Sliced Inverse Regression for Dimension Reduction: Comment

John T. Kent

Journal of the American Statistical Association, Vol. 86, No. 414. (Jun., 1991), pp. 336-337.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199106%2986%3A414%3C336%3ASIRFDR%3E2.0.CO%3B2-P

Journal of the American Statistical Association is currently published by American Statistical Association.

Comment

JOHN T. KENT*

1. OVERVIEW

Sliced inverse regression (SIR) looks like a novel and fascinating method for the analysis of multivariate data. In my discussion, I would like to look at the method more geometrically and to bring out some links with other well established methods of statistical analysis. It will be useful to split the discussion into two parts: a restatement of the theoretical model and a description of the numerical calculations.

First though, some notation from linear algebra is needed. If $U \subset \mathbf{R}^p$ is a K-dimensional subspace, let $\mathbf{P}_U$ denote the $p \times p$ symmetric idempotent matrix representing orthogonal projection onto U. In particular, for all vectors $\mathbf{z} \in \mathbf{R}^p$ (regarding $\mathbf{z}$ as a column vector), $P_U\mathbf{z} \in U$, and, further, if $\mathbf{z} \in U$, then $P_U\mathbf{z} = \mathbf{z}$. Next, let $U_\perp$ denote the complementary orthogonal subspace to U in $\mathbf{R}^p$. Then $\mathbf{P}_U + \mathbf{P}_{U_\perp} = \mathbf{I}_p$, the $p \times p$ identity matrix, and $\mathbf{z} = \mathbf{P}_U\mathbf{z} + \mathbf{P}_{U_\perp}\mathbf{z}$ represents the unique decomposition of $\mathbf{z}$ into a sum of an element of U plus an element of $U_\perp$.

2. THEORETICAL MODEL BEHIND SIR

Let $(y, \mathbf{x})$ be a $(p + 1)$-dimensional random vector, with $y \in R^1$, and $\mathbf{x} \in R^p$. Let Σ denote the covariance matrix of $\mathbf{x}$ and set $\mathbf{z} = \Sigma^{-1/2}\mathbf{x}$, so that the covariance matrix of $\mathbf{z}$ is $\mathbf{I}_p$. With some slight simplification, Li's model involves the following assumptions. (a) The standardized random vector $\mathbf{z}$ is spherically distributed about its mean. (b) For some subspace $U \subset R^p$ of dimension K say, $\mathbf{y}$ is conditionally independent of $\mathbf{P}_{U_\perp}\mathbf{z}$, given $\mathbf{P}_U\mathbf{z}$. (This is the underlying property enabling dimension reduction.) (c) The span of $E(\mathbf{z} \mid y)$, as y varies, equals all of U. (d) The conditional expectation $E(\mathbf{z} \mid y)$ is continuous in y.

Li's key observation is that under assumptions (a) and (b), $E(\mathbf{z} \mid y) \in U$. Assumption (c) is a strengthened form of this result. Assumption (d) forms the basis of his numerical procedure for estimating this conditional expectation in practice.

Assumption (b) can be rephrased in terms of the original $\mathbf{x}$ variables. Indeed it can be argued that the subspaces $\Sigma^{1/2}U = U^*$, say, and $\Sigma^{1/2}U_\perp = U_\perp^*$, say, in terms of the original $\mathbf{x}$ variables, are more natural than $\mathbf{U}$ and $\mathbf{U}_\perp$. Some care is needed, however, because U^* and $U_\perp^*$ will no longer necessarily be orthogonal.

If $(y, \mathbf{x})$ follows a $(p + 1)$-dimensional normal distribution, with y not independent of $\mathbf{x}$, then $E(\mathbf{z} \mid y)$ will be linear in y (i.e., $K = 1$). Thus for $K > 1$, SIR can be regarded as a method for detecting certain sorts of nonlinearity in the regression.

To compare two K-dimensional subspaces in z space, U and V say, a natural distance can be given in terms of the projection matrices,

$$d^2(U, V) = \mathrm{tr}(\mathbf{P}_U - \mathbf{P}_V)^2 = 2(K - \mathrm{tr}\mathbf{P}_U\mathbf{P}_V).$$

Furthermore, it can be shown that $K^{-1}\mathrm{tr}\mathbf{P}_U\mathbf{P}_V$ reduces to the trace correlation mentioned by Li in Section 2.

3. NUMERICAL CALCULATIONS IN SIR

Let $(y_i, \mathbf{x}_i)$ $(i = 1, \ldots, n)$ be a set of data from the above model. Li suggests "slicing" the y values into H groups, say. Then the SIR algorithm is the same as a classical multivariate discriminant analysis on the x variables using these H groups; see for example, Mardia et al. (1979, chaps. 11 and 12). This point of view is hinted at in Section 4. Thus let $\mathbf{W}_x$, $\mathbf{B}_x$, and $\mathbf{T}_x$ denote the usual $p \times p$ matrices for the "within-groups," "between-groups," and "total" sum of squares and products matrices for the x variables, with $\mathbf{W}_x + \mathbf{B}_x = \mathbf{T}_x$. If K is known, the subspace U^* is estimated by the span of the eigenvectors corresponding to the K largest eigenvalues of $\mathbf{T}_x^{-1}\mathbf{B}_x$, or equivalently of $\mathbf{W}_x^{-1}\mathbf{B}_x$. The linear combinations of $\mathbf{x}$ defined by these eigenvectors are known as canonical variates.

