Sliced Inverse Regression for big data analysis

Li Kevin

Sliced Inverse Regression for big data analysis

Kevin B. Li

Beijing Jiaotong University

Abstract

Modem advances in computing power have greatly widened scientists' scope in gathering and investigating information from many variables. We describe sliced inverse regression (SIR), for reducing the dimension of the input variable x without going through any parametric or nonparametric model-fitting process. This method explores the simplicity of the inverse view of regression. Instead of regressing the univariate output variable y against the multivariate x, we regress x against y. Forward regression and inverse regression are connected by a theorem that motivates this method. The theoretical properties of SIR are investigated under a model of the form, $y = f(\beta_1' x, \beta_2' x, \ldots, \beta_K' x, \epsilon)$ where the β's are unknown vectors. This model looks like a nonlinear regression, except for the crucial difference that the functional form off is completely unknown. For effectively reducing the dimension, one only needs to estimate the effective dimension reduction (e.d.r.) space generated by the β's. If the distribution of x has been standardized to have the zero mean and the identity covariance, the inverse regression curve falls into the e.d.r. space. Hence a principal component analysis on the covariance matrix for the estimated inverse regression curve can be conducted to locate its main orientation, yielding our estimates for e.d.r. directions. Furthermore, a simple step function can be used to estimate the inverse regression curve.

Regression is a popular way of studying the relationship between a response variable y and its explanatory variable x, a p-dimensional vector. Quite often, a parametric model is used to guide the analysis. When the model is parsimonious, standard estimation techniques such as the maximum likelihood or the least squares method have proved to be successful. In most applications, however, any parametric model is at best an approximation to the true one, and the search for an adequate model is not easy. When there are no persuasive models available, nonparametric regression techniques emerge as promising alternatives that offer the needed flexibility in modeling. A common theme of nonparametric regression is the idea of local smoothing, which explores only the continuity or differentiability property of the true regression function. The success of local smoothing hinges on the presence of sufficiently many data points around each point of interest in the design space to provide adequate information. For one-dimensional problems, many smoothing techniques are available. As the dimension of x increases, however, the total number of observations needed for local smoothing grows exponentially. Unless one has a huge sample, standard methods, such as kernel estimates or nearest-neighbor estimates, break down quickly because of the sparseness of the data points in any region of interest. To challenge the curse of dimensionality, one hope that statisticians may capitalize on is that interesting features of high-dimensional data are retrievable from low-dimensional projections. For regression problems, the following model describes such an ideal situation:

