Contents lists available at ScienceDirect

Journal of Econometrics

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

Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative Lasso

Mehmet Caner [a], Anders Bredahl Kock [b,c,*]

[a] Department of Economics, Translational Data Analytics, Department of Statistics, Ohio State University, 452 Arps Hall, OH 43210, USA
[b] University of Oxford, Department of Economics, Manor Road, Oxford, OX1, 3UQ, UK
[c] Aarhus University, Department of Economics and Business Economics, CREATES, Denmark

ARTICLE INFO

Article history:
Received 9 November 2015
Received in revised form 10 November 2017
Accepted 11 November 2017
Available online xxxx

JEL classification:
C12
C13
C21

Keywords:
Conservative Lasso
Honest inference
High-dimensional data
Uniform inference
Confidence intervals
Tests

ABSTRACT

In this paper we consider the conservative Lasso which we argue penalizes more correctly than the Lasso and show how it may be desparsified in the sense of van de Geer et al. (2014) in order to construct asymptotically honest (uniform) confidence bands. In particular, we develop an oracle inequality for the conservative Lasso only assuming the existence of a certain number of moments. This is done by means of the Marcinkiewicz–Zygmund inequality. We allow for heteroskedastic non-subgaussian error terms and covariates. Next, we desparsify the conservative Lasso estimator and derive the asymptotic distribution of tests involving an increasing number of parameters. Our simulations reveal that the desparsified conservative Lasso estimates the parameters more precisely than the desparsified Lasso, has better size properties and produces confidence bands with superior coverage rates.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

In recent years we have seen a burgeoning literature on high-dimensional problems where the number of parameters is much greater than the sample size. Statistical inference in the sense of constructing tests and confidence bands in the high-dimensional linear regression model were considered in a seminal series of papers by Belloni et al. (2010, 2012, 2011b, 2014, 2011a). These authors showed how a cleverly constructed (double) post selection estimator can be used to construct uniformly valid confidence intervals for the parameter of interest in instrumental variable and treatment effect models allowing for imperfect model selection in the first step. Also Fan et al. (2015) show how to set up test statistics in high dimensions with power enhancing components against sparse alternatives. Nickl and van de Geer (2013) consider honest adaptive inference when $p > n$. This can be obtained as long as the rate of sparse estimation does not exceed $n^{-1/4}$. Hoffmann and Nickl (2011) consider the existence of honest adaptive confidence bands for an unknown density function. They show that this is possible if the non-parametric hypotheses for the null and alternative are asymptotically consistently distinguishable. Berk et al. (2013) propose a conservative post selection inference method. The idea is simultaneous inference in all models' submodels and this results in very wide confidence intervals. Taylor and Tibshirani (2015) discuss a practical way of taking into account the model selection's effect on post selection inference. Tibshirani (2011) provides a nice summary of developments in the literature while Lockhart et al. (2014) provide a computation based significance test for Lasso estimators. Also Zou and Li (2008) and Fan et al. (2014a) used adaptive weights in Lasso type estimators that enhance model selection properties.

The paper closest in spirit to ours is van de Geer et al. (2013, 2014) who cleverly showed how the classical Lasso estimator may be *desparsified* to construct asymptotically valid confidence bands for a low-dimensional subset of a high-dimensional parameter vector. This paper in turn is related to Zhang and Zhang (2014) and Javanmard and Montanari (2013, 2014). The idea behind desparsification is to remove the bias introduced by shrinkage via desparsifying the estimator using a cleverly constructed approximate inverse of the non-invertible empirical Gram matrix. Furthermore,

[*] Corresponding author.
E-mail addresses: caner.12@osu.edu (M. Caner), anders.kock@economics.ox.ac.uk (A.B. Kock).

https://doi.org/10.1016/j.jeconom.2017.11.005
0304-4076/© 2017 Elsevier B.V. All rights reserved.

these confidence bands do not suffer from the critique of Pötscher (2009) regarding the overly large size of confidence bands based on variable selection consistent estimators. By using the desparsified Lasso to construct confidence bands and tests, van de Geer et al. (2014) strike a middle ground between classical low dimensional inference, which relies heavily on testing, and Lasso-type techniques which perform estimation and variable selection in one step without any testing.

In the framework of the high-dimensional linear regression model and inspired by the work of van de Geer et al. (2014) we study the so-called conservative Lasso. The important observation here is that, in the presence of an oracle inequality on the plain Lasso, the penalty of the conservative Lasso on the non-zero parameters will be no larger than the one for the Lasso while the penalty on the zero parameters will be the same as the one induced by the plain Lasso. Hence, the conservative Lasso may be expected to deliver more precise parameter estimates (in finite samples) than the Lasso. And indeed, our theoretical results and simulations strongly indicate that this is the case. Also note that recently Fan et al. (2014b) proposed a weighted ℓ_1 penalized estimator with very similar weights. Their focus is on strong oracle optimality and we show that a variant of our conservative Lasso possesses the strong oracle optimality property.

We provide an oracle inequality for the conservative Lasso estimator and use the method of desparsification introduced in van de Geer et al. (2014). This approach has the advantage that the zero and non-zero coefficients do not have to be well-separated (no $\beta_{\min}$-condition is imposed) in order to conduct valid inference. We only assume the existence of r moments as opposed to the classical sub-gaussianity assumption. The oracle inequalities rely on the use of the Marcinkiewicz–Zygmund inequality which we argue delivers slightly more precise estimates than Nemirovski's inequality.

