Qin, Shanshan; **Ding, Hao**; **Wu, Yuehua**; **Liu, Feng**

High-dimensional sign-constrained feature selection and grouping. (English) Zbl 1469.62301

Ann. Inst. Stat. Math. 73, No. 4, 787-819 (2021).

Summary: In this paper, we propose a non-negative feature selection/feature grouping (nnFSG) method for general sign-constrained high-dimensional regression problems that allows regression coefficients to be disjointly homogeneous, with sparsity as a special case. To solve the resulting non-convex optimization problem, we provide an algorithm that incorporates the difference of convex programming, augmented Lagrange and coordinate descent methods. Furthermore, we show that the aforementioned nnFSG method recovers the oracle estimate consistently, and that the mean-squared errors are bounded. Additionally, we examine the performance of our method using finite sample simulations and applying it to a real protein mass spectrum dataset.

MSC:

62J05 Linear regression; mixed models
62H30 Classification and discrimination; cluster analysis (statistical aspects)
62P10 Applications of statistics to biology and medical sciences; meta analysis
90C25 Convex programming

Keywords:

difference convex programming; feature grouping; feature selection; high-dimensional regression; protein mass spectrum dataset; malaria vaccine data

Software:

nnls; NITPICK; CVXR; CRAN

Full Text: DOI

References:

[1] Arnold, TB; Tibshirani, RJ, Efficient implementations of the generalized lasso dual path algorithm, Journal of Computational and Graphical Statistics, 25, 1, 1-27 (2016)· doi:10.1080/10618600.2015.1008638

[2] Esser, E.; Lou, YF; Xin, J., A method for finding structured sparse solutions to nonnegative least squares problems with applications, SIAM Journal on Imaging Sciences, 6, 4, 2010-2046 (2013) · Zbl 1282.90239 · doi:10.1137/13090540X

[3] Fan, J.; Li, R., Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American Statistical Association, 96, 456, 1348-1360 (2001) · Zbl 1073.62547 · doi:10.1198/016214501753382273

[4] Frank, LE; Friedman, JH, A statistical view of some chemometrics regression tools, Technometrics, 35, 2, 109-135 (1993) · Zbl 0775.62288 · doi:10.1080/00401706.1993.10485033

[5] Friedman, J.; Hastie, T.; Simon, N.; Tibshirani, R., Lasso and elastic-net regularized generalized linear models, R-Package Version, 2, -5, 2016 (2016)

[6] Fu, A., Narasimhan, B., Boyd, S. (2017). CVXR: An R package for disciplined convex optimization. arXiv:1711.07582.

[7] Goeman, JJ, \(L_1\) penalized estimation in the Cox proportional hazards model, Biometrical Journal, 52, 1, 70-84 (2010) · Zbl 1207.62185

[8] Hu, Z.; Follmann, DA; Miura, K., Vaccine design via nonnegative lasso-based variable selection, Statistics in Medicine, 34, 10, 1791-1798 (2015)· doi:10.1002/sim.6452

[9] Huang, J.; Ma, S.; Xie, H.; Zhang, CH, A group bridge approach for variable selection, Biometrika, 96, 2, 339-355 (2009) · Zbl 1163.62050 · doi:10.1093/biomet/asp020

[10] Itoh, Y.; Duarte, MF; Parente, M., Perfect recovery conditions for non-negative sparse modeling, IEEE Transactions on Signal Processing, 65, 1, 69-80 (2016) · Zbl 1414.94272 · doi:10.1109/TSP.2016.2613067

[11] Jang, W.; Lim, J.; Lazar, N.; Loh, JM; McDowell, J.; Yu, D., Regression shrinkage and equality selection for highly correlated predictors with HORSES, Biometrics, 64, 1-23 (2011)

[12] Koike, Y.; Tanoue, Y., Oracle inequalities for sign constrained generalized linear models, Econometrics and Statistics, 11, 145-157 (2019)· doi:10.1016/j.ecosta.2019.02.001

[13] Luenberger, DG; Ye, Y., Linear and nonlinear programming (2015), New York: Springer, New York · Zbl 1207.90003

[14] Mandal, BN; Ma, J., \(l_1\) regularized multiplicative iterative path algorithm for non-negative generalized linear models, Computational Statistics and Data Analysis, 101, 289-299 (2016) · Zbl 1466.62156 · doi:10.1016/j.csda.2016.03.009

[15] Meinshausen, N., Sign-constrained least squares estimation for high-dimensional regression, Electronic Journal of Statistics, 7, 1607-1631 (2013) · Zbl 1327.62422 · doi:10.1214/13-EJS818

[16] Mullen, K. M., van Stokkum, I. H. (2012). The Lawson-Hanson algorithm for nonnegative least squares (NNLS). CRAN: R package. https://cran.r-project.org/web/packages/nnls/nnls.pdf.

[17] Rekabdarkolaee, HM; Boone, E.; Wang, Q., Robust estimation and variable selection in sufficient dimension reduction, Computational Statistics and Data Analysis, 108, 146-157 (2017) · Zbl 1466.62182 · doi:10.1016/j.csda.2016.11.007

[18] Renard, BY; Kirchner, M.; Steen, H.; Steen, JA; Hamprecht, FA, NITPICK: Peak identification for mass spectrometry data, BMC Bioinformatics, 9, 1, 355 (2008)· doi:10.1186/1471-2105-9-355

[19] Shadmi, Y., Jung, P., Caire, G. (2019). Sparse non-negative recovery from biased sub-Gaussian measurements using NNLS. arXiv:1901.05727.

