Author/s:
Prendergast, LA;Garnham, AL

Title:
Response and predictor folding to counter symmetric dependency in dimension reduction

Date:
2016-12-01

Citation:
Prendergast, L. A. & Garnham, A. L. (2016). Response and predictor folding to counter symmetric dependency in dimension reduction. Australian and New Zealand Journal of Statistics, 58 (4), pp.515-532. https://doi.org/10.1111/anzs.12170.

Persistent Link:
https://hdl.handle.net/11343/292157

 doi: 10.1111/j.1467-842X.XXX

# RESPONSE AND PREDICTOR FOLDING TO COUNTER SYMMETRIC DEPENDENCY IN DIMENSION REDUCTION

L. A. PRENDERGAST[1*] AND A. L. GARNHAM[2]

*La Trobe University, Walter and Eliza Hall Institute of Medical Research*



**Summary**

In the regression setting, dimension reduction allows for complicated regression structures to be detected via visualisation in a low-dimensional framework. However, some popular dimension reduction methodologies fail to achieve this aim when faced with a problem often referred to as symmetric dependency. In this paper we show how vastly superior results can be achieved when carrying out response and predictor transformations for methods such as least squares and sliced inverse regression. These transformations are simple to implement and utilise estimates from other dimension reduction methods that are not faced with the symmetric dependency problem. We highlight the effectiveness of our approach via simulation and an example. Furthermore, we show that ordinary least squares can effectively detect multiple dimension reduction directions. Methods robust to extreme response values are also considered.

*Key words:* cumulative slicing estimation; ordinary least squares, principal Hessian directions, robust $M$-estimation, sliced inverse regression, sliced average variance estimates

## 1. Introduction

Let $Y \in \mathbb{R}$ denote a random univariate response and $\boldsymbol{x} \in \mathbb{R}^p$ a random $p$-dimensional vector of predictors. Li & Duan (1989) considered the model

$$Y = f(\boldsymbol{\beta}^\top \boldsymbol{x}, \varepsilon) \tag{1}$$

where $\boldsymbol{\beta}$ is an unknown $p$-dimensional vector of predictor coefficients, $f$ is the unknown link function and $\varepsilon$ is the error term that is assumed to be independent of $\boldsymbol{x}$. Of interest is the regression function $\mathrm{E}(Y|\boldsymbol{x})$ where, ideally, a plot of $Y$ versus $\boldsymbol{x}$ can reveal the form of $f$. However, we are limited in this sense when $p$ is large due to our inability to visualise objects in high dimensions. Importantly, $Y$ depends on $\boldsymbol{x}$ only through $\boldsymbol{\beta}^\top \boldsymbol{x}$ so that if we could determine $\boldsymbol{\beta}$ then we could replace the $p$-dimensional $\boldsymbol{x}$ with the one-dimensional $\boldsymbol{\beta}^\top \boldsymbol{x}$.

*Author to whom correspondence should be addressed.
[1]Department of Mathematics and Statistics, La Trobe University, Melbourne, Australia, 3086. luke.prendergast@latrobe.edu.au
[2] Division of Bioinformatics, Walter and Eliza Hall Institute of Medical Research
Acknowledgements. The authors are very thankful to two anonymous referees and an Associate Editor for their useful comments and suggestions that lead to a clearer and more accessible manuscript.

Our ability to explore possibilities for $f$ would then be enhanced due to the resulting lower-dimensional framework.

In the sample setting, let $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$ be $n$ sample realisations of $Y$ and $\boldsymbol{x}$ where the relationship between $Y$ and $\boldsymbol{x}$ is assumed to be of the form given in (1). Suppose that $\boldsymbol{\beta}$ can be estimated and let this estimate be denoted $\widehat{\boldsymbol{\beta}}$. Then the $y_i$s can be plotted against the $\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}_i$s to visually determine $f$. Such a plot is called an *Estimated Sufficient Summary Plot* (ESSP, see, e.g., Cook 1998b). The focus of our work here will be to obtain good ESSPs in settings for which estimation of $\boldsymbol{\beta}$ is difficult.

