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

Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables

Benhuai Xie, *University of Minnesota*
Wei Pan, *University of Minnesota*
Xiaotong Shen, *University of Minnesota*

Abstract

Clustering analysis is one of the most widely used statistical tools in many emerging areas such as microarray data analysis. For microarray and other high-dimensional data, the presence of many noise variables may mask underlying clustering structures. Hence removing noise variables via variable selection is necessary. For simultaneous variable selection and parameter estimation, existing penalized likelihood approaches in model-based clustering analysis all assume a common diagonal covariance matrix across clusters, which however may not hold in practice. To analyze high-dimensional data, particularly those with relatively low sample sizes, this article introduces a novel approach that shrinks variance together with mean parameters, in a more general situation with cluster-specific (diagonal) covariance matrices. Furthermore, selection of grouped variables via inclusion or exclusion of a group of variables altogether is permitted by a specific form of penalty, which facilitates incorporating subject-matter knowledge, such as gene functions in clustering microarray samples for disease subtype discovery. For implementation, EM algorithms are derived for parameter estimation, in which the M-steps clearly demonstrate the effects of shrinkage and thresholding. Numerical examples, including an application to acute leukemia subtype discovery with microarray gene expression data, are provided to demonstrate the utility and advantage of the proposed method.

AMS 2000 subject classifications: Primary 62H30.

Keywords: BIC, EM algorithm, High-dimension but low-sample size, L_1 penalization, Microarray gene expression, Mixture model, Penalized likelihood.

Full Text: PDF

Xie, Benhuai, Pan, Wei, Shen, Xiaotong, Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables, Electronic Journal of Statistics, 2, (2008), 168-212 (electronic). DOI: 10.1214/08-EJS194.

References

[1] Alaiya, A.A. et al. (2002). Molecular classification of borderline ovarian tumors using hierarchical cluster analysis of protein expression profiles. Int. J. Cancer, 98, 895–899.

[2] Antonov AV, Tetko IV, Mader MT, Budczies J, Mewes HW. (2004). Optimization models for cancer classification: extracting gene interaction information from microarray expression data. Bioinformatics, 20, 644–652.

[3] Baker, Stuart G. and Kramer, Barnett S. (2006). Identifying genes that contribute most to good classification in microarrays. BMC Bioinformatics, Sep 7;7:407.

[4] Bardi E, Bobok I, Olah AV, Olah E, Kappelmayer J, Kiss C. (2004). Cystatin C is a suitable marker of glomerular function in children with cancer Pediatric Nephrology, 19, 1145–1147.

[5] Bickel P.J., Levina E. (2004). Some theory for Fisher's linear discriminant function, " naive Bayes", and some alternatives when there are many more variables than observations. Bernoulli, 10, 989–1010. MR2108040

[6] Dempster AP, Laird NM, Rubin DB. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). JRSS-B 39, 1–38. MR0501537

[7] Dudoit S, Fridlyand J, Speed T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc., 97, 77–87. MR1963389

[8] Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. Annals of Applied Statistics. 1, 107–129.

[9] Efron B, Hastie T, Johnstone I, Tibshirani R. (2004). Least angle regression. Annals of Statistics 32, 407–499. MR2060166

[10] Eisen M, Spellman P, Brown P and Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. PNAS 95, 14863–14868.

[11] Friedman, J.H. and Meulman, J.J. (2004). Clustering objects on subsets of attributes (with discussion), J. Royal Statist. Soc. B 66, 1–25. MR2102467

[12] Fraley, C. and Raftery, A.E. (2006). MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering. Technical Report no. 504, Department of Statistics, University of Washington.

[13] Ghosh D, Chinnaiyan, AM. (2002). Mixture modeling of gene expression data from microarray experiments. Bioinformatics, 18, 275–286.

[14] Gnanadesikan, R., Kettenring, J.R. and Tsao, S.L. (1995). Weighting and selection of variables for cluster analysis. Journal of Classification, 12, 113–136.

[15] Golub T et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286, 531–537.

[16] Gu, C. and Ma, P. (2005). Optimal smoothing in nonparametric mixed-effect models. Ann. Statist., 33, 377–403. MR2195638

[17] Hoff, P.D. (2004). Discussion of 'Clustering objects on subsets of attributes,' by J. Friedman and J. Meulman. Journal of the Royal Statistical Society, Series B, 66, 845. MR2102467

[18] Hoff P.D. (2006). Model-based subspace clustering. Bayesian Analysis, 1, 321–344. MR2221267

[19] Huang, X. and Pan, W. (2002). Comparing three methods for variance estimation with duplicated high density oligonucleotide arrays. Functional & Integrative Genomics, 2, 126–133.

[20] Huang, D. and Pan, W. (2006). Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. Bioinformatics, 22, 1259–1268.

[21] Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance selection and estimation via penalised normal likelihood. Biometrika, 93, 85–98. MR2277742

[22] Hubert, L. and Arabie, P. (1985). Comparing partitions. Journal of Classification, 2, 1993–218.

[23] Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res., 28, 27–30.

