



## ldr: An R Software Package for Likelihood-Based Sufficient Dimension Reduction

**Kofi P. Adragani**  
University of Maryland,  
Baltimore County

**Andrew M. Raim**  
University of Maryland,  
Baltimore County

---

### Abstract

In regression settings, a sufficient dimension reduction (SDR) method seeks the core information in a  $p$ -vector predictor that completely captures its relationship with a response. The reduced predictor may reside in a lower dimension  $d < p$ , improving ability to visualize data and predict future observations, and mitigating dimensionality issues when carrying out further analysis. We introduce **ldr**, a new R software package that implements three recently proposed likelihood-based methods for SDR: covariance reduction, likelihood acquired directions, and principal fitted components. All three methods reduce the dimensionality of the data by projection into lower dimensional subspaces. The package also implements a variable screening method built upon principal fitted components which makes use of flexible basis functions to capture the dependencies between the predictors and the response. Examples are given to demonstrate likelihood-based SDR analyses using **ldr**, including estimation of the dimension of reduction subspaces and selection of basis functions. The **ldr** package provides a framework that we hope to grow into a comprehensive library of likelihood-based SDR methodologies.

*Keywords:* dimension reduction, inverse regression, principal components, Grassmann manifold, variable screening.

---

## 1. Introduction

Consider a regression or classification problem of a univariate response  $Y$  on a  $p \times 1$  vector  $X$  of continuous predictors. We assume that  $Y$  is discrete or continuous. When  $p$  is large, it is always worthwhile to find a reduction  $R(X)$  of dimension less than  $p$  that captures all regression information of  $Y$  on  $X$ . Replacing  $X$  by a lower dimensional function  $R(X)$  is called dimension reduction. When  $R(X)$  retains all the relevant information about  $Y$ , it is referred to as a *sufficient reduction*. Advantages of reducing the dimension of  $X$  include visualization

of data, mitigation of dimensionality issues in estimating the mean function  $E(Y|X)$ , and better prediction of future observations.

Most sufficient reductions consist of linear combinations of the predictors, which can be represented conveniently in terms of the projection  $P_{\mathcal{S}}X$  of  $X$  onto a subspace  $\mathcal{S}$  of dimension  $d$  in  $\mathbb{R}^p$ . The subspace  $\mathcal{S}$  is called a sufficient dimension reduction subspace if  $Y$  is independent of  $X$  given  $P_{\mathcal{S}}X$ . Under some conditions, the intersection of all dimension-reduction subspaces is also a dimension-reduction subspace which is referred to as the *central subspace* (Cook 1998). The central subspace is the parameter of interest in the sufficient dimension reduction framework.

Dimension reduction has been ubiquitous in the applied sciences with the use of principal components as the method of choice to reduce the dimensionality of data. However, principal components do not yield a sufficient reduction as the components are obtained with no use of the response. Numerous sufficient dimension reduction methods have been proposed in the statistics literature since the seminal paper of Li (1991) on sliced inverse regression (SIR). Among these methods are the sliced average variance estimation (SAVE, Cook and Weisberg 1991), principal Hessian directions (PHD, Li 1992), directional regression (Li and Wang 2007), and the inverse regression estimation (IRE, Cook and Ni 2005). These methods do not make any assumptions on the distribution of the predictors. Some recently developed methods adopt a likelihood-based approach. These methods assume a conditional distribution of the predictor vector  $X$  given the response  $Y$ , and include covariance reduction (CORE, Cook and Forzani 2008a), likelihood acquired directions (LAD, Cook and Forzani 2009), principal fitted components (PFC, Cook 2007; Cook and Forzani 2008b), and the envelope model (Cook, Li, and Chiaromonte 2010). The R (R Core Team 2014) package **ldr** (Adragni and Raim 2014) described in this article implements CORE, LAD and PFC.

At least two packages for sufficient dimension reduction can be found in the literature. One is the R package **dr** (Weisberg 2002) which implements SIR, SAVE, PHD, and IRE. The other package is **LDR**, a MATLAB (The MathWorks, Inc. 2012) package by Cook, Forzani, and Tomassi (2012) which implements CORE, LAD, PFC and the envelope model. The package introduced in this paper is the R counterpart of the MATLAB **LDR** package.

For the likelihood-based methods, except in some PFC model setups, there is no closed-form solution to the maximum likelihood estimator of the central subspace, which is the parameter of interest in dimension reduction framework. The maximum likelihood estimation is carried out over the set of all  $d$ -dimensional subspaces in  $\mathbb{R}^p$ , referred to as a Grassmann manifold, and denoted by  $\mathcal{G}_{(d,p)}$ . Optimization on the Grassmann manifold adds a layer of difficulty to the estimation procedure. Fortunately, a recent R package called **GrassmannOptim** (Adragni, Cook, and Wu 2012) has been made available for such optimization, which allows the current implementation of **ldr** in R. The implementation of **GrassmannOptim** uses gradient-based algorithms and embeds a stochastic gradient method for global search.

The package is built to use the typical command line interface of R. It has a hierarchical structure, similar to the **dr** package. The top command is **ldr**, which handles the maximum likelihood estimation of all parameters including the central subspace. The call of the **ldr** function, as described in the manual, is

```
ldr(X, y = NULL, fy = NULL, Sigmas = NULL, ns = NULL, numdir = NULL,
    nslices = NULL, model = c("core", "lad", "pfc"), numdir.test = FALSE, ...)
```

There are also functions which are specifically dedicated to each of the three models. Their calls are the following:

```
core(X, y, Sigmas = NULL, ns = NULL, numdir = 2, numdir.test = FALSE, ...)
lad(X, y, numdir = NULL, nslices = NULL, numdir.test = FALSE, ...)
pfc(X, y, fy = NULL, numdir = NULL, structure = c("iso", "aniso",
  "unstr", "unstr2"), eps_aniso = 1e-3, numdir.test = FALSE, ...)
```

For PFC and LAD, the reduction data-matrix of the predictor is also provided, and can be readily plotted. Available procedures allow the choice of the dimension of the central subspace using information criteria such as AIC and BIC (Burnham and Anderson 2002) and the likelihood ratio test.

The remainder of this article is organized essentially in two main sections. The concept of sufficient dimension reduction and a brief description of the three aforementioned likelihood-based methods are provided in Section 2. Section 3 provides an overview of the package, along with examples on the use of the R package **ldr** to estimate a sufficient reduction. The package is available from the Comprehensive R Archive Network (CRAN) at <http://CRAN.R-project.org/package=ldr>.

## 2. Likelihood-based sufficient dimension reduction

We describe the following three sufficient dimension reduction methods: covariance reduction (Cook and Forzani 2008a), likelihood acquired directions (Cook and Forzani 2009), and principal fitted components (Cook 2007; Cook and Forzani 2008b). These methods are likelihood-based inverse regression methods. They use the randomness of the predictors and assume that the conditional distribution of  $X$  given  $Y = y$ , denoted by  $X_y$ , follows a multivariate normal distribution with mean  $\mu_y = E(X_y)$  and variance  $\Delta_y = \text{VAR}(X_y)$ . In some regressions  $R(X)$  may be a nonlinear function of  $X$ , but most often we encounter multi-dimensional reductions consisting of several linear combinations  $R(X) = \Gamma^\top X$ , where  $\Gamma$  is an unknown  $p \times d$  matrix,  $d \leq p$ , that must be estimated from the data. Linear reductions are often imposed to facilitate progress. If  $\Gamma^\top X$  is a sufficient linear reduction then so is  $(\Gamma M)^\top X$  for any  $d \times d$  full rank matrix  $M$ . Consequently, only  $\text{span}(\Gamma)$ , the subspace spanned by the columns of  $\Gamma$ , can be identified. The subspace  $\text{span}(\Gamma)$  is called a *dimension reduction subspace*.