While our initial focus will be on the model in (1), later we will also look at a further generalised model studied by Li (1991) and given as

$$Y = f(\boldsymbol{\beta}_1^\top \boldsymbol{x}, \ldots, \boldsymbol{\beta}_K^\top \boldsymbol{x}, \varepsilon) \tag{2}$$

which allows for more than one vector of predictor coefficients.

Li & Duan (1989) extended earlier works by Brillinger (1977, 1983) to show that ordinary least squares (OLS), and robust versions, can be used to estimate the direction of $\boldsymbol{\beta}$ when the model is of the form (1) and under some mild conditions on $\boldsymbol{x}$. We will provide a brief review of these results in Section 2. However, for some forms of $f$ OLS is not expected to find $\boldsymbol{\beta}$. Consequently we also discuss another approach, Principal Hessian Directions (PHD) introduced by Li (1992), which may often be more suitable for models with these forms of $f$. In Section 3 we propose a simple transformation of the response based on an initial PHD estimate that can be used to ensure that OLS can provide a good ESSP. Simulation results, described in Section 4, demonstrate that this approach can be used to obtain vastly superior estimates. Extensions to other approaches are discussed in Section 5. These have applications to the multi-index model given in (2). Finally, an example is provided in Section 6 and the paper is concluded with a discussion in Section 7.

## 2. Methods

In the context of the multi-index model in (2), Li (1991) defined the *Linear Design Condition* to be:

**Condition 1.** *For any* $\boldsymbol{c} \in \mathbb{R}^p$, $E(\boldsymbol{c}^\top \boldsymbol{x} | \boldsymbol{\beta}_1^\top \boldsymbol{x}, \ldots, \boldsymbol{\beta}_K^\top \boldsymbol{x})$ *is linear in* $\boldsymbol{\beta}_1^\top \boldsymbol{x}, \ldots, \boldsymbol{\beta}_K^\top \boldsymbol{x}$.

While this condition holds when $\boldsymbol{x}$ belongs to the family of elliptically symmetric distributions, Hall & Li (1993) showed that Condition 1 often approximately holds in practice when $p$ is large. One also has the possibility to utilise predictor transformations to ensure that it approximately holds (see, e.g., Fox & Weisberg 2011).

Li & Duan (1989) had earlier considered Condition 1 in the context of the model in (1) (i.e. when $K = 1$ in (2)). This will be our focus for the remainder of this section and we will revisit the case of general $K$ again later.

### 2.1. Least squares and similar approaches

When Condition 1 and the model in (1) hold, Li & Duan (1989) showed that the OLS slope vector, denoted $\boldsymbol{b} = \text{var}(\boldsymbol{x})^{-1}\text{Cov}(\boldsymbol{x}, Y)$, is equal to $c\boldsymbol{\beta}$ for some $c \in \mathbb{R}$. Consequently, OLS can recover the direction of $\boldsymbol{\beta}$ when $c \neq 0$ and a plot of $Y$ versus $\boldsymbol{b}^\top \boldsymbol{x}$ can be used to seek $f$. It should be pointed out that any non-zero $c$ is adequate since any $\boldsymbol{b}$ in the direction of $\boldsymbol{\beta}$ is suitable for finding an appropriate link function. In practice , OLS can be used to obtain $\widehat{\boldsymbol{b}}$, the usual OLS slope estimate, and an ESSP created using the $y_i$s and the $\widehat{\boldsymbol{b}}^\top \boldsymbol{x}_i$s. While OLS is one simple approach (e.g. Brillinger 1983), Li and Duan generalised Brillinger's results to include estimators satisfying

$$\underset{a,\boldsymbol{b}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho(a + \boldsymbol{b}^\top \boldsymbol{x}_i, y_i) \tag{3}$$

provided that the function $\rho$ is convex in its first argument and that a solution exists. Hence, other possibilities can be robust estimators such as $M$-estimators with the Huber weight function (Huber 1973). While the temptation would be to consider a robust approach only when possible errors are present in the data set, Prendergast & Sheather (2013) showed that for some models the robust estimators can outperform OLS even when data is sampled without error. Similarly, Prendergast (2008) used trimming of influential observations to also improve estimates.

