Implementable confidence sets in high dimensional regression

Alexandra Carpentier
Statistical Laboratory, Center for Mathematical Sciences - University of Cambridge

Abstract

We consider the setting of linear regression in high dimension. We focus on the problem of constructing *adaptive and honest confidence sets* for the sparse parameter θ, i.e. we want to construct a confidence set for θ that contains θ with high probability, and that is as small as possible. The l_2 diameter of a such confidence set should depend on the sparsity S of θ - the larger S, the wider the confidence set. However, in practice, S is unknown. This paper focuses on constructing a confidence set for θ which contains θ with high probability, whose diameter is adaptive to the unknown sparsity S, and which is implementable in practice.

1 Introduction

We consider the regression model in dimension p, with n observations,

$$Y = X\theta + \epsilon, \tag{1.1}$$

where Y is the n-dimensional observation vector, θ is the p dimensional unknown parameter, X is the $n \times p$ design matrix, and ϵ is the n dimensional white noise (see Section 3 for a complete presentation of the setting). We focus on the high dimensional setting where $n \ll p$.

Models such as the one described in Equation (1.1) have received very much attention recently. In particular, finding good estimates of θ when p is very large has many important applications (see [Starck et. al.(2010), Moriaka and Satoh(2010), Kavukcuoglu et. al.(2009)]). Solving this problem in a satisfying way is nevertheless impossible in general, since it is ill-posed. For this reason, and also because it is an assumption that holds in many concrete cases, it is usual in this setting to focus on the case where θ is a sparse parameter. Let $l_0(S)$ be the l_0 "ball" of radius S, i.e. the set of vectors that have less than S non-zero coordinates (that are S-sparse). It has been proved that in some specific cases, namely when the design matrix X satisfies some desirable conditions for $\bar{p}$ sparse vectors (e.g. null space property, R.I.P., restricted eigenvalue condition, etc, see e.g. [Candès et. al.(2006), Koltchinskii(2009), Donoho and Stark(1989), Candès(2008), Foucart and Lai(2009), Bickel et. al.(2009)]), it is possible to construct an estimate $\hat{\theta}(X,Y)$ of the parameter θ such that

$$\sup_{S: S\leq \bar{p}} \sup_{\theta \in l_0(S)} \mathbb{P}_\theta\Big(\|\hat{\theta} - \theta\|_2^2 \geq E\frac{S\log(p/\delta)}{n}\Big) < \delta, \tag{1.2}$$

where $E > 0$ is a constant, where for any vector u, $\|u\|_2^2 = \sum_j u_j^2$ is the usual l_2 norm, and where $\mathbb{P}_\theta$ is the probability according to the noise ϵ when the true parameter is θ. This result is minimax-optimal over $S-$sparse vectors for any $S \leq \bar{p}$, see e.g. [Raskutti et al(2011)]. Moreover, this bound on the accuracy of $\hat{\theta}$ scales with the true sparsity S of θ that is not available to the learner : the estimate $\hat{\theta}$ is adaptive. Some key references for the existence of such an adaptive estimate are [Zou(2006), Candès and Tao(2007), Bickel et. al.(2009), Bühlmann and van de Geer(2011)]. Although this problem is a difficult combinatorial problem, there exist some computationally feasible techniques, under stronger assumptions on the design X, for instance the thresholding procedures, the orthogonal matching pursuit, the Lasso, the Dantzig Selector etc. For more references on the techniques and the associated bounds and design assumptions, see e.g. [Donoho and Stark(1989), Tibshirani(1994), Donoho(2006), Candès et. al.(2006), Lee et. al.(2013)].

Another important problem is the one of confidence statements for the parameter θ, i.e. of quantifying the precision of an estimate of θ. Constructing confidence sets in high dimensional regression was studied e.g. in the papers [Abbasi-Yadkori et. al.(2012), Javanmard and Montanari(2013), Beran and Dümbgen(1998), Baraud(2004), Nickl and van de Geer(2013)] and it is, with the problem of estimating θ, the second fundamental problem of inference in this setting. The objective is to construct

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

a set C_n that contains θ with high probability, and also that is as small as possible, i.e. that is such that its l_2 diameter is as small as possible. One can deduce from the lower and upper bounds for the estimation problem (see [Raskutti et al(2011), Nickl and van de Geer(2013), Javanmard and Montanari(2013)]), that the optimal l_2-width of a confidence interval for the sparse vector θ should depend on its sparsity S - it should be of order $\sqrt{S\log(p)/n}$. For $\delta > 0$, if θ is S sparse and *S is known*, and if $\hat{\theta}$ is an estimator of θ that satisfies Equation (1.2), a l_2-confidence interval C_n of coverage $1-\delta$ should ideally be of the form

$$C_n = \{u \in \mathbb{R}^p : \|u - \hat{\theta}\|_2 \leq \sqrt{E\frac{S\log(p/\delta)}{n}}\}.$$

On the other hand, if the sparsity S of the parameter θ is unknown, which is the case in real applications since θ is unknown, one cannot construct directly this optimally sized confidence interval.

For the problem of estimating θ, *computationally feasible* techniques that are *adaptive to the unknown sparsity S* of θ exist (see Equation (1.2) and associated references). Do similar results hold for the problem of constructing a confidence set for θ? In particular, can one construct, in a non computationally extensive way, a confidence set for the adaptive estimate of Equation (1.2) that contains θ with high probability and whose diameter is adaptive to the unknown sparsity S of θ? This is the problem that this paper targets. We would like to insist on the importance of the construction of confidence sets. Indeed, most of the sequential learning algorithms rely on such confidence sets. For instance, in the papers [Abbasi-Yadkori et. al.(2012), Carpentier and Munos(2012), Desphandes and Montanari(2012)] that are on the topic of linear bandit in high dimensions, the authors assume that an upper bound $\bar{S}$ on the sparsity S is known, and they consider a large confidence interval for θ whose diameter depends on $\bar{S}$. The final bounds on the regret of their bandit algorithms then depend on the chosen upper bound $\bar{S}$, and not on the correct sparsity S of θ. In this setting, it would be quite useful to have an adaptive and honest confidence set for θ, that adapts to the sparsity S. The bound on the regret would then depend on S and not on $\bar{S}$.