[24] Kaufman, L. and Rousseeuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York. MR1044997

[25] Kim, S., Tadesse, M.G. and Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. Biometrika, 93, 877–893. MR2285077

[26] Koo, J. Y., Sohn, I., Kim, S., and Lee, J. (2006). Structured polychotomous machine diagnosis of multiple cancer types using gene expression. Bioinformatics, 22, 950–958.

[27] Li H. and Hong F. (2001). Cluster-Rasch models for microarray gene expression data. Genome Biology 2, research0031.1-0031.13.

[28] Liao, J.G. and Chin, K.V. (2007). Logistic regression for disease classification using microarray data: model selection in a large p and small n case. Bioinformatics, 23, 1945–1951.

[29] Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. JRSS-B, 61, 381–400. MR1680318

[30] Liu JS, Zhang JL, Palumbo MJ, Lawrence CE. (2003). Bayesian clustering with variable and transformation selection (with discussion). Bayesian Statistics 7, 249–275. MR2003177

[31] Ma, P., Castillo-Davis, C.I., Zhong, W. and Liu, J.S. (2006). A data-driven clustering method for time course gene expression data. Nucleic Acids Research, 34, 1261–1269.

[32] Mangasarian, OL, Wild EW. (2004). Feature selection in k-median clustering. Proceedings of SIAM International Conference on Data Mining, Workshop on Clustering High Dimensional Data and its Applications, April 24, 2004, La Buena Vista, FL, pages 23–28.

[33] McLachlan, G.J., Bean, R.W. and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. Bioinformatics, 18, 413–422.

[34] McLachlan, G.J. and Peel, D. (2002). Finite Mixture Model. New York, John Wiley & Sons, Inc. MR1789474

[35] McLachlan, G.J., Peel, D. and Bean, R.W. (2003). Modeling high-dimensional data by mixtures of factor analyzers. Computational Statistics and Data Analysis, 41, 379–388. MR1973720

[36] Newton, M.A., Quintana, F.A., den Boon, J.A., Sengupta, S. and Ahlquist, P. (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. Annals of Applied Statistics, 1, 85–106.

[37] Pan, W. (2006). Incorporating gene functional annotations in detecting differential gene expression. Applied Statistics, 55, 301–316. MR2224227

[38] Pan W. (2006b). Incorporating gene functions as priors in model-based clustering of microarray gene expression data. Bioinformatics, 22, 795–801.

[39] Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. Journal of Machine Learning Research, 8, 1145–1164.

[40] Pan, W., Shen, X., Jiang, A., Hebbel, R.P. (2006). Semi-supervised learning via penalized mixture model with application to microarray sample classification. Bioinformatics 22, 2388–2395.

[41] Raftery AE, Dean N. (2006). Variable selection for model-based clustering. Journal of the American Statistical Association, 101, 168–178. MR2268036

[42] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. JASA, 66, 846–850.

[43] Tadesse, M.G., Sha, N. and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. Journal of the American Statistical Association, 100, 602–617. MR2160563

[44] Thalamuthu A., Mukhopadhyay I., Zheng X. and Tseng G.C. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics, 22, 2405–2412.

[45] Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. (2005). Discovering statistically significant pathways in expression profiling studies. PNAS 102, 13544–13549.

[46] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. JRSS-B, 58, 267–288. MR1379242

[47] Tibshirani R, Hastie T, Narasimhan B, Chu G. (2003). Class prediction by nearest shrunken centroids, with application to DNA microarrays. Statistical Science 18, 104–117. MR1997067

[48] Tycko, B., Smith, S.D. and Sklar, J. (1991). Chromosomal translocations joining LCK and TCRB loci in human T cell leukemia. Journal of Experimental Medicine, 174, 867–873.

[49] Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KFX, Mewes HW. (2005). Gene selection from microarray data for cancer classification -a machine learning approach. Comput Biol Chem, 29, 37–46.

[50] Wang, S. and Zhu, J. (2008). Variable Selection for Model-Based High-Dimensional Clustering and Its Application to Microarray Data. To appear in Biometrics.

[51] Wright, D.D., Sefton, B.M. and Kamps, M.P. (1994). Oncogenic activation of the Lck protein accompanies translocation of the LCK gene in the human HSB2 T-cell leukemia. Mol Cell Biol., 14, 2429–2437.

[52] Xie, B, Pan, W. and Shen, X. (2008). Variable selection in penalized model-based clustering via regularization on grouped parameters. To appear in Biometrics. Available at http://www.biostat.umn.edu./rrs.php as Research Report 2007–018, Division of Biostatistics, University of Minnesota.

[53] Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. (2001). Model-based clustering and data transformations for gene expression data. Bioinformatics 17, 977–987.

[54] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. JRSS-B, 68, 49–67. MR2212574

[55] Yuan, M. and Lin, Y. (2007), Model selection and estimation in the Gaussian graphical model. Biometrika, 94, 19–35.

[56] Zhao, P., Rocha, G., Yu, B. (2006). Grouped and hierarchical model selection through composite absolute penalties. Technical Report, Dept of Statistics, UC-Berkeley.

[57] Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. JASA, 101, 1418–1429. MR2279469

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

[59] Zou H, Hastie T, Tibshirani R. (2004). On the “Degrees of Freedom” of the Lasso. To appear Ann. Statistics. Available at http://stat.stanford.edu/~hastie/pub.htm. MR2363967