Elastic-net regularized high-dimensional negative binomial regression: consistency and weak signal detection.

Summary: We study a sparse negative binomial regression (NBR) for count data by showing the non-asymptotic advantages of using the elastic-net estimator. Two types of oracle inequalities are derived for the NBR’s elastic-net estimates by using the Compatibility Factor Condition and the Stabil Condition. The second type of oracle inequality is for the random design and can be extended to many $\ell_1 + \ell_2$ regularized M-estimations, with the corresponding empirical process having stochastic Lipschitz properties. We derive the concentration inequality for the suprema empirical processes for the weighted sum of negative binomial variables to show some high-probability events. We apply the method by showing the sign consistency, provided that the nonzero components in the true sparse vector are larger than a proper choice of the weakest signal detection threshold. In the second application, we show the grouping effect inequality with high probability. Third, under some assumptions for a design matrix, we can recover the true variable set with a high probability if the weakest signal detection threshold is larger than the turning parameter up to a known constant. Lastly, we briefly discuss the de-biased elastic-net estimator, and numerical studies are given to support the proposal.

MSC:

62–XX Statistics

Keywords:
de-biased elastic-net; empirical processes; high-dimensional count data regressions; oracle inequalities; sign consistency; stochastic Lipschitz condition

Software:
catdata; logcondiscr; COUNT

Full Text: DOI arXiv

References:

