Electronic Journal of Statistics > Vol. 2 (2008) open journal systems

Honest variable selection in linear and logistic regression models via ell_1 and $ell_1 + ell_2$ penalization

Florentina Bunea, *FSU*

Abstract

This paper investigates correct variable selection in finite samples via ell_1 and $ell_1 + ell_2$ type penalization schemes. The asymptotic consistency of variable selection immediately follows from this analysis. We focus on logistic and linear regression models. The following questions are central to our paper: given a level of confidence $1 - delta$, under which assumptions on the design matrix, for which strength of the signal and for what values of the tuning parameters can we identify the true model at the given level of confidence? Formally, if $widehat{I}$ is an estimate of the true variable set I^*, we study conditions under which $mathbb {P}(widehat{I} = I^*) geq 1 - delta$, for a given sample size n, number of parameters M and confidence $1 - delta$. We show that in identifiable models, both methods can recover coefficients of size $frac{1}{sqrt{n}}$, up to small multiplicative constants and logarithmic factors in M and $frac{1}{delta}$. The advantage of the $ell_1 + ell_2$ penalization over the ell_1 is minor for the variable selection problem, for the models we consider here. Whereas the former estimates are unique, and become more stable for highly correlated data matrices as one increases the tuning parameter of the ell_2 part, too large an increase in this parameter value may preclude variable selection.

AMS 2000 subject classifications: Primary 62J07; secondary 62J02, 62G08.

Keywords: Lasso, elastic net, ℓ_1 and $\ell_1 + \ell_2$ regularization, penalty, sparse, consistent, variable selection, regression, generalized linear models, logistic regression, high dimensions.

Full Text: PDF

Bunea, Florentina, Honest variable selection in linear and logistic regression models via ell_1 and $ell_1 + ell_2$ penalization, Electronic Journal of Statistics, 2, (2008), 1153-1194 (electronic). DOI: 10.1214/08-EJS287.

References

[1] Armitage, P. (1955) Tests for linear trends in proportions and frequencies. Biometrics, 11 (3), 375–386.

[2] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. Simultaneous analysis of Lasso and Dantzig Selector. The Annals of Statistics: To appear.

[3] Bunea, F. (2008) Consistent selection via the Lasso for high dimensional approximating regression models. The IMS Collections 3 122–137.

[4] Bunea, F., Tsybakov, A. B. and Wegkamp, M. H. (2007) Aggregation for Gaussian regression. The Annals of Statistics 35 (4), 1674–1697. MR2351101

[5] Bunea, F., Tsybakov, A. B. and Wegkamp, M. H. (2007). Sparsity oracle inequalities for the Lasso. The Electronic Journal of Statistics 1, 169–194. MR2312149

[6] Candés, E. J. and Plan, Y. (2007) Near-ideal model selection by ℓ_1 minimization.

Technical Report, Caltech.

[7] Devroye, L. and Lugosi, G. (2001) Combinatorial methods in density estimation Springer-Verlag. MR1843146

[8] Donoho, D. L., Elad, M. and Temlyakov, V. N. (2006) Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Trans. Inform. Theory 52 (1), 6–18. MR2237332

[9] Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96(456), 1348–1360. MR1946581

[10] Greenshtein, E. (2006) Best subset selection, persistence in high-dimensional statistical learning and optimization under l1 constraint. The Annals of Statistics 34(5), 2367–2386. MR2291503

[11] Koltchinskii, V. Sparsity in penalized empirical risk minimization. Technical report, School of Mathematics, Georgia Tech.

[12] Lounici, K. (2008) Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. The Electronic Journal of Statistics 2, 90–102. MR2386087

[13] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. The Annals of Statistics 34 (3), 1436–1462. MR2278363

[14] Meinshausen, N. and Yu, B. (2007) Lasso-type recovery of sparse representations for high dimensional data. To appear in the Annals of Statistics.

[15] Osborne, M.R., Presnell, B. and Turlach, B.A (2000a). On the lasso and its dual. Journal of Computational and Graphical Statistics 9, 319–337. MR1822089

[16] Osborne, M.R., Presnell, B. and Turlach, B.A (2000b). A new approach to variable selection in least squares problems. IMA Journal of Numerical Analysis 20(3), 389–404.

[17] Ravikumar, P., Wainwright, M. J. and Lafferty, J. (2008) High-dimensional graphical model selection using ℓ_1-regularized logistic regression. Technical Report, UC Berkeley, Dept of Statistics.

[18] Steinwart, I. (2007) How to compare different loss functions and their risks. Constructive Approximation 26, 225–287. MR2327600

[19] van de Geer, S. (2008) High-dimensional generalized linear models and the Lasso. The Annals of Statistics 36 (2), 614–645. MR2396809

[20] Zhang, T. (2007) Some sharp performance bounds for least squares regression with l1 regularization. Technical report, Rutgers University.

[21] Zhao, P. and Yu, B. (2007). On model selection consistency of Lasso. Journal of Machine Learning Research 7, 2541–2567. MR2274449

[22] Zou, H. (2006) The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429. MR2279469

[23] Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B 67 (2) 301–320. MR2137327

[24] Wainwright, M. J. (2007). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. Technical Report, UC Berkeley, Department of Statistics.

[25] Wasserman, L. and Roeder, K. (2007). High dimensional variable selection. Technical Report, Carnegie Mellon University, Department of Statistics.

[26] Wegkamp, M. H. (2007) Lasso type classifiers with a reject option. Electronic

Journal of Statistics 1, 155–168. MR2312148

Electronic Journal of Statistics. ISSN: 1935-7524