If  $\text{span}(\Gamma)$  is a dimension reduction subspace, then so is  $\text{span}(\Gamma, \Gamma_1)$  for any matrix  $p \times d_1$  matrix  $\Gamma_1$ . If  $\text{span}(\Gamma_1)$  and  $\text{span}(\Gamma_2)$  are both dimension reduction subspaces, then under mild conditions so is their intersection  $\text{span}(\Gamma_1) \cap \text{span}(\Gamma_2)$  (Cook 1996, 1998). Consequently, the inferential target in sufficient dimension reduction is often taken to be the central subspace  $\mathcal{S}_{Y|X}$ , defined as the intersection of all dimension reduction subspaces (Cook 1994, 1998). A *minimal sufficient linear reduction* is then of the form  $R(X) = \Gamma^\top X$ , where the columns of  $\Gamma$  now form a basis for  $\mathcal{S}_{Y|X}$ .

Let  $d = \dim(\mathcal{S}_{Y|X})$  and let  $\eta$  denote a  $p \times d$  semi-orthogonal matrix whose columns form a basis for  $\mathcal{S}_{Y|X}$ , so  $Y|X \sim Y|\eta^\top X$ . In other words,  $Y|X$  has the same distribution as  $Y|\eta^\top X$  so that  $\eta^\top X$  is sufficient to characterize the relationship between  $Y$  and  $X$ . The three aforementioned likelihood-based methodologies are designed to estimate  $\mathcal{S}_{Y|X} = \text{span}(\eta)$  under different structures for the mean function  $\mu_y$  and the covariance function  $\Delta_y$ . Essentially, covariance reduction (CORE) is concerned with the problem of characterizing the covariance

matrices  $\Delta_y, y = 1, \dots, h$ , of  $X$  observed in each of  $h$  normal populations and does not utilize the population means  $\mu_y$ . The likelihood acquired directions (LAD) method considers both a general conditional covariance  $\Delta_y$  and a general mean  $\mu_y$ . The principal fitted components (PFC) method considers a model for the mean function  $\mu_y$ , and while  $\Delta_y = \Delta$  is assumed for all  $y$ , several structures of  $\Delta$  are involved. Each of these three methods is summarized in the sections that follow.

## 2.1. Covariance reduction

Covariance reducing models were proposed by Cook and Forzani (2008a) for studying the sample covariance matrices of a random vector observed in different populations. The models allow the reduction of the sample covariance matrices to an informational core that is sufficient to characterize the variance heterogeneity among the populations. It does not address the population means.

Let  $\tilde{\Delta}_y, y = 1, \dots, h$  be the sample covariance matrices of the random vector  $X$  observed in each of  $h$  normal populations, and let  $S_y = (n_y - 1)\tilde{\Delta}_y$ . The goal is to find a semi-orthogonal matrix  $\Gamma \in \mathbb{R}^{p \times d}$ ,  $d \leq p$ , with the property that for any two populations  $j$  and  $k$ ,

$$S_j | (\Gamma^\top S_j \Gamma = B, n_j = m) \sim S_k | (\Gamma^\top S_k \Gamma = B, n_k = m).$$

That is, given  $\Gamma^\top S_y \Gamma$  and  $n_y$ , the conditional distribution of  $S_y$  must not depend on  $y$ . Thus  $\Gamma^\top S_y \Gamma$  is sufficient to account for the heterogeneity among the population covariance matrices. The central subspace is  $\mathcal{S}_{Y|X} = \text{span}(\Gamma)$ .

The methodology is built upon the assumption that the  $S_y$ 's follow a Wishart distribution and its development allows the covariance matrix for the population  $y$  to be modeled as

$$\Delta_y = \Gamma_0 M_0 \Gamma_0^\top + \Gamma M_y \Gamma^\top,$$

where  $\Gamma_0$  is an orthogonal completion of  $\Gamma$ , and  $M_0$  and  $M_y$  are symmetric positive definite matrices. En route to obtain the central subspace  $\mathcal{S}_{Y|X}$ , let  $\Delta_y = \text{VAR}(X_y)$  and  $\Delta = \text{E}(\Delta_Y)$ , and define  $P_{\Gamma(\Delta_y)} = \Gamma(\Gamma^\top \Delta_y \Gamma)^{-1} \Gamma^\top \Delta_y$  as the projection operator that projects onto  $\mathcal{S}_{Y|X}$  relative to  $\Delta_y$ . The central subspace  $\text{span}(\Gamma)$  is redefined to be the smallest subspace satisfying

$$\Delta_y = \Delta + P_{\Gamma(\Delta_y)}^\top (\Delta_y - \Delta) P_{\Gamma(\Delta_y)}.$$

This equality means that the structure of the covariance  $\Delta_y$  can be decomposed in terms of the average structure of  $\Delta$  and a structure that is not shared with the other covariances. The maximum likelihood estimator of  $\mathcal{S}_{Y|X}$  is the subspace that maximizes over  $\mathcal{G}_{(d,p)}$  the log-likelihood function

$$L_d(\mathcal{S}) = c - \frac{n}{2} \log |\tilde{\Delta}| + \frac{n}{2} \log |P_{\mathcal{S}} \tilde{\Delta} P_{\mathcal{S}}| - \sum_{y=1}^h \frac{n_y}{2} \log |P_{\mathcal{S}} \tilde{\Delta}_y P_{\mathcal{S}}|, \quad (1)$$

where  $P_{\mathcal{S}}$  indicates the projection onto the subspace  $\mathcal{S}$  in the usual inner product,  $c$  is a constant that does not depend on  $\mathcal{S}$ , and  $\tilde{\Delta} = \sum_{y=1}^h \tilde{\Delta}_y$ . The maximum likelihood estimators of  $M_0$  and  $M_y$  are  $\widehat{M}_0 = \widehat{\Gamma}_0^\top \tilde{\Delta} \widehat{\Gamma}_0$  and  $\widehat{M}_y = \widehat{\Gamma}^\top \tilde{\Delta}_y \widehat{\Gamma}$ , where  $\widehat{\Gamma}$  is an orthogonal basis of the maximum likelihood estimator of  $\mathcal{S}_{Y|X}$  and  $\widehat{\Gamma}_0$  is an orthogonal completion of  $\widehat{\Gamma}$ .

The dimension  $d$  of the central subspace must be estimated. A likelihood ratio test and information criterion were proposed to help estimate  $d$ . The likelihood ratio test  $\Lambda(d_0) = 2(L_p - L_{d_0})$  was used in a sequential manner for the hypothesis  $H_0 : d = d_0$  against  $H_a : d = p$ , to choose  $d$ , starting with  $d = 0$ . The first hypothesized value of  $d$  that is not rejected is the estimated dimension  $d$ . The term  $L_p$  denotes the value of the maximized log-likelihood for the full model with  $d_0 = p$  and  $L_{d_0}$  is the maximum value of the log-likelihood function in Equation 1. Under the null hypothesis  $\Lambda(d_0)$  is distributed asymptotically as a chi-squared random variable with degrees of freedom  $(p - d)(h - 1)(p + 1) + (h - 3)d/2$ , for  $h \geq 2$  and  $d < p$ .

With information criterion, the dimension  $d$  is selected to minimize over  $d_0$  the function  $IC(d_0) = -2L_{d_0} + h(n)g(d_0)$ , where  $g(d_0)$  is the number of parameters to be estimated, and  $h(n)$  is equal to  $\log n$  for the Bayesian criterion and 2 for Akaike's.

## 2.2. Likelihood acquired directions

The likelihood acquired directions method estimates the central subspace using both the conditional mean  $\mu_y$  and conditional variance function  $\Delta_y$ . It assumes that  $X_y$  is normally distributed with mean  $\mu_y$  and variance  $\Delta_y$ , where  $y \in S_Y = \{1, \dots, h\}$ . Continuous response can be sliced into finite categories to meet this condition. It furthermore assumes that the data consist of  $n_y$  independent observations of  $X_y$ ,  $y \in S_Y$ , and the total sample size is  $n = \sum_{y=1}^h n_y$ .

