext}(A,a,e)$ measures how far the set A extends from the point a into the direction given by e.

The following result shows that confidence sets based on a sparse estimator are necessarily large.

Theorem 1 *Suppose the statistical experiment given in (1) satisfies the above contiguity assumption. Let $\hat{\theta}_n$ be a sparse estimator sequence and let C_n be a sequence of random sets based on the estimator $\hat{\theta}_n$ in the sense of (3). Assume that C_n is a confidence set for θ with asymptotic infimal coverage probability δ, i.e.,*

$$\delta = \liminf_{n\to\infty} \inf_{\theta\in\mathbb{R}^k} P_{n,\theta}\left(\theta \in C_n\right).$$

Then for every $t \geq 0$ and every $e \in \mathbb{R}^k$ of length 1 we have

$$\liminf_{n\to\infty} \sup_{\theta\in\mathbb{R}^k} P_{n,\theta}\left(\sqrt{n}\operatorname{ext}(C_n,\hat{\theta}_n,e) \geq t\right) \geq \delta. \tag{5}$$

In particular, we have for every $t \geq 0$

$$\liminf_{n\to\infty} \sup_{\theta\in\mathbb{R}^k} P_{n,\theta}\left(\sqrt{n}\operatorname{diam}(C_n) \geq t\right) \geq \delta. \tag{6}$$

[If the set inside of the probability in (5) or (6) is not measurable, the probability is to be replaced by inner probability.]

Proof. Since obviously $\operatorname{diam}(C_n) \geq \operatorname{ext}(C_n, \hat{\theta}_n, e)$ holds, it suffices to prove (5). Now, for every sequence $\theta_n \in \mathbb{R}^k$ we have in view of (3)

$$\begin{aligned} \delta &= \liminf_{n\to\infty} \inf_{\theta\in\mathbb{R}^k} P_{n,\theta}\left(\theta \in C_n\right) \leq \liminf_{n\to\infty} P_{n,\theta_n}\left(\theta_n \in C_n\right) \qquad (7) \\ &= \liminf_{n\to\infty} \Big\{ P_{n,\theta_n}\left(\theta_n \in C_n, \hat{\theta}_n \in C_n, \hat{\theta}_n = 0\right) \\ &\quad + P_{n,\theta_n}\left(\theta_n \in C_n, \hat{\theta}_n \neq 0\right) \Big\}. \end{aligned}$$

Sparsity implies

$$\lim_{n\to\infty} P_{n,0}\left(\hat{\theta}_n \neq 0\right) = 0,$$

and hence for $\theta_n = \gamma/\sqrt{n}$ the contiguity assumption implies

$$\limsup_{n\to\infty} P_{n,\theta_n}\left(\theta_n \in C_n, \hat{\theta}_n \neq 0\right) \leq \lim_{n\to\infty} P_{n,\theta_n}\left(\hat{\theta}_n \neq 0\right) = 0.$$

Consequently, we obtain from (7) for $\theta_n = \gamma/\sqrt{n}$ with $\gamma \neq 0$

$$\begin{aligned} \delta &\leq \liminf_{n\to\infty} P_{n,\theta_n}\left(\theta_n \in C_n, \hat{\theta}_n \in C_n, \hat{\theta}_n = 0\right) \\ &\leq \liminf_{n\to\infty} P_{n,\theta_n}\left(\sqrt{n}\operatorname{ext}(C_n, \hat{\theta}_n, \gamma/\left\|\gamma\right\|) \geq \left\|\gamma\right\|\right) \qquad (8) \end{aligned}$$

because of the obvious inclusion

$$\left\{\theta_n \in C_n, \hat{\theta}_n \in C_n, \hat{\theta}_n = 0\right\} \subseteq \left\{\operatorname{ext}(C_n, \hat{\theta}_n, \theta_n/\left\|\theta_n\right\|) \geq \left\|\theta_n\right\|\right\}.$$

Since γ was arbitrary, the result (5) follows from (8) upon identifying t and $\|\gamma\|$. ■

Corollary 2 *Suppose the assumptions of Theorem 1 are satisfied and C_n is a confidence 'interval' of the form (4). Then for every $i = 1, \ldots, k$ and every $t \geq 0$*

$$\liminf_{n\to\infty} \sup_{\theta\in\mathbb{R}^k} P_{n,\theta}\left(\sqrt{n}a_{n,i} \geq t\right) \geq \delta$$

and

$$\liminf_{n\to\infty} \sup_{\theta\in\mathbb{R}^k} P_{n,\theta}\left(\sqrt{n}b_{n,i} \geq t\right) \geq \delta$$

hold, where $a_{n,i}$ and $b_{n,i}$ denote the i-th coordinate of a_n and b_n, respectively. In particular, if a_n and b_n are nonrandom,

$$\liminf_{n\to\infty} \sqrt{n}a_{n,i} = \liminf_{n\to\infty} \sqrt{n}b_{n,i} = \infty$$

holds for every $i = 1, \ldots, k$, provided that $\delta > 0$.

