LASSO FOR INSTRUMENTAL VARIABLE SELECTION

Author(s): MARTIN SPINDLER

Source: *Journal of Applied Econometrics*, March 2016, Vol. 31, No. 2 (March 2016), pp. 450-454

Published by: Wiley

Stable URL: https://www.jstor.org/stable/10.2307/26609619

JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **31**: 450–454 (2016)
Published online 26 January 2015 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.2432

LASSO FOR INSTRUMENTAL VARIABLE SELECTION: A REPLICATION STUDY

MARTIN SPINDLER*
Massachusetts Institute of Technology, Cambridge, USA; and Max Planck Society, Munich, Germany

SUMMARY

Recently, Lasso methods have been applied to economic questions. In a seminal paper, Belloni *et al.* (*Econometrica* 2012; **80**(6): 2369–2429) make use of (post-)Lasso for instrumental variable selection in a setting where the number of instruments p is large or might even exceed the number of observations n—a situation which is prevalent in many current applications. We replicate their simulation study with the statistical package R (R Development Core Team (2008)) and, moreover, analyze in more detail the importance of the choice of the penalization parameter, a crucial component in applications. Copyright © 2015 John Wiley & Sons, Ltd.

Received 12 December 2013; Revised 26 June 2014

1. INTRODUCTION

Lasso, an acronym for 'least absolute shrinkage and selection operator', was introduced in Tibshirani (1996) in order to guarantee both prediction accuracy and interpretable models by shrinking some of the coefficient to zero. The basic idea of Lasso is to minimize the residual sum of squares (RSS) subject to the constraint that the sum of the absolute value of the coefficients is less than a given constant. This is equivalent to minimizing the RSS and adding some penalization term for the absolute sum of the coefficients. Because of the nature of this constraint, it tends to produce some coefficients that are exactly zero and hence gives rise to interpretable models while maintaining predictive accuracy. Given a model of the form

$$y_i = x_i'\beta_0 + \varepsilon_i$$

where x_i denotes a fixed p-dimensional regressor and ε are independent disturbances with $\mathbb{E}\left[\varepsilon_i^2 | x_i\right] = \sigma^2$, $i = 1, \ldots, n$, the Lasso estimate of β_0 is defined as

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 + \frac{\lambda}{n} \|\beta\|_1$$

with $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$. The regressors might also include transformations of variables, like polynomials. A very important aspect is how one chooses the penalization parameter λ. λ is often chosen by cross-validation, but in a high-dimensional setting no or only very few results concerning the properties are known.[1] An important quantity for describing the choice of λ is the so-called score:

* Correspondence to: Martin Spindler, Department of Economics, Massachusetts Institute of Technology, 77 Massachusetts Avenue E19-750, Cambridge, MA 02142, USA. E-mail: mspindl@mit.edu

[1] For example, Homrighausen and McDonald (2013) show that Lasso estimators remain risk consistent relative to their linear oracles even in the case of cross-validated tuning parameter choice. Feng and Yu (2013) propose a new procedure, called consistent cross-validation (CCV), for selecting the optimal tuning parameter.

$$S = 2\frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i$$

which describes the noise in the problem. One would like to choose the smaller penalty level so that

$$\lambda \geq cn||S||_\infty \quad \text{with probability at least } 1-\alpha$$

where α should be close to zero and $c > 1$ is a constant. Two choices that are grounded on those theoretical considerations are given by

$$\text{X-independent penalty} \quad \lambda := 2c\sigma\sqrt{n}\Phi^{-1}(1-\alpha/2p)$$

$$\text{X-dependent penalty} \quad \lambda := 2c\sigma\Lambda(1-\alpha|X)$$

where $\alpha \in (0,1)$, $c > 1$ constant, Φ^{-1} the inverse of the cumulative normal distribution, and $\Lambda(1-\alpha|X)$ the $(1-\alpha)-$quantile of $n\|S/(2\sigma)\|_\infty$ conditional on $X = (x_1, \ldots, x_n)'$. As σ and $\Lambda(\cdot)$ are unknown in applications they must be estimated. $\Lambda(\cdot)$ can be determined by simulation, and σ can be estimated consistently (Belloni and Chernozhukov, 2011, section 6).