2 Literature review

The problem of constructing a confidence set, when the support of the parameter or its sparsity S is known, is a problem that has received attention recently, see e.g.[Abbasi-Yadkori et. al.(2012), Javanmard and Montanari(2013), Javanmard and Montanari(2014), Beran and Dümbgen(1998), Baraud(2004), Lee et. al.(2014), van de Geer and Bühlmann(2014)]. The papers [Javanmard and Montanari(2013), Javanmard and Montanari(2014), Lee et. al.(2014), van de Geer and Bühlmann(2014)] are concerned with finding the limiting distribution of an estimate of θ, and using it to build tests and confidence sets for fixed, low dimensional sub models (fixed subsets of coordinates of θ). This approach does not provide an optimally sized confidence set for the parameter θ itself, since its support (i.e. the correct low dimensional model of interest) is not known. On the other hand, the problem of constructing an adaptive and honest confidence interval for θ when S is unknown, has only recently started to receive attention. It is an important problem, since there is no reason why the low dimensional support of the parameter has to be known before hand : therefore, the low dimensional model of interest cannot be chosen efficiently in a non data driven way. This problem is related to the problem of estimating the sparsity S of the parameter θ, as explained in various related settings in the papers [Hoffmann and Lepski(2002), Juditsky and Lambert-Lacroix(2003), Giné and Nickl(2010), Nickl and van de Geer(2013)]. Indeed, if a good estimate $\hat{S}$ of S is available, then one could consider the confidence interval

$$C_n = \{u \in \mathbb{R}^p : \|u - \hat{\theta}\|_2^2 \leq E\frac{\hat{S}\log(p/\delta)}{n}\}.$$

Let us first consider a simpler instance of this problem that will enlighten its difficulties. It is the case where one wants to be adaptive to two sparsities $S_0 < S_1$ (and not to any sparsity S). The objective is to construct a confidence set C_n that is *adaptive and honest*, i.e. that contains θ with high probability, and whose diameter is of order $\sqrt{S_0\log(p)/n}$ if θ is of sparsity S_0 or below, or of order $\sqrt{S_1\log(p)/n}$ if the sparsity of θ is between S_0 and S_1. In other words, the objective is to construct a set C_n based on the data such that for $\mathcal{P} = l_0(S_1)$, for $\mathcal{I} = \{S_0, S_1\}$, and for $\delta > 0$,

$$\max_{S\in\mathcal{I}} \sup_{\theta\in l_0(S)\bigcap\mathcal{P}} \mathbb{P}_\theta(\theta \in C_n) \geq 1-\delta, \text{ and}$$
$$\max_{S\in\mathcal{I}} \sup_{\theta\in l_0(S)\bigcap\mathcal{P}} \mathbb{P}_\theta\Big(|C_n|_2 \geq \sqrt{E'\frac{S\log(p/\delta)}{n}}\Big) \leq \delta, \tag{2.1}$$

where $E' > 0$ is a constant and where $|C_n|_2 = \sup\{\|u-v\|_2, (u,v) \in C_n^2\}$ is the diameter of C_n. There is however a serious obstruction to the creation of a such confidence set. It is possible to prove (see e.g. [Baraud(2004), Nickl and van de Geer(2013)]) that in many situations, there exists no adaptive and honest confidence sets on the entire parameter space $\mathcal{P} = l_0(S_1)$. The problem is that there are some parameters that are not S_0 sparse, but that are very close to S_0 sparse vectors, and for which it is impossible to detect that one needs a confidence set of diameter $\sqrt{E'S_1\log(p/\delta)/n}$ (since a confidence set of diameter $\sqrt{E'S_0\log(p/\delta)/n}$ won't provide enough coverage). A reasonable and important question is then to provide a confidence set that is adaptive and honest on the largest possible model $\mathcal{P} \subset l_0(S_1)$. Intuitively, this model $\mathcal{P}$ should be the set $l_0(S_1)$ where the problematic parameters that are not S_0 sparse, but are very close to $l_0(S_0)$ have been removed.

There has been recently some very important advances on this problem in the paper [Nickl and van de Geer(2013)]. They define the separated set $\tilde{l_0}(S_1, \rho)$ for a constant $\rho > 0$ as

$$\tilde{l_0}(S_1, \rho) = \{u \in l_0(S_1) : \|u - l_0(S_0)\|_2 \geq \rho\}, \quad (2.2)$$

where, if $A \subset \mathbb{R}^p$, we write for $u \in \mathbb{R}^p$, $\|u - A\|_2 = \inf_{v \in A} \|u - v\|_2$. They then define

$$\mathcal{P} := \mathcal{P}(\rho) := l_0(S_0) \bigcup \tilde{l_0}(S_1, \rho).$$

This new model excludes vectors that are not S_0 sparse, but at a distance that is less than ρ from S_0 sparse vectors. The smaller ρ, the more similar $\mathcal{P}$ is to $l_0(S_1)$ (equality holds when $\rho = 0$). The restriction to the model $\mathcal{P}$ can be seen as a margin condition : the $\rho-$margin condition is satisfied if the true parameter θ belongs to a sub-model where the two classes of sparsity are $\rho-$away of each other, i.e. if θ belongs either to $l_0(S_0)$, or to $\tilde{l_0}(S_1, \rho)$. This margin condition is necessary for being able to distinguish between the two sets of sparsity S_0 and S_1.

The objective is then to characterize the smallest possible ρ for which a such confidence set exists, and then to construct this confidence set. Table 1 summarizes the minimax-optimal order of $\rho := \rho_n$ (with lower and upper bounds) such that an adaptive and honest confidence set for θ exists in $\mathcal{P}(\rho_n)$, in function of S_0, S_1 (see [Nickl and van de Geer(2013)]). All upper bounds and lower bounds match in the three cases summarized in the table, and $\lim_{n\to\infty} \rho_n = 0$ at a rate which depends on S_0, S_1, which implies that $\mathcal{P}(\rho_n)$ converges to $l_0(S_1)$. The minimax-optimal rates at which ρ_n can go to 0 are known for this problem, but an important issue that remain on the existence of *computationally feasible* adaptive and honest confidence intervals in cases (ii) and (iii) of Table 1. Indeed, the procedure in [Nickl and van de Geer(2013)] consists on computing a quantity of the form $\inf_{u \in l_0(S_0)} |t_n(u)|$ where $t_n(u)$ is some quadratic functional of the data and u, then testing if this quantity exceeds a threshold, and finally using the output of this test for constructing the confidence interval. Computing this statistic is a computationally extensive problem, since its computational complexity is of order $p^{S_0}n$. It is thus not proper for concrete applications. Also, they assume that their design matrix is a random Gaussian matrix, which is quite restrictive. Another aspect that moreover prevents the use of this method in practical applications is the fact that sparsity is unlikely to hold exactly in practice, and that results involving approximate sparsity are more appealing.