Proof. Follows immediately from the previous theorem upon observing that (4) implies $\operatorname{ext}(C_n, \hat{\theta}_n, -e_i) = a_{n,i}$ and $\operatorname{ext}(C_n, \hat{\theta}_n, e_i) = b_{n,i}$ where e_i denotes the i-th standard basis vector. ■

It is instructive to compare with standard confidence sets. For example, in a normal linear regression model $\sqrt{n}$ times the diameter of the standard confidence ellipsoid is stochastically bounded uniformly in θ. In contrast, Theorem 1 tells us that any confidence set C_n based on sparse estimators with $\sqrt{n}\operatorname{diam}(C_n)$ being stochastically bounded uniformly in θ necessarily has infimal coverage probability equal to zero.

Remark 3 *(Nuisance parameters) Suppose that the sequence of statistical experiments is of the form* $\left\{P_{n,\theta,\tau} : \theta \in \mathbb{R}^k, \tau \in T\right\}$ *where* θ *is the parameter of interest and* τ *is now a (possibly infinite dimensional) nuisance parameter, and assume that the contiguity condition and sparsity condition are satisfied for any* $\tau \in T$*. Suppose further that we are again interested in confidence sets for* θ *based on* $\hat{\theta}_n$ *(in the sense that* $P_{n,\theta,\tau}\left(\hat{\theta}_n \in C_n\right) = 1$ *for all* $\theta \in \mathbb{R}^k, \tau \in T$*) that have infimal (over* θ *and* τ*) coverage probability* δ*. Applying Theorem 1 for any given* $\tau \in T$*, gives result analogous to (5) and (6) with the supremum extending now over* $\mathbb{R}^k \times T$*.*

Remark 4 *(Confidence sets for linear functions of* θ*) Suppose that a statistical experiment as before and a sparse estimator* $\hat{\theta}_n$ *is given but that we are interested in setting a confidence set for* $\vartheta = A\theta$ *that is based on* $\hat{\vartheta}_n = A\hat{\theta}_n$*, where* A *is a given* $q \times k$ *matrix. Without loss of generality assume that* A *has full row rank. [In particular, this covers the case where we have a sparse estimator for* θ*, but are interested in confidence sets for a subvector only.] Suppose* C_n *is a confidence set for* ϑ *that is based on* $\hat{\vartheta}_n$ *(in the sense that* $P_{n,\theta}\left(\hat{\vartheta}_n \in C_n\right) = 1$ *for all* $\theta \in \mathbb{R}^k$*) and that has asymptotic infimal coverage probability* δ*. Then essentially the same proof as for Theorem 1 shows that for every* $t \geq 0$ *and every* $e \in \mathbb{R}^q$ *of length* 1 *we have*

$$\liminf_{n\to\infty} \sup_{\theta\in\mathbb{R}^k} P_{n,\theta}\left(\sqrt{n}\operatorname{ext}(C_n, \hat{\vartheta}_n, e) \geq t\right) \geq \delta \tag{9}$$

and consequently also the analogue of (6) holds.

Remark 5 *The contiguity assumption together with the sparsity of the estimator was used in the proof of Theorem 1 to imply* $\lim_{n\to\infty} P_{n,\theta_n}\left(\hat{\theta}_n \neq 0\right) = 0$ *for all sequences of the form* $\theta_n = \gamma/\sqrt{n}$*. For some classes of sparse estimators this relation can even be established for all sequences of the form* $\theta_n = \gamma/v_n$ *where* v_n *is a sequence that satisfies* $v_n \to \infty$ *and* $v_n/\sqrt{n} \to 0$ *(cf. Leeb and Pötscher (2005)). Inspection of the proof of Theorem 1 shows that then a stronger result follows, namely that (5) and (6) hold even with* $\sqrt{n}$ *replaced by* v_n*. This shows that in such a case confidence sets based on sparse estimators are even larger than what is predicted by Theorem 1. This simple extension immediately applies mutatis mutandis also to the other results in the paper (with the exception of Theorem 8, an extension of which would require a separate analysis). The example discussed in Section 3 nicely illustrates the phenomenon just described.*

