Journal of Multivariate Analysis 146 (2016) 248–260

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Worst possible sub-directions in high-dimensional models

Sara van de Geer

Seminar for Statistics, ETH Zürich, Switzerland

ARTICLE INFO

Article history:
Received 14 December 2014
Available online 3 October 2015

AMS 2000 subject classifications:
62J07
62H12

Keywords:
De-sparsifying
Graphical Lasso
Irrepresentable condition
Lasso
Oracle rates
Sub-direction

ABSTRACT

We examine the rate of convergence of the Lasso estimator of lower dimensional components of the high-dimensional parameter. Under bounds on the ℓ_1-norm on the worst possible sub-direction these rates are of order $\sqrt{|J|\log p/n}$ where p is the total number of parameters, n is the number of observations and $J \subset \{1,\ldots,p\}$ represents a subset of the parameters. We also derive rates in sup-norm in terms of the rate of convergence in ℓ_1-norm. The irrepresentable condition on a set J requires that the ℓ_1-norm of the worst possible sub-direction is sufficiently smaller than one. In that case sharp oracle results can be obtained. Moreover, if the coefficients in J are small enough the Lasso will put these coefficients to zero. By de-sparsifying one obtains fast rates in supremum norm without conditions on the worst possible sub-direction. The results are extended to M-estimation with ℓ_1-penalty for generalized linear models and exponential families. For the graphical Lasso this leads to an extension of known results to the case where the precision matrix is only approximately sparse. The bounds we provide are non-asymptotic but we also present asymptotic formulations for ease of interpretation.

1. Introduction

We consider estimation bounds for parameters of interest in high-dimensional models. We apply the M-estimation procedure with ℓ_1-penalty. Let p be the number of parameters and n the number of observations. We show that under certain conditions, for a subset $J \subset \{1,\ldots,p\}$, the group of parameters with index in J can be estimated with rate $\sqrt{|J|\log p/n}$ in ℓ_2-norm. For this to happen, the "worst possible sub-direction" is required to have a bounded ℓ_1-norm. If this ℓ_1-norm is less than one, we obtain oracle rates $\sqrt{|J\cap S|\log p/n}$ where S is the set of active parameters or a sparse approximation thereof. Taking J to be the complement of the set S gives variable selection results under an irrepresentable condition.

By de-sparsifying one obtains fast rates (of order $1/\sqrt{n}$ under certain conditions) for one-dimensional parameters without conditions on the ℓ_1-norm of the worst possible sub-direction. The de-sparsified estimator can moreover be used for the construction of asymptotic confidence intervals for parameters of interest. This study is an intermediate step towards this end. We investigate the rates and conditions for remainder terms to be negligible. Global convergence (e.g. in ℓ_1-norm) is generally sufficient for the latter. However, in high-dimensional models which are not very sparse, global convergence does not happen. For example, when estimating a $p\times p$ precision matrix (where there are actually $p(p-1)/2$ parameters), the global rate in ℓ_1-norm will not be faster than $p\sqrt{\log p/n}$. To handle such cases, we show that the irrepresentable condition on the set of small coefficients yield rates in the sup-norm.

The high-dimensional models studied in this paper are close relatives of the infinite-dimensional models occurring in functional data analysis. We refer for the latter to the books Silverman and Ramsay [24], Ferraty and Vieu [12], Horváth and Kokoszka [13] and Bongiorno et al. [9] and to Cuevas [11] for a recent overview. Functional regression can sometimes

E-mail address: geer@stat.math.ethz.ch.

http://dx.doi.org/10.1016/j.jmva.2015.09.018
0047-259X/

be discretized making the model formally equivalent to a high-dimensional one. However, functional regression often has a different structure for the coefficients i.e., the structure is not directly described by sparsity. We refer to Kneip and Sarda [17], Aneiros and Vieu [1,2] for important results on variable selection in functional regression. The paper Koltchinskii and Minsker [19] on the other hand studies the functional regression model with sparsity. It will be an interesting topic for future research to combine these various results and further unify functional and high-dimensional theory.