Belloni *et al.* (2012) allow non-Gaussian, heteroscedastic disturbances via the use of a data-weighted l_1-penalty function. For each regressor different penalty loadings are employed. The idea behind these data-driven penalty loadings is to first self-normalize the first-order condition of the Lasso problem and then to enable the application of moderate deviation theory to bound the maximal element of the score vector. An algorithm for calculating the loadings is given in their Appendix A.1.

2. MONTE CARLO SIMULATION

Belloni *et al.* (2012) use the following simple instrumental variables model as their data-generating process (DGP):

$$y_i = \beta d_i + e_i,$$

$$d_i = z_i'\Pi + v_i,$$

$$(e_i, v_i) \sim N\left[0, \begin{pmatrix} \sigma_e^2 & \sigma_{ev} \\ \sigma_{ev} & \sigma_v^2 \end{pmatrix}\right] \text{i.i.d.}$$

where $\beta = 1$ is the parameter of interest, and $z_i = (z_{i1}, z_{i2}, \ldots, z_{i100})' \sim N(0, \Sigma_Z)$ is a 100×1 vector with $E[z_{ih}^2] = \sigma_z^2$ and correlation $\text{corr}(z_{i1}, z_{ij}) = 0.5^{[j-h]}$. In all simulations, $\sigma_e^2 = 1$, $\sigma_z^2 = 1$ and $\text{corr}(e, v) = 0.6$. $\sigma_v^2 = 1 - \Pi'\Sigma_z\Pi$, so that the unconditional variance of the endogenous variable equals one.

For the other parameters, different settings are considered. Both the sample size ($n = 100, 250$) and the strength of instruments via the concentration parameter μ^2 ($\mu^2 = 30, 180$) are varied. C, defined below, is chosen by solving $\mu^2 = \frac{nC^2\tilde{\Pi}\Sigma_z\tilde{\Pi}}{1-C^2\tilde{\Pi}\Sigma_z\tilde{\Pi}}$, with $\tilde{\Pi}$ also defined below.

For the choice of parameters, Belloni *et al.* (2012) propose an 'exponential' and 'cut-off' ($s = 5, 50$) design. In the exponential design the parameters are given by $\Pi = C\tilde{\Pi} = C(1, 0.7, 0.7^2, \ldots, 0.7^{98}, 0.7^{99})'$, and in the cut-off case by $\Pi = C\tilde{\Pi} = C(\iota_s, 0_{n-s})'$, where ι_s is a $1 \times s$ vector of ones and 0_{n-s} is a $1 \times n - s$ vector of zeros.