Remark 6 *The assumption that the parameter space indexing the statistical experiment, Θ say, is an entire Euclidean space is not essential as can be seen from the proofs. The results equally well hold if, e.g., Θ is a subset of Euclidean space that contains a ball with center at zero (simply put $\theta_n(\gamma) = \gamma/\sqrt{n}$ if this belongs to Θ, and set $\theta_n(\gamma) = 0$ otherwise). In fact, Θ could even be allowed to depend on n and to "shrink" to zero at a rate slower than $n^{-1/2}$. [In that sense the results are of a "local" rather than of a "global" nature.]*

2.1 Confidence sets based on partially sparse estimators

Suppose that in the framework of (1) the parameter vector θ is partitioned as $\theta = (\alpha', \beta')'$ where α is $(k - k_\beta) \times 1$ and β is $k_\beta \times 1$ $(0 < k_\beta < k)$. Furthermore, suppose that the estimator $\hat{\theta}_n = (\hat{\alpha}_n', \hat{\beta}_n')'$ is 'partially' sparse in the sense that it finds the zeros in β with probability approaching 1 (but not necessarily the zeros in α). That is, for every $\theta \in \mathbb{R}^k$ and $i = 1, \ldots, k_\beta$

$$\lim_{n\to\infty} P_{n,\theta}\left(\hat{\beta}_{n,i} = 0\right) = 1 \quad \text{holds whenever} \quad \beta_i = 0. \tag{10}$$

E.g., $\hat{\theta}_n$ could be a post-model-selection estimator based on a consistent model selection procedure that only subjects the elements in β to selection, the elements in α being 'protected'.

If we are now interested in a confidence set for β that is based on $\hat{\beta}_n$, we can immediately apply the results obtained sofar: By viewing α as a 'nuisance' parameter, we can use Remark 3 to conclude that Theorem 1 applies mutatis mutandis to this situation. Combining Remarks 3 and 4, we can then immediately obtain a result of the form (9) for confidence sets for $A\beta$ that are based on $A\hat{\beta}_n$, A being an arbitrary matrix of full row rank.

The above results, however, do not cover the case where one is interested in a confidence set for θ based on a partially sparse estimator $\hat{\theta}_n$, or more generally the case of confidence sets for $A\theta$ based on $A\hat{\theta}_n$, where the linear function $A\theta$ is also allowed to depend on α. For this case we have the following result.

Theorem 7 *Suppose the statistical experiment given in (1) is such that for some $\alpha \in \mathbb{R}^{k-k_\beta}$ the sequence $P_{n,(\alpha,\gamma/\sqrt{n})'}$ is contiguous w.r.t. $P_{n,(\alpha,0)'}$ for every $\gamma \in \mathbb{R}^{k_\beta}$. Let $\hat{\theta}_n$ be an estimator sequence that is partially sparse in the sense of (10). Let A be a $q \times k$ matrix of full row rank, which is partitioned conformably with θ as $A = (A_1, A_2)$, and that satisfies* $\operatorname{rank} A_1 < q$. *Let C_n be a sequence of random sets based on $A\hat{\theta}_n$ (in the sense that $P_{n,\theta}\left(A\hat{\theta}_n \in C_n\right) = 1$ for all $\theta \in \mathbb{R}^k$). Assume that C_n is a confidence set for $A\theta$ with asymptotic infimal coverage probability δ, i.e.,*

$$\delta = \liminf_{n\to\infty} \inf_{\theta\in\mathbb{R}^k} P_{n,\theta}\left(A\theta \in C_n\right).$$

Then for every $t \geq 0$ we have

$$\liminf_{n\to\infty} \sup_{\theta\in\mathbb{R}^k} P_{n,\theta}\left(\sqrt{n}\operatorname{diam}(C_n) \geq t\right) \geq \delta. \tag{11}$$

[If the set inside of the probability in (11) is not measurable, the probability is to be replaced by inner probability.]

Proof. Consider sequences $\theta_n = (\alpha, \gamma/\sqrt{n})' \in \mathbb{R}^k$ where α is as in the theorem. Then similar as in the proof of Theorem 1 exploiting partial sparsity and contiguity we arrive at

$$\begin{aligned}\delta &\leq \liminf_{n\to\infty} P_{n,\theta_n}\left(A\theta_n \in C_n\right) \\ &\leq \liminf_{n\to\infty} P_{n,\theta_n}\left(A\theta_n \in C_n, A\hat{\theta}_n \in C_n, \hat{\beta}_n = 0\right) \\ &\leq \liminf_{n\to\infty} P_{n,\theta_n}\left(\operatorname{diam}(C_n) \geq \left\|A((\alpha-\hat{\alpha}_n)', \gamma'/\sqrt{n})'\right\|\right). \qquad (12)\end{aligned}$$