If an estimator is expected to find the direction of $\boldsymbol{\beta}$, then it is required that $c \neq 0$. Some discussion of the case $c = 0$ can be found in Li (1991) and Cook & Weisberg (1991). While these discussions are for a different method they can similarly be applied to OLS. That is, when the link function $f$ is symmetric about the mean of $\boldsymbol{\beta}^\top \boldsymbol{x}$, then $\boldsymbol{b} = \boldsymbol{0}$. To highlight this, we consider two simulated examples. The first does not have the symmetric dependency issue while the second does. The models we use are

**Model 1.** $Y = \sin(0.7\boldsymbol{\beta}^\top \boldsymbol{x}) + 0.35\varepsilon$

**Model 2.** $Y = \cos(0.7\boldsymbol{\beta}^\top \boldsymbol{x}) + 0.35\varepsilon$

where, for both models, $\boldsymbol{x}$ is a 10 dimensional random vector which is distributed as $\text{N}(\boldsymbol{0}, \boldsymbol{I})$, $\varepsilon \sim \text{N}(0, 1)$ and $\boldsymbol{\beta} = (1, -2, 0, \ldots, 0)^\top$.

******** Figure 1 about here ********

In Figure 1 we provide true views (where the $y_i$s are plotted against the ideally dimension reduced $\boldsymbol{x}_i$s, i.e. the $\boldsymbol{\beta}^\top \boldsymbol{x}_i$s) and ESSPs where OLS has been used as the estimator. Plots A and B are for Model 1 and Plots C and D for Model 2 where, in both cases, $n = 200$ observations have been randomly generated. If OLS has performed well, then we would expect the ESSP to look similar to the true views, although with possible differences in scale on the horizontal axis, since OLS is targeting $c\boldsymbol{\beta}$ for a $c$ that is not necessarily one. For Model 1, we can see that OLS has performed exceptionally well, producing an excellent ESSP as seen in Plot B. However, OLS has failed for Model 2 with an ESSP in Plot D that does not provide any evidence of a relationship between the responses and dimension reduced predictors. The true view shows that there is certainly something to find. Recall that Model 2 exhibits symmetric dependency and OLS is trying to estimate $0 \times \boldsymbol{\beta}$.

### 2.2. Principal Hessian Directions

Li (1992) introduced Principal Hessian Directions (PHD), a method that does not suffer from the symmetric dependency problem and one that is also capable of finding multiple vectors of predictor coefficients. That is, the model can be assumed to be of the form shown in (2) where $K \geq 1$, in which case it is desirable to find a basis for the span of the $\boldsymbol{\beta}_k$s.

Consider the following condition required by PHD.

**Condition 2.** $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

When Condition 2 holds, imposing normality of the predictor, Condition 1 also holds. As a consequence, if PHD is applicable due to this condition being met, then so too is OLS. There are slightly weaker conditions needed for PHD, namely that $\text{var}(\boldsymbol{x}|\boldsymbol{\beta}^\top \boldsymbol{x})$ is constant in conjunction with assuming that Condition 1 holds. Recent work by Leeb (2013) showed that this will often hold approximately in practice.