$$y = f(\beta_1' x, \beta_2' x, \ldots, \beta_K' x, \epsilon).$$

Here the β's are unknown vectors, ϵ is independent of x, and f is an arbitrary unknown function. When this model holds, the projection of the p-dimensional explanatory variable x onto the K dimensional subspace, $(\beta_1' x, \beta_2' x, \ldots, \beta_K' x)$ captures all the information about y. When K is small, one may achieve the goal of data reduction by estimating the β's efficiently. For convenience, we shall refer to any linear combination of the β's as an effective dimension-reduction (e.d.r.) direction, and to the linear space B generated by the β's as the e.d.r. space. The main focus is on the estimation of the e.d.r. directions, leaving questions such as how to estimate main features of f for further investigation. Intuitively speaking, after estimating the e.d.r. directions, standard smoothing techniques can be more successful because the dimension has been lowered. On the other hand, during the exploratory stage of data analysis, one often wants to view the data directly. Many graphical tools are available but plotting y against every combination of x within a reasonable amount of time is impossible. So, to use the scatterplot-matrix techniques, one often focus on coordinate variables only. Likewise, 3D rotating plots can handle only one two-dimensional projection of x at a time (the third dimension is reserved for y). Therefore, to take full advantage of modem graphical tools, guidance on how to select the projection directions is clearly called for. A good estimate of the e.d.r. directions can lead to a good view of the data. Our method of estimating the e.d.r. directions is based on the idea of inverse regression. Instead of regressing y against x (forward regression) directly, x is regressed against y (inverse regression). The immediate benefit for exchanging the roles of y and x is that one can overcome the dimensionality problem. This comes out because inverse regression can be carried out by regressing each coordinate of x against y. Thus, one essentiallys deal with a one-dimension to one-dimension regression problem, rather than the high-dimensional forward regression. As y varies, $E(x|y)$ draws a curve, called the inverse regression curve. This curve typically hovers around a K- dimensional affine subspace. At one extreme, the inverse regression curve actually falls into a K-dimensional affine subspace determined by the e.d.r. directions. If x is standardized x to have mean 0 and the identity covariance, then this subspace coincides with the e.d.r. space. Exploring the simplicity of inverse regression, a simple algorithm is proposed, called sliced inverse regression (SIR), for estimating the e.d.r. directions. After standardizing x, SIR proceeds with a crude estimate of the inverse regression curve $E(x|y)$, which is the slice mean of x after slicing the range of y into several intervals and partitioning the whole data into several slices according to the y value. A principal component analysis is then applied to these slice means of x, locating the most important K-dimensional subspace for tracking the inverse regression curve $E(x|y)$. The output of SIR is these components after an affine re- transformation back to the original scale. Besides offering estimates of e.d.r. directions, the outputs of SIR are themselves interesting descriptive statistics containing useful information about the inverse regression curve. As a sharp contrast to most nonparametric techniques that require intensive computation, SIR is very simple to implement. Moreover, the sampling property of SIR is easy to understand, another advantage over other methods. Thus it is possible to assess the effectiveness of SIR by using the companion output eigenvalues at the principal component analysis step. These eigenvalues provide valuable information for assessing the number of components in the data. Finally, selection of the number of slices for SIR is less crucial than selection of the smoothing parameter for typical nonparametric regression problems. In view of these virtues, however, SIR is not intended to replace other computer-intensive methods. Rather it can be used as a simple tool to aid other methods; for instance, it provides a good initial estimate for many methods based on the forward regression viewpoint.

References

[1] Aragon, Y. (1997). A Gauss implementation of multivariate sliced inverse regression. *Computational Statistics*, 12, 355–372.

[2] Aragon, Y. and Saracco, J. (1997). Sliced Inverse Regression (SIR): an appraisal of small sample alternatives to slicing. *Computational Statistics*, 12, 109–130.

[3] Azaïs, R., Gégout-Petit, A., and Saracco, J. (2012). Optimal quantization applied to sliced inverse regression. *Journal of Statistical Planning and Inference*, 142, 481–492.

[4] Barreda, L., Gannoun, A., and Saracco, J. (2007). Some extensions of multivariate SIR. *Journal of Statistical Computation and Simulation*, 77, 1–17.

[5] Bentler, P.M., Xie, J., (2000). Corrections to test statistics in principal Hessian directions. *Statistics and Probability Letters*, 47, 381-389.

[6] Bernard-Michel, C., Gardes, L. and Girard, S. (2008). A Note on Sliced Inverse Regression with Regularizations, *Biometrics*, 64, 982–986.

[7] Bernard-Michel, C., Douté, S., Fauvel, M., Gardes, L. and Girard, S. (2009). Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression. *Journal of Geophysical Research - Planets*, 114, E06005.

[8] Bernard-Michel, C., Gardes, L. and Girard, S. (2009). Gaussian Regularized Sliced Inverse Regression, *Statistics and Computing*, 19, 85–98.

[9] Bura, E. and Cook, R. D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society, Series B*, 63, 393–410.

[10] Chavent, M., Kuentz, V., Liquet, B., and Saracco, J. (2011). A sliced inverse regression approach for a stratified population. *Communications in statistics - Theory and methods*, 40, 1–22.

[11] Chavent, M., Girard, S., Kuentz, V., Liquet, B., Nguyen, T.M.N. and Saracco, J. (2014). A sliced inverse regression approach for data stream. *Computational Statistics*, 29, 1129–1152.

[12] Chen, C.-H. and Li, K.-C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica*, 8, 289–316.

[13] Cook, R.D., Weisberg, S., (1991). Discussion of 'sliced inverse regression for dimension reduction'. *Journal of the American Statistical Association*, 86, 328-332.

[14] Cook, R.D., (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. Proceedings of the Section on Physical and Engineering Sciences. Alexandria, VA: American Statistical Association. 18-25.

[15] Cook, R.D., 1996. Graphics for regressions with a binary response. *Journal of the American Statistical Association*, 91, 983-992.

[16] Cook, R. D.(1998). Principal hessian directions revisited (with discussion). *Journal of the American Statistical Association*, 93, 84–100.