Let  $\mu = \mathbf{E}(X)$  and  $\Sigma = \mathbf{VAR}(X)$  denote the marginal mean and variance of  $X$  and let  $\Delta = \mathbf{E}(\Delta_Y)$  denote the average covariance matrix. The subspace  $\text{span}(\Gamma)$  is the central subspace if and only if  $\text{span}(\Gamma)$  is the minimum subspace with the conditions

$$\begin{aligned} (i.) \quad & \text{span}(\Gamma) \subset \Delta^{-1} \text{span}(\mu_y - \mu) \\ (ii.) \quad & \Delta_y = \Delta + P_{\Gamma(\Delta_y)}^\top (\Delta_y - \Delta) P_{\Gamma(\Delta_y)} \end{aligned}$$

The first condition (i.) implies that  $\mathcal{S}_{\Delta\Gamma}$ , the subspace spanned by  $\Delta\Gamma$ , falls into the subspace spanned by the translated conditional means  $\mu_y - \mu$ . The second condition is identical to that of CORE.

Let  $\tilde{\Sigma}$  denote the sample covariance matrix of  $X$ , and let  $\tilde{\Delta}_y$  denote the sample covariance matrix for the data with  $Y = y$ . The maximum likelihood estimate of the  $d$ -dimensional central subspace  $\mathcal{S}_{Y|X}$  maximizes over  $\mathcal{G}_{(d,p)}$  the log-likelihood function

$$L_d(\mathcal{S}) = c - \frac{n}{2} \log |\tilde{\Sigma}| + \frac{n}{2} \log |P_{\mathcal{S}} \tilde{\Sigma} P_{\mathcal{S}}|_0 - \frac{1}{2} \sum_{y=1}^h n_y \log |P_{\mathcal{S}} \tilde{\Delta}_y P_{\mathcal{S}}|_0, \quad (2)$$

where  $|A|_0$  indicates the product of the non-zero eigenvalues of a positive semi-definite symmetric matrix  $A$ ,  $P_{\mathcal{S}}$  indicates the projection onto the subspace  $\mathcal{S}$  in the usual inner product, and  $c$  is a constant independent of  $\mathcal{S}$ . The desired reduction is then  $\hat{\Gamma}^\top X$ , where the columns of  $\hat{\Gamma}$  form an orthogonal basis for the maximum likelihood estimator of  $\mathcal{S}_{Y|X}$ .

There is a striking resemblance between Equation 1 of CORE and Equation 2 of LAD. The difference is essentially that CORE extracts the sufficient information using the conditional variance only, while LAD uses both the conditional mean and the conditional variance.

The estimation of  $d$  is carried out as for CORE, using either the sequential likelihood ratio test or information criteria AIC and BIC. The statistic  $\Lambda(d_0)$  is distributed asymptotically as a chi-squared random variable with degrees of freedom  $(p-d_0)\{(h-1)(p+1)+(h-3)d_0+2(h-1)\}/2$ , for  $h \geq 2$  and  $d_0 < p$ . For the information criteria,  $IC(d_0) = -2L_{d_0} + h(n)g(d_0)$ ,  $g(d_0) = p + (h-1)d_0 + d_0(p-d_0) + (h-1)d_0(d_0+1)/2 + p(p+1)/2$ , and  $h(n)$  is equal to  $\log n$  for BIC and 2 for AIC.

### 2.3. Principal fitted components

Principal fitted components methodology was proposed by Cook (2007) and estimates the central subspace by modeling the inverse mean function  $E(X|Y = y)$  while assuming that the conditional variance function  $\Delta$  is independent of  $Y$ . The response may be discrete or continuous. A generic PFC model can be written as

$$X_y = \mu + \Gamma\beta f_y + \Delta^{1/2}\varepsilon, \quad (3)$$

where  $\mu = E(X)$ ,  $\beta$  is an unconstrained  $d \times r$  matrix,  $\varepsilon \sim N(0, I)$ , and  $f_y \in \mathbb{R}^r$  is a known vector-valued function of  $y$  with linearly independent elements, referred to as basis functions. The basis function is a user-selected function; it is supposed to adequately help capture the dependency between the predictors and the response.

Several basis functions have been suggested in conjunction with PFC models (Cook 2007; Adraghi and Cook 2009; Adraghi 2009), including polynomial, piecewise polynomial, and Fourier bases. To gain an insight into the choice of the appropriate basis function, inverse response plots (Cook 1998) of  $X_j$  versus  $y$ ,  $j = 1, \dots, p$  could be used. In some regressions there may be a natural choice for  $f_y$ . For example, suppose for instance that  $Y$  is categorical, taking values  $\{1, \dots, h\}$ . We can then set the  $k$ th coordinate of  $f_y$  to be

$$f_{y_k} = J(y = k) - n_k/n, \quad k = 1, \dots, h,$$

where  $J(\cdot)$  is the indicator function and  $n_k$  is the number of observations falling into category  $k$ .

When graphical guidance is not available, Cook (2007) suggested constructing a piecewise constant basis. With a continuous  $Y$ , an option consists of ‘‘slicing’’ the observed values into  $h$  bins (categories) and then specifying the  $k$ th coordinate of  $f_y$  as for the case of a categorical  $Y$ . This has the effect of approximating each conditional mean  $E(X_j|Y = y)$  as a step function of  $Y$  with  $h$  steps.

With a continuous  $Y$ , piecewise polynomial bases, including piecewise linear, piecewise quadratic and piecewise cubic polynomial bases, are constructed by slicing the observed values of  $Y$  into bins  $C_k$ ,  $k = 1, \dots, h$ . Let  $\tau_0, \tau_1, \dots, \tau_h$  denote the end-points or knots of the slices. For example,  $(\tau_0, \tau_1)$  are the end-points of  $C_1$ ;  $(\tau_1, \tau_2)$  are the end-points of  $C_2$ , and so on. We give here the description of the piecewise linear discontinuous basis  $f_y \in \mathbb{R}^{2h-1}$  by its coordinates

$$\begin{aligned} f_{2i-1}(y) &= J(y = i), & i &= 1, 2, \dots, h-1 \\ f_{2i}(y) &= J(y = i)(y - \tau_{i-1}), & i &= 1, 2, \dots, h-1 \\ f_{2h-1}(y) &= J(y = h)(y - \tau_{h-1}). \end{aligned}$$

The number of components of  $f_y$  is to be small enough compared to  $n$  to avoid modeling random variation rather than the overall shape between the response and individual predictor.

A number of structures of  $\Delta$  have been investigated. When the predictors are conditionally independent,  $\Delta$  has a diagonal structure; the *isotropic* PFC corresponds to the case with identical diagonal elements, while the *anisotropic* (initially called ‘*diagonal*’ in [Adragni and Cook 2009](#)) is for the case of non-identical diagonal elements. The *unstructured* PFC refers to the case with a general structure for  $\Delta$  which allows conditional dependencies between the predictors. *Extended* structures are more evolved and involve expressions of  $\Delta$  that depend on  $\Gamma$ . We discuss each of these PFC models in the next sections.

### *Isotropic principal fitted components*

The isotropic PFC ([Cook 2007](#)) assumes that  $\Delta = \sigma^2 I_p$ . The central subspace is obtained as  $\mathcal{S}_{Y|X} = \text{span}(\Gamma)$ . To provide the estimate of  $\mathcal{S}_{Y|X}$ , let  $\mathbf{X} = (X_1 - \bar{X}, \dots, X_n - \bar{X})^\top$  denote the  $n \times p$  data matrix of the predictors, and let  $\mathbf{F} = (f_{y_1} - \bar{f}, \dots, f_{y_n} - \bar{f})^\top$  be the  $n \times r$  data matrix of the basis function where  $\bar{f} = (1/n) \sum_{i=1}^n f_{y_i}$ . Let  $\tilde{\Sigma}_{\text{fit}} = \mathbf{X}^\top P_{\mathbf{F}} \mathbf{X} / n$ , where  $P_{\mathbf{F}}$  denotes the projection operator onto the subspace spanned by the columns of  $\mathbf{F}$ . The term  $\tilde{\Sigma}_{\text{fit}}$  can be seen as the sample covariance matrix of the fitted vectors  $P_{\mathbf{F}} \mathbf{X}$  from the multivariate linear regression of  $\mathbf{X}$  on  $\mathbf{f}$ , including an intercept. The maximum likelihood estimator for  $\text{span}(\Gamma)$  is the subspace spanned by the first  $d$  eigenvectors of  $\tilde{\Sigma}_{\text{fit}}$  corresponding to its largest eigenvalues.