DOI: 10.1002/jae

Table I. Simulation results: Lasso

	Exponential				$S = 5$				$S = 50$			
	Median				Median				Median			
Estimator	N(0)	Bias	MAD	rp(0.05)	N(0)	Bias	MAD	rp(0.05)	N(0)	Bias	MAD	rp(0.05)
A. Concentration parameter = 30, $n = 100$												
2SLS(100)		0.524	0.523	1.000		0.525	0.525	1.000		0.521	0.521	1.000
FULL(100)		0.381	0.977	0.388		0.444	1.069	0.376		0.282	0.945	0.366
Post-LASSO (paper)	481	0.102	0.217	0.014	488	0.121	0.218	0.010	500	0.348	0.375	0.000
Post-LASSO-F (paper)	481	0.102	0.217	0.014	488	0.121	0.218	0.014	500	0.348	0.375	0.000
Post-LASSO (X-dep.)	112	0.101	0.215	0.158	113	0.113	0.212	0.144	379	0.347	0.374	0.126
Post-LASSO-F (X-dep.)	112	0.112	0.214	0.158	113	0.124	0.213	0.140	379	0.349	0.375	0.128
Post-LASSO (X-indep.)	104	0.105	0.216	0.170	105	0.115	0.212	0.148	367	0.348	0.374	0.136
Post-LASSO-F (X-indep.)	104	0.112	0.215	0.170	105	0.125	0.213	0.146	367	0.351	0.378	0.138
Post-LASSO (CV)	7	0.292	0.318	0.654	3	0.297	0.320	0.660	20	0.381	0.387	0.806
Post-LASSO-F (CV)	7	0.242	0.272	0.500	3	0.246	0.277	0.470	20	0.337	0.343	0.646
Post-LASSO (Ridge)	500	0.103	0.218	0.000	500	0.120	0.217	1.000	500	0.348	0.375	0.000
Post-LASSO-F (Ridge)	500	0.103	0.218	0.000	500	0.120	0.217	1.000	500	0.348	0.375	0.000
B. Concentration parameter = 30, $n = 250$												
2SLS(100)		0.485	0.485	1.000		0.492	0.492	1.000		0.490	0.490	1.000
FULL(100)		0.026	0.364	0.102		−0.033	0.392	0.112		−0.027	0.402	0.112
Post-LASSO (paper)	398	0.090	0.188	0.056	400	0.092	0.191	0.066	497	0.358	0.376	0.014
Post-LASSO-F (paper)	398	0.093	0.190	0.058	400	0.094	0.192	0.072	497	0.358	0.376	0.014
Post-LASSO (X-dep.)	56	0.094	0.184	0.166	57	0.091	0.182	0.172	324	0.362	0.380	0.210
Post-LASSO-F (X-dep.)	56	0.103	0.184	0.162	57	0.101	0.180	0.176	324	0.365	0.382	0.218
Post-LASSO (X-indep.)	55	0.093	0.184	0.166	55	0.091	0.183	0.176	328	0.363	0.383	0.210
Post-LASSO-F (X-indep.)	55	0.102	0.183	0.160	55	0.101	0.181	0.178	328	0.366	0.376	0.220
Post-LASSO (CV)	4	0.278	0.293	0.678	1	0.279	0.293	0.666	29	0.370	0.379	0.792
Post-LASSO-F (CV)	4	0.220	0.243	0.468	1	0.218	0.241	0.462	29	0.329	0.337	0.652
Post-LASSO (Ridge)	500	0.090	0.189	0.000	500	0.092	0.191	0.000	500	0.358	0.376	0.000
Post-LASSO-F (Ridge)	500	0.090	0.189	0.000	500	0.092	0.191	0.000	500	0.338	0.376	0.000
C. Concentration parameter = 180, $n = 100$												
2SLS(100)		0.358	0.358	0.976		0.359	0.359	1.000		0.355	0.355	0.986
FULL(100)		0.054	0.837	0.392		0.090	0.773	0.330		0.082	0.652	0.338
Post-LASSO (paper)	102	0.028	0.115	0.096	138	0.036	0.114	0.126	496	0.204	0.262	0.006
Post-LASSO-F (paper)	102	0.025	0.115	0.088	138	0.031	0.112	0.104	496	0.204	0.262	0.006
Post-LASSO (X-dep.)		0.025	0.111	0.108		0.036	0.107	0.152	94	0.197	0.246	0.242
Post-LASSO-F (X-dep.)		0.021	0.110	0.102		0.031	0.106	0.138	94	0.199	0.244	0.230
Post-LASSO (X-indep.)		0.024	0.111	0.112		0.037	0.107	0.156	78	0.193	0.242	0.246
Post-LASSO-F (X-indep.)		0.021	0.110	0.104		0.031	0.106	0.140	78	0.195	0.239	0.236
Post-LASSO (CV)		0.150	0.165	0.420		0.153	0.171	0.438		0.251	0.251	0.760
Post-LASSO-F (CV)		0.091	0.130	0.260		0.094	0.135	0.248		0.159	0.170	0.404
Post-LASSO (Ridge)	500	0.033	0.129	0.000	500	0.043	0.129	1.000	500	0.204	0.263	0.000
Post-LASSO-F (Ridge)	500	0.033	0.129	0.000	500	0.043	0.129	1.000	500	0.204	0.263	0.000