In this paper, we provide a computationally feasible method for constructing an honest and adaptive confidence set on a maximal model $\mathcal{P}$ (in the minimax-optimal sense of Table 1), in a more general setting, i.e. in the setting of approximate sparsity for the parameter θ, and under more general assumptions for the design matrix X, namely that its condition number is not too large for S_1 sparse vectors. The confidence set we propose is actually trivial to implement and its computational complexity is of order $O(pn)$ (provided that an adaptive estimate satisfying Equation 1.2 is available). We first provide a method in the case of two sparsity indexes (that achieves adaptivity to two sparsities S_0, S_1), and we then extend it to the more general setting of many sparsity indexes $\mathcal{I}$. In particular, the method we propose applies to constructing a confidence set that is adaptive and honest for all the sparsities smaller than a large sparsity index $\bar{p}$. It is minimax optimal also in this setting, adaptive, and implementable in practice. We test our method with numericals simulations and all proofs are in the supplementary material [Carpentier(2015)].

3 Setting

Let n and p be two integers with $n \ll p$. Consider the model defined in Equation (1.1).

3.1 Assumption on the noise ϵ

We state the following assumption about the noise ϵ, namely that its entries are independent and sub-Gaussian.

Assumption 3.1. The entries of ϵ are independent. Moreover, $\forall i \leq n, \mathbb{E}\epsilon_i = 0, \mathbb{V}\epsilon_i = \sigma^2$ and $\forall i \leq n, \forall \lambda > 0$, there exists $c > 0$ such that

$$\mathbb{E}\exp(-\lambda\epsilon_i) \leq \exp(-\lambda^2 c^2/2).$$

For instance bounded random variables and Gaussian random variables satisfy this Assumption.

3.2 Assumption on the design matrix X

We make the following assumption about the design matrix X.

Assumption 3.2. Let $\bar{p} > 0$. The matrix X is such that there exists two constants $C_M > c_m > 0$ such that for any u that is $\bar{p}-$sparse, we have

$$c_m\|u\|_2 \leq \|\frac{1}{\sqrt{n}}Xu\|_2 \leq C_M\|u\|_2.$$

This assumption is a relaxation of the celebrated R.I.P. condition (see [Foucart and Lai(2009)] for another paper in

Values of S_0 and S_1	Case (i) $S_0 = o(n^{1/2}/\log(p))$ $S_1 = o(n^{1/2}/\log(p))$	Case (ii) $S_0 = o(n^{1/2}/\log(p))$ $S_1 = o(n/\log(p))$	Case (iii) $S_0 = o(n/\log(p))$ $S_1 = o(n/\log(p))$
UB on ρ	$\sqrt{\frac{S_1 \log(p)}{n}}$	$n^{1/4}$	0
LB on ρ	$\sqrt{\frac{S_1 \log(p)}{n}}$	$n^{1/4}$	0
Computational complexity	np	np^{S_0}	np^{S_0}

Table 1: Upper and lower bounds on the parameter ρ^2 for the problem of constructing an honest and adaptive confidence set. UP stands for Upper Bound, LB stands for Lower Bound.

this setting). This condition makes sense, when $n \ll p$, only if $\bar{p}$ is actually smaller than n, i.e. $\bar{p} = O(n/\log(p))$. In this case, for instance, random Fourier matrices, and more generally RIP matrices, satisfy this condition, with c_m and C_M close to 1. More generally, e.g. random matrices with sub-Gaussian entries whose variance-covariance matrix has bounded condition number satisfy this condition for $\bar{p} = O(n/\log(p))$ and c_m and C_M depending on the condition number of the variance-covariance matrix.

3.3 The set of vectors of approximate sparsity

We are interested in situations where θ is *approximately* S-sparse. More specifically, we focus on vectors θ that have less than S "large" components, but that can have up to $\bar{p}$ "small" components such that their l_2 norm is not too large.

Definition 3.1. We define the following sets of approximately $S-$sparse vectors, for $B, C, \bar{p}, \delta > 0$, as

$$\mathcal{S}_S(C, B, \bar{p}, \delta) = \Big\{u \in l_0(\bar{p}), \|u\|_2 \leq B : \|u - l_0(S)\|_2^2 \leq \frac{CS\log(p/\delta)}{n}\Big\},$$

where for a vector u, $\|u - l_0(S)\|_2^2 = \inf_{v \in l_0(S)} \|u - v\|_2^2 = \sum_{j=S+1}^{p} u_{(j)}^2$, where $u_{(.)}$ is the ordered version of $|u|$, i.e. is such that $|u_{(1)}| \geq |u_{(2)}| \geq ... \geq |u_{(p)}|$.

The vectors in these sets have at most S "large" components, and the $p - S$ remaining components have small l_2 norm. An important property of $\mathcal{S}_S(C, B, \bar{p}, \delta)$ is that it contains all $S-$sparse vectors whose l_2 norm is bounded by B, and is an enlargement of the set of sparse vector to "approximately" sparse vectors.

Let $0 < S_0 < S_1$ be two sparsities. Similarly to what is proposed in Equation (2.2), we define the separated set as

$$\begin{aligned}\tilde{\mathcal{S}}_{S_1}(C, B, \bar{p}, \delta, \rho) &:= \tilde{\mathcal{S}}_{S_1, S_0}(C, B, \bar{p}, \delta, \rho)\\ &= \{u \in \mathcal{S}_{S_1}(C, B, \bar{p}, \delta) : \|u - l_0(S_0)\|_2 \geq \rho\}. \qquad (3.1)\end{aligned}$$

These sets are such that, between $\mathcal{S}_{S_0}$ and $\tilde{\mathcal{S}}_{S_1}$, there is a ρ-margin condition.