We also show that hypotheses involving an increasing number of parameters can be tested (we are considering a *fixed* sequence of hypotheses) which generalizes the results on hypotheses involving a bounded number of parameters in van de Geer et al. (2014). Furthermore, we allow for heteroskedastic error terms and provide a uniformly consistent estimator of the high-dimensional asymptotic covariance matrix. This is an important generalization in practical problems as heteroskedasticity is omniscient in econometrics and statistics. A similar approach could be of interest in large linear panel data models under strict exogeneity.

The simulations show that vast improvements can be obtained by using the desparsified conservative Lasso as opposed to the plain desparsified Lasso. To be precise, the true parameter β_0 is in general estimated much more precisely and χ^2-tests based on the desparsified conservative Lasso have much better size properties (and often also higher power) than their counterparts based on the desparsified Lasso.

When implementing Lasso-type estimators the choice of tuning parameter is important. Thus, in Theorem 5 in the appendix, we show how the method of Fan and Tang (2013) can be used to choose the tuning parameter of the variant of the conservative Lasso when the objective is consistent model selection in high dimensions.

The rest of the paper is organized as follows. Section 2 introduces the model and the conservative Lasso. Section 3 introduces nodewise regression, desparsification, and the approximate inverse to the empirical Gram matrix. Section 4 introduces inference and establishes honest confidence intervals and shows that they contract at the optimal rate. The simulations can be found in Section 5. Section 6 concludes the paper. All proofs are deferred to the appendix.

2. The model

Before stating the model setup we introduce some notation used throughout the paper.

2.1. Notation

For any real vector x, we let $\|x\|_q$ denote the ℓ_q-norm. We will primarily use the ℓ_1-, ℓ_2-, and the ℓ_∞-norm. For any $m \times n$ matrix A, we define $\|A\|_\infty = \max_{1\le i\le m, 1\le j\le n}|A_{i,j}|$. Occasionally we shall also use the induced ℓ_∞-norm. This will be denoted by $\|A\|_{\ell_\infty}$ and equals the maximum absolute row sum of A. For any symmetric matrix B, let $\phi_{\min}(B)$ and $\phi_{\max}(B)$ denote the smallest and largest eigenvalue of B, respectively. If $x \in \mathbb{R}^n$ and S is a subset of $\{1, \dots, n\}$ we let x_S be the subvector of x that picks out only those elements indexed by S.

For any set S, let $|S|$ denote its cardinality and for $x \in \mathbb{R}^n$ its prediction norm is defined as $\|x\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n x_i^2}$. $\xrightarrow{d}$ will indicate convergence in distribution and $o_p(a_n)$ as well as $O_p(b_n)$ are used in their usual meaning for sequences a_n and b_n. $a_n \asymp b_n$ means that these sequences differ at most by strictly positive multiplicative constants.

2.2. The model

We consider the model

$$Y = X\beta_0 + u, \tag{1}$$

where X is the $n \times p$ matrix of explanatory variables and u is a vector of error terms. β_0 is the $p \times 1$ population regression coefficient which we shall assume to be sparse. However, the location of the non-zero coefficients is unknown and potentially p could be much greater than n. The sparsity assumption can be replaced by a weak sparsity assumption as we shall make precise after Theorem 1 below. We assume that the explanatory variables are exogenous and precise assumptions will be made in Assumption 1 below. Let $S_0 = \{j : \beta_{0,j} \neq 0\}$ and $s_0 = |S_0|$. For later purposes define X_j as the jth column of X and X_{-j} as all columns of X except for the jth one.

2.3. The conservative Lasso and comparison to (adaptive) Lasso

The conservative Lasso is a two-step estimator defined as the weighted Lasso

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}}\{\|Y - X\beta\|_n^2 + 2\lambda_n \sum_{j=1}^{p} \hat{w}_j|\beta_j|\} \tag{2}$$

with weights $\hat{w}_j = \frac{\lambda_{prec}}{|\hat{\beta}_{L,j}| \vee \lambda_{prec}}$ where $\hat{\beta}_L$ is the plain Lasso estimator which is used to construct the weights $\hat{w}_j$. The plain Lasso corresponds to $w_j = 1$ for $j = 1, \dots, p$ in (2). Here λ_n and λ_{prec} are positive non-random quantities chosen by the researcher which we shall be specific about shortly. In Lemma A.7 and the simulation section we show that λ_{prec} can be chosen as an estimable multiple of λ_n. Hence, the only tuning parameter is λ_n. We choose λ_n by either BIC or the Generalized Information Criterion (GIC) of Fan and Tang (2013). Details are provided in the Monte Carlo section. A theorem tying GIC to model selection consistency of a variant of our conservative Lasso (which will be described in the next subsection) is at the end of Appendix B.

As opposed to the adaptive Lasso, the conservative Lasso gives variables that were excluded by the first step initial Lasso estimator a second chance — even if $|\hat{\beta}_{L,j}| = 0$ one has $\hat{w}_j = 1$ instead of an "infinitely" large penalty. Hence, the name "conservative" Lasso. The adaptive Lasso usually performs its worst when a relevant variable has been left out by the initial Lasso estimator. The

Download English Version:

https://daneshyari.com/en/article/7358101

Download Persian Version:

https://daneshyari.com/article/7358101

Daneshyari.com