While OLS returns an estimate of $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{xy}$ (where $\boldsymbol{\Sigma}_{xy}$ is the covariance vector between $\boldsymbol{x}$ and $Y$) as an estimate of the direction of $\boldsymbol{\beta}$, PHD instead carries out an eigen-decomposition of an estimate of $\overline{\boldsymbol{H}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{yxx}\boldsymbol{\Sigma}^{-1}$, where

$$\boldsymbol{\Sigma}_{yxx} = \text{E}\left\{(Y - \text{E}(Y))(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top\right\}.$$

When Condition 2 holds, for *many models* satisfying (1), the rank of $\overline{\boldsymbol{H}}$ is one and the eigenvector corresponding to the non-zero eigenvalue is in the same direction as $\boldsymbol{\beta}$. We have emphasised "many models" here since PHD will not be able to find the direction of $\boldsymbol{\beta}$ when there is an odd symmetric dependency between $Y$ and the mean of $\boldsymbol{\beta}^\top \boldsymbol{x}$. Here odd symmetric dependency refers to the type of symmetry seen for Model 1 (see Figure 1). Li (1991) also noted that $Y$ can be replaced by the OLS residual without changing $\overline{\boldsymbol{H}}$ where, notationally,

we replace $\boldsymbol{\Sigma}_{yxx}$ by $\boldsymbol{\Sigma}_{rxx}$ to distinguish between the two approaches. While $\overline{\boldsymbol{H}}$ does not change, the estimator is influenced. Empirical and theoretical results have suggested that this residual-based PHD approach is often a better estimator of $\boldsymbol{\beta}$ (Cook 1998a; Prendergast & Smith 2010). Consequently the residual-based PHD will be our method of choice.

Since $\boldsymbol{\Sigma}_{rxx}$ is moment-based, estimation is straightforward. Let $r_1, \ldots, r_n$ denote the usual OLS residuals for the regression of the $y_i$s on the $\boldsymbol{x}_i$s and also let $\bar{\boldsymbol{x}}$ be the sample mean of the $\boldsymbol{x}_i$s. Then the estimate of $\boldsymbol{\Sigma}_{rxx}$ is

$$\widehat{\boldsymbol{\Sigma}}_{rxx} = \frac{1}{n}\sum_{i=1}^{n} r_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top.$$

Lue (2001) has previously shown that, similarly to OLS estimation, trimming can improve PHD estimation.

## 3. Predictor and response transformations to remove symmetric dependency

Garnham & Prendergast (2013) showed that response transformations can greatly improve OLS estimates. Their results, however, do not solve the issue of the symmetric dependency that troubles OLS. The aim here is to introduce two transformation functions, one for the response and the other for the predictor, that can be useful in the symmetric dependency setting.

### 3.1. Theory

We start by introducing the transformations by discussing their motivation. The response transformation that we will focus on is

$$t_y(Y;\boldsymbol{v}) = \begin{cases} Y, & \boldsymbol{v}^\top\boldsymbol{x} > \boldsymbol{v}^\top\boldsymbol{\mu} \\ Y - 2(Y - \mathrm{E}(Y|\boldsymbol{v}^\top\boldsymbol{x} = \boldsymbol{v}^\top\boldsymbol{\mu})), & \boldsymbol{v}^\top\boldsymbol{x} \leq \boldsymbol{v}^\top\boldsymbol{\mu} \end{cases} \tag{4}$$

where $\boldsymbol{v}$ needs to be chosen. If $\boldsymbol{v} = c_1\boldsymbol{\beta}$ for a non-zero scalar $c_1$, then $t_y(Y;\boldsymbol{v})$ and $\boldsymbol{x}$ still satisfy the model in (1) and Condition 1. An estimator of $\boldsymbol{\beta}$ is then the OLS slope vector estimator for the regression of $t_y(Y;\boldsymbol{v})$ on $\boldsymbol{x}$. Soon we will show that, in the empirical setting, good estimates of $\boldsymbol{\beta}$ used in the transformation function can generate vastly improved results. That is, if we can find a good estimate of $\boldsymbol{\beta}$, then we can use this estimate to transform the response and obtain a vastly improved estimate of $\boldsymbol{\beta}$ for some models. We also consider the following predictor transformation function

$$t_x(\boldsymbol{x} - \boldsymbol{\mu};\boldsymbol{v}) = \mathrm{sign}(\boldsymbol{v}^\top\boldsymbol{x} - \boldsymbol{v}^\top\boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu}). \tag{5}$$