By the assumption on A there exists a vector γ_0 such that $A_2\gamma_0$ is non-zero and is linearly independent of the range space of A_1. Consequently, $\Pi A_2\gamma_0 \neq 0$, where Π denotes the orthogonal projection on the orthogonal complement of the range space of A_1. Set $\gamma = c\gamma_0$ for arbitrary c. Then

$$\begin{aligned}\left\|A((\alpha-\hat{\alpha}_n)', \gamma'/\sqrt{n})'\right\|^2 &= \left\|A_1(\alpha-\hat{\alpha}_n) + A_2\gamma/\sqrt{n}\right\|^2 \\ &\geq n^{-1}c^2\left\|\Pi A_2\gamma_0\right\|^2.\end{aligned}$$

Combined with (12), this gives

$$\delta \leq \liminf_{n\to\infty} P_{n,\theta_n}\left(\sqrt{n}\operatorname{diam}(C_n) \geq |c|\left\|\Pi A_2\gamma_0\right\|\right).$$

Since $\|\Pi A_2\gamma_0\| > 0$ by construction and since c was arbitrary, the result (11) follows upon identifying t and $|c|\,\|\Pi A_2\gamma_0\|$. ■

The condition on A in the above theorem is, for example, satisfied when considering confidence sets for the entire vector θ as this corresponds to the case $A = I_k$ (and $q = k$). [The condition is also satisfied in case $A = (0_{k\times k_\beta}, I_{k_\beta})$ which corresponds to setting confidence sets for β. However, in this case already Theorem 1 applies as discussed prior to Theorem 7.]

Theorem 7 does not cover the case where a confidence set is desired for α only (i.e., $A = (I_{k-k_\beta}, 0_{(k-k_\beta)\times k_\beta})$). In fact, without further assumptions on the estimator $\hat{\theta}_n$ no result of the above sort is in general possible in this case (to see this consider the case where $\hat{\alpha}_n$ and $\hat{\beta}_n$ are independent and $\hat{\alpha}_n$ is a well-behaved estimator). However, under additional assumptions, results that show that confidence sets for α are also necessarily large will be obtained next. We first present the result and subsequently discuss the assumptions.

Theorem 8 *Suppose the statistical experiment given in (1) is such that for some $\alpha \in \mathbb{R}^{k-k_\beta}$ the sequence $P_{n,(\alpha,\gamma/\sqrt{n})'}$ is contiguous w.r.t. $P_{n,(\alpha,0)'}$ for every $\gamma \in \mathbb{R}^{k_\beta}$. Let $\hat{\theta}_n$ be an estimator sequence that is partially sparse in the sense of (10). Suppose that there exists a $(k-k_\beta)\times k_\beta$-matrix D such that for every γ the random vector $n^{1/2}(\hat{\alpha}_n - \alpha)$ converges in $P_{n,(\alpha',\gamma'/\sqrt{n})'}$-distribution to a $N(D\gamma, \Sigma)$-distributed random vector. Let A be a $q\times k$ matrix of full row*

rank, which is partitioned conformably with θ as $A = (A_1, A_2)$, and assume that $A_1 D - A_2 \neq 0$. Let C_n be a sequence of random sets based on $A\hat{\theta}_n$ (in the sense that $P_{n,\theta}\left(A\hat{\theta}_n \in C_n\right) = 1$ for all $\theta \in \mathbb{R}^k$). Assume that C_n is a confidence set for $A\theta$ with asymptotic infimal coverage probability δ, i.e.,

$$\delta = \liminf_{n\to\infty} \inf_{\theta\in\mathbb{R}^k} P_{n,\theta}\left(A\theta \in C_n\right).$$

Then for every $t \geq 0$ we have

$$\liminf_{n\to\infty} \sup_{\theta\in\mathbb{R}^k} P_{n,\theta}\left(\sqrt{n}\,\mathrm{diam}(C_n) \geq t\right) \geq \delta. \tag{13}$$

[If the set inside of the probability in (13) is not measurable, the probability is to be replaced by inner probability.]

