Bull. Aust. Math. Soc. **103** (2021), 517–519
doi:10.1017/S0004972720001161

INFLUENCE FUNCTIONS FOR DIMENSION REDUCTION METHODS

JODIE ANN SMITH

(Received 11 August 2020; first published online 30 October 2020)

2020 *Mathematics subject classification*: primary 62F35; secondary 62H12, 62H30.

Keywords and phrases: dimension reduction, eigen-subspace, influence analysis, influence functions, influential observations, linear discriminant analysis, principal Hessian directions, sliced inverse regression.

The capacity of the general-purpose computer to store information has approximately doubled every 40 months since circa 1980 [7]. With near-constant increases in computational capacity and power it is no surprise that there is an overabundance of high-dimensional data. When high-dimensional data sets contain hundreds, even thousands, of explanatory variables, the visualisation and identification of complex regression structures that may be contained within the mountains of data can be difficult. In fact, visualisation and inference already becomes a difficult task when the data exceeds only the third dimension. Some regression information may be obtained by projecting high-dimensional data onto lower-dimensional spaces but, in general, this does not provide useful inference [5]. And these are not the only issues raised by the existence of high-dimensional data. The 'curse of dimensionality' [1, 4] becomes a problem when the number of observations in the sample is not sufficiently high with respect to the dimensions of the data. As a result of this curse, traditional regression methodologies may fail to identify underlying regression structures.

The above issues have given rise to the popularity of dimension reduction methodologies, the purpose of which is to capture the important information contained within any data set in just a few summary measures with minimal loss of information. Therefore, without requiring a prior parametric model, the regression information is contained within the reduced data, which in turn can be explored and viewed via what are called sufficient summary plots [2]. This has resulted in an increased acceptance of dimension reduction methods and thus brings about the need for further understanding with regards to the sensitivity of the associated estimators. For some dimension reduction methods, a consequence is the lack of diagnostics that can be used to detect influential observations.

Thesis submitted to La Trobe University in September 2019; degree approved on 25 February 2020; principal supervisor Luke Prendergast, co-supervisor Robert Staudte.

This thesis consists of three publications, the purpose of which is to use the influence function (IF) [6] to study the robustness properties of dimension reduction methods and to use those IFs to create efficient influence diagnostics.

In the first publication [10], a sensitivity comparison is made for two competing versions of the dimension reduction method applied to the principal Hessian directions (pHd) [9]. These comparisons consider the effects of small perturbations on the estimation of the dimension reduction subspace via the IF. This highlights that the two versions of pHd can behave completely differently in the presence of certain observations. The results also provide evidence that outliers in the traditional sense may or may not be highly influential in practice. Since influential observations may be hidden within typical data, the IF is also considered in the empirical setting for the efficient detection of influential observations.

The second publication [11] provides a sensitivity analysis of pHd, which is then used to highlight how such an analysis can provide valuable insight into the behaviour of the methods. Such insight includes reasons as to why pHd can sometimes return informative results when it is not expected to do so, and why many prefer a residuals-based pHd method over its response-based counterpart. A new influence measure, based on average squared canonical correlations, is then introduced that is applicable to many other dimension reduction methods. This new IF has the advantage that it considers sensitivity on the dimension-reduced predictors and not just the directions that define them. A sample version of the measure is considered, with respect to pHd, and is applied to two example data sets.

And, finally, the third publication [12] extends the new IF from the second publication to the method of linear discriminant analysis (LDA) [3]. Whilst influence functions for LDA have been found for a single discriminant when dealing with two groups, until now these have not been derived in the setting of a general number of groups. The relationship between sliced inverse regression (SIR) [8] and LDA is explored and exploited to develop IFs for LDA for robustness analysis. Efficient sample diagnostics are introduced that are able to detect influential observations in practice, and these are applied to a real data set in order to illustrate their usefulness.

References

[1] R. E. Bellman, *Adaptive Control Processes* (Princeton University Press, Princeton, NJ, 1961).

[2] R. D. Cook, *Regression Graphics: Ideas for Studying Regressions through Graphics* (John Wiley, New York, 1998).

[3] R. A. Fisher, 'The use of multiple measurements in taxonomic problems', *Ann. Eugen.* **7**(2) (1936), 179–188.

[4] J. H. Friedman, 'An overview of computational learning and function approximation', *From Statistics to Neural Networks. Theory and Pattern Recognition Applications* (eds. V. Cherkassky, J. H. Friedman and H. Wechsler) (Springer, Berlin, 1994), 1–61.

[5] U. Gather and C. Becker, 'The curse of dimensionality—a challenge for mathematical statistics', *Jahresber. Dtsch. Math.-Ver.* **103**(1) (2001), 19–36.

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

https://doi.org/10.1017/S0004972720001161 Published online by Cambridge University Press

[7] M. Hilbert and P. López, ‘The world’s technological capacity to store, communicate, and compute information’, *Science* **332** (2011), 60–65.

[8] K. C. Li, ‘Sliced inverse regression for dimension reduction’, *J. Amer. Statist. Assoc.* **86** (1991), 316–342. With comments and a rejoinder by the author.

[9] K. C. Li, ‘On principal Hessian directions for data visualization and dimension reduction: another application of Stein’s lemma’, *J. Amer. Statist. Assoc.* **87** (1992), 1025–1039.

[10] L. A. Prendergast and J. A. Smith, ‘Sensitivity of principal Hessian direction analysis’, *Electron. J. Stat.* **1** (2007), 253–267.

[11] L. A. Prendergast and J. A. Smith, ‘Influence functions for dimension reduction methods: an example influence study of principal Hessian direction analysis’, *Scand. J. Stat.* **37**(4) (2009), 588–611.

[12] L. A. Prendergast and J. A. Smith, ‘Influence functions for linear discriminant analysis: sensitivity analysis and efficient influence diagnostics’, Preprint, 2019, arXiv:1909.13479.

JODIE ANN SMITH, School of Engineering and Mathematical Sciences,
La Trobe University, Melbourne, VIC 3086, Australia
e-mail: ja12smith@gmail.com

https://doi.org/10.1017/S0004972720001161 Published online by Cambridge University Press