Turning to the estimation of the dimension  $d$  of the central subspace  $\mathcal{S}_{Y|X}$ , the likelihood ratio test  $\Lambda(d_0)$ , and information criteria AIC and BIC used previously for CORE and LAD are still used.

To get the expression of the maximized log-likelihood, let  $\hat{\lambda}_i^{\text{fit}}, i = 1, \dots, d$  be the eigenvalues of  $\tilde{\Sigma}_{\text{fit}}$  such that  $\hat{\lambda}_1^{\text{fit}} > \dots > \hat{\lambda}_d^{\text{fit}}$  and let  $\hat{\lambda}_i, i = 1, \dots, p$  be the eigenvalues of  $\tilde{\Sigma}$ , the sample covariance of  $\mathbf{X}$ . The estimate of  $\sigma^2$  is  $\hat{\sigma}^2 = (\sum_{i=1}^p \hat{\lambda}_i - \sum_{i=1}^d \hat{\lambda}_i^{\text{fit}}) / p$ . The maximized log-likelihood is

$$L_d = -\frac{np}{2}(1 + \log(2\pi)) - \frac{np}{2} \log \hat{\sigma}^2.$$

For the information criteria  $IC(d_0) = -2L_{d_0} + h(n)g(d_0)$ , we have  $g(d_0) = (p - d_0)(r - d_0)$ , and  $h(n)$  is equal to  $\log n$  for BIC and 2 for AIC.

### *Anisotropic principal fitted components*

The anisotropic PFC supposes that  $\Delta = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$ . The central subspace is obtained as  $\mathcal{S}_{Y|X} = \Delta^{-1} \mathcal{S}_\Gamma$ . There is no closed-form expression of the maximum likelihood estimator of  $\Delta$  and  $\Gamma$ . An alternating algorithm is used to carry the maximum likelihood estimation. The idea is to notice that, letting

$$Z = \Delta^{-1/2} \mu + \Delta^{-1/2} \Gamma \beta \mathbf{f} + \varepsilon, \quad (4)$$

that we can estimate  $\text{span}(\Gamma)$  as  $\Delta^{1/2}$  times the estimate  $\tilde{\Gamma}$  of  $\Delta^{-1/2} \Gamma$  from the isotropic model (4). With  $\hat{\mu} = \bar{X}$ , alternating between these two steps leads to the following algorithm:

1. Fit the isotropic PFC model to the original data, getting initial estimates  $\hat{\Gamma}_{(1)}$  and  $\hat{\beta}_{(1)}$ .
2. For some small  $\epsilon > 0$ , repeat for  $j = 1, 2, \dots$  until  $\text{Tr}\{(\hat{\Delta}_{(j)} - \hat{\Delta}_{(j+1)})^2\} < \epsilon$

- (a) Calculate  $\hat{\Delta}_{(j)} = \text{diag}\{(\mathbf{X} - \mathbf{F} \hat{\beta}_{(j)}^\top \hat{\Gamma}_{(j)}^\top)^\top (\mathbf{X} - \mathbf{F} \hat{\beta}_{(j)}^\top \hat{\Gamma}_{(j)}^\top)\}$ ,

- (b) Transform  $Z = \hat{\Delta}_{(j)}^{-1/2} X$ .
- (c) Fit the isotropic PFC model to  $Z$ , yielding estimates  $\tilde{\Gamma}$ , and  $\tilde{\beta}$ .
- (d) Back-transform the estimates to the original scale  $\hat{\Gamma}_{(j+1)} = \hat{\Delta}_{(j)}^{1/2} \tilde{\Gamma}$ ,  $\hat{\beta}_{(j+1)} = \tilde{\beta}$ .

At convergence, the estimators  $\hat{\Gamma}$  and  $\hat{\Delta}$  are obtained to form the estimator of  $\mathcal{S}_{Y|X}$  as  $\hat{\Delta}^{-1} \mathcal{S}_{\hat{\Gamma}}$ , where  $\mathcal{S}_{\hat{\Gamma}}$  is the subspace spanned by the columns of  $\hat{\Gamma}$ . There is no closed-form expression of the maximized log-likelihood  $L_d$  under the anisotropic model, but it can be numerically evaluated. The estimation of the dimension  $d$  is carried out similarly as described for the isotropic case.

#### *Unstructured principal fitted components*

Unstructured PFC (Cook and Forzani 2008b) assumes that the covariance matrix  $\Delta$  has a general structure, positive definite, and independent of  $y$ . The central subspace is  $\mathcal{S}_{Y|X} = \Delta^{-1} \text{span}(\Gamma)$ .

To present the maximum likelihood estimator of the central subspace, let  $\tilde{\Sigma}_{\text{res}} = \tilde{\Sigma} - \tilde{\Sigma}_{\text{fit}}$ , and let  $\hat{\omega}_1 > \dots \geq \hat{\omega}_p$  and  $\hat{V} = (\hat{v}_1, \dots, \hat{v}_p)$  be the eigenvalues and the corresponding eigenvectors of  $\tilde{\Sigma}_{\text{res}}^{-1/2} \tilde{\Sigma}_{\text{fit}} \tilde{\Sigma}_{\text{res}}^{-1/2}$ . Finally, let  $\hat{K}$  be a  $p \times p$  diagonal matrix, with the first  $d$  diagonal elements equal to zero and the last  $p - d$  diagonal elements equal to  $\hat{\omega}_{d+1}, \dots, \hat{\omega}_p$ . Then the maximum likelihood estimator of  $\Delta$  is  $\hat{\Delta} = \tilde{\Sigma}_{\text{res}}^{1/2} \hat{V} (I + \hat{K}) \hat{V}^\top \tilde{\Sigma}_{\text{res}}^{1/2}$  (Theorem 3.1, Cook and Forzani 2008b). The maximum likelihood estimator of  $\mathcal{S}_{Y|X}$  is then obtained as the span of  $\hat{\Delta}^{-1}$  times the first  $d$  eigenvectors of  $\tilde{\Sigma}_{\text{fit}}$ .

The estimation of the dimension  $d$  is carried out as specified in the previous methods. The maximized value of the log-likelihood is given by

$$L_d = \frac{np}{2} (1 + \log(2\pi)) - \frac{n}{2} \log |\hat{\Sigma}_{\text{res}}| - \frac{n}{2} \sum_{i=d+1}^p \log(1 + \hat{\omega}_i).$$

The likelihood ratio test  $\Lambda(d_0)$  follows a chi-squared distribution with  $(p - d_0)(r - d_0)$  degrees of freedom. For the information criteria  $IC(d_0) = -2L_{d_0} + h(n)g(d_0)$ , we have  $g(d_0) = p(p + 3)/2 + rd_0 + d_0(p - d_0)$ , and  $h(n)$  is equal to  $\log n$  for BIC and 2 for AIC.

#### *Extended principal fitted components*

The heterogenous PFC (Cook 2007) arises from a structure of  $\Delta$  as a function of  $\Gamma$ . Let  $\Gamma_0$  be the orthogonal complement of  $\Gamma$  such that  $(\Gamma, \Gamma_0)$  is a  $p \times p$  orthogonal matrix, and let  $\Omega \in \mathbb{R}^{d \times d}$  and  $\Omega_0 \in \mathbb{R}^{(p-d) \times (p-d)}$  be two positive definite matrices. The model can be written as

$$X_y = \mu + \Gamma \beta f_y + \Gamma \Omega^{1/2} \varepsilon + \Gamma_0 \Omega_0^{1/2} \varepsilon_0,$$

where  $(\varepsilon^\top, \varepsilon_0^\top)^\top \sim N(0, I)$ . The conditional covariance  $\Delta$  can be expressed as

$$\Delta = \Gamma \Omega \Gamma^\top + \Gamma_0 \Omega_0 \Gamma_0^\top.$$

Then the subsequent PFC model is referred to as an extended PFC (Cook 2007). The central subspace is obtained as  $\mathcal{S}_{Y|X} = \text{span}(\Gamma)$ . The maximum likelihood estimator of  $\mathcal{S}_{Y|X}$



maximizes over  $\mathcal{G}_{(d,p)}$  the log-likelihood function

$$L(\mathcal{S}) = -\frac{n}{2} \log |\tilde{\Sigma}| - \frac{n}{2} \log |P_{\mathcal{S}}^{\top} \tilde{\Sigma}^{-1} P_{\mathcal{S}}| - \frac{n}{2} \log |P_{\mathcal{S}} \tilde{\Sigma}_{\text{res}} P_{\mathcal{S}}|. \quad (5)$$

This expression (5) assumes that both  $\Omega$  and  $\Omega_0$  are unstructured. The estimation of  $d$  is carried out as in the previous cases, using either a sequential likelihood ratio test or the information criteria.