Proof. Consider sequences $\theta_n = (\alpha, \gamma/\sqrt{n})' \in \mathbb{R}^k$ where α is as in the theorem. Then for every $t \geq 0$ we have

$$\begin{aligned}
\delta &\leq \liminf_{n\to\infty} P_{n,\theta_n}\left(A\theta_n \in C_n\right) = \liminf_{n\to\infty} P_{n,\theta_n}\left(A\theta_n \in C_n, A\hat{\theta}_n \in C_n\right) \\
&\leq \liminf_{n\to\infty} P_{n,\theta_n}\left(A\theta_n \in C_n, A\hat{\theta}_n \in C_n, n^{1/2}\left\|A(\hat{\theta}_n - \theta_n)\right\| \geq t\right) \\
&\quad + \limsup_{n\to\infty} P_{n,\theta_n}\left(n^{1/2}\left\|A(\hat{\theta}_n - \theta_n)\right\| < t\right) \\
&\leq \liminf_{n\to\infty} P_{n,\theta_n}\left(n^{1/2}\,\mathrm{diam}(C_n) \geq t\right) \\
&\quad + \limsup_{n\to\infty} P_{n,\theta_n}\left(n^{1/2}\left\|A(\hat{\theta}_n - \theta_n)\right\| < t\right).
\end{aligned} \tag{14}$$

Exploiting partial sparsity and contiguity we get

$$\begin{aligned}
&\limsup_{n\to\infty} P_{n,\theta_n}\left(n^{1/2}\left\|A(\hat{\theta}_n - \theta_n)\right\| < t\right) \\
\leq\ &\limsup_{n\to\infty} P_{n,\theta_n}\left(\hat{\beta}_n = 0, n^{1/2}\left\|A(\hat{\theta}_n - \theta_n)\right\| < t\right) \\
&+ \limsup_{n\to\infty} P_{n,\theta_n}\left(\hat{\beta}_n \neq 0\right) \\
=\ &\limsup_{n\to\infty} P_{n,\theta_n}\left(\hat{\beta}_n = 0, n^{1/2}\left\|A(\hat{\theta}_n - \theta_n)\right\| < t\right) \\
\leq\ &\limsup_{n\to\infty} P_{n,\theta_n}\left(n^{1/2}\left\|A_1(\hat{\alpha}_n - \alpha) - A_2\gamma/\sqrt{n}\right\| < t\right) \\
=\ &\limsup_{n\to\infty} P_{n,\theta_n}\left(\|X_n + (A_1 D - A_2)\gamma\| < t\right) \\
\leq\ &\limsup_{n\to\infty} P_{n,\theta_n}\left(\|X_n\| > \|(A_1 D - A_2)\gamma\| - t\right)
\end{aligned} \tag{15}$$

where X_n converges to $N(0,\Sigma)$ in P_{n,θ_n}-distribution. Since $A_1 D - A_2 \neq 0$ by assumption, we can find a γ such that $\|(A_1 D - A_2)\gamma\| - t$ is arbitrarily large,

making the far right-hand side of (15) arbitrarily small. This, together with (14), establishes the result. ■

Note that the case where a confidence set for α is sought, that is, $A = (I_{k-k_\beta}, 0_{(k-k_\beta)\times k_\beta})$, which was not covered by Theorem 7, is covered by Theorem 8 except in the special case where $D = 0$.

The weak convergence assumption in the above theorem merits some discussion: Suppose $\hat{\theta}_n$ is a post-model-selection estimator based on a model selection procedure that consistently finds the zeroes in β and then computes $\hat{\theta}_n$ as the restricted maximum likelihood estimator $\hat{\theta}_n(R)$ under the zero-restrictions in β. Under the usual regularity conditions, the restricted maximum likelihood estimator $\hat{\alpha}_n(R)$ for α will then satisfy that $n^{1/2}(\hat{\alpha}_n(R) - \alpha)$ converges to a $N(D\gamma, \Sigma)$-distribution under the sequence of local alternatives $\theta_n = (\alpha', \gamma'/\sqrt{n})'$. Since $\lim_{n\to\infty} P_{n,\theta_n}\left(\hat{\beta}_n = 0\right) = 1$ by partial sparsity and contiguity, the estimators $\hat{\alpha}_n$ and $\hat{\alpha}_n(R)$ coincide with P_{n,θ_n}-probability approaching one. This shows that the assumption on $\hat{\alpha}_n$ will typically be satisfied for such post-model-selection estimators. [For a precise statement of such a result in a simple example see Leeb and Pötscher (2005, Proposition A.2).] While we expect that this assumption on the asymptotic behavior of $\hat{\alpha}_n$ is also shared by many other partially sparse estimators, this remains to be verified on a case by case basis.