******** Figure 2 about here ********

In Figure 2 we show the effects of the transformations on the curves defining $Y$ in Model 2 under the assumption of zero error. In Plot A the original curve for Model 2 is displayed as the thick grey line. It is clear that this curve is symmetric around $\boldsymbol{\beta}^\top \boldsymbol{\mu} = 0$, which is marked by the vertical line. We can also see that $\mathrm{E}(Y|\boldsymbol{\beta}^\top \boldsymbol{x} = 0) = 1$, and this is shown as the horizontal line on the plot. The black line is the curve resulting from the transformation in (4). The curve in the lower-left quadrant is folded up to the top-left quadrant thus removing the symmetric dependency. Similarly, in Plot B we show the effect of employing the predictor transformation in (5). Here, the curve in the lower-left quadrant is folded over to the bottom-right quadrant. Again, the symmetric dependency is removed. For further clarity, in Plots C and D we show how the transformations work with respect to 500 observations generated according to Model 2. The original data are the grey points and the transformed data are the black points. As can be seen in these plots, the transformed data do not suffer from symmetric dependency

In Theorem 1 below, we show that for certain choices of $\boldsymbol{v}$, the transformation in (5) can be useful for finding directions hidden by symmetric dependency. The result is provided in the more general context of the multi-index model in (2). The proof is in the Appendix.

**Theorem 1.** *Consider the predictor transformation given in (5) and let $\boldsymbol{v} \in \mathcal{S}$ where $\mathcal{S}$ is the column space of the $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$. Under the model in (2) and Condition 1,*

$$var(\boldsymbol{x})^{-1} cov\left(t_x(\boldsymbol{x} - \boldsymbol{\mu}; \boldsymbol{v}), Y\right) \in \mathcal{S} \tag{6}$$

$$var(t_x(\boldsymbol{x} - \boldsymbol{\mu}; \boldsymbol{v}))-1 cov\left(t_x(\boldsymbol{x} - \boldsymbol{\mu}; \boldsymbol{v}), Y\right) \in \mathcal{S}. \tag{7}$$

The estimator in (7) is simply the OLS slope from the regression of $Y$ on $t_x(\boldsymbol{x} - \boldsymbol{\mu}; \boldsymbol{v})$. The estimator in (6) is similar although it utilises the variance estimator for the original $\boldsymbol{x}$.

### 3.2. Application in practice

Recall that our sample of $n$ observations is denoted $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$. Throughout let $\bar{y}$, $\bar{\boldsymbol{x}}$, $\boldsymbol{S}_x$ and $\boldsymbol{S}_{xy}$ denote the sample mean of the $y_i$s, sample mean of the $\boldsymbol{x}_i$s, sample covariance matrix of the $\boldsymbol{x}_i$s and the sample covariance between the $\boldsymbol{x}_i$s and $y_i$s respectively. Also let $\boldsymbol{X}$ denote the $n \times p$ design matrix whose $i$-th row is $\boldsymbol{x}_i$.

There are two points that need clarification prior to application in practice. Firstly, how to choose an appropriate vector $\boldsymbol{v}$? Our simulations indicate that for many models, OLS is a better estimator of $\boldsymbol{\beta}$ in (1) than PHD. However, PHD is preferred when symmetric dependency is evident, in which case OLS can struggle to find $\boldsymbol{\beta}$. Consequently, in practice we propose to set $\boldsymbol{v} = \widehat{\boldsymbol{b}}_{phd}$ - the PHD estimate of $\boldsymbol{\beta}$. In the next section, our results show

that reasonable PHD estimates of $\boldsymbol{\beta}$ can lead to much improved estimates of $\boldsymbol{\beta}$ when OLS is employed following the transformations in Section 3.1.