For the same value of ρ, these sets are strictly larger than the sets presented in Equation (2.2) with bounded l_2 norm, which are actually the sets considered in paper [Nickl and van de Geer(2013)]. Indeed, they correspond to the vectors in the enlarged sets $\mathcal{S}_{S_1}(C, B, \bar{p}, \delta)$ that are at least ρ-away from $l_0(S_0)$.

3.4 Adaptive and honest confidence sets

We now provide the definition of adaptive and honest confidence sets in the wider model of approximately sparse vectors. It is an extension of the definition provided in Equation (2.1) to the larger set of approximately sparse vectors (it demands that the second equation in Definition (3.2) holds also for approximately sparse vectors).

Definition 3.2. Let $\delta, C, B, \bar{p} > 0$. A set C_n is a $\delta-$adaptive and honest confidence set for $\mathcal{P} \subset \mathcal{S}_{S_1}(C, B, \bar{p}, \delta)$ and for $\mathcal{I} \subset \{1, \ldots, \bar{p}\}$ if

$$\max_{S \in \mathcal{I}} \sup_{\theta \in \mathcal{S}_S(C,B,\bar{p},\delta) \bigcap \mathcal{P}} \mathbb{P}_\theta(\theta \in C_n) \geq 1 - \delta, \quad \text{and}$$

$$\max_{S \in \mathcal{I}} \sup_{\theta \in \mathcal{S}_S(C,B,\bar{p},\delta) \bigcap \mathcal{P}} \mathbb{P}_\theta\Big(|C_n|_2 \geq \sqrt{E' \frac{S\log(p/\delta)}{n}}\Big) \leq \delta,$$

where $E' > 0$ is a constant.

4 Adaptive estimation of θ on the enlarged sets

We are first going to prove that on these enlarged sets $\mathcal{S}_S(C, B, \bar{p}, \delta)$, adaptive inference remains possible, i.e. that it is possible to build an estimate of θ that satisfies results similar to what is described in Equation (1.2). More precisely, we prove that if the design is not too correlated (c_m and C_M not too far from 1 in Assumption 3.2) then the lasso estimator will provide good results on the enlarged sets.

Theorem 4.1 (Adaptive Lasso on the enlarged sets)**.** *Let Assumptions 3.1, and 3.2 be satisfied for $c, c_m, c_M, 66\bar{p}$ such that $c > 0, c_m > 2/3, C_M < 4/3$ and $\bar{p} > 0$. Let $B > 0$ and $C > 0$. Let $\delta > 0$. The solution $\hat{\theta}$ of l_1 minimization or lasso*

$$\hat{\theta} = \arg\min_u \Big[\|Y - Xu\|_2^2 + \kappa\sqrt{\log(p/\delta)n}\|u\|_1\Big],$$

where $\kappa > 4\max(c, \sqrt{C}/3, c^2, C/9)$ *is such that we have* $\forall 0 < S \leq \bar{p}$

$$\sup_{\theta \in \mathcal{S}_S(C,B,\bar{p},\delta)} \mathbb{P}_\theta\Big(\|\hat{\theta} - \theta\|_2^2 \geq \Big(12(36\kappa + 36)^2 + C^2\Big)\frac{S\log(p/\delta)}{n}\Big) \leq \delta.$$

The proof of this theorem is in the Supplementary Material [Carpentier(2015)] (see Appendix B). Proving this bound for the lasso on the enlarged sets is actually very similar to proving it on the set of exactly sparse vectors. An important remark is that the lasso's computational complexity is not high and is well defined, see [Tibshirani(1994)]. As usual, the lasso does not work on too correlated designs, it can be applied when $c_m > 2/3$ and $C_M < 4/3$. When this is not satisfied, other techniques have to be considered, see e.g. [Foucart and Lai(2009)]. It is actually possible to prove that for any $0 < c_m < C_M$, there exists an estimate that satisfies a result similar to the one in Theorem 4.1, see Theorem D.1 in the Supplementary Material [Carpentier(2015)] Appendix D. This estimate is however the result of l_0 minimization, and is thus computationally extensive (the computational complexity is np^{S_0}), and is in practice not implementable whenever p, S_0 are too large.

The really nice feature of such a result is that it provides an estimate whose l_2-risk is adaptive *uniformly* to the sparsity of *any* vector of the enlarged sparsity class, for any sparsity smaller than $\bar{p}$. The estimate is data driven, but it needs an upper bound C on the amount of which θ deviates from the sparsity S, and also it needs an upper bound c on the parameter that bounds the sub-Gaussian tail of the noise.

5 Adaptive and honest confidence sets for θ

We now propose a method that is computationally feasible for constructing an adaptive and honest confidence set for θ. We fist present this method in the case of adaptivity to only two sparsities $S_0 < S_1$ (two sparsity indexes method), and then explain how to extend these results to larger sets of sparsities (multi sparsity indexes method).

5.1 Presentation of the confidence set for two sparsity indexes $S_0 < S_1$

Construction of the confidence set Let $S_0 < S_1 < \bar{p}$. The algorithm for the two sparsity indexes, Algorithm 1, contains two main steps. The first step is to construct a test Ψ_n for deciding whether θ is S_0 approximately sparse, or S_1 approximately sparse. The test consists in first computing an adaptive estimate $\hat{\theta}$ of θ, and then on thresholding all non-significant components. Then, the testing decision is based on two factors (i) testing whether the number of non-zero entries of the thresholded estimate is larger than S_0 and (ii) testing whether the squared residuals $\|\hat{r}\|_2^2$ are significant or not. If both these quantities are small enough, the test is accepted, otherwise it is rejected. The outcome of this is the test Ψ_n. The second step is to use this information to construct the confidence set C_n. Based on Ψ_n, the confidence interval C_n will be of diameter of order $\sqrt{\frac{S_0\log(p)}{n}}$ (if $\Psi_n = 0$), or $\sqrt{\frac{S_1\log(p)}{n}}$ (if $\Psi_n = 1$). The procedure is explained in Algorithm 1.

