Electronic Journal of Statistics > Vol. 3 (2009) open journal systems

Penalized orthogonal-components regression for large *p* small *n* data

Dabao Zhang, *Purdue University*
Yanzhu Lin, *Purdue University*
Min Zhang, *Purdue University*

Abstract

Here we propose a penalized orthogonal-components regression (POCRE) for large p small n data. Orthogonal components are sequentially constructed to maximize, upon standardization, their correlation to the response residuals. A new penalization framework, implemented via empirical Bayes thresholding, is presented to effectively identify sparse predictors of each component. POCRE is computationally efficient owing to its sequential construction of leading sparse principal components. In addition, such construction offers other properties such as grouping highly correlated predictors and allowing for collinear or nearly collinear predictors. With multivariate responses, POCRE can construct common components and thus build up latent-variable models for large p small n data.

AMS 2000 subject classifications: Primary 62J05; secondary 62H20, 62J07.

Keywords: Empirical Bayes thresholding, Latent-variable model, p ≫ n data, POCRE, Sparse predictors, Supervised dimension reduction.

Full Text: PDF

Zhang, Dabao, Lin, Yanzhu, Zhang, Min, Penalized orthogonal-components regression for large *p* small *n* data, Electronic Journal of Statistics, 3, (2009), 781-796 (electronic). DOI: 10.1214/09-EJS354.

References

[1] Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006). Prediction by supervised principal components. Journal of the American Statistical Association, 101 119–137. MR2252436

[2] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. Annals of Statistics, 24 2350–2383. MR1425957

[3] Cao, K.-A. L., Rossouw, D., Robert-Granie, C. and Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. Statistical Applications in Genetics and Molecular Biology, 7 Article 35.

[4] Chun, H. and Keles, S. (2007). Sparse partial least squares for simultaneous dimension reduction and variable selection. http://www.stat.wisc.edu /~keles/Papers/SPLS_Nov07.pdf.

[5] Cook, R. D. (2007). Fisher lecture: dimensional reduction in regression. Statistical Science, 22 1–26. MR2408655

[6] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. Biometrika, 81 425–455. MR1311089

[7] Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null

hypothesis. Journal of the American Statistical Association, 99 96–104. MR2054289

[8] Garthwaite, P. H. (1994) An Interpretation of Partial Least Squares. Journal of the American Statistics Association, 89 122–127. MR1266290

[9] Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D. and Brown, P. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biology, 1 research0003.1–research0003.21.

[10] James, W. and Stein, C. (1961). Estimation with quadratic loss. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1 361–379. University of California Press, Berkeley. MR0133191

[11] Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequence. The Annals of Statistics, 32, 1594–1649. MR2089135

[12] Johnstone, I. M. and Silverman, B. W. (2005). EbayesThresh: R programs for empirical Bayes thresholding. Journal of Statistical Software, 12 1–38.

[13] Kramer, R. (1998). Chemometric Techniques for Quantitative Analysis. Marcel-Dekker.

[14] Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., Mui, E. T., Flowers, M. T., Schueler, K. L., Manly, K. F., Williams, R. W., Kendziorski, K., and Attie, A. D. (2006). Combined expression trait correlations and expression quantitative trait locus mapping. PLoS Genetics, 2 e6.

[15] Nguyen, D. V. and Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. Bioinformatics, 18 39–50.

[16] Park, T. and Casella, G. (2008). The Bayesian lasso. Journal of the American Statistical Association, 103 681–686.

[17] Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis, 99 1015–1034. MR2419336

[18] Stewart, G. W. (1974). Introduction to Matrix Computations. New York: Academic Press. MR0458818

[19] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of Royal Statistical Society B, 58 267–288. MR1379242

[20] Tibshirani, R., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. Journal of Royal Statistical Society B, 67 91–108. MR2136641

[21] Wold, H. (1975). Soft modelling by latent variables: the nonlinear iterative partial least squares approach. In Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett (eds J. Gani). London: Academic Press. MR0394782

[22] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of Royal Statistical Society B, 68 49–67. MR2212574

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

[24] Zou, H., Hastie, H. and Tibshirani, R. (2006). Sparse Principal Component Analysis. Journal of Computational and Graphical Statistics, 15 265–286. MR2252527

Electronic Journal of Statistics. ISSN: 1935-7524