What is SIR/PHD?

The need for statistical guidance on informative profiles for plotting.

Despite of the revolutionary accomplishment in how to plot, it has long been recognized within the statistics community that a greater challenge in dealing the large dimensional data lies on how to find informative profiles for plotting. For example, imagine that one has 10 variables (say, housing price, crime rate, number of rooms, lot size, and so on) in a data set of 400 houses randomly collected in a metropolitan study and wishes to study their relationship by 3-D plotting. Then there are already $\binom{10}{3} = 1200$ such plots to examine. This is quite a daunting task for visualization even though the data set is no more than a tiny fraction of a megabyte. Clearly some statistical profiles are more informative than others. But the question is how to find the right ones to "see" first?

Example: A simulated data set.

Consider the function:

$$y = f(x, \beta_1, \beta_2) = \frac{\beta_1' x}{2 + (\beta_2' x + 3)^2} \tag{1}$$

Here $x = (x_1, \cdots, x_p)'$ denotes p input variables. Suppose a data set with a total of 11 variables $y, x_1, \cdots, x_{10}$ and 400 cases is generated. Each case is created by first generating $p = 10$ random numbers from the standard normal distribution $N(0,1)$ independently, recording them as $x_1, \cdots, x_{10}$, and then plugging them in (1) with $\beta_1 = (1,1,1,1,0,0,0,0,0,0)', \beta_2 = (0,0,0,0,0,1,1,1,1,0)'$ to obtain the corresponding y value for that case. Plotting of y against any pair of coordinate variables, say x_1 and x_2, can only give a picture which is too shadowy to reveal the nonlinear structure in the data; see Figure 2.1, first row. So a critical question is how to find the better projection angles without knowing where the data were generated from.

The directions of β_1 and β_2 are called effective dimension reduction (e.d.r.) directions because the relationship between y and x hinges entirely on the associated projections $\beta_1' x$ and $\beta_2' x$. For visualization purpose, they are the most informative statistical profiles to find the nonlinear structure of the data . If these two directions are given, then we can visualize perfectly the functional relationship between y and x by plotting y against $\beta_1' x$ and $\beta_2' x$; see Figure 2.1, second row. Note that any linear combinations of β_1 and β_2 can also be called e.d.r. directions because one can rewrite equation (1) in terms of these vectors. However, the space spanned by the e.d.r. directions, called the e.d.r. space, is unique.

Finding the unknown β vectors from the data set is not difficult at all, *if we are given the true model* f. This can be done by the traditional least squares method; namely to minimize $\sum_{i=1}^{400}(y_i - f(x_i, \beta_1, \beta_2))^2$. However, the rationale behind such a procedure, which imposes strict model conditions, has limited the scope of data visualization. One can argue that if an imposed model is not a good approximation, then the directions obtained by least squares fitting may not be very meaningful (see Li and Duan 1989, however). On the other hand, if the model is known to be a reasonable approximation a priori, then the need for visualization is no longer that pressing - because what patterns that come out would be more or less anticipated by the model.

SIR and PHD

The development of SIR (sliced inverse regression) and PHD (principal Hessian directions) originates from the need for finding a domain-free methodology for discovering nonlinear structure in noisy data with several dimensions. In Li(1991), a dimension reduction protocol is introduced for modeling the relationship between a response variable y and the multi-dimensional input variable x :

$$y = f(\beta_1' x, \cdots, \beta_k' x, \epsilon) \tag{2}$$

Here the variable ϵ introduces a source for random noise. This model includes (1) as a special case. The distinctive feature of this model is that the structural relationship between dependent variable y and the p dimensional regressor x is completely unspecified because f is unknown and so is the distribution of the independent error term ϵ. What is assumed is only the existence of a small number k of directions, $\beta_1, \cdots, \beta_k$ to reduce the dimension of x. The space spanned by these beta vectors is called *effective dimension reduction* (e.d.r.) space. Since k can also be estimated from the data, (1) is completely general. Unlike other curve-fitting models, it does not impose any scope-limiting conditions on the regression surface (Cook 1998b). Further investigation on the concept of e.d.r. space is pursued by Cook (1994a).

SIR and PHD offer very simple and efficient ways of estimating e.d.r. directions; Figures 2.1 third row. It agrees very well with the most informative plot. SIR involves only a simple slicing step on y to be followed by a principal component analysis type of eignevalue decomposition. SIR requires no iteration. There are no model fitting or searching steps involved. The algorithm of PHD takes a similar form of eigenvalue decomposition.

The SIR algorithm.

Step 1. Sort the data by y and divide the data set into H slices so that the cases within the same slice have similar y values. Let n_h be the number of cases in slice h. The number of slices H is a user-specified parameter.

Step 2. Within each slice, compute the sample mean of x, $\bar{x}_h = n_h^{-1} \sum_{(i) \in \text{slice}\, h} x_{(i)}$.

Step 3. Compute the covariance matrix for the slice means of x, weighted by the slice sizes:

$$\hat{\Sigma}_\eta = n^{-1} \sum_{h=1}^{H} n_h (\bar{x}_h - \bar{x})(\bar{x}_h - \bar{x})'$$

Here $\bar{x}$ denotes sample mean of $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$.

Step 4 . Compute the sample covariance for x_i's, $\hat{\Sigma}_x = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'$.

Step 5. Find the SIR directions by conducting the eigenvalue decomposition of $\hat{\Sigma}_\eta$ with respect to $\hat{\Sigma}_x$:

$$\hat{\Sigma}_\eta \hat{\beta}_i = \hat{\lambda}_i \hat{\Sigma}_x \hat{\beta}_i$$

$$\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$$

The i-th eigenvector $\hat{\beta}_i$ is called the i-th SIR direction. The first few SIR directions can be used for dimension reduction.

Step 6. Project x along the SIR directions; that is, use each SIR direction to form a linear combination of x. We shall call $\hat{\beta}_1' x$ the first SIR variate, $\hat{\beta}_2' x$ the second SIR variate, and so on.

Step 7. Plot y against the SIR variates.