### 3. Overview and use of the package

The following three functions, `core`, `lad`, and `pfc` mainly carry all the computations of the maximum likelihood estimation of the reduction subspaces and all other parameters for CORE, LAD, and PFC models. These functions could be called using the function `ldr` with a `model` argument for the appropriate model to use. For any of the models, the predictor data-matrix must be provided as an  $n \times p$  matrix `X`. For `core`, covariance matrices may be provided in lieu of the predictor data-matrix.

Covariance reduction, LAD, and extended PFC rely on an external R package called **GrassmannOptim** (Adraghi *et al.* 2012) for the optimization. Additional optional arguments may be provided for the optimization, especially arguments related to simulated annealing for global optimization. It should be pointed out that two identical calls of the optimization function in **GrassmannOptim** may produce different numerical results if no random seed is set. This is due to the fact that subspaces are estimated, and they are not uniquely determined by a matrix. In addition, the optimization has a random search process embedded.

The dimension of the reduction subspace `numdir` must be provided. The boolean argument `numdir.test` determines the type of output with respect to the test on the dimension of the central subspace. When `numdir.test` is `FALSE`, no estimation of the dimension of the central subspace is carried out. Instead, the reduction subspace is estimated at the provided `numdir`. With `numdir.test` being `TRUE`, a summary of the fit provides the AIC and BIC, along with the LRT for all values of the dimension less than or equal to the provided `numdir`. A default value of `numdir` is used when it is not provided by the user.

In the following sections, we provide some examples on the use of the `ldr` package. These examples are related to `core`, `lad` and `pfc`. In all cases, the goal is to estimate the central subspace.

#### 3.1. An example with CORE

We replicate the results in Cook and Forzani (2008a) by applying CORE methodology to the covariance matrices of the garter snakes. The matrices are genetic covariance for six traits of two female garter snake populations, one from a coastal site and the other from an inland site in northern California. The data set was initially studied by Phillips and Arnold (1999). It is of interest to compare the structures of the two covariances and determine whether they share similarities.

We fit the model as follows:

```
R> set.seed(1810)
R> library("ldr")
R> data("snakes", package = "ldr")
R> fit <- ldr(Sigmas = snakes[-3], ns = snakes[[3]], numdir = 4,
+   model = "core", numdir.test = TRUE, verbose = TRUE, sim_anneal = TRUE,
+   max_iter = 100, max_iter_sa = 100)
R> summary(fit)
```

The optimization was carried out using a simulated annealing procedure that is optional within **GrassmannOptim**, the Grassmann manifold optimization procedure. The simulated annealing helps to avoid convergence to local maxima in the pursuit of a global maximum. The summary output below was obtained using the command `summary(fit)`. It provides the results for the estimation of  $d$ , the dimension of the central subspace.

Estimated Basis Vectors for Central Subspace:

	[,1]	[,2]	[,3]	[,4]
[1,]	0.1362	0.0246	0.1705	-0.2998
[2,]	-0.2690	-0.0088	-0.0910	0.3418
[3,]	-0.0340	-0.8347	-0.5050	-0.2105
[4,]	0.8899	-0.1334	-0.0145	0.4236
[5,]	0.2776	-0.0711	0.3548	-0.6930
[6,]	0.1973	0.5289	-0.7626	-0.2989

Information Criterion:

	d=0	d=1	d=2	d=3	d=4
aic	104.6859	66.02803	59.89242	57.83486	62.44468
bic	176.7941	158.73852	169.77152	181.44885	196.35984

Large sample likelihood ratio test

	Stat	df	p.value
0D vs >= 1D	78.24122	18	1.737392e-09
1D vs >= 2D	27.58335	12	6.362675e-03
2D vs >= 3D	11.44774	7	1.202462e-01
3D vs >= 4D	1.39018	3	7.078382e-01

The sequential likelihood ratio test rejects the null hypotheses for  $d = 0$ ,  $d = 1$ , and fails to reject  $d = 2$ , thus the LRT procedure estimates  $d$  to be  $\hat{d} = 2$ . The smallest AIC value is attained for  $d = 3$ , but the value of the objective function is close for  $d = 2$  and  $d = 3$ . The smallest value of BIC is at  $d = 1$ , although the value of the objective function at  $d = 1$  and  $d = 2$  are close. Following the result from the LRT procedure we have reached the same conclusion as [Cook and Forzani \(2008a\)](#) in selecting  $\hat{d} = 2$ , so two linear combinations of the traits are needed to explain differences in variation. The maximum likelihood estimator of the central subspace is  $\text{span}(\hat{\Gamma})$ , that is  $\text{span}(\text{fit}\$Gammahat[[3]])$ . The index `[[3]]` refers to the third element of the list corresponding to the estimates when  $d = 2$ , while `[[1]]` is for  $d = 3$ , and `[[2]]` is for  $d = 1$ . The sample covariances can be written as  $\tilde{\Delta}_y = \hat{\Gamma}_0 \hat{M}_0 \hat{\Gamma}_0^\top + \hat{\Gamma} \hat{M}_y \hat{\Gamma}^\top$ , for  $y = 1, 2$ .

We obtain the estimates  $\widehat{M}_1$  and  $\widehat{M}_2$  of  $M_1$  and  $M_2$  as follows

```
R> Gammahat <- fit$Gammahat[[3]]
R> Sigmahat1 <- fit$Sigmashat[[3]][[1]]
R> Sigmahat2 <- fit$Sigmashat[[3]][[2]]
R> Mhat1 <- t(Gammahat) %*% Sigmahat1 %*% Gammahat
R> Mhat2 <- t(Gammahat) %*% Sigmahat2 %*% Gammahat
```

The estimated directions vectors that span the estimated central subspace are

$$\hat{\gamma}_1 = (-0.0302, -0.0267, 0.8041, 0.4313, 0.0475, -0.4044)^\top,$$

$$\hat{\gamma}_2 = (-0.0113, -0.0606, -0.4414, 0.8881, -0.0925, 0.0635)^\top.$$

These estimates are numerically different from the reported estimates from [Cook and Forzani \(2008a\)](#) but are quite close. The third element of  $\hat{\gamma}_1$  and the fourth of  $\hat{\gamma}_2$  are the largest entries for the two directions of the reduction matrix. We reach the same conclusions as [Cook and Forzani \(2008a\)](#) that the third and fourth traits are largely responsible for the differences in the covariance matrices.

### 3.2. Examples with LAD

We present two examples using the LAD procedure. The first example has a categorical response while the second has a continuous response. We use the flea dataset ([Lubischew 1962](#)) where the response is categorical. The data set contains 74 observations on six variables regarding three species of flea-beetles: *concinna*, *heptapotamica*, and *heikertingeri*. The measurements on the six variables are of continuous type. The goal is to find the sufficient directions that best separate the three species. This is essentially a discriminant analysis. We fit the following:

```
R> data("flea", package = "ldr")
R> fit <- ldr(X = flea[, -1], y = flea[, 1], model = "lad",
+   numdir.test = TRUE)
```

The summary result of the fit obtained using the method `summary` for the `fit` object gives the following output:

Estimated Basis Vectors for Central Subspace:

	Dir1	Dir2
tars1	0.2628	-0.3004
tars2	-0.1374	0.2772
head	-0.3617	-0.2636
aede1	-0.2079	0.8167
aede2	0.8526	0.2477
aede3	-0.1051	0.1876

Information Criterion:

	d=0	d=1	d=2
aic	2843.332	2641.641	2535.783

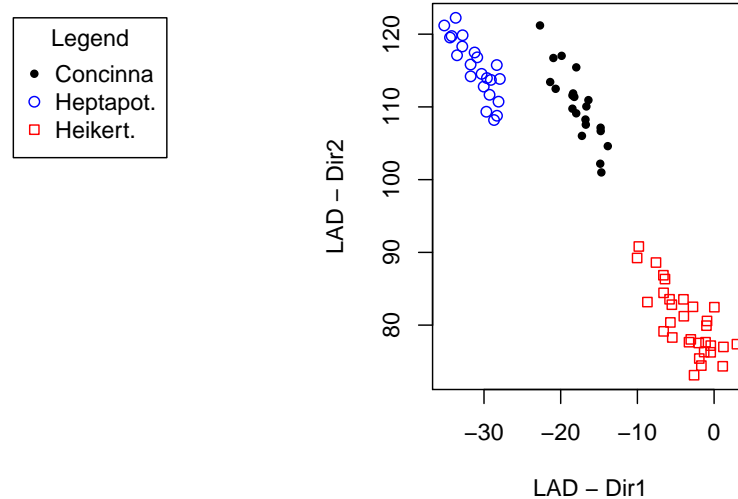
**Scatterplot of the components of the sufficient reduction**

Figure 1: Plot of the two estimated directions of the sufficient reduction of the flea data using LAD.

```
bic 2905.542 2724.587 2639.466
```

```
Large sample likelihood ratio test
```

	Stat	df	p.value
0D vs >= 1D	343.5494	18	0
1D vs >= 2D	123.8577	9	0

This result suggests that the dimension of the central subspace is  $\hat{d} = 2$ . The two directions are obtained as the sufficient reduction of the data that suffices to separate the three species. The method `plot` for the fit object plots the two directions as on Figure 1 using the code

```
R> plot(fit)
```

The second example uses the BigMac data set (Enz 1991). The data set gives the average values in 1991 on several economic indicators for 45 world cities, and contains ten continuous variables ( $X$ ). The response ( $Y$ ), which is also continuous, is the minimum labor to purchase one BigMac in US dollars. The interest is in the regression of  $Y$  on  $X$  and we seek a sufficient reduction of  $X$  such that  $Y|X \sim Y|\alpha^\top X$ , with  $\alpha \in \mathbb{R}^{p \times d}$  where  $d < p$ . Once  $\alpha$  is estimated as a basis of the central subspace, the  $d$  components of  $\alpha^\top X$  can be used in place of  $X$ . The following code lines are used to fit a LAD model.

```
R> data("bigmac", package = "ldr")
R> fit <- ldr(X = bigmac[, -1], y = bigmac[, 1], nslices = 3,
+   model = "lad", numdir.test = TRUE)
R> summary(fit)
```

## Components of the reduction vs Response

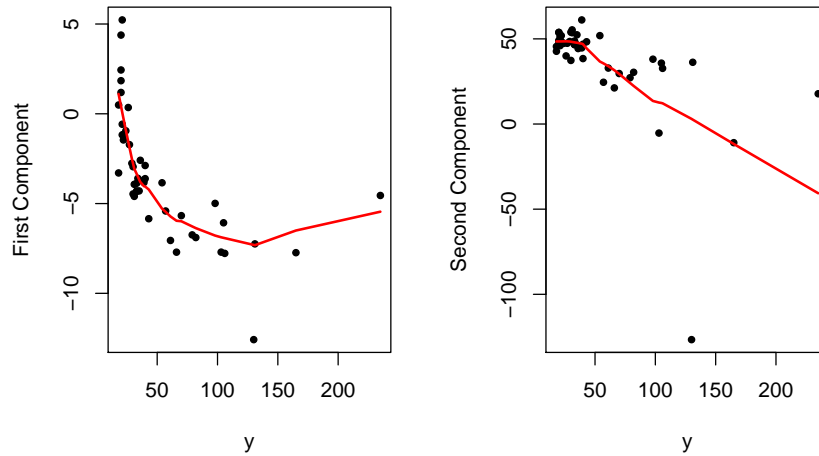


Figure 2: Plot of the two LAD estimated components of the sufficient reduction of the BigMac predictors against the response.

Since the LAD methodology requires a discrete response, we provided the arguments `nslices` to be used to discretize the continuous response. We used three slices (`nslices = 3`) so that each slice has nine observations. The summary of the fit gives the following output:

Estimated Basis Vectors for Central Subspace:

	Dir1	Dir2
Bread	-0.0329	-0.8087
BusFare	-0.9585	-0.0999
EngSal	-0.0064	0.2559
EngTax	-0.0039	-0.0346
Service	-0.0040	0.0083
TeachSal	0.2482	-0.2418
TeachTax	-0.0572	0.2806
VacDays	-0.1234	0.3628
WorkHrs	-0.0008	0.0212

Information Criterion:

	d=0	d=1	d=2
aic	3180.861	3127.153	3051.53
bic	3278.421	3246.393	3192.45

Large sample likelihood ratio test

	Stat	df	p.value
0D vs >= 1D	177.33065	24	0.000000e+00
1D vs >= 2D	99.62326	12	6.661338e-16

With three slices, the reduction would have at most two directions. The likelihood ratio test

and the information criteria agree on  $\hat{d} = 2$ . The plot of the fitted object `fit` is obtained by

```
R> plot(fit)
```

which produces Figure 2 of the two components of the reduction of the predictors against the response. The plot shows a nonlinear trend between the response and the first component of the sufficient reduction, with perhaps an influential observation. These two components are sufficient to characterize the regression of the response on the nine predictors.

### 3.3. An example with PFC

The performance of principal fitted components in estimating the central subspace may depend on the choice of the basis function  $f_y$ . It is suggested to start with a graphical exploration of the data to have an insight into the appropriate degrees of the polynomial. If quadratic relationships are suspected, it might be best to use a cubic basis  $f_y = (y, y^2, y^3)^\top$ . Piecewise constant and piecewise linear bases among others could be considered. The choice of  $f_y$  is set by the function

```
bf(y, case = c("poly", "categ", "fourier", "pcont", "pdisc"),
    degree = 1, nslices = 1, scale = FALSE)
```

The following bases are implemented: polynomial basis (`poly`), categorical response (`categ`), Fourier basis (`fourier`), piecewise polynomial continuous (`pcont`), and piecewise polynomial discontinuous (`pdisc`). The argument `degree` specifies the degree of the polynomial, and `nslices` gives the number of slices when using piecewise polynomial bases. The columns of the data-matrix can optionally be scaled to have unit variance by specifying `scale = TRUE`.

Once the basis function is determined, the structure of  $\Delta$  should be investigated. With conditionally independent predictors, an isotropic or an anisotropic PFC might be suggested, otherwise an unstructured model should be used, assuming that the sample size is larger than the number of predictors. For a given dimension  $d$  for `numdir`, the structure may be determined by a likelihood ratio test.

We reconsider the BigMac data set (Enz 1991) for an example. An initial exploration suggests using a cubic polynomial basis function. We fit three models with isotropic, anisotropic and unstructured  $\Delta$  with the following code.

```
R> fy <- bf(bigmac[, 1], case = "pdisc", degree = 0, nslices = 5)
R> fit1 <- pfc(X = bigmac[, -1], y = bigmac[, 1], fy = fy,
+   structure = "iso")
R> fit2 <- pfc(X = bigmac[, -1], y = bigmac[, 1], fy = fy,
+   structure = "aniso")
R> fit3 <- pfc(X = bigmac[, -1], y = bigmac[, 1], fy = fy,
+   structure = "unstr")
R> structure.test(fit1, fit3)
```

This output of this test has the results of the likelihood ratio test, and also of AIC and BIC.

```
Likelihood Ratio Test:
NH: structure= iso
```

```
AH: structure= unstr
      Stat df p.value
1502.188 44      0
```

```
Information Criteria:
      iso    unstr
aic 4581.696 3167.509
bic 4664.803 3330.108
```

These results provide strong evidence against the isotropic structure. Similarly, we found strong evidence against the anisotropic structure in a test against the unstructured  $\Delta$ . Thus, a model with unstructured covariance matrix is suggested. With such covariance structure, we then proceed to estimate the dimension of the central subspace. Using the following code with `numdir.test = TRUE`,

```
R> fit <- pfc(X = bigmac[, -1], y = bigmac[, 1], fy = fy,
+   structure = "unstr", numdir.test = TRUE)
```

the summary output `summary(fit)` gives results about the test on the dimension.