Algorithm 1 Two sparsity indexes confidence set

Parameters: $\delta, S_0, S_1, \sigma^2$

set the following quantities, computed on the first half of the samples only, as $\hat{B}^2 := 3/2\Big(n^{-1}\sum_{i\leq n} Y_i^2(1 + 2\log(1/\delta)) + 2\log(1/\delta)\Big)$, and $\tau_n := 14|\hat{B}|\sqrt{n^{-1/2}\log(1/\delta)} + 381|\hat{B}|\sqrt{\frac{S_0\log(p/\delta)}{n}}$, and $\tau_n' := 330|\hat{B}|\sqrt{\frac{S_1\log(p/\delta)}{n}}$, and let $\hat{\theta}$ be the lasso estimate as in Theorem 4.1. All these quantities are computed on the first half of the sample, and from now on we only use the second half of the sample.

set the residual $\hat{r} = Y - X\hat{\theta}$

set the test statistic $R_n = \|\hat{r}\|_2^2 - n\sigma^2$

set the test $\Psi_n = 1 - \mathbf{1}\{R_n \leq \tau_n^2\}\mathbf{1}\{\sum_{j=S_0+1}^p \hat{\theta}_{(j)}^2 \leq (\tau_n')^2\}$, where $\hat{\theta}_{(.)}$ is the ordered version of $|\hat{\theta}|$

set the confidence interval

$$C_n := \Big\{u \in \mathcal{S}_{S_1} : \|u - \hat{\theta}\|_2 \leq 650\sqrt{\frac{S_0\log(p)}{n}}\mathbf{1}\{\Psi_n = 0\} + 650\sqrt{\frac{S_1\log(p)}{n}}\mathbf{1}\{\Psi_n = 1\}\Big\}.$$

return C_n

The parameter σ^2 car be replaced by a consistent estimator of the variance of the noise ϵ (e.g. a Bootstrap estimate, a cross validation estimate, etc).

Main result The following theorem states that this confidence interval is adaptive and honest (in the sense of Definition (3.2)) over a large model $\mathcal{P}$.

Theorem 5.1. *Assume that the noise is either Gaussian of variance less than* 1*, or bounded by* 1*, and assume that the assumptions for convergence of the adaptive lasso, stated in Theorem 4.1, are satisfied, and that* $S_1 \leq \bar{p}$. *Then the confidence set presented in Algorithm 1 is* δ *adaptive and honest for* $\mathcal{I} = \{S_0, S_1\}$ *and over the model*

$$\mathcal{P}(\rho_n) = \mathcal{S}_{S_0}(32, \infty, \bar{p}, \delta)\bigcup \tilde{\mathcal{S}}_{S_1}(32, \infty, \bar{p}, \delta, \rho_n),$$

where

$$\rho_n = |\hat{B}|\min\Big(54\sqrt{\log(1/\delta)}n^{-1/4}, 460\sqrt{\frac{S_1\log(p/\delta)}{n}}\Big).$$

By definition of the enlarged set, this implies in particular that the confidence set is δ *adaptive and honest for* $\mathcal{I} =$

$\{S_0, S_1\}$ *and over the model*

$$\mathcal{P}(\rho_n) = l_0(S_0) \bigcup \tilde{l}_0(S_1, \tilde{\rho}_n), \quad \text{with} \quad \tilde{\rho}_n = 2\rho_n.$$

This theorem is actually a corollary of a more general result, presented in the Supplementary Material [Carpentier(2015)], Appendix A. The proof of this theorem is in the Supplementary Material [Carpentier(2015)] (Appendix C).

The confidence set is adaptive and honest under the same assumptions that ensure consistency of the lasso estimate. It is quite reasonable that it is so, since the creation of adaptive and honest confidence sets is a strictly more difficult problem than the problem of estimating the parameter (indeed, any point of the confidence set is a good estimate of the parameter). Also, since $\lim_{n\to\infty} \rho_n = 0$, for any θ and for n large enough, the confidence set contains θ with high probability, and its diameter adapts to the sparsity of θ. It is adaptive to the two sparsities $\{S_0, S_1\}$ only and not to the whole spectrum of sparsities, but it already allows to improve many existing learning algorithms by diminishing the size of the confidence interval (by not always setting it to $\sqrt{S_1 \log(p)/n}$ independently of θ). Moreover, it computational complexity is of order np, which is linear.

Comparison with results in paper [Nickl and van de Geer(2013)]. Our results imply all the upper bounds of [Nickl and van de Geer(2013)] in all cases (i), (ii) and (iii) of Table 1, i.e. imply the upper bounds on exactly sparse sets (this is illustrated in Theorem 5.1). Also, our confidence set is adaptive and honest in all three cases (i), (ii) and (iii), and we do not need to change the construction method as in paper [Nickl and van de Geer(2013)]. Our assumptions on the design of X are weaker than in the paper [Nickl and van de Geer(2013)], where the authors consider Gaussian design which a fortiori satisfies Assumption 3.2 with high probability. Finally, the confidence set is, as we saw, computationally feasible, since its computational complexity is of order np. As mentioned in the introduction, the procedure in the paper [Nickl and van de Geer(2013)] boils down to minimizing over the set of S_0- sparse vectors a quadratic quantity, which has complexity of order $p^{S_0}n$. This implies that our procedure is computationally efficient on a set that is as large as possible in a minimax sense, as illustrated by the lower bounds in Figure 1.

5.2 Adaptive and honest confidence sets for multiple sparsities

Construction of the confidence set In the last subsection, we restricted ourselves to constructing a confidence set that is adaptive to only two sparsities S_0, S_1. Although it is already useful with respect to existing techniques that are not adaptive at all, it is only a first step toward a more global result where all sparsity indexes $\mathcal{I} = \{1, \dots, \bar{p}\}$ are considered. Algorithm 2 solves this problem.