3 An Example: A confidence set based on a hard-thresholding estimator

Suppose the data $y_1, \ldots, y_n$ are independent identically distributed as $N(\theta, 1)$, $\theta \in \mathbb{R}$. Let the hard-thresholding estimator $\hat{\theta}_n$ be given by

$$\hat{\theta}_n = \bar{y}\mathbf{1}(|\bar{y}| > \eta_n)$$

where the threshold η_n is a positive number and $\bar{y}$ denotes the maximum likelihood estimator, i.e., the arithmetic mean of the data. Of course, $\hat{\theta}_n$ is nothing else than a post-model-selection estimator following a t-type test of the hypothesis $\theta = 0$ versus the alternative $\theta \neq 0$. It is well-known and easy to see that $\hat{\theta}_n$ satisfies the sparsity condition if $\eta_n \to 0$ and $n^{1/2}\eta_n \to \infty$ (i.e., the underlying model selection procedure is consistent); in this case then $n^{1/2}(\hat{\theta}_n - \theta)$ converges to a standard normal distribution if $\theta \neq 0$, whereas it converges to pointmass at zero if $\theta = 0$. Note that $\hat{\theta}_n$ – with such a choice of the threshold η_n – is an instance of Hodges' estimator. In contrast, if $\eta_n \to 0$ and $n^{1/2}\eta_n \to e$, $0 \leq e < \infty$, the estimator $\hat{\theta}_n$ is a post-model-selection estimator based on a conservative model selection procedure. See Pötscher and Leeb (2007) for further discussion and references.

In the consistent model selection case the estimator possesses the "oracle" property suggesting as a confidence interval the "naive" interval given by $C_n^{naive} = \{0\}$ if $\hat{\theta}_n = 0$ and by $C_n^{naive} = [\hat{\theta}_n - z_{(1-\delta)/2}, \hat{\theta}_n + z_{(1-\delta)/2}]$ otherwise, where δ is the nominal coverage level and $z_{(1-\delta)/2}$ is the $1-(1-\delta)/2$-quantile of

the standard normal distribution. This interval satisfies $P_{n,\theta}(\theta \in C_n^{naive}) \to \delta$ for every θ, but – as discussed in the introduction and as follows from the results in Section 2 – it is not honest and, in fact, has infimal coverage probability converging to zero. A related, but infeasible, construction is to consider the intervals $C_n^* = [\hat{\theta}_n - c_n(\theta), \hat{\theta}_n + c_n(\theta)]$ where $c_n(\theta)$ is chosen as small as possible subject to $P_{n,\theta}(\theta \in C_n^*) = \delta$ for every θ. [Note that C_n^{naive} can be viewed as being obtained from C_n^* by replacing $c_n(\theta)$ by the limits $c_\infty(\theta)$ for $n \to \infty$, where $c_\infty(\theta) = 0$ if $\theta = 0$ and $c_\infty(\theta) = z_{(1-\delta)/2}$ if $\theta \neq 0$, and then by replacing θ by $\hat{\theta}_n$ in $c_\infty(\theta)$.] An obvious idea to obtain a feasible and honest interval is now to use $c_n = \max_{\theta \in \mathbb{R}} c_n(\theta)$ as the half-length of the interval, i.e. $C_n = [\hat{\theta}_n - c_n, \hat{\theta}_n + c_n]$. From Theorem 1 we know that $\sqrt{n} c_n \to \infty$ in the case where $\eta_n \to 0$ and $n^{1/2}\eta_n \to \infty$ (and if $\delta > 0$), but it is instructive to study the behavior of C_n in more detail.

We therefore consider now confidence intervals C_n for θ of the form $C_n = [\hat{\theta}_n - a_n, \hat{\theta}_n + b_n]$ with nonnegative constants a_n and b_n (thus removing the symmetry restriction on the interval). Note that the subsequent result is a finite-sample result and hence does not involve any assumptions on the behavior of η_n.

Proposition 9 *For every $n \geq 1$, the interval $C_n = [\hat{\theta}_n - a_n, \hat{\theta}_n + b_n]$ has a infimal coverage probability satisfying*

$$\begin{aligned} & \inf_{\theta \in \mathbb{R}} P_{n,\theta}(\theta \in C_n) \\ = & \begin{cases} \Phi(n^{1/2}(a_n - \eta_n)) - \Phi(-n^{1/2} b_n) & \text{if} \quad \eta_n \leq a_n + b_n \quad \text{and} \quad a_n \leq b_n \\ \Phi(n^{1/2} a_n) - \Phi(n^{1/2}(-b_n + \eta_n)) & \text{if} \quad \eta_n \leq a_n + b_n \quad \text{and} \quad a_n \geq b_n \\ 0 & \text{if} \quad \eta_n > a_n + b_n \end{cases}, \end{aligned}$$

where Φ denotes the standard normal cumulative distribution function.