[17] Cook, R.D., (1998b). Regression Graphics, Ideas for Studying Regressions through Graphics. Wiley, New York.

[18] Cook, R.D., Lee, H., (1999). Dimension-reduction in binary response regression. *Journal of the American Statistical Association*, 94, 1187-1200.

[19] Cook, R. D. (2000). SAVE: a method for dimension reduction and graphics in regression. *Communications in statistics - Theory and methods*, 29, 2109–2121.

[20] Cook, R.D., Critchley, F., (2000). Identifying regression outliers and mixtures graphically. *Journal of the American Statistical Association*, 95, 781-794.

[21] Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30, 450–474.

[22] Cook, R.D., Ni, L., (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, 100, 410–428.

[23] Coudret, R., Girard, S. and Saracco, J. (2014). A new sliced inverse regression method for multivariate response regression, *Computational Statistics and Data Analysis*, 77, 285–299.

[24] Duan, N. and Li, K.-C. (1991). Slicing regression: a link-free regression method. *The Annals of Statistics*, 19, 505–530.

[25] Dunia, R. and Joe Qin, S. (1998). Subspace approach to multidimensional fault identification and reconstruction. *AIChE Journal*, 44, 1813–1831.

[26] Ferré, L. (1998). Determining the dimension in sliced inverse regression and related methods. *Journal of the American Statistical Association*, 93, 132–140.

[27] Gannoun, A., Girard, S., Guinot, C. and Saracco, J. (2002). Reference ranges based on nonparametric quantile regression, *Statistics in Medicine*, 21, 3119-3135.

[28] Gannoun, A., Girard, S., Guinot, C. and Saracco, J. (2004). Sliced inverse regression in reference curves estimation. *Computational Statistics and Data Analysis*, 46, 103–122.

[29] Gannoun, A. and Saracco, J. (2003). An asymptotic theory for SIR_α method. *Statistica Sinica*, 13, 297–310.

[30] Hall, P. and Li, K. C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, 21, 867-889.

[31] Hsing, T. (1999). Nearest neighbor inverse regression. *The Annals of Statistics*,27, 697–731.

[32] Hsing, T. and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *The Annals of Statistics*, 20, 1040–1061.

[33] Kuentz, V., Liquet, B., and Saracco, J. (2010). Bagging versions of sliced inverse regression. *Communications in statistics - Theory and methods*, 39, 1985–1996.

[34] Kuentz, V. and Saracco, J. (2010). Cluster-based sliced inverse regression. *Journal of the Korean Statistical Society*, 39, 251–267.

[35] Li, K.C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86, 316-342.

[36] Li, K.C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Steins lemma. *Journal of the American Statistical Association*, 87, 1025–1039.

[37] Li, K.-C., Aragon, Y., Shedden, K., and Agnan., C. T. (2003). Dimension reduction for multivariate response data. *Journal of the American Statistical Association*, 98, 99–109.

[38] Li, L. and Nachtsheim, C. J. (2006). Sparse sliced inverse regression. *Technometrics*, 48, 503–510.

[39] Li, L., Cook, R. D. and Tsai, C. L. (2007) Partial inverse regression. *Biometrika*, 94, 615–625.

[40] Li, B., Wang, S., (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102, 997-1008.

[41] Li, L. and Yin, X. (2008). Sliced inverse regression with regularizations, *Biometrics*, 64, 124–131.

[42] Li, B. and Wang, S. (2007). On directional regression for dimension reduction, *Journal of the American Statistical Association*, 102, 997–1008.

[43] Li, K.B. (2013). Invariance properties of Sliced Inverse Regression, `http://hal.archives-ouvertes.fr/hal-00805491`

[44] Li, K.B. (2013). A review on Sliced Inverse Regression, `http://hal.archives-ouvertes.fr/hal-00803698`

[45] Li, K.B. (2013b). Some limitations of Sliced Inverse Regression, `http://hal.archives-ouvertes.fr/hal-00803602`

[46] Liquet, B. and Saracco, J. (2008). Application of the bootstrap approach to the choice of dimension and the α parameter in the SIR_α method. *Communications in statistics - Simulation and Computation*, 37, 1198–1218.