[1] Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. Electronic Journal of Probability 13, 1000-1034. · Zbl 1190.60010
[2] Balabdaoui, F., Jankowski, H., Rufibach, K. and Pavlides, M. (2013). Asymptotics of the discrete log-concave maximum likelihood estimator and related applications. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75, 769-790. · Zbl 1411.62119
[3] Bickel, P. J., Ritov, Y. A. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics 37, 1705-1732. · Zbl 1173.62022
[4] Blazere, M., Loubes, J. M. and Gamboa, F. (2014). Oracle inequalities for a group Lasso procedure applied to generalized linear models in high dimension. IEEE Transactions on Information Theory 60, 2303-2318. · Zbl 1360.94123
[5] Bühlmann, P. and van de Geer, S. A. (2011). Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer. · Zbl 1273.62015
[6] Bunea, F. (2008). Honest variable selection in linear and logistic regression models via 1 and 1+ 2 penalization. Electronic Journal of Statistics 2, 1153-1194. · Zbl 1320.62170
[7] Cameron, A. C. and Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. Journal of Econometrics 46, 347-364.
[8] Chi, Z. (2010). A local stochastic Lipschitz condition with application to Lasso for high dimensional generalized linear models. arXiv preprint arXiv:1009.1052.
[9] Hilbe, J. M. (2011). Negative Binomial Regression. 2nd Edition. Cambridge University Press. · Zbl 1269.62063
[10] Huang, J., Sun, T., Ying, Z., Yu, Y. and Zhang, C. H. (2013). Oracle inequalities for the Lasso in the Cox model. Annals of Statistics 41, 1142-1163. · Zbl 1292.62135
[11] Janková, J. and van de Geer, S. (2016). Confidence regions for high-dimensional generalized linear models under sparsity.
[12] Johnson, N. L., Kemp, A. W. and Kotz, S. (2005). Univariate Discrete Distributions. 3rd Edition. Wiley. – Zbl 1092.62010
[13] Jia, J. and Yu, B. (2010). On model selection consistency of the Elastic Net when p ~ n.
[14] Statistica Sinica 20, 595-611.
[15] Kakade, S., Shamir, O., Sindharan, K. and Tewari, A. (2010). Learning exponential families in high-dimensions: Strong convexity and sparsity. In International Conference on Artificial Intelligence and Statistics, 381-388.
[16] Lv, S., You, M., Lin, H., Lian, H. and Huang, J. (2018). On the sign consistency for the Lasso for the high-dimensional Cox model. Journal of Multivariate Analysis 167, 79-96. – Zbl 1395.62009
[17] Massaro, T. J. (2016). Variable Selection via Penalized Regression and the Genetic Algorithm Using Information Complexity, with Applications for High-Dimensional-Omics Data. PhD Dissertations, University of Tennessee.
[18] Nielsen, F. and Nock, R. (2009). Sided and symmetrized Bregman centroids. IEEE Transactions on Information Theory 55, 2882-2904. – Zbl 1367.94138
[19] Qin, Y., Chen, S. X. and Nettleton, D. (2018). Detecting rare and faint signals via thresholding maximum likelihood estimators. The Annals of Statistics 46, 895-923. – Zbl 1392.62163
[20] Rauschenberger, A., Jonker, M. A., van de Wiel, M. A. and Menezes, R. X. (2016). Testing for association between RNA-Seq and high-dimensional data. BMC Bioinformatics 17, 118.
[21] Rigollet, P. (2012). Kullback-Leibler aggregation and misspecified generalized linear models. The Annals of Statistics 40, 639-665. – Zbl 1274.62298
[22] Sørensen, O., Hellton, K. H., Frigessi, A. and Thoresen, M. (2018). Covariate selection in high-dimensional generalized linear models with measurement error. Journal of Computational and Graphical Statistics 27, 739-749.
[23] Sellers, K. F. and Shmueli, G. (2008). A flexible regression model for count data. The Annals of Applied Statistics 4, 943-961. – Zbl 1194.62091
[24] Tutz, G. (2011). Regression for Categorical Data. Cambridge University Press. – Zbl 1305.62259
[25] Tibshirani, R. (2006). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58, 267-288. – Zbl 1092.62010
[26] van de Geer, S. A. (2007). The deterministic Lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich.
[27] van de Geer, S. A. (2008). High-dimensional generalized linear models and the Lasso. The Annals of Statistics 36, 614-645. – Zbl 1138.62324
[28] van de Geer, S. A., Bühlmann, P., Ritov, Y. A. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. The Annals of Statistics 42, 1166-1202. – Zbl 1305.62259
[29] van der Vaart, A. W. (1998). Asymptotic Statistics. 3rd Edition. Cambridge University Press. – Zbl 0910.62001
[30] Wang, Z., Ma, S., Zappitelli, M., Parikh, C., Wang, C. Y. and Devarajan, P. (2016). Penalized count data regression with application to hospital stay after pediatric cardiac surgery. Statistical Methods in Medical Research 25, 2685-2703.
[31] Weißbach, R. and Radloff, L. (2020). Consistency for the negative binomial regression with fixed covariates. Metrika 83, 627-641. – Zbl 1442.62124
[32] Ye, F. and Zhang, C. H. (2010). Rate Minimaxity of the Lasso and Dantzig Selector for the lq Loss in lr Balls. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72, 288-311. – Zbl 1197.62126
[33] Ye, F. and Zhang, C. H. (2010). Rate Minimaxity of the Lasso and Dantzig Selector for the lq Loss in lr Balls. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72, 288-311. – Zbl 1197.62126
[34] Yu, Y., Bradic, J. and Samworth, R. J. (2021). Confidence intervals for high-dimensional Cox models. Statistica Sinica 31, 243-267. – Zbl 1464.62351
[35] Zhang, C. H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. Huiming Zhang School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing 100871, P.R. China. Department of Mathematics, Faculty of Science and Technology, University of Macau, Taipa, Macau, P.R. China.
[36] Zhang, H., Liu, Y. and Li, B. (2014). Notes on discrete compound Poisson model with applications to risk theory. Insurance: Mathematics and Economics 59, 325-336. – Zbl 1306.60050
[37] Zhang, H. and Chen, S. X. (2021). Concentration inequalities for statistical inference. Communications in Mathematical Research 37, 1-85. – Zbl 1488.60028
[38] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. Journal of Machine learning research 7, 2541-2563. – Zbl 1222.62008
[39] Zhou, D. X. (2013). On grouping effect of elastic net. Statistics & Probability Letters 83, 2108-2112. – Zbl 1242.62074
[40] Zhou, D. X. (2013). On grouping effect of elastic net. Statistics & Probability Letters 83, 2108-2112. – Zbl 1242.62074
[41] Zhu, J. (2012). Rate Minimaxity of the Lasso and Dantzig Selector for the lq Loss in lr Balls. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74, 288-311. – Zbl 1250.62074
[42] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. Journal of Machine learning research 7, 2541-2563. – Zbl 1222.62008
[43] Zhu, J. (2012). Rate Minimaxity of the Lasso and Dantzig Selector for the lq Loss in lr Balls. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74, 288-311. – Zbl 1250.62074
[44] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. Journal of Machine learning research 7, 2541-2563. – Zbl 1222.62008
[45] Zhou, D. X. (2013). On grouping effect of elastic net. Statistics & Probability Letters 83, 2108-2112. – Zbl 1250.62074
[46] Huiming Zhang School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing 100871, P.R. China. Department of Mathematics, Faculty of Science and Technology, University of Macau, Taipa, Macau, P.R. China.
[47] Jinzhao Jia School of Public Health and Center for Statistical Science, Peking University, Beijing 100191, P.R. China.

This reference list is based on information provided by the publisher or from digital mathematics libraries. Its items are heuristically matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original paper as accurately as possible without claiming the completeness or perfect precision of the matching.