Algorithm 2 Multi sparsity indexes confidence set, second version

Parameters: δ, σ^2

set using only the first half of the $2n$ samples,

$$\hat{B}^2 := 3/2\Big(n^{-1}\sum_{i\le n} Y_i^2(1 + 2\log(1/\delta)) + 2\log(1/\delta)\Big),$$

and $\tau_n(S) := 14|\hat{B}|\sqrt{n^{-1/2}\log(1/\delta)} + 381|\hat{B}|\sqrt{\frac{S\log(p/\delta)}{n}}$, and $\tau'_n(S) := 330|\hat{B}|\sqrt{\frac{(S+1)\log(p/\delta)}{n}}$, and let $\hat{\theta}$ be the lasso estimate as in Theorem 4.1.

set the residual $\hat{r} = Y - X\hat{\theta}$

set the statistic $R_n = \|\hat{r}\|_2^2 - \sigma^2$

set for any $S \le p$ the statistic $R'_n(S) := \sum_{j=S+1}^{p} \hat{\theta}^2_{(j)}$, where $\hat{\theta}_{(.)}$ is the ordered version of $|\hat{\theta}|$.

set $\hat{S}$ as the smallest $S \le p$ such that $R_n \le \tau_n(S)^2$, and, $R'_n(S) \le (\tau'_n(S))^2$.

set the confidence interval

$$C_n := \{u \in \mathbb{R}^p : \|u - \hat{\theta}\|_2 \le 650\sqrt{\frac{\hat{S}\log(p/\delta)}{n}}.$$

return C_n

The parameter σ^2 car be replaced by a consistent estimator of the variance of the noise ϵ (e.g. a Bootstrap estimate, a cross validation estimate, etc).

The following theorem holds in this case (it is a direct consequence of Theorem 5.1).

Theorem 5.2. *Assume that the noise is either Gaussian of variance less than* 1*, or bounded by* 1*, and assume that the assumptions for convergence of the adaptive lasso, stated in Theorem 4.1, are satisfied, and that for $S_1 \le \bar{p}$. Then the confidence set presented in Algorithm 2 is δ adaptive and honest for $\mathcal{I} = \{1, \dots, \bar{p}\}$ and over the model*

$$\mathcal{P} := \mathcal{S}_1(32, \infty, \bar{p}, \delta) \bigcup \bigcup_{S=2}^{\bar{p}} \tilde{\mathcal{S}}_{S,S-1}(32, \infty, \bar{p}, \delta, \rho_n(S)).$$

where

$$\tilde{\rho}_n(S) = |\hat{B}| \min\Big(50\sqrt{\log(1/\delta)}n^{-1/4}, 460\sqrt{\frac{(S+1)\log(p/\delta)}{n}}\Big).$$

A more general procedure is presented in the Supplementary Material [Carpentier(2015)], Appendix A.

Discussion This result is minimax optimal from the lower bounds in Figure 1, and it is also computationally feasible. The resulting confidence interval is adaptive and honest for all indexes $\mathcal{I}$ over $\mathcal{P}$. Moreover, $\mathcal{P}$ is significantly larger than the set of "detectable" parameters such that all non-zero component are larger than $\sqrt{\log(p/\delta)/n}$. For this reason, this method is more efficient than the naive method of counting the number of non-zero entries in a thresholded adaptive estimate, and using this number for constructing the confidence set.

6 Experimental results

In this section, we present some simulations and also some applications on images.

6.1 Simulations

In order to illustrate the efficiency of our method, we apply it to simulated data. We consider a problem in dimension $p = 10000$, and where $n = 1000$ (the sampling rate is 10%). Let $0 < S_0 < S_1$ be the two approximate sparsity levels. We are going to define three types of distributions (priors) on the set of parameter θ:

- $\theta \sim \Theta_1$: (i) S_0 random coordinates of θ are $\mathcal{N}(0,1)$ and (ii) the remaining coordinates are $\mathcal{N}(0,\sigma_0^2)$ where $\sigma_0^2 = \frac{S_0\log(p)}{np}$. With high probability, $\theta \in \mathcal{S}_{S_0}(C,B,\bar{p},\delta)$.
- $\theta \sim \Theta_2$: (i) S_1 random coordinates of θ are $\mathcal{N}(0,1)$ and (ii) the remaining coordinates are $\mathcal{N}(0,\sigma_0^2)$. With high probability, $\theta \in \tilde{\mathcal{S}}_{S_1}(C,B,\bar{p},\delta,\rho_n^2)$ where $\rho_n^2 = O(\frac{S_1\log(p)}{n} + \frac{S_0\log(p)}{n})$.
- $\theta \sim \Theta_3$: (i) S_0 random coordinates of θ are $\mathcal{N}(0,1)$ and (ii) the remaining coordinates are $\mathcal{N}(0,\sigma_1^2)$ with $\sigma_1^2 = C\left(\frac{1}{n^{1/2}p} + \frac{S_0\log(p)}{np}\right)$. With high probability, $\theta \in \tilde{\mathcal{S}_{S_1}}(C,B,\bar{p},\delta,\rho_n^2)$, where $\rho_n^2 = O(n^{-1/2} + \frac{S_0\log(p)}{n})$.

See Figure 1 for an illustration of this.

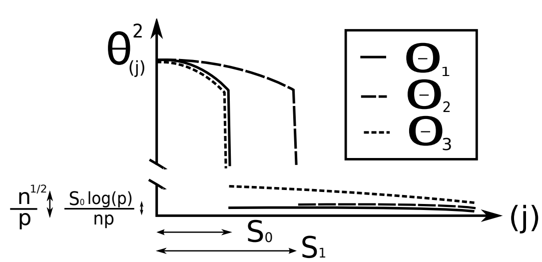

Figure 1: Mean of the square of the re-ordered coordinates $\theta^2_{(j)}$ of θ sampled according to priors Θ_1, Θ_2, and Θ_3.

The distributions correspond to two extremal cases in which the sampled vector θ is not approximately sparse, i.e. either the norm of the tail coefficients is large, or the number of detectable coefficients is larger than S_0. For a given $\theta \sim \Theta_k$ ($k \in \{1,2,3\}$), we write Ψ for its "class", i.e. if the hypothesis in test (C.1) to which they belong with high probability. This means that $\Psi = 0$ for $\theta \sim \Theta_1$ and $\Psi = 1$ for $\theta \sim \Theta_2$ or $\theta \sim \Theta_3$.

For all distributions Θ_k, we do 10000 experiments (corresponding to trying our methods over 10000 samples of θ sampled according to Θ_k) where we perform the test we described in Section 5 for the testing problem (C.1), infer the sparsity, compute adaptive confidence interval, and compute the risk of the estimate. The sampling matrix X is composed of Gaussian random variables, and the noise ϵ is i.i.d. Gaussian with variance 1. In this design, $\bar{p} = O(n/\log(p))$. We compute an estimate of θ via hard thresholding, which happens in this orthogonal setting to be equivalent to lasso, on the first half of the samples. We then construct the test on the second half of the sample. We summarise the results in Table 2.

