Cook, R. D., and Forzani, L., "Likelihood-Based Sufficient Dimension Reduction," *Journal of the American Statistical Association*, 104, 197–208.

Cook and Forzani (2009), hereafter CF09, propose a likelihood-based method for dimension reduction that advances existing methods such as SIR (Li 1991) and SAVE (Cook and Weisberg 1991). CF09 acknowledge that their likelihood-based method is related to the method of Zhu and Hastie (2003), hereafter ZH03, but add that "[their] full maximum likelihood estimator (MLE) of the central subspace under normality would prove to have advantages over ... the Zhu–Hastie method under the same assumptions." In this letter, we explain more fully the relationship between the methods in CF09 and ZH03.

In CF09, continuous responses are first discretized into several groups. Thus, we assume without loss of generality that the response variable y is categorical, say $y = 1, 2, \ldots, h$. The relevant notation from CF09 we need includes: $\tilde{\boldsymbol{\Sigma}} = \widehat{\mathrm{Var}}(\mathbf{x})$; $\tilde{\boldsymbol{\Delta}}_y = \widehat{\mathrm{Var}}(\mathbf{x}|y)$; $\mathbf{P}_{\mathcal{S}}$ = projection onto subspace $\mathcal{S}$; $|\mathbf{A}|_0$ = products of nonzero eigenvalues of $\mathbf{A}$; n = total sample size; and n_y = number of observations in the subgroup indexed by y.

Let $\mathcal{S}$ be a dimension-reduction subspace, and suppose $\dim(\mathcal{S}) = d$ is given. CF09 propose estimating $\mathcal{S}$ by maximizing

$$L_d^*(\mathcal{S}) = \log|\mathbf{P}_{\mathcal{S}}\tilde{\boldsymbol{\Sigma}}\mathbf{P}_{\mathcal{S}}|_0 - \sum_{y=1}^{h}\frac{n_y}{n}\log|\mathbf{P}_{\mathcal{S}}\tilde{\boldsymbol{\Delta}}_y\mathbf{P}_{\mathcal{S}}|_0 \qquad (1)$$

with respect to $\mathcal{S}$. Our Equation (1) above is CF09's equation (1), except that additive terms not depending on $\mathcal{S}$ are removed and the remaining terms are scaled by a factor of $2/n$. Neither difference alters the maximization problem.

In ZH03, we proposed sequentially maximizing a likelihood-ratio criterion. This we did for general densities, but also used the Gaussian as a special illustrative example. In the special case that $\mathbf{x}|y$ is normally distributed, our method amounts to maximizing

$$\mathrm{LR}(\boldsymbol{\alpha}) = \sum_{y=1}^{h}\left(\frac{n_y}{n}\right)(\log\boldsymbol{\alpha}^T\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha} - \log\boldsymbol{\alpha}^T\tilde{\boldsymbol{\Delta}}_y\boldsymbol{\alpha}) \qquad (2)$$

sequentially over unit vector $\boldsymbol{\alpha}$—after the first maximizing solution is obtained, say $\hat{\boldsymbol{\alpha}}_1$, we maximize (2) again, adding an extra constraint such as $\boldsymbol{\alpha} \perp \hat{\boldsymbol{\alpha}}_1$, and so on. Equation (2) above is equation (4.3) in ZH03, expressed here using the notation of CF09. Noting that $n = \sum_{y=1}^{h} n_y$ by definition, it is easy to see that, if $d = 1$ in Equation (1), then Equations (1) and (2) are identical. The case of $d > 1$ was not considered in ZH03 because we took a *sequential* approach to optimize (2). In other words, the basis vectors of $\mathcal{S}$ were obtained one at a time, so it sufficed to consider only the one-dimensional case of the objective function. In short, the underlying proposals in CF09 and ZH03 are the same; the main difference lies in joint optimization versus sequential optimization.

Although the subspace resulting from d steps of our sequential algorithm would be suboptimal with respect to (1), the approach might be seen to be more practical: the sequence of subspaces would be nested. On the other hand, the $(d-1)$-dimensional solution to (1) will not in general be nested in the d-dimensional solution. Nesting is desirable if we want to estimate a suitable value for d.