[47] Liquet, B. and Saracco, J. (2012). A graphical tool for selecting the number of slices and the dimension of the model in SIR and SAVE approaches. *Computational Statistics*, 27, 103–125.

[48] Lue, H.-H. (2009). Sliced inverse regression for multivariate response regression. *Journal of Statistical Planning and Inference*, 139, 2656–2664.

[49] Naik, P. and Tsai, C. L. (2000) Partial least squares estimator for single-index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 763–771.

[50] Nkiet, G.-M. (2008). Consistent estimation of the dimensionality in sliced inverse regression. *Annals of the Institute of Statistical Mathematics*, 60, 257–271.

[51] Prendergast, L. A. (2005). Influence functions for sliced inverse regression. *Scandinavian Journal of Statistics*, 32, 385–404.

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

[53] Saracco, J. (1997). An asymptotic theory for Sliced Inverse Regression. *Communications in statistics - Theory and methods*, 26, 2141–2171.

[54] Saracco, J. (1999). Sliced inverse regression under linear constraints. *Communications in statistics - Theory and methods*, 28, 2367–2393.

[55] Saracco, J. (2001). Pooled slicing methods versus slicing methods. *Communications in statistics - Simulation and Computation*, 30, 489–511.

[56] Saracco, J. (2005). Asymptotics for pooled marginal slicing estimator based on SIR_α approach. *Journal of Multivariate Analysis*, 96 117–135.

[57] Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association*, 89, 141–148.

[58] Scrucca, L. (2007). Class prediction and gene selection for DNA microarrays using regularized sliced inverse regression. *Computational Statistics & Data Analysis*, 52, 438–451.

[59] Setodji, C. M. and Cook, R. D. (2004). K-means inverse regression. *Technometrics*, 46, 421–429.

[60] Shao, Y., Cook, R.D., Weisberg, S., (2007). Marginal tests with sliced average variance estimation. *Biometrika*, 94, 285-296.

[61] Shao, Y., Cook, R. D., and Weisberg, S. (2009). Partial central subspace and sliced average variance estimation. *Journal of Statistical Planning and Inference*, 139, 952–961.

[62] Szretter, M. E. and Yohai, V. J. (2009). The sliced inverse regression algorithm as a maximum likelihood procedure. *Journal of Statistical Planning and Inference*, 139, 3570–3578.

[63] Weisberg, S., (2002). Dimension reduction regression in R. *Journal of Statistical Software*, Available from `http://www.jstatsoft.org`

[64] Wen, X., Cook, R.D., (2007). Optimal sufficient dimension reduction in regressions with categorical predictors. *Journal of Statistical Inference and Plannin*, 137, 1961-1979.

[65] Wu, H.-M. (2008). Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics*, 17, 590–610.

[66] Ye, Z. and Yang, J. (2010). Sliced inverse moment regression using weighted chi-squared tests for dimension reduction. *Journal of Statistical Planning and Inference*, 140, 3121–3131.

[67] Ye, Z., Weiss, R.E., (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98, 968-979.

[68] Yin, X., Cook, R.D., (2002). Dimension reduction for the conditional kth moment in regression. *Journal of the Royal Statistical Society*, Ser. B. 64, 159-175.

[69] Yin, X., Cook, R.D., (2003). Estimating central subspaces via inverse third moments. *Biometrika*, 90, 113-125.

[70] Yin, X. and Bura, E. (2006). Moment-based dimension reduction for multivariate response regression. *Journal of Statistical Planning and Inference*, 136, 3675–3688.

[71] Zhong, W., Zeng, P., Ma, P., Liu, J.S. and Zhu, Y. (2005). RSIR: Regularized Sliced Inverse Regression for motif discovery. *Bioinformatics*, 21, 4169–4175.

[72] Zhu, L., Miao, B., and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, 101, 630–643.

[73] Zhu, L.-P. and Yu, Z. (2007). On spline approximation of sliced inverse regression. *Science in China Series A: Mathematics*, 50, 1289–1302.

[74] Zhu, L. X. and Fang, K. T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics*, 24, 1053–1068.

[75] Zhu, L. X. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statistica Sinica*, 5, 727–736.

[76] Zhu, L. X., Ohtaki, M., and Li, Y. (2007). On hybrid methods of inverse regression-based algorithms. *Computational Statistics*, 51, 2621–2635.