A first general remark is that the test we consider manages to distinguish efficiently between H_0 and H_1, for many different configuration of sparsity. The adaptive and honest confidence sets we built using this test are also quite efficient. The probability that the true parameter belongs to the adaptive confidence set is very close to the probability of correctly inferring the class of θ. The strength of these sets is to be adaptive to the sparsity of the problem, i.e. they contain θ with high probability and do not have the same width depending on the complexity of the true parameter (S_0 or S_1). As a matter of fact, the width of the adaptive confidence set is close to the value of the risk of the adaptive estimate, which is exactly what is wanted. It is particularly interesting, since as expected, the risk is much larger under H_1 than under H_0.

In the case of distribution Θ_3, it is interesting to remark that although the sparsity of $\hat{\theta}$ is close to S_0 in expectation, it does not prevent our test to efficiently classify it as H_1. This is actually quite important in terms of confidence sets since we can observe that, for each configuration of sparsity, the risk and thus the width of the adaptive confidence interval, is much larger for Θ_3 than for Θ_1. A test only based on the inferred sparsity (i.e. on $\|\hat{\theta}\|_0$) would not have detected this since the inferred sparsity is approximately similar in these two cases.

6.2 Application of the method to images

We consider now a more concrete setting, where we apply our method to images. We focus on black and white drawings[1]. The particularity of such images is that only a few

[1] Note that a pre-treatment, like differentiation, can be applied to regular images to transform them into drawing-like images.

Prior	$S_1 = 5$	and	$S_2 = 10$	$S_1 = 5$	and	$S_2 = 10^3$	$S_1 = 50$	and	$S_2 = 10^3$
	Θ_1	Θ_2	Θ_3	Θ_1	Θ_2	Θ_3	Θ_1	Θ_2	Θ_3
$\mathbb{P}_\theta(\Psi_n \neq \Psi)$	$5\ 10^{-2}$	$1\ 10^{-2}$	X	$1\ 10^{-2}$	$8\ 10^{-7}$	$8\ 10^{-2}$	$1\ 10^{-3}$	$2\ 10^{-5}$	$6\ 10^{-3}$
$\mathbb{E}\|\hat{\theta}\|_0$	4.8	9.7	X	4.8	$2.1\ 10^3$	4.8	48.7	$2.3\ 10^3$	49.3
$\mathbb{P}_\theta(\theta \notin C_n)$	$1.2\ 10^{-1}$	$9\ 10^{-2}$	X	$5\ 10^{-2}$	9.10^{-7}	$8.1\ 10^{-2}$	$3.2\ 10^{-3}$	$4\ 10^{-5}$	$7.8\ 10^{-3}$
$\mathbb{E}\|C_n\|$	$6.16\ 10^{-2}$	$1.8\ 10^{-1}$	X	$7.76\ 10^{-2}$	$9.8\ 10^3$	$5.7\ 10^{-1}$	27.6	$9.9\ 10^3$	$9.7\ 10^3$
$\mathbb{E}\|\hat{\theta} - \theta\|_2^2$	$3.37\ 10^{-2}$	$1.1\ 10^{-1}$	X	$3.7\ 10^{-2}$	$9.6\ 10^3$	$3.2\ 10^{-1}$	15.04	$9.7\ 10^3$	$9.4\ 10^3$

Table 2: Expected results for the test, risk and adaptive and honest confidence set for the three priors.

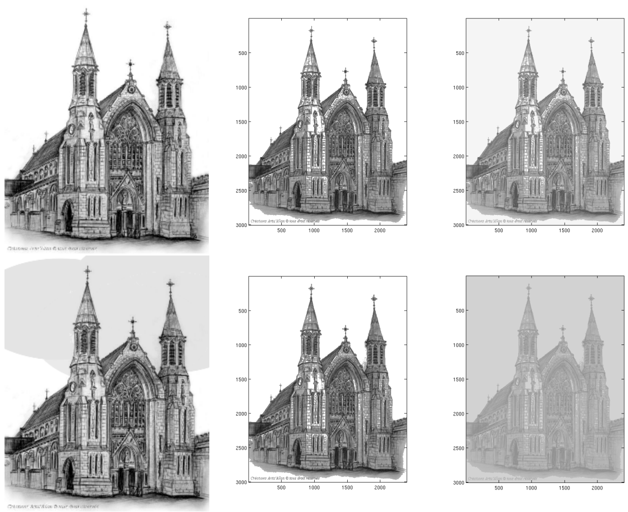

Figure 2: First column = original image. Second column= reconstructed image. Third column: extremal point of the confidence set that minimises the contrast. The test of $3\%p$ sparsity is accepted for the first image but rejected for the other.

pixels are non-zero, i.e. if we align the columns of such an image, it can be considered as a sparse vector θ of dimension p where p is the number of lines times the number of columns of the image. Such an image θ can easily be compressed by conserving n Fourier coefficient of this vector, chosen at random frequencies. We write X the $n \times p$ matrix that represents this convolution. The model is then again $Y = X\theta + \epsilon$, where ϵ is some noise to the compression, due for instance to transmission. An interesting question when observing the compression Y of an image is to infer what is the quality of the reconstruction, i.e. if the image is indeed quite sparse or not.

When one then observes such a compression, it is possible to reconstruct the image θ by, as we explained in Section 4. Also, the test and the confidence sets are build in the same way as what was done in Section 5.

We consider here an image with $p = 7200000$ pixels. We consider $n = 5\%p = 36000$ Fourier coefficients obtained by FFT (Fast Fourier Transform). We consider 2 images that we display in Figure 2 (first column), and write $\theta^{(k)}$ where $k \in \{1, 2\}$. The first image is a black and white drawing of a cathedral, and the second one is the same drawing but with a background (a cloud): the first one will be approximately sparse and the second one not. It is possible to compress these images by considering, instead of $\theta^{(k)}$, the vector $Y^{(k)}$. The quality of the compression, i.e. the proximity between the image reconstructed trough $Y^{(k)}$ and the image $\theta^{(k)}$, will depend however very much on the sparsity of the image, as we saw in Theorem 4.1. We can use the results of Section 5 to test whether the image is a least $3\%p$ approximately sparse or not, and then build confidence sets around it. Figure 2 (two last columns, the first one containing the reconstructed images and the second one an extremal point of the confidence sets) illustrates this. More precisely, we display, for each image, the estimate of $\theta^{(k)}$ (i.e. the reconstructed image), and an extreme points of the confidence sets that we choose as being the image that minimises the contrast.