We think that the formulation in ZH03, of which (2) is a special case, is more direct and intuitive. In addition, it does not make any parametric assumption about the distribution of $\mathbf{x}|y$, and thus is more general than CF09. The general formulation of ZH03 sequentially maximizes

$$\mathrm{LR}(\boldsymbol{\alpha}) = \log\frac{\prod_{y=1}^{h}\prod_{\mathbf{x}_j\in C_y}\hat{p}_y^{(\alpha)}(\boldsymbol{\alpha}^T\mathbf{x}_j)}{\prod_{y=1}^{h}\prod_{\mathbf{x}_j\in C_y}\hat{p}^{(\alpha)}(\boldsymbol{\alpha}^T\mathbf{x}_j)}$$

over unit vector $\boldsymbol{\alpha}$, where C_y refers to the subgroup indexed by y; $\hat{p}_y^{(\alpha)}$ is the (nonparametric) MLE for the conditional distribution of $\mathbf{x}|y$ in the direction of $\boldsymbol{\alpha}$; and $\hat{p}^{(\alpha)}$ is the (nonparametric) MLE of the marginal distribution of $\mathbf{x}$ in the direction of $\boldsymbol{\alpha}$ regardless of subgroup membership. As shown in ZH03, this is a generalization of Fisher's LDA problem that seeks directions to maximize between/within variance, or equivalently between/total variance. Here the sequential approach is even more compelling, since it requires only one-dimensional density estimation.

Mu ZHU
University of Waterloo
Waterloo, Ontario N2L 3G1
Canada
m3zhu@math.uwaterloo.ca

Trevor J. HASTIE
Stanford University
Stanford, CA 94305
hastie@stanford.edu

REFERENCES

Cook, R. D., and Weisberg, S. (1991), Comment on "Sliced Inverse Regression for Dimension Reduction," by K. C. Li, *Journal of the American Statistical Association*, 86, 328–332. [880]

Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction" (with discussion), *Journal of the American Statistical Association*, 86, 316–342. [880]

Zhu, M., and Hastie, T. J. (2003), "Feature Extraction for Nonparametric Discriminant Analysis," *Journal of Computational and Graphical Statistics*, 12 (1), 101–120. [880]

Journal of the American Statistical Association
June 2010, Vol. 105, No. 490, Letters to the Editor
DOI: 10.1198/jasa.2010.tm09295

REPLY TO ZHU AND HASTIE

In CF09 we used the method of maximum likelihood to derive an estimator—called LAD—of the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ that can dominate other dimension reduction methods like IRE, SIR, SAVE, and DR. Computation of the LAD estimator, $\widehat{\mathcal{S}}_{Y|\mathbf{X}} = \arg\max L_d^*(\mathcal{S})$, is correctly represented by (1) in Zhu and Hastie's letter, hereafter ZH. Reasoning by analogy from Fisher's linear discriminant, ZH03 proposed a sequential nonparametric algorithm for discriminant analysis. Although CF09 and ZH03 differ on their starting points, goals, and levels of theoretical development, the estimators do match in a special parametric case. We discuss that parametric case first and then turn to the nonparametric setting.

As stated in CF09 near the end of section 4.3 and restated in ZH, the CF09 and ZH03 estimators are the same when $\mathbf{X}|Y$ is normal and $d = \dim(\mathcal{S}_{Y|\mathbf{X}}) = 1$; that is, $\arg\max L_1^*(\mathcal{S}) = \text{span}\{\arg\max \text{LR}(\boldsymbol{\alpha})\}$, where $\text{LR}(\boldsymbol{\alpha})$ is as given in (2) of ZH. The estimators differ when $d > 1$, with ZH03 proceeding sequentially and CF09, following the dictates of the likelihood function, using full optimization. This is a distinction that can have a great deal of difference.

To illustrate, we applied the sequential algorithm as described in ZH's Equation (2) to the bird–planes–cars data used in CF09. A plot showing the discriminatory information contained in the first two sequential linear combinations is given in Figure 1. For visual clarity, we removed one point from the plot, but not from the analysis. This figure can be compared directly to figures 4 and 5 in CF09. Cars are well separated in Figure 1, but birds and planes are largely over-plotted, as they were for some of the methods represented in figure 4 of CF09. LAD and the sequential method provide very different representation of the data, with LAD's results being clearly preferable. This difference is likely a consequence of the suboptimality of the sequential method. It suggests that the sequential method could require a larger value of d to encapsulate the data structure found by LAD.

