Confidence Sets Based on Sparse Estimators Are Necessarily Large

Benedikt M. Pötscher

Department of Statistics, University of Vienna

Sparse Estimators and the "Oracle" Property

Given is a parametric statistical model indexed by a parameter $\theta \in \mathbb{R}^k$. An estimator $\hat{\theta}_n$ for θ is said to be *sparse* if for every $\theta \in \mathbb{R}^k$ and $i = 1, \ldots, k$

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

Examples of sparse estimators (that are also consistent for θ):

- Post-model-selection estimators based on a consistent model selection procedure.

- Thresholding estimators with suitable choice of threshold c_n (typically $c_n \to 0$, $n^{1/2} c_n \to \infty$).

Sparse Estimators and the "Oracle" Property (cont'd)

- Various penalized maximum likelihood (least squares) estimators (e.g., SCAD, LASSO, adaptive LASSO, certain Bridge estimators) for an appropriate choice of the regularization parameter.

For many (but not all) estimators, sparsity implies the so-called **"oracle" property**: That is, their (pointwise) asymptotic distribution coincides with the distribution of an infeasible "estimator" (the "oracle") that makes use of the zero restrictions holding for the true parameter vector θ. I.e., the estimator "adapts" to the unknown zero restrictions.

A Simple Example

$Y_1, \ldots, Y_n$ iid $N(\theta, 1)$ and $\hat{\theta}_n = \bar{Y}\mathbf{1}(\left|\bar{Y}\right| > c_n)$ with $c_n \to 0$ and $n^{1/2}c_n \to \infty$. This is Hodges' estimator. It is a post-model-selection estimator (hard-thresholding) based on consistent selection between the unrestricted model $M_U = \mathbb{R}$ and the restricted model $M_R = \{0\}$. Then $\hat{\theta}_n$ is consistent for θ and satisfies the sparsity property:

$$\lim_{n\to\infty} P_{n,\theta}\left(\hat{\theta}_n = 0\right) = 1 \qquad \text{whenever } \theta = 0,$$

as well as the "oracle" (superefficiency) property

$$n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{d} \begin{cases} N(0,1) & \theta \neq 0 \\ N(0,0) & \theta = 0 \end{cases},$$

the "oracle" being the unrestricted MLE $\hat{\theta}(U) = \bar{Y}$ if $\theta \neq 0$, and the restricted MLE $\hat{\theta}(R) = 0$ if $\theta = 0$. This seems to say that $\hat{\theta}_n$ is as good as the unrestricted MLE if $\theta \neq 0$ and as good as the restricted MLE if $\theta = 0$.

A Simple Example (cont'd)

The "oracle" property suggests the following confidence interval for θ

$$C_n = \begin{cases} (\hat{\theta}_n - n^{-1/2} z_{1-\alpha/2}, \hat{\theta}_n + n^{-1/2} z_{1-\alpha/2}) & \text{if} \quad \hat{\theta}_n \neq 0 \\ \{0\} & \text{if} \quad \hat{\theta}_n = 0 \end{cases}.$$

That, is C_n chooses between the standard confidence intervals based on the unrestricted and restricted MLE, respectively, depending on whether the model selection procedure underlying $\hat{\theta}_n$ chooses the unrestricted model $M_U = \mathbb{R}$ or the restricted model $M_R = \{0\}$. Due to the "oracle" property, C_n satisfies

$$\lim_{n \to \infty} P_{n,\theta}(\theta \in C_n) = \left\{ \begin{array}{cc} 1 - \alpha & \text{for } \theta \neq 0 \\ 1 & \text{for } \theta = 0 \end{array} \right\} \geq 1 - \alpha \quad \text{for every } \theta \in \mathbb{R}.$$

Comments on the "Oracle" Property

A selection of recent papers establishing the "oracle" property for a variety of estimators in (semi)parametric models:

Bunea (AS 2004), Bunea & McKeague (JMVA 2005)

Fan & Li (JASA 2001, AS 2002, JASA 2004), Zou (JASA 2006)

Wang & Leng (JASA 2007), Li & Liang (AS 2007)

Wang, G. Li, & Tsai (JRSS B 2007), Zhang & Li (BA 2007)

Wang, R. Li, & Tsai (BA 2007), Zou & Yuan (AS 2008), etc.

Comments on the "Oracle" Property (cont'd)

This literature views the "oracle" property as a desirable property of an estimator as the "oracle" property seems to lead to a gain in efficiency and to a gain in the size of confidence sets.

Zou & Yuan (AS 2008) call the "oracle" property a "gold standard for evaluating variable selection and coefficient estimation procedures".

Comments on the "Oracle" Property (cont'd)

However, nothing could be farther from the truth: Bad minimax risk behavior of Hodges' estimator has been known for decades (e.g., Lehmann & Casella (1998)). Furthermore, the "confidence" set C_n constructed above, although satisfying

$$\lim_{n\to\infty} P_{n,\theta}(\theta \in C_n) = \left\{ \begin{array}{cc} 1-\alpha & \text{for } \theta \neq 0 \\ 1 & \text{for } \theta = 0 \end{array} \right\} \geq 1-\alpha \quad \text{for every } \theta \in \mathbb{R},$$

is **dishonest** in the sense that its minimal coverage probability satisfies

$$\lim_{n\to\infty} \inf_{\theta\in\mathbb{R}} P_{n,\theta}(\theta \in C_n) = 0$$

as pointed out by Beran (1992) and Kabaila (1995).

We establish general results of this sort for *arbitrary* confidence sets based on *arbitrary* sparse estimators in general (semi)parametric models.

Comments on the "Oracle" Property (cont'd)

These results complement results on bad minimax risk behavior of sparse estimators in Yang (BA 2005) and Leeb &Pötscher (JE 2008); earlier minimax risk results can be found in Hosoya (1984), Shibata (AIM 1986), Foster & George (AS 1994).

Results

Assume the statistical experiment $\left\{P_{n,\theta} : \theta \in \mathbb{R}^k\right\}$ satisfies for every $\gamma \in \mathbb{R}^k$

$$P_{n,\gamma/\sqrt{n}} \text{ is contiguous w.r.t. } P_{n,0}. \tag{1}$$

Let C_n be a random set in $\mathbb{R}^k$ "based" on the sparse estimator $\hat{\theta}_n$ in the sense that

$$P_{n,\theta}(\hat{\theta}_n \in C_n) = 1 \quad \text{for every } \theta \in \mathbb{R}^k. \tag{2}$$

E.g., $C_n = [\hat{\theta}_n - a_n, \hat{\theta}_n + b_n]$ is a k-dimensional box centered at $\hat{\theta}_n$ with a_n, b_n possessing only nonnegative coordinates.

Results (cont'd)

Theorem 1: Suppose Assumption (1) is satisfied, $\hat{\theta}_n$ is sparse, and C_n satisfies (2). Let δ denote the asymptotic minimal coverage probability of C_n, i.e.,

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

Then for every $t \geq 0$

$$\liminf_{n\to\infty} \sup_{\theta\in\mathbb{R}^k} P_{n,\theta}(\sqrt{n}\,\mathsf{diam}(C_n) \geq t) \geq \delta. \tag{3}$$

More generally, for every $t \geq 0$ and every unit vector $e \in \mathbb{R}^k$

$$\liminf_{n\to\infty} \sup_{\theta\in\mathbb{R}^k} P_{n,\theta}(\sqrt{n}\,\mathsf{ext}(C_n, \hat{\theta}_n, e) \geq t) \geq \delta \tag{4}$$

where $\mathsf{ext}(C_n, \hat{\theta}_n, e) = \mathsf{sup}\{\lambda \geq 0 : \lambda e + \hat{\theta}_n \in C_n\}$.

Results (cont'd)

- Any confidence set C_n based on a sparse estimator that has positive asymptotic minimal coverage probability is necessarily larger by an order of magnitude than the classical MLE based confidence set which has diameter $\sim n^{-1/2}$. (If $\mathsf{diam}\, C_n$ is nonrandom, then $\sqrt{n}\,\mathsf{diam}\, C_n \to \infty$.)

- Confidence sets C_n based on sparse estimators and constructed from the "oracle" property, like the interval in the Hodges' estimator example, have bounded $\sqrt{n}\,\mathsf{diam}\, C_n$. Hence, they have asymptotic minimal coverage probability 0.

Results (cont'd)

- Extension to semiparametric models $\left\{P_{n,\theta,\tau} : \theta \in \mathbb{R}^k, \tau \in T\right\}$ and to confidence sets for linear functions $A\theta$ is simple.

- For particular classes of sparse estimators the results in (3) and (4) can be strengthened.

- Assumption $\Theta = \mathbb{R}^k$ not essential. Results hold as long as 0 is an interior point of Θ.

Partially Sparse Estimators

Suppose now $\theta = (\alpha', \beta')'$ where β is $k_\beta \times 1$, and the estimator $\hat{\theta}_n$ for θ is *partially sparse* in the sense that for every $\theta \in \mathbb{R}^k$ and $i = 1, \ldots, k_\beta$

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

If C_n is a confidence set for β based on $\hat{\beta}_n$, Theorem 1 (extended to semiparametric models) can be immediately applied to give a similar result. This is not so if confidence sets for θ or $A\theta$ (with this linear function also depending on α) are considered.

Partially Sparse Estimators (cont'd)

Theorem 2: Suppose 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 partially sparse. Let $A = (A_1, A_2)$ be a $q \times k$ matrix of full row-rank satisfying $\mathsf{rank}\, A_1 < q$. Suppose C_n is based on $A\hat{\theta}_n$ (i.e., $P_{n,\theta}(A\hat{\theta}_n \in C_n) = 1$ for every θ). Let δ denote the asymptotic minimal coverage probability of C_n, i.e.,

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

Then for every $t \geq 0$

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

Partially Sparse Estimators (cont'd)

The condition $\operatorname{rank} A_1 < q$ in Theorem 2 is, e.g., satisfied if $A = I_k$ or $A = (0, I_{k_\beta})$. It is not satisfied if $A = (I_{k-k_\beta}, 0)$. In this case a similar result can be obtained under an additional condition on the estimator.

Summary

- Confidence sets based on sparse estimators are necessarily larger then standard MLE based confidence sets by an order of magnitude. This results hold under very weak conditions on the (semi)parametric model. Similar results hold for partially sparse estimators.

- Sparse estimators also have bad minimax risk properties (Lehmann & Casella (1998), Yang (2005), Leeb &Pötscher (2008)).

- Hence, despite its appeal at first sight, the sparsity property and the closely related "oracle" property have detrimental consequences for an estimator and associated confidence sets. This downside of sparse estimators is not visible in the pointwise asymptotic framework underlying the "oracle" property concept of Fan & Li (2001) and others.