Proof. Elementary calculations and the fact that $n^{1/2}(\bar{y} - \theta)$ is standard normally distributed give for the coverage probability $p_n(\theta) = P_{n,\theta}(\theta \in C_n)$

$$\begin{aligned} p_n(\theta) &= P_{n,\theta}\left(-n^{1/2} b_n \leq n^{1/2}(\hat{\theta}_n - \theta) \leq n^{1/2} a_n\right) \\ &= \Pr\left(-n^{1/2} b_n \leq Z \leq n^{1/2} a_n, \left|Z + n^{1/2}\theta\right| > n^{1/2}\eta_n\right) \\ &\quad + \Pr\left(-b_n \leq -\theta \leq a_n, \left|Z + n^{1/2}\theta\right| \leq n^{1/2}\eta_n\right) \\ &= A + B, \end{aligned}$$

where Z is a standard normally distributed random variable and Pr denotes a generic probability. Simple, albeit tedious computations give the coverage probability as follows. If $\eta_n > a_n + b_n$

$$p_n(\theta) = \begin{cases} \Phi(n^{1/2} a_n) - \Phi(-n^{1/2} b_n) & \text{if } \theta < -a_n - \eta_n \text{ or } \theta > b_n + \eta_n \\ \Phi(n^{1/2}(-\theta - \eta_n)) - \Phi(-n^{1/2} b_n) & \text{if } -a_n - \eta_n \leq \theta < b_n - \eta_n \\ 0 & \text{if } b_n - \eta_n \leq \theta < -a_n \text{ or } b_n < \theta \leq -a_n + \eta_n \\ \Phi(n^{1/2}(-\theta + \eta_n)) - \Phi(n^{1/2}(-\theta - \eta_n)) & \text{if } -a_n \leq \theta \leq b_n \\ \Phi(n^{1/2} a_n) - \Phi(n^{1/2}(-\theta + \eta_n)) & \text{if } -a_n + \eta_n < \theta \leq b_n + \eta_n \end{cases}.$$

Hence, the infimal coverage probability in this case is obviously zero. Next, if $(a_n + b_n)/2 \leq \eta_n \leq a_n + b_n$ then

$$p_n(\theta) = \begin{cases} \Phi(n^{1/2}a_n) - \Phi(-n^{1/2}b_n) & \text{if } \theta < -a_n - \eta_n \text{ or } \theta > b_n + \eta_n \\ \Phi(n^{1/2}(-\theta - \eta_n)) - \Phi(-n^{1/2}b_n) & \text{if } -a_n - \eta_n \leq \theta < -a_n \\ \Phi(n^{1/2}(-\theta + \eta_n)) - \Phi(-n^{1/2}b_n) & \text{if } -a_n \leq \theta < b_n - \eta_n \\ \Phi(n^{1/2}(-\theta + \eta_n)) - \Phi(n^{1/2}(-\theta - \eta_n)) & \text{if } b_n - \eta_n \leq \theta \leq -a_n + \eta_n \\ \Phi(n^{1/2}a_n) - \Phi(n^{1/2}(-\theta - \eta_n)) & \text{if } -a_n + \eta_n < \theta \leq b_n \\ \Phi(n^{1/2}a_n) - \Phi(n^{1/2}(-\theta + \eta_n)) & \text{if } b_n < \theta \leq b_n + \eta_n \end{cases},$$

and if $\eta_n < (a_n + b_n)/2$

$$p_n(\theta) = \begin{cases} \Phi(n^{1/2}a_n) - \Phi(-n^{1/2}b_n) & \text{if } \theta < -a_n - \eta_n \text{ or } \theta > b_n + \eta_n \\ & \text{or } -a_n + \eta_n \leq \theta \leq b_n - \eta_n \\ \Phi(n^{1/2}(-\theta - \eta_n)) - \Phi(-n^{1/2}b_n) & \text{if } -a_n - \eta_n \leq \theta < -a_n \\ \Phi(n^{1/2}(-\theta + \eta_n)) - \Phi(-n^{1/2}b_n) & \text{if } -a_n \leq \theta < -a_n + \eta_n \\ \Phi(n^{1/2}a_n) - \Phi(n^{1/2}(-\theta - \eta_n)) & \text{if } b_n - \eta_n < \theta \leq b_n \\ \Phi(n^{1/2}a_n) - \Phi(n^{1/2}(-\theta + \eta_n)) & \text{if } b_n < \theta \leq b_n + \eta_n \end{cases}.$$

Inspection shows that in both cases the function does not have a minimum, but the infimum equals the smaller of the left-hand side limit $p_n(-a_n-)$ and the right-hand side limit $p_n(b_n+)$, which shows that the infimum of $p_n(\theta)$ equals $\min[\Phi(n^{1/2}(a_n - \eta_n)) - \Phi(-n^{1/2}b_n), \Phi(n^{1/2}a_n) - \Phi(n^{1/2}(-b_n + \eta_n))]$. ■

As a point of interest we note that the coverage probability $p_n(\theta)$ has exactly two discontinuity points (jumps), one at $\theta = -a_n$ and one at $\theta = b_n$.