3. SIMULATION RESULTS

We replicate the simulation study of Belloni *et al.* (2012) in R (version 3.0.1). For estimation of the Lasso coefficients we apply the functions from the package lars (Efron and Hastie, 2013) and glmnet (Friedman *et al.*, 2010) and also an implementation of the shooting Lasso (Fu, 1998). Although one uses different algorithms for the calculations, all three functions give identical results.[2]

As mentioned above, the choice of λ is critical in applications. We apply several different rules: the iterative approach of Belloni *et al.* (2012) with different loadings for each regressor as described in their Appendix A ('paper'), 10-fold cross-validation (CV) and the X-dependent and X-independent rule as described in (Belloni and Chernozhukov, 2011) with plug-in estimates for σ (both without iteration; 'X-dep.' and 'X-indep.' For the X-dependent approach we use 500 simulations.

[2] The penalization parameters are defined differently, and therefore care is necessary.

DOI: 10.1002/jae

Table 1. *Continued*

	Exponential				$S = 5$				$S = 50$			
		Median				Median				Median		
Estimator	N(0)	Bias	MAD	rp(0.05)	N(0)	Bias	MAD	rp(0.05)	N(0)	Bias	MAD	rp(0.05)
D. Concentration parameter = 180, $n = 250$												
2SLS(100)		0.281	0.281	0.982		0.286	0.286	0.990		0.278	0.278	0.988
FULL(100)		0.004	0.095	0.090		−0.004	0.099	0.084		−0.003	0.102	0.094
Post-LASSO (paper)		0.024	0.082	0.138		0.027	0.082	0.116		0.029	0.084	0.126
Post-LASSO-F (paper)		0.018	0.081	0.122		0.020	0.080	0.092		0.022	0.083	0.112
Post-LASSO (X-dep.)		0.026	0.083	0.140		0.025	0.081	0.011		0.029	0.084	0.126
Post-LASSO-F (X-dep.)		0.019	0.081	0.120		0.016	0.080	0.094		0.020	0.083	0.110
Post-LASSO (X-indep.)		0.026	0.082	0.138		0.025	0.081	0.112		0.028	0.084	0.124
Post-LASSO-F (X-indep.)		0.019	0.081	0.120		0.016	0.080	0.094		0.020	0.082	0.108
Post-LASSO (CV)		0.124	0.128	0.464		0.124	0.134	0.438		0.129	0.140	0.472
Post-LASSO-F (CV)		0.070	0.102	0.254		0.069	0.096	0.218		0.074	0.103	0.232
Post-LASSO (Ridge)	243	0.047	0.100	0.109	292	0.057	0.104	0.115	295	0.054	0.107	0.128
Post-LASSO-F (Ridge)	243	0.046	0.100	0.099	292	0.056	0.104	0.099	295	0.052	0.106	0.122

Note: Results are based on 500 simulation replications and 100 instruments. Column labels indicate the structure of the first-stage coefficients as described in Belloni *et al.* (2012). 2SLS(100) and FULL(100) are, respectively, the 2SLS and Fuller estimators using all 100 potential instruments. Post-LASSO and Post-LASSO-F, respectively, correspond to 2SLS and Fuller using the instruments selected from Lasso variable selection among the 100 instruments, with inference based on the asymptotic normal approximation; in cases where no instruments are selected, the procedure switches to using the sup-Score test for inference. sup-Score provides the rejection frequency for a weak identification robust procedure that is suited to situations with more instruments than observations. Post-LASSO (Ridge) and Post-LASSO-F (Ridge) are defined as Post-LASSO and Post-LASSO-F but augment the instrument set with a fitted value obtained via ridge regression. We report the number of replications in which Lasso selected no instruments ($N(0)$), median bias (Med. Bias), median absolute deviation (MAD) and rejection frequency for 5% level tests (rp(0.05)). In cases where Lasso selects no instruments, Med. Bias and MAD are based on 2SLS using the single instrument with the largest sample correlation to the endogenous variable and rp(0.05) is based on the sup-Score test.