Although image 1 and 2 are different images, their estimates are very close. The test reveals the fact that they are not the same, and that in particular the reconstruction of image 1 will be good while the reconstruction of image 2 will be bad (although they seem similar from their reconstruction). The confidence sets also show how much the true image could actually be different from the reconstructed image. In particular, the extremal point of the confidence set that minimises the contrast implies that although it is rather unlikely that there is a background in image 1, image 2 might well have one. For these images, the notion of approximate sparsity is very important since even the first image is not at all sparse (not even $40\%p$ sparse). It is however less than $3\%p$ approximately sparse. Because of the background, however, the second image is not even close to $3\%p$ approximately sparse.

Conclusion In this paper, we developed a computationally feasible, adaptive and honest confidence interval, first in the two sparsity indexes case, and then in the general setting of multi sparsity indexes. The method we propose is efficient on a maximal set (in a minimax-optimal sense), and is implementable, which is a novelty with respect to the existing results. The assumptions we make are also less restrictive than what was previously required.We also provided an experimental validation of this results by simulations, and also an application on images.

References

[Abbasi-Yadkori et. al.(2012)] Y. Abbasi-Yadkori, D. Pal, C. Szepesvari. Online-to-Confidence-Set Conversions and Application to Sparse Stochastic Bandits. *Artificial Intelligence and Statistics*, 2012.

[Baraud(2004)] Y. Baraud. Confidence balls in gaussian regression. *Annals of statistics*, 528–551, 2004.

[Beran and Dümbgen(1998)] R. Beran and L. Dümbgen. Modulation of estimators and confidence sets. *Annals of Statistics*, 1826–1856,1998.

[Bickel et. al.(2009)] P. Bickel, Y. Ritov, A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[Bühlmann and van de Geer(2011)] P. Bühlmann and S.A. van de Geer. Statistics for high-dimensional data: : Methods, Theory and Applications. *Springer*, 2011.

[Candès et. al.(2006)] E.J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[Candès and Tao(2007)] E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

[Candès(2008)] E.J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9-10):589–592, 2008.

[Carpentier and Munos(2012)] A. Carpentier and R. Munos. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. *Artificial Intelligence and Statistics*, 2012.

[Carpentier(2015)] A. Carpentier. Implementable confidence sets in high dimensional regression (Supplementary Material). *arXiv preprint*, arXiv:1501.04467, 2015.

[Desphandes and Montanari(2012)] Y. Desphandes, and A. Montanari. Linear Bandits in High Dimension and Recommendation Systems. *arXiv preprint*, 2013.

[Donoho and Stark(1989)] D.L. Donoho and P.B. Stark. Uncertainty principles and signal recovery. *SIAM Journal on Applied Mathematics*, 49(3):906–931, 1989.

[Donoho(2006)] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[Foucart and Lai(2009)] S. Foucart and M.J. Lai. Sparsest solutions of underdetermined linear systems via lq-minimization for $0 < q < p$. *Applied and Computational Harmonic Analysis*, 26(3):395–407, 2009.

[Giné and Nickl(2010)] E. Giné and R. Nickl. Confidence bands in density estimation. *Annals of Statistics*, 38(2):1122–1170, 2010.

[Hoffmann and Lepski(2002)] M. Hoffmann and O. Lepski. Random rates in anisotropic regression. *Annals of statistics*,38(2):325–358, 2002.

[Juditsky and Lambert-Lacroix(2003)] A. Juditsky and S. Lambert-Lacroix. Nonparametric confidence set estimation. *Mathematical Methods of Statistics*, 12(4): 410–428, 2003.

[Javanmard and Montanari(2013)] A. Javanmard, and A. Montanari. Confidence Intervals and Hypothesis Testing for High-Dimensional statistical model. *Advances in Neural Information Processing System*, 1187–1195, 2013.

[Javanmard and Montanari(2014)] A. Javanmard, and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning research*, to appear, 2014.

[Kavukcuoglu et. al.(2009)] K. Kavukcuoglu, M. Ranzato, R. Fergus and Y. Le-Cun. Learning invariant features through topographic filter maps. *IEEE Conference on Computer Vision and Pattern Recognition.*, 1605–1612, 2009.

[Koltchinskii(2009)] V. Koltchinskii. The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828, 2009.

[Lee et. al.(2013)] Lee, Jason D and Sun, Dennis L and Sun, Yuekai and Taylor, Jonathan E Exact inference after model selection via the Lasso. arXiv preprint arXiv:1311.6238, 2013.

[Lee et. al.(2014)] Lee, Jason D and Taylor, Jonathan E Exact Post Model Selection Inference for Marginal Screening. arXiv preprint arXiv:1402.5596, 2014.

[Moriaka and Satoh(2010)] N. Morioka and S. Satoh. Generalized Lasso based Approximation of Sparse Coding for Visual Recognition. *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor and R.S. Zemel and P. Bartlett and F.C.N. Pereira and K.Q. Weinberger, 181–189, 2011.

[Nickl and van de Geer(2013)] R. Nickl and S. van de Geer. Confidence Sets in Sparse Regression. *Annals of Statistics*, 41(6):2852–2876, 2013.

[Raskutti et al(2011)] Raskutti, Garvesh and Wainwright, Martin J and Yu, Bin. Minimax Rates of Estimation for High-Dimensional Linear Regression Over l_q balls. *Information Theory, IEEE Transactions on*, vol 57, number 10 pages 6979-6994, 2011.

[Starck et. al.(2010)] J. -L. Starck, F. Murtagh and J. Fadili. Sparse image and signal processing: wavelets, curvelets, morphological diversity. *Cambridge University Press*, 2010.

[Tibshirani(1994)] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[van de Geer and Bühlmann(2014)] S.A. van de Geer and P. Bühlmann . On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 2014.

[Zou(2006)] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101 476:1418–1429, 2006.