Estimated Basis Vectors for Central Subspace:

```
      Dir1   Dir2   Dir3   Dir4
[1,] -0.0031  0.0011 -0.0623 -0.1791
[2,]  0.9598 -0.0476  0.2630 -0.0589
[3,]  0.0617  0.1399 -0.0246  0.5565
[4,] -0.0340  0.3177 -0.0646 -0.2327
[5,]  0.0054  0.0022 -0.0119  0.0000
[6,] -0.2467 -0.4602  0.7578 -0.2421
[7,]  0.0395 -0.7402 -0.2804  0.4345
[8,]  0.1064 -0.3424 -0.5188 -0.5938
[9,]  0.0021 -0.0132 -0.0108 -0.0241
```

Information Criterion:

```
      d=0    d=1    d=2    d=3    d=4
aic 3180.861 3148.350 3148.373 3156.211 3167.509
bic 3278.421 3267.589 3285.679 3307.970 3330.108
```

Large sample likelihood ratio test

```
      Stat df    p.value
0D vs >= 1D 85.3519652 36 6.900197e-06
1D vs >= 2D 28.8407314 24 2.261826e-01
2D vs >= 3D  8.8639289 14 8.396672e-01
3D vs >= 4D  0.7018758  6 9.944505e-01
```

```
      Dir1   Dir2   Dir3   Dir4
Eigenvalues 2.5107 0.5588 0.1989 0.0157
R^2(OLS|pfc) 0.3690 0.5013 0.5514 0.5517
```

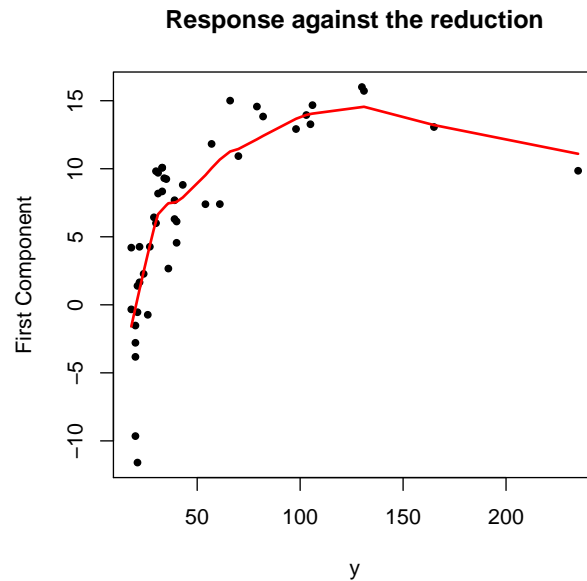


Figure 3: Plot of the sufficient reduction of the predictors against the response using PFC on the BigMac data.

The likelihood ratio test, AIC and BIC all point to the dimension of the central subspace to be  $\hat{d} = 1$ . The summary output also provides the eigenvalues of the fitted covariance matrix in decreasing order corresponding to the successive estimated reduction directions. In addition, linear regressions of the response on the components of the reduction are fitted. The coefficients of determination  $R^2$  are provided for each fit.

The final fit with `numdir = 1` is obtained with

```
R> fit <- pfc(X = bigmac[, -1], y = bigmac[, 1], fy = fy,
+   structure = "unstr", numdir = 1)
```

The plot of `fit` on Figure 3 is obtained by `plot(fit)`. It shows a clear nonlinear relationship between the reduction of the predictors and the response. This single component reduction retains all the regression information of the response that is contained in the nine predictors.

The call to `ldr` provides a number of outputs including the estimates of the parameters. The list of outputs is provided by printing the `fit` object

```
R> fit
```

which gives the following list of components that are described in the package manual.

```
[1] "R"           "Muhat"      "Betahat"   "Gammahat"
[5] "Deltahat"   "evalues"    "loglik"    "aic"
[9] "bic"        "numpar"     "numdir"    "model"
[13] "structure"  "y"          "fy"        "Xc"
[17] "call"       "numdir.test"
```



## 4. Variable screening with PFC

When dealing with a very large number of predictors and a substantial number of them do not contain any regression information about the response, it is often worth screening out these irrelevant predictors. The package contains a function called `screen.pfc` for that task. The methodology is based on the principal fitted components model. We provide a brief description and show an example of its use on a data set.

### 4.1. Inverse regression screening

We consider the normal inverse regression model (3) where  $\Delta$  is assumed independent of  $Y$ . Let  $X$  be partitioned into the vector of active predictors  $X_1$  and the vector of inactive predictors  $X_2$ . Assuming that  $X_2 \perp\!\!\!\perp X_1|Y$ , let us partition  $\Gamma$  according to the partition of  $X$  as  $(\Gamma_1^\top, \Gamma_2^\top)^\top$ . Then  $X_2 \perp\!\!\!\perp Y$  if and only if  $\Gamma_2 = 0$ .

Screening the predictors by testing whether  $\Gamma_2 = 0$  can be done by considering individual rows of  $\Gamma$ . A predictor  $X_j$  is relevant or active when the corresponding row  $\gamma_j$  of  $\Gamma$  is nonzero. We can thus test whether  $\gamma_j = 0, j = 1, \dots, p$ . Equivalently, letting  $\phi_j = \beta^\top \gamma_j$ , the screening procedure is based upon whether  $\phi_j = 0$  for each predictor  $X_j$ . The hypothesis

$$H_0 : \phi_j = 0 \quad \text{against} \quad H_a : \phi_j \neq 0 \quad (6)$$

can be tested for each of the  $p$  predictors at a specified significance level  $\alpha$ . The predictor  $X_j$  will be set as *active* or *relevant* when  $H_0$  is reject, and *inactive* or *irrelevant* otherwise. The model could be written for individual predictors as

$$X_{jy} = \mu_j + \phi_j^\top \mathbf{f}_y + \delta_j \epsilon, \quad \epsilon \sim N(0, 1), \quad (7)$$

The relevance of a predictor  $X_j$  is assessed by determining whether the mean function  $E(X_{jy}) = \mu_j + \phi_j^\top \mathbf{f}_y$  depends on the outcome  $y$ . Model (7) is a forward linear regression model where the ‘‘predictor’’ vector is  $\mathbf{f}_y$  and the ‘‘response’’ is  $X_j$ . An  $F$  test statistic can be obtained for the hypotheses (6). The predictor  $X_j$  is relevant if the model yields an  $F$  statistic larger than a user-specified cutoff value. The cutoff may correspond to a significance level  $\alpha$  such as 0.1 or 0.05 for example. The methodology was initially described in [Adraghi and Cook \(2008\)](#).

### 4.2. Illustration

We applied the screening method to a chemometrics data set. The data is from a study to predict the functional hydroxyl group *OH* activity of compounds from molecular descriptors. The response (*act*) is the activity level of the compound, and the predictors are the nearly 300 descriptors. The response and predictors are continuous variables and 719 observations were recorded. The data set was provided by Tomas Oberg. We screened the descriptors at a 5% significance level.

The choice of the basis function determines the type of screening. For example, a correlation screening is obtained with a one degree polynomial basis function. A correlation screening selects the predictors linearly related to the response. The following example gives the correlation screening.

Descriptor	<i>p</i> value CS	<i>p</i> value PL
SssCH2	0.07863	0.000
tets2	0.05993	0.000
ka2	0.09536	0.000
jy	0.89416	0.000
jb	0.68736	0.000
bpc4	0.61782	0.000
spc4	0.33737	0.000
sc3	0.05593	0.000

Table 1: Screening *p* values under the correlation (CS) and the piecewise linear discontinuous basis (PL).

```
R> data("OH", package = "ldr")
R> X <- OH[, -c(1, 295)]; y <- OH[, 295]
R> out <- screen.pfc(X, fy = bf(y, case = "poly", degree = 1), cutoff = 0.05)
```

The output shows an *F* statistic and *p* value for the test of dependence between the predictors and the response, where the latter is captured through a basis function. The **Index** corresponds to the index of the predictors. The predictors are sorted in the decreasing order of their *F* value. The following output is obtained with `head(out)`.