2SLS(100) and FULL(100) denote the 2SLS and Fuller estimator based on all available instruments. For all choices of λ both the two stage least squares (2SLS) and Fuller ('-F') version are estimated. The regressors selected in the first stage are used for the second stage, so that the estimation can be described as 'post-Lasso'. For each estimator, median bias (Med. Bias), median absolute deviation (MAD) and rejection frequencies for 5% level tests (rp(0.05)) are reported. For computing rejection frequencies, conventional, homoscedastic 2SLS standard errors are used for 2SLS(100) and post-Lasso and the many-instrument robust standard errors of Hansen *et al.* (2008), which rely on homoscedasticity, are estimated for FULL(100) and post-Lasso-F. Moreover, the number of cases in which Lasso selected no instruments is given in the column labeled N(0).

The results are shown in Table I. We can replicate the results in Belloni *et al.* (2012), and the only difference concerns the Ridge-based estimator. The Ridge-based estimator uses a combination of Ridge regression, Lasso and sample splitting. The MAD is comparable, but we estimate a lower bias. As the Ridge-based estimator almost never selects a variable, it is of minor importance for our comparison. The main focus is on the comparison of the different choices for the penalization parameter which extends the simulation study of Belloni *et al.* 2012). We see that the cross-validated estimates tend to select too many regressors in all cases, which is well known in the literature (see, among others, Feng and Yu, 2013). The parameter setting with $\mu = 30$ is the only parameter constellation where no variables are selected at all, but this occurs only in very few cases. The results concerning the rejection frequencies of the level tests are poor. Although CV is very common when using Lasso, in our application it leads to very poor estimates, resulting in unsatisfactory test results, i.e. high rejection frequencies. Both the X-independent and X-dependent rule give comparable results and therefore there seems to be no need to use the simulation-based, X-dependent rule in applications. An interesting observation is that both rules deliver comparable results to the iterative procedure ('paper') in the

DOI: 10.1002/jae

setting with strong instruments ($\mu = 180$). In the case of weak instruments ($\mu = 30$) they give similar results concerning the bias and MAD, but give higher rejection rates. Another important observation is that the implementation of the choice of λ in Belloni *et al.* (2012) produces sparser results than the other rules, e.g. more often than any other rule it selects none of the instruments.

ACKNOWLEDGEMENTS

I thank Victor Chernozhukov for his most valuable comments and Christian Hansen for answering questions concerning the original implementation. I am grateful to the editor Badi Baltagi and three anonymous referees for their comments, which helped to improve the paper. Furthermore, I thank MIT for its great hospitality, where parts of the paper were written, and gratefully acknowledge financial support from the Fritz Thyssen Stiftung.

REFERENCES

Belloni A, Chernozhukov V. 2011. High dimensional econometric models: an introduction. *MIT Working Paper* **11–17**.

Belloni A, Chen D, Chernozhukov V, Hansen C. 2012. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80**(6): 2369–2429.

Efron B, Hastie T. 2013. lars: Least angle regression, lasso and forward stagewise. R package version 1.2. Available: http://CRAN.R-project.org/package=lars [14 December 2014].

Feng Y, Yu Y. 2013. Consistent cross-validation for tuning parameter selection in high-dimensional variable selection. Available: http://arxiv.org/pdf/1308.5390v1.pdf [14 December 2014].

Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1): 1–22.

Fu WJ. 1998. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7**: 397–416.

Hansen C, Hausman J, Newey WK. 2008. Estimation with many instrumental variables. *Journal of Business and Economic Statistics* **26**: 398–422.

Homrighausen D, McDonald D. 2013. Risk consistency of cross-validation with lasso-type procedures. Available: http://arxiv.org/pdf/1308.0810v1.pdf [14 December 2014].

R Development Core Team. 2008. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B* **58**(1): 267–288.

DOI: 10.1002/jae