1.1. Related work

The literature on a semi-parametric approach to confidence intervals and testing in high dimensions is expanding quickly. An important reference is Zhang and Zhang [35] and further work can be found in [15,16,32] and the papers Belloni et al. [6,3,4]. Our work presents rates in sup-norm for Lasso estimators and is in that aspect related to Lounici [20], Ye and Zhang [34] although our conditions are based on worst possible sub-directions instead of incoherence or invertibility factors. Also related is Wainwright [33] but our work does not rely on irrepresentable conditions, i.e., the ℓ_1-norm of the worst possible sub-direction is allowed to be larger than one (but if it is smaller we reproduce variable selection results). Irrepresentable conditions for variable selection were introduced in [21,36]. Our formulation shows these are conditions on worst possible sub-directions. We moreover extend the situation to models which are only approximately sparse.

1.2. Organization of the paper

The paper is organized as follows. In Section 2 we consider the linear model with fixed design and the Lasso estimator. The section stays close to results in literature. Sections 2.1 and 2.2 contain results for single variables (and their implications for sup-norms). We consider de-sparsifying the Lasso in Section 2.3, and discuss thresholding yielding a re-sparsified estimator.

Section 3 generalizes the results for single variables to results for groups. Sharp oracle inequalities as well as variable selection results are derived. This leads to a further refinement in Section 3.1 where we prove that under certain irrepresentable conditions the Lasso will estimate small coefficients as being zero. In Section 3.2, we present results for a de-sparsified estimator of a group of variables.

In the remainder of the paper worst possible sub-directions are in terms of theoretical (unknown) quantities, which mean they do not immediately lead to a de-sparsifying procedure. We remark that de-sparsifying is nevertheless possible (see also [32,14]) but a full discussion goes beyond the scope of this paper. Section 4 studies general loss functions. In Section 5 we discuss the remainder term, for the linear model with random design (Section 5.1), the generalized linear model (Section 5.2) and exponential families (Section 5.3). Then we move to Brouwer's fixed point theorem for deriving rates for estimators defined as solution of a system of equations. This theorem provides a way to handle the situation where the global rate is not fast enough to deal with the remainder term. We apply this in Section 7 to derive rates in sup-norm from the KKT conditions. Finally, we examine in Section 8 the remainder term of the graphical Lasso as an example. The approach there is as in [23] but with the extension to models which are only approximately sparse. All proofs are in the supplement van de Geer [31].

The results in this paper are presented in a non-asymptotic form. To simplify their interpretation, we present a separate asymptotic formulation at various stages, where we assume "standard" asymptotic scenarios.

2. The linear model with fixed design: results for single variables

Let Y be an n-vector of response variables and X a fixed $n \times p$ design matrix and consider the model

$$Y = X\beta^0 + \epsilon, \tag{1}$$

where ϵ is unobservable noise and β^0 is a p-vector of unknown coefficients. We assume throughout that $\mathbb{E}\epsilon = 0$ and $\mathbb{E}\epsilon\epsilon^T = \sigma_\epsilon^2 I$, where in asymptotic formulations $\sigma_\epsilon = \mathcal{O}(1)$. For a vector $v \in \mathbb{R}^n$ we write (with some abuse of notation) $\|v\|_n^2 := v^T v/n$. The Lasso estimator [25] is

$$\hat\beta := \hat\beta(\lambda) := \min_{\beta\in\mathbb{R}^p}\left\{\|Y - X\beta\|_n^2 + 2\lambda\|\beta\|_1\right\}. \tag{2}$$

Here, λ is a tuning parameter which may be chosen data-dependent (e.g. when using the square root Lasso introduced in [5]). Typically, λ is chosen of order $\sqrt{\log p/n}$ and proportional to some estimate of the noise level $\sigma_\epsilon := (\mathbb{E}\|\epsilon\|_n^2)^{1/2}$.

The estimator $\hat\beta$ satisfies the Karush–Kuhn–Tucker or KKT conditions

$$-X^T(Y - X\hat\beta)/n + \lambda\hat z = 0 \tag{3}$$

where $\hat z_j = \text{sign}(\hat\beta_j)$ if $\hat\beta_j \neq 0$ and $\|\hat z\|_\infty \leq 1$. Thus, $\hat\beta^T\hat z = \|\hat\beta\|_1$ and

$$Y^T(Y - X\hat\beta)/n = \|Y - X\hat\beta\|_n^2 + \lambda\|\hat\beta\|_1. \tag{4}$$

These equalities will play a key role in our proofs.