[20] She, Y., Sparse regression with exact clustering, Electronic Journal of Statistics, 4, 1055-1096 (2010) · Zbl 1329.62327 · doi:10.1214/10-EJS578

[21] Shen, X.; Huang, HC; Pan, W., Simultaneous supervised clustering and feature selection over a graph, Biometrika, 99, 4, 899-914 (2012) · Zbl 1452.62467 · doi:10.1093/biomet/ass038

[22] Shen, X.; Pan, W.; Zhu, Y., Likelihood-based selection and sharp parameter estimation, Journal of the American Statistical Association, 107, 497, 223-232 (2012) · Zbl 1261.62020 · doi:10.1080/01621459.2011.645783

[23] Shen, X.; Pan, W.; Zhu, Y.; Zhou, H., On constrained and regularized high-dimensional regression, Annals of the Institute of Statistical Mathematics, 65, 5, 807-832 (2013) · Zbl 1329.62307 · doi:10.1007/s10463-012-0396-3

[24] Slawski, M., Hein, M. (2010). Sparse recovery for protein massspectrometry data. In NIPS workshop on practical applications of sparse modelling.

[25] Slawski, M.; Hein, M., Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization, Electronic Journal of Statistics, 7, 3004-3056 (2013) · Zbl 1280.62086 · doi:10.1214/13-EJS868

[26] Slawski, M.; Hussong, R.; Tholey, A.; Jakoby, T.; Gregorius, B.; Hildebrandt, A.; Hein, M., Isotope pattern deconvolution for peptide mass spectrometry by non-negative least squares/least absolute deviation template matching, BMC Bioinformatics, 13, 1, 291 (2012)· doi:10.1186/1471-2105-13-291

[27] Tibshirani, R., Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological), 58, 1, 267-288 (1996) · Zbl 0850.62538

[28] Tibshirani, R.; Wang, P., Spatial smoothing and hot spot detection for CGH data using the fused lasso, Biostatistics, 9, 1, 18-29 (2008) · Zbl 1274.62886 · doi:10.1093/biostatistics/kxm013

[29] Tibshirani, R.; Saunders, M.; Rosset, S.; Zhu, J.; Knight, K., Sparsity and smoothness via the fused lasso, Journal of the Royal Statistical Society. Series B (Statistical Methodology), 67, 1, 91-108 (2005) · Zbl 1060.62049 · doi:10.1111/j.1467-9868.2005.00490.x

[30] Tibshirani, RJ; Taylor, J., The solution path of the generalized lasso, The Annals of Statistics, 39, 3, 1335-1371 (2011) · Zbl 1234.62107 · doi:10.1214/11-AOS878

[31] Wen, YW; Wang, M.; Cao, Z.; Cheng, X.; Ching, WK; Vassiliadis, VS, Sparse solution of nonnegative least squares problems with applications in the construction of probabilistic Boolean networks, Numerical Linear Algebra with Applications, 22, 5, 883-899 (2015) · Zbl 1349.65140 · doi:10.1002/nla.2001

[32] Wu, L.; Yang, Y., Nonnegative elastic net and application in index tracking, Applied Mathematics and Computation, 227, 541-552 (2014) · Zbl 1364.91156 · doi:10.1016/j.amc.2013.11.049

[33] Wu, L.; Yang, Y.; Liu, H., Nonnegative-lasso and application in index tracking, Computational Statistics and Data Analysis, 70, 116-126 (2014) · Zbl 1471.62220 · doi:10.1016/j.csda.2013.08.012

[34] Xiang, S.; Shen, X.; Ye, J., Efficient nonconvex sparse group feature selection via continuous and discrete optimization, Artificial Intelligence, 224, 28-50 (2015) · Zbl 1343.68210 · doi:10.1016/j.artint.2015.02.008

[35] Yang, S., Yuan, L., Lai, Y. C., Shen, X., Wonka, P., Ye, J. (2012). Feature grouping and selection over an undirected graph. Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 922-930). ACM. New York.

[36] Yang, Y.; Wu, L., Nonnegative adaptive lasso for ultra-high dimensional regression models and a two-stage method applied in financial modeling, Journal of Statistical Planning and Inference, 174, 52-67 (2016) · Zbl 1353.62076 · doi:10.1016/j.jspi.2016.01.011

[37] Yuan, M.; Lin, Y., Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68, 1, 49-67 (2006) · Zbl 1141.62030 · doi:10.1111/j.1467-9868.2005.00532.x

[38] Zhang, CH, Nearly unbiased variable selection under minimax concave penalty, The Annals of Statistics, 38, 2, 894-942 (2010) · Zbl 1183.62120 · doi:10.1214/09-AOS729

[39] Zhu, Y.; Shen, X.; Pan, W., Simultaneous grouping pursuit and feature selection over an undirected graph, Journal of the American Statistical Association, 108, 502, 713-725 (2013) · Zbl 06195973 · doi:10.1080/01621459.2013.770704

[40] Zou, H., The adaptive lasso and its oracle properties, Journal of the American Statistical Association, 101, 476, 1418-1429 (2006) · Zbl 1171.62326 · doi:10.1198/016214506000000735

[41] Zou, H.; Hastie, T., Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67, 2, 301-320 (2005) · Zbl 1069.62054 · doi:10.1111/j.1467-9868.2005.00503.x

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.