Geometric insight into these canonical variates can be enhanced by a judicious transformation. Let $\mathbf{z}_i = \mathbf{T}_x^{-1/2}\mathbf{x}_i$, so that $\mathbf{T}_z = \mathbf{I}_p$, where in an obvious notation $\mathbf{T}_z$ denotes the total sum of squares and products matrix in the z variables. This transformation, up to a proportionality constant, is the sample counterpoint to the standardization used above in the population model. In terms of the z variables, the desired subspace is determined by the first K eigenvectors of $\mathbf{B}_z$. A rotation to these eigenvectors in the z variables is essentially the same as that used by Li in his graphical displays of Section 6.3.

On the other hand, in discriminant analysis it is more usual to make a transformation to $\mathbf{v}_i = \mathbf{W}_x^{-1/2}\mathbf{x}_i$, for which $\mathbf{W}_\mathrm{v} = I_\mathrm{p}$. Again, the desired subspace is found by taking dominant eigenvectors, this time of $\mathbf{B}_\mathrm{v}$. The eigenvectors are the same as in the preceding paragraph but with different eigenvalues. Hence a plot of the canonical variates is the same as before but with different scalings for the axes. It would be interesting to know whether plots of the data in the v variables offer any extra insight beyond plots in the z variables. It might also be helpful to mark the H groups separately in the plots.

In conventional discriminant analysis, it is assumed that the population within-group covariance matrix is the same for all groups. In the present context this is the same as assuming that $\mathrm{cov}(\mathbf{z} \mid y)$ does not depend on y. Li points out that this homogeneity assumption is not needed for SIR,

* John T. Kent is Professor, Department of Statistics, University of Leeds, Leeds LS2 9JT, England.

Journal of the American Statistical Association
June 1991, Vol. 86, No. 414, Theory and Methods

and indeed he suggests that any heterogeneity might be used to help estimate U when the span of $E(\mathbf{z} \mid y)$ is not all of U. It would be interesting to see how these ideas work out in practice.

4. OPEN QUESTIONS

Let me finish with some questions about the likely behavior of SIR in practice and some issues that need more careful study.

1. How heavily does the performance of SIR depend on the sphericity assumption on $\mathbf{z}$? Is a violation of sphericity likely to be a problem in practice?

2. What is the effect of changing the number of slices H? Clearly a large H will cut down the variability in $\mathbf{B}_x$, whereas a small value of H will cut down the variability in $\mathbf{W}_x$. The pleasing results from the simulation study may be due merely to the relatively large sample sizes. I suspect some normal theory calculations might be able to offer some quantitative insight into an optimal choice of H.

3. The conditional expectation $E(\mathbf{z} \mid y)$ may be of inherent interest, and it should be plotted along with the other data summaries. SIR essentially fits a piecewise constant function to this conditional expectation as y varies. Other fits would also be of interest, such as splines. Indeed something like a spline fit might be used to generalize the whole SIR procedure.

Lastly, I look forward with interest to seeing some real examples where the use of SIR has enhanced the interpretation of the data.

Rejoinder

KER-CHAU LI

First, I would like to thank the discussants for their thought-provoking comments. I appreciate their support on SIR, as evidenced by the richness of their discussions in highlighting some obscure facets of SIR, in demonstrating SIR's power, and in proposing several extensions. I agree with them that this article is just the beginning of something that might evolve into routine practice in data analysis. There is much to be done to reach that point. Since the idea of SIR was conceived, I have gathered a string of related ideas and results. I am pleased to find some of these in agreement with key suggestions from the discussants.

For example, the connection with classical discriminant analysis suggested by Kent was addressed in Li (1989). Chun-Houh Chen is now working on SIR's application in the classification tree context. He is also working with me on SIRII, second-moment based SIR, which appears to have a good deal of overlap with the SAVE suggested by Cook and Weisberg. The proposals by Härdle and Tsybakov based on a different viewpoint are stimulating in building up a better theory for dimension reduction and data visualization.

Another shortcoming of this article, the application of SIR to real data, was remedied by several examples in Cook and Weisberg's discussion. To further ease the reader's mind on the applicability of SIR, let me briefly comment on my own efforts in this vein, reported elsewhere. For instance, the Boston housing data (Harrison and Rubinfeld 1978) are treated in Li (1989), where, with SIR, we reduced the number of regressors from thirteen to three and found a slide- (or helix-) looking data cloud. In Li (1990a), a six-variable function describing the voltage level of a push–pull circuit in television manufacturing was visualized by SIR. Li (1990b) demonstrated how SIR could be applied to the residual analysis for the Los Angeles ozone data (Breiman and Friedman 1985). Regarding small data sets, the worsted yarn data (Box and Cox 1964), which has 27 observations for a 3^3 factorial design, was reanalyzed with SIR, recovering the logarithm transformation of y well.

In the following, I will first concentrate on three major issues raised by the discussants: (1) design condition, (2) second moment SIR (SIRII), and (3) distribution of eigenvalues. After that, I will respond to each discussant separately. The last section is added to address Brillinger's discussion, which arrived late.

1. DESIGN CONDITION

I agree with all discussants that the most controversial condition in this article is (3.1). As Cook and Weisberg have explicitly pointed out, in order to guarantee this condition *before analyzing the data,* we need to check if $\mathbf{x}$ is elliptically symmetric. I would like to reemphasize, however, that (3.1) is in fact much weaker than the elliptic symmetry because the linear conditional expectation only needs to hold for the β_k's that are in the e.d.r. space. Thus if we are lucky, we can still have (3.1) without elliptic symmetry. Cook and Weisberg gave a nice illustration of how this might happen. But, of course, the first question is how often can we be so lucky? The next question is what to do if we are not. Both will be discussed here.

Journal of the American Statistical Association
June 1991, Vol. 86, No. 414, Theory and Methods