Secondly, the response transformation requires estimation of $\mathrm{E}(Y|\boldsymbol{v}^\top\boldsymbol{x} = \boldsymbol{v}^\top\boldsymbol{\mu})$. We propose to find an approximation to this estimate as

$$\bar{y}(\boldsymbol{v}) = \widehat{f}(\boldsymbol{v}^\top\bar{\boldsymbol{x}}) \tag{8}$$

where $\widehat{f}$ is a fitted spline for the $y_i$s versus the $\boldsymbol{v}^\top\boldsymbol{x}_i$s. We use a cubic smoothing spline (for details see, e.g., Faraway 2006). If computational issues arise with fitting the spline, such as can happen when $n$ is small, then a simple option is to use $\sum_{j\in I_m} y_j/m$ with, for example, $m = 10$ and where $I_m$ is the set of indices for the closest $m$ $\boldsymbol{v}^\top\boldsymbol{x}_i$s to $\boldsymbol{v}^\top\bar{\boldsymbol{x}}$.

The transformations we employ are then

$$y_i^* = t_{y_i}(y_i;\widehat{\boldsymbol{b}}_{phd}) = \begin{cases} y_i, & \widehat{\boldsymbol{b}}_{phd}^\top\boldsymbol{x}_i > \widehat{\boldsymbol{b}}_{phd}^\top\bar{\boldsymbol{x}} \\ y_i - 2(y_i - \bar{y}(\widehat{\boldsymbol{b}}_{phd})), & \widehat{\boldsymbol{b}}_{phd}^\top\boldsymbol{x}_i \le \widehat{\boldsymbol{b}}_{phd}^\top\bar{\boldsymbol{x}} \end{cases} \tag{9}$$

and

$$\boldsymbol{x}_i^* = t_{x_i}(\boldsymbol{x}_i - \bar{\boldsymbol{x}};\widehat{\boldsymbol{b}}_{phd}) = \mathrm{sign}(\widehat{\boldsymbol{b}}_{phd}^\top\boldsymbol{x}_i - \widehat{\boldsymbol{b}}_{phd}^\top\bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \tag{10}$$

where we use the notation $y_i^*$ and $\boldsymbol{x}_i^*$ for convenience. The methods we use are:

**Method 1.** *The OLS slope vector for the regression of the $y_i^*$s on the $\boldsymbol{x}_i$s.*

**Method 2.** *The OLS slope vector for the regression of the $y_i$s on the $\boldsymbol{x}_i^*$s.*

Potentially, a combination of the transformations could also be used. However, our simulations revealed that better results are achieved by using only one at a time. Another possibility is to use $\boldsymbol{S}_x^{-1}\boldsymbol{S}_{xy}^*$ . However, our simulations also revealed that this approach was very typically inferior to the other two. For brevity, we therefore do not consider this approach further.

### 3.3. An iterative approach

A potential problem with the transformation methods is that the initial estimated direction is poor. However, one approach to alleviate this is to apply an iterative scheme which starts with the initial estimate, obtains a new estimate after transformation and iteratively uses the new estimate as the initial estimate until convergence. A general algorithm for this approach is:

**Step 0.1:** Estimate the direction of $\boldsymbol{\beta}$ using PHD and denote this as $\widehat{\boldsymbol{b}}^{(1)}$.

**Step 0.2:** Set $i = 1$ and tol.met = FALSE.

**Step $i$:** While tol.met is FALSE do

**Step $i$.1:** Apply **Method j** using $\widehat{\boldsymbol{b}}^{(i)}$ as the direction for transformation and obtain a new estimated direction $\widehat{\boldsymbol{b}}^{(i+1)}$.