An immediate consequence of the above proposition is that $n^{1/2}\operatorname{diam}(C_n) = n^{1/2}(a_n + b_n)$ is not less than $n^{1/2}\eta_n$, provided the infimal coverage probability is positive. Hence, in case that $\eta_n \to 0$ and $n^{1/2}\eta_n \to \infty$, i.e., in case that $\hat{\theta}_n$ is sparse, we see that $n^{1/2}\operatorname{diam}(C_n) \to \infty$, which of course just confirms the general result obtained in Theorem 1 above. [In fact, this result is a bit stronger as only the infimal coverage probabilities need to be positive, and not their limes inferior.]

If the interval is symmetric, i.e., $a_n = b_n$ holds, and $a_n \geq \eta_n/2$ is satisfied, the infimal coverage probability becomes $\Phi(n^{1/2}a_n) - \Phi(n^{1/2}(-a_n + \eta_n))$. Since this expression is zero if $a_n = \eta_n/2$, and is strictly increasing to one as a_n goes to infinity, any prescribed infimal coverage probability less than one is attainable. Suppose $0 < \delta < 1$ is given. Then the (shortest) confidence interval C_n of the form $[\hat{\theta}_n - a_n, \hat{\theta}_n + a_n]$ with infimal coverage probability equal to δ has to satisfy $a_n \geq \eta_n/2$ and

$$\Phi(n^{1/2}a_n) - \Phi(n^{1/2}(-a_n + \eta_n)) = \delta.$$

If now $\eta_n \to 0$ and $n^{1/2}\eta_n \to \infty$, i.e., if $\hat{\theta}_n$ is sparse, it follows that $n^{1/2}a_n \to \infty$ and

$$n^{1/2}(-a_n + \eta_n) \to \Phi^{-1}(1 - \delta)$$

or in other words that $a_n \geq \eta_n/2$ has to satisfy

$$a_n = \eta_n - n^{-1/2}\Phi^{-1}(1 - \delta) + o(n^{-1/2}). \tag{16}$$

Conversely, any $a_n \geq \eta_n/2$ satisfying (16) generates a confidence interval with asymptotic infimal coverage probability equal to δ. We observe that (16) shows that $\kappa_n \operatorname{diam}(C_n) = 2\kappa_n a_n \to \infty$ for any sequence that satisfies $\kappa_n \eta_n \to \infty$, which includes sequences that are $o(n^{1/2})$ by the assumptions on η_n. Hence, this result is stronger than what is obtained from applying Theorem 1 (or its Corollary) to this example, and illustrates the discussion in Remark 5.

References

[1] Beran, R. (1992): The radial process for confidence sets. Probability in Banach spaces, 8 (Brunswick, ME, 1991), 479–496, *Progress in Probability* 30, Birkhäuser Boston, Boston, MA.

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

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

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

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

[6] 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.

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

[8] Kabaila, P. (1998): Valid confidence intervals in regression after variable selection. *Econometric Theory* 14, 463–482.

[9] Kabaila, P. & H. Leeb (2006): On the large-sample minimal coverage probability of confidence intervals after model selection. *Journal of the American Statistical Association* 101, 619-629.

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

[11] Leeb, H. & B. M. Pötscher (2007): Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics*, doi:10.1016/j.jeconom.2007.05.017.

[12] Li, R. & H. Liang (2007): Variable selection in semiparametric regression modeling. *Annals of Statistics*, forthcoming.

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

[14] Pötscher, B. M. (1995): Comment on 'The effect of model selection on confidence regions and prediction regions'. *Econometric Theory* 11, 550–559.

[15] Pötscher, B. M. & H. Leeb (2007): On the distribution of penalized maximum likelihood estimators: the LASSO, SCAD, and thresholding. Working Paper, Department of Statistics, University of Vienna.

[16] Wang, H. & C. Leng (2006): Unified LASSO estimation via least squares approximation. Working paper.

[17] Wang, H., Li, G. & C. L. Tsai (2007): Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B* 69, 63-78.

[18] Wang, H., Li, R. & C. L. Tsai (2007): Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 94, 553-568.

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

[20] Zhang, H. H. & W. Li (2007): Adaptive lasso for Cox's proportional hazards model. *Biometrika* 94, 691-703.

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