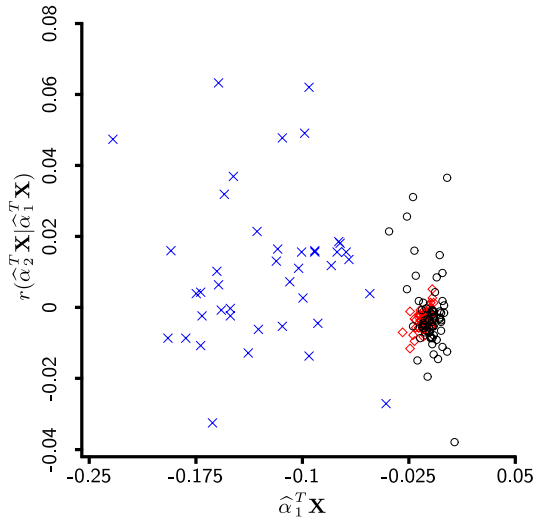

Figure 1. Plot of the first sequential predictor $\widehat{\boldsymbol{\alpha}}_1^T\mathbf{X}$ versus the residuals $r(\widehat{\boldsymbol{\alpha}}_2^T\mathbf{X}|\widehat{\boldsymbol{\alpha}}_1^T\mathbf{X})$ from the ordinary least squares fit of the second sequential predictor $\widehat{\boldsymbol{\alpha}}_2^T\mathbf{X}$ on the first. Birds, red ◇'s; planes, black ○'s; cars, blue ×'s.

Nesting and the ability to estimate $\mathcal{S}_{Y|\mathbf{X}}$ sequentially do have some intuitive appeal. They hold in classical dimension reduction methods like SIR and SAVE that are based on spectral analyses of kernel matrices, but not in recent methods like IRE and LAD. However, using sequential optimization in lieu of full optimization is suboptimal and can have a significant inferential cost, as illustrated in Figure 1. Although LAD does not produce nested estimators of $\mathcal{S}_{Y|\mathbf{X}}$, it is likelihood-based and thus we developed methods for inferring about d and testing predictors using familiar likelihood procedures. We know of no similar inference procedures for the sequential method. Recognizing that a proper likelihood has its own intuitive appeal, the advantages of nesting and sequential estimation are for us not sufficient to compensate for their downside.

Turning to the nonparametric setting, the main point of ZH03 was perhaps their sequential algorithm for discriminant analysis, where the density of $\mathbf{X}|Y$ was not specified and was largely unrestricted. Comments on relative performance in this context are not relevant because the methods have different goals and typically estimate different quantities. Assuming that the nonparametric algorithm of ZH03 is estimating an identified subspace, that subspace is generally not equal to the central subspace but it might equal the central discriminant subspace (Cook and Yin 2001).

In principle we prefer full optimization over sequential optimization in nonparametric settings as well. However, added complications arise because full optimization can be infeasible due to the difficulty in estimating multidimensional densities nonparametrically. Sequential estimation then takes on added importance, not because of its intuitive appeal, but because there may be no workable joint alternatives. Yin, Li, and Cook (2008) studied sequential nonparametric methods for estimating $\mathcal{S}_{Y|\mathbf{X}}$ in regressions where $(Y, \mathbf{X})$ has a density. They demonstrated that sequential estimation can be consistent for $\mathcal{S}_{Y|\mathbf{X}}$ under the conditions that the predictors are elliptically distributed and that certain densities are directionally identified. Reasoning from their findings, we judge that further investigation is required to gain a necessary appreciation for the operating characteristics of the nonparametric sequential algorithm of ZH03.

R. Dennis Cook
University of Minnesota
Minneapolis, MN 55455
dennis@stat.umn.edu

Liliana Forzani
Universidad Nacional del Litoral
Santa Fe, Argentina
liliana.forzani@gmail.com

ADDITIONAL REFERENCES

Cook, R. D., and Yin, X. (2001), "Dimension Reduction and Visualization in Discriminant Analysis," *Australian & New Zealand Journal of Statistics*, 43, 901–931. [881]

Yin, X., Li, B., and Cook, R. D. (2008), "Successive Direction Extraction for Estimating the Central Subspace in a Multiple-Index Regression," *Journal of Multivariate Analysis*, 99, 1733–1757. [881]

This article has been cited by: