A note on sensitivity of principal component subspaces and the efficient detection of influential observations in high dimensions

Luke A. Prendergast, *La Trobe University*

Abstract

In this paper we introduce an influence measure based on second order expansion of the RV and GCD measures for the comparison between unperturbed and perturbed eigenvectors of a symmetric matrix estimator. Example estimators are considered to highlight how this measure compliments recent influence analysis. Importantly, we also show how a sample based version of this measure can be used to accurately and efficiently detect influential observations in practice.

AMS 2000 subject classifications: Primary 62F35; secondary 62H12.

Keywords: distance between subspaces, influential observations, perturbation, principal component analysis.

Full Text: PDF

Prendergast, Luke A., A note on sensitivity of principal component subspaces and the efficient detection of influential observations in high dimensions, Electronic Journal of Statistics, 2, (2008), 454-467 (electronic). DOI: 10.1214/08-EJS201.

References

[1] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A.J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA. (1999), 96 6745–6750.

[2] Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A. and Sorensen, D. LAPACK Users' Guide. 3rd Ed. Society for Industrial and Applied Mathematics: Philadelphia, PA. (1999)

[3] Bénasséni, J., Sensitivity coefficients for the subspaces spanned by principal components. Commun. Statist.-Theory Meth. 19, (1990) 2021–2034. MR1086218

[4] Brillinger, D. R., The identification of a particular nonlinear time series system. Biometrika. 64, (1977) 509–515. MR0483236

[5] Brillinger, D. R. A Generalized Linear Model with "Gaussian" Regressor Variables. In: A Festschrift for Erich L. Lehmann, Wadsworth International Group, Belmont, California, (1983) pp. 97–114. MR0689741

[6] Cook, R. D. and Weisberg, S., Discussion of "Sliced Inverse Regression for Dimension Reduction". J. Amer. Statist. Assoc. 86, (1991) 328–332.

[7] Cook, R. D. Regression Graphics: Ideas for Studying Regressions Through Graphics. Wiley, New York. (1998) MR1645673

[8] Cook, R. D. Principal Hessian directions revisited. J. Amer. Statist. Assoc. 93, (1998)

84–94. MR1614584

[9] Critchley, F. Influence in principal components analysis. Biometrika 72, (1985) 627–636. MR0817577

[10] Croux, C. and Haesbroeck, G. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. J. Mult. Anal. 71, (1999) 161–190.

[11] Croux, C. and Haesbroeck, G. Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. Biometrika 87, (2000) 603–618.

[12] Davies, P. L. Asymptotic behaivior of S-estimates of multivariate location parameters and dispersion matrices. Ann. Statist. 15, (1987) 1269–1292. MR0902258

[13] Devlin, S. J., Gnanadesikan, R. and Kettenring, J. R., Robust estimation and Outlier Detection with Correlation Coefficients. Biometrika 62, 531–545.

[14] Enguix-González, A. and Muñoz-Pichardo, J. M and Moreno-Rebollo, J. L. and Pino-Mejías, R. Influence Analysis in Principal Component Analysis through power-series expansions. Commun. Statist.-Theory Meth. 34, (2007) 2025–2046.

[15] Escoufier, Y. Le traitement des variables vectorielles. Biometrics 29, (1973) 751–760. MR0334416

[16] Hampel, F. R., The influence curve and its role in robust estimation. J. Amer. Statist. Assoc. 69, (1974) 383–393. MR0362657

[17] Hampel, F. R, Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A., Robust Statistics: The Approach Based on Influence Functions, New York: Wiley (1986) MR0829458

[18] Li, K.-C., Sliced Inverse Regression for Dimension Reduction (with discussion). J. Amer. Statist. Assoc. 86, (1991) 316–342. MR1137117

[19] Li, K.-C., On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. J. Amer. Statist. Assoc. 87, (1992) 1025–1039. MR1209564

[20] Lopuhaä, H. P., On the relation between S-estimators and M-estimators of multivariate location and covariance. Ann. Statist. 17, (1989) 1662–1683. MR1026304

[21] Lopuhaä, H. P., Asymptotics of reweighted estimators of multivariate location and scatter. Ann. Statist. 27, (1999) 1638–1665. MR1742503

[22] Li, K.-C. and Duan, N., Regression analysis under link violation. Ann. Statist. 17, (1989) 1009–1052.

[23] Prendergast, L. A, Detecting influential observations in Sliced Inverse Regression analysis. Aust. N. Z. J. Stat. 48, (2006) 285–304. MR2328519

[24] Prendergast, L. A., Implications of influence function analysis for sliced inverse regression and sliced average variance estimation. Biometrika. 94, (2007) 585–601.

[25] Prendergast, L. A. and Smith, J. A., Sensitivity of principal Hessian direction analysis. Electronic Journal of Statistics 1, (2007) 253–267 (electronic).

[26] Rellich, F. Perturbation theory of eigenvalue problems. Gordon and Breach (1969) MR0240668

[27] Robert, P. and Escoufier, Y. A unifying tool for linear multivariate statistical methods: the RV coefficient. Appl. Statist. 25, (1976) 257–265.

[28] Rousseeuw, P. J., Multivariate estimation with high breakdown point, In: Mathematical Statistics and Applications. Eds: Grossman, W., Pflug, G., Vincze, I. and

Wertz, W. Vol. B, Reidel: Dordrecht, pp. 283–297 (1985) MR0851060

[29] Rousseeuw, P. J. and Leroy, A. M., Robust regression and outlier detection. Wiley: New York. (1987)

[30] Rousseeuw, P. J. and Yohai, V. J. Robust regression by means of S-estimators. Robust and Nonlinear Time Series Analysis. Lecture Notes in Statist. 26 Springer: New York (1984).

[31] Tanaka, Y. and Castaño-Tostado, E., Quadratic perturbation expansions of certain functions of eigenvalues and eigenvectors and their application to sensitivity analysis in multivariate methods. Commun. Statist.-Theory Meth. 19, (1990) 2943–2965.

[32] Yanai, H. Unification of various techniques of multivariate analysis by means of generalized coefficient of determination (G.C.D.). Behaviour metrics 1, (1974) 45–54.

Electronic Journal of Statistics. ISSN: 1935-7524