	F	P-value	Index
SdsCH	226.84789	0	231
SHCsatu	141.52605	0	275
SHvin	130.34194	0	276
SHother	97.05707	0	220
numHBa	81.54723	0	273
SHHBA	80.56008	0	281

The application of the correlation screening captured a set of 203 descriptors related to the response. However using a piecewise linear discontinuous basis function with two slices, the screening yielded a set of 235 descriptors, including 41 descriptors not detected by the correlation screening. Table 1 is a short list of some of the 41 descriptors captured by the piecewise linear discontinuous basis that were not selected by the correlation screening at the 5% significance level.

This example shows that, as expected, when a relevant predictor is not linearly related to the response, a correlation screening may fail to capture it. However, with the appropriate choice of the basis function, all relevant predictors would be detected. Graphical investigations showed non-linear relationships between the response and some predictors listed on Table 1. Figure 4 plots  $\log(\text{act})$  against compounds `sc3` and `tets2` as two examples of predictors which were not selected by the correlation screening. A clear nonlinear relationship between the predictors and the log-transformed response is displayed. We used the log-transformation of the response to help see the graphical evidence of nonlinear relationship between the response and the mentioned predictors. The solid line is a loess smooth curve.

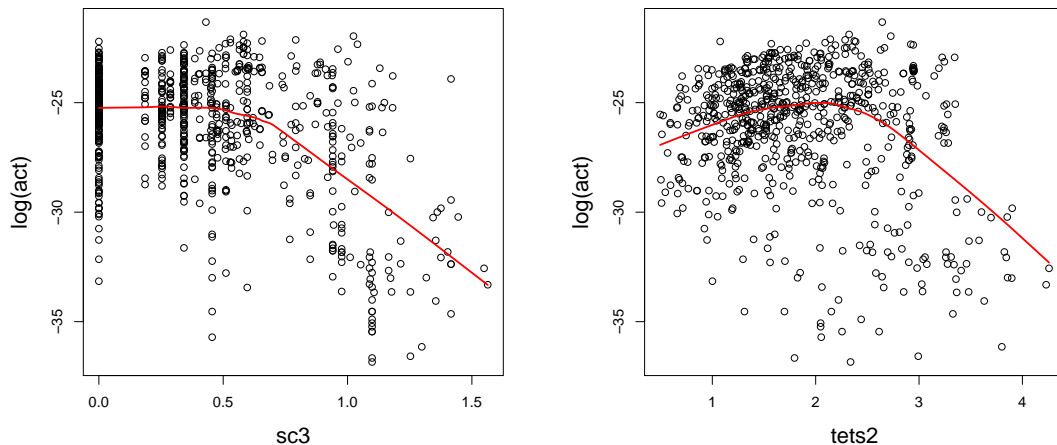


Figure 4: Scatter-plots of `sc3` and `tets2` against the log-transformed response `act` showing nonlinear relationships.

## 5. Conclusions

We have presented `ldr`, an R package for dimension reduction in regression. The package implements three recent likelihood-based methods for sufficient dimension reduction and a variable screening methodology. It can be considered as the likelihood-based addition to the R package `dr`. The package presents an architecture that allows the implementation of other sufficient dimension reduction methods such as the envelope model of Cook *et al.* (2010), which will be added in the near future. We foresee continuing the development of the package into a comprehensive library of tools and methods for likelihood-based sufficient dimension reduction in R.

## Acknowledgements

The authors thank Tomas Oberg for the OH data set, and also thank the reviewers for their thorough review of the manuscript and the software.

## References

- Adraghi K, Raim A (2014). *ldr: Methods for Likelihood-Based Dimension Reduction in Regression*. R package version 1.3.3, URL <http://CRAN.R-project.org/package=ldr>.
- Adraghi KP (2009). *Dimension Reduction and Prediction in Large  $p$  Regressions*. Ph.D. thesis, University of Minnesota, Minneapolis.
- Adraghi KP, Cook RD (2008). “Discussion on the “Sure Independence Screening for Ultrahigh Dimensional Feature Space” by Jianqing Fan and Jinchi Lv (2007).” *Journal of the Royal Statistical Society B*, **70**(5), 893.
- Adraghi KP, Cook RD (2009). “Sufficient Dimension Reduction and Prediction in Regression.” *Philosophical Transactions of the Royal Society A*, **367**(1906), 4385–4405.

- Adraghi KP, Cook RD, Wu S (2012). “**GrassmannOptim**: An R Package for Grassmann Manifold Optimization.” *Journal of Statistical Software*, **50**(5), 1–18. URL <http://www.jstatsoft.org/v50/i05/>.
- Burnham K, Anderson D (2002). *Model Selection and Multimodel Inference*. John Wiley & Sons.
- Cook RD (1994). “Using Dimension-Reduction Subspaces to Identify Important Inputs in Models of Physical Systems.” In *Proceedings of the Section on Physical and Engineering Sciences*, pp. 18–25. American Statistical Association.
- Cook RD (1998). *Regression Graphics*. John Wiley & Sons.
- Cook RD (2007). “Fisher Lecture: Dimension Reduction in Regression.” *Statistical Science*, **22**(1), 1–26.
- Cook RD, Forzani L (2008a). “Covariance Reducing Models: An Alternative to Spectral Modelling of Covariance.” *Biometrika*, **95**(4), 799–812.
- Cook RD, Forzani L (2008b). “Principal Fitted Components in Regression.” *Statistical Science*, **23**(4), 485–501.
- Cook RD, Forzani L (2009). “Likelihood-Based Sufficient Dimension Reduction.” *Journal of the American Statistical Association*, **104**(485), 197–208.
- Cook RD, Forzani L, Tomassi DR (2012). “**LDR**: A Package for Likelihood-Based Sufficient Dimension Reduction.” *Journal of Statistical Software*, **39**(3), 1–20. URL <http://www.jstatsoft.org/v39/i03/>.
- Cook RD, Li B, Chiaromonte F (2010). “Envelope Models for Parsimonious and Efficient Multivariate Linear Regression.” *Statistica Sinica*, **20**(3), 927–1010.
- Cook RD, Ni L (2005). “Sufficient Dimension Reduction via Inverse Regression: A Minimum Discrepancy Approach.” *Journal of the American Statistical Association*, **100**(470), 927–1010.
- Cook RD, Weisberg S (1991). “Discussion of “Sliced Inverse Regression for Dimension Reduction” by Ker-Chau Li.” *Journal of the American Statistical Association*, **86**(414), 328–332.
- Enz R (1991). “Prices and Earnings around the Globe.” Published by the Union Bank of Switzerland.
- Li B, Wang S (2007). “On Directional Regression for Dimension Reduction.” *Journal of American Statistical Association*, **102**(479), 997–1008.
- Li KC (1991). “Sliced Inverse Regression for Dimension Reduction.” *Journal of the American Statistical Association*, **86**(414), 316–342.
- Li KC (1992). “On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein’s Lemma.” *Journal of the American Statistical Association*, **87**(420), 1025–1039.

- Lubischew AA (1962). “On the Use of Discriminant Functions in Taxonomy.” *Biometrics*, **18**(4), 455–477.
- Phillips PC, Arnold SJ (1999). “Hierarchical Comparison of Genetic Variance-Covariance Matrix Using the Flury Hierarchy.” *Evolution*, **53**(5), 1506–1515.
- R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- The MathWorks, Inc (2012). *MATLAB – The Language of Technical Computing, Version 8 (R2012b)*. Natick, Massachusetts. URL <http://www.mathworks.com/products/matlab/>.
- Weisberg S (2002). “Dimension Reduction Regression in R.” *Journal of Statistical Software*, **7**(1), 1–22. URL <http://www.jstatsoft.org/v07/i01/>.

**Affiliation:**

Kofi Placid Adragani  
Department of Mathematics & Statistics  
University of Maryland, Baltimore County  
1000 Hilltop Circle  
Baltimore, MD 21250, United States of America  
Phone: +1/410/455-2406  
Fax: +1/410/455-1066  
E-mail: [kofi@umbc.edu](mailto:kofi@umbc.edu)  
URL: <http://www.math.umbc.edu/~kofi/>