**Step $i$.2:** If $1 - \text{cor}^2(\boldsymbol{X}\widehat{\boldsymbol{b}}^{(i)}, \boldsymbol{X}\widehat{\boldsymbol{b}}^{(i+1)}) <$ tol then set tol.met to TRUE.

**Step $i$.3:** Increment $i = i + 1$.

**Step $i + 1$:** Return $\widehat{\boldsymbol{b}}^{(i)}$ as the final estimate to the direction of $\boldsymbol{\beta}$.

We have used $\text{cor}^2(\boldsymbol{X}\widehat{\boldsymbol{b}}^{(i)}, \boldsymbol{X}\widehat{\boldsymbol{b}}^{(i+1)})$ in the criterion for exiting the iterating loop since, when there is correlation present amongst the columns of $\boldsymbol{X}$, notably different directions can result in very similar values of the ESSP, which is the targeted estimate. However, substantial differences in the squared correlation will similarly be accompanied by changes in the ESSP. Such an assessment is often used; for example, Li (1991) used the squared trace correlation, which is a multi-index version, to compare collections of estimated directions in dimension reduction.

## 4. Simulations

In this section we consider the performance of the transformation approaches defined earlier as Methods 1 and 2. Comparisons are also made with standard OLS and PHD estimation before we consider other methods in the next section.

******** Table 1 about here ********

In Table 1 we provide simulated average squared correlations (with standard deviations in italics) for Model 2 over 10,000 runs between $\boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{X}\widehat{\boldsymbol{b}}$, the true and estimated dimension reduced predictors. The estimators considered are OLS, PHD and the transformation approaches. Both $p = 10$ and $p = 20$ were considered. As expected OLS performs poorly due the symmetric dependency evident in the model, while PHD performs well. However, Methods 1 and 2 perform exceptionally well, having successfully drawn on the good PHD estimates to remove the symmetric dependency. The iterative estimation methods also provide improved estimates. We exited the iterative procedure when $\text{cor}^2(\boldsymbol{X}\widehat{\boldsymbol{b}}^{(i)}, \boldsymbol{X}\widehat{\boldsymbol{b}}^{(i+1)}) \geq 0.999$ or when ten iterations were reached. The small standard deviations for the methods indicate consistently excellent results.

Prendergast & Sheather (2013) found that robust least squares regression methods can provide improved single-index model estimates even when the data are well-behaved in the sense that they were sampled from a single index model and with a normal $\boldsymbol{x}$. Similarly, Prendergast (2008) found that trimming observations in the estimation step can also improve outcomes. We further explore this by considering the following model:

**Model 3.** $Y = 1/|\boldsymbol{\beta}^\top \boldsymbol{x}| + 0.2\varepsilon$ *with* $p = 20$ *and* $\boldsymbol{\beta} = (1, -2, 0, \ldots, 0)^\top$.

For this model we will assume that $\boldsymbol{x}$ is a random vector of dimension 10, distributed as $\mathrm{N}(\mathbf{0}, \boldsymbol{I})$ so that this model also includes the symmetric dependency that troubles OLS. Moreover data simulated from this model can result in exceptionally extreme responses, since the denominator in the first term on the right hand side can be very close to zero.

******** Table 2 about here ********

In Table 2 we report the average squared correlations between the true and estimated dimension reduced predictors for 10,000 simulated runs from Model 3, with standard deviations in italics. For PHD, the very large response values often generated can result in extremely poor results. However, for OLS Garnham & Prendergast (2013) showed that using the rank of the response instead of the response itself could provide improved results. Consequently, we used the rank of the response values for PHD and this approach provides vast improvements. Therefore the PHD results presented in this table are based on this estimation approach. As well as employing OLS, PHD based on ranks and Methods 1 and 2, we also consider other variations that can be used to limit the influence of very large response values. RR, RM1 and RM2 refer to the usual OLS, Methods 1 and 2 but where OLS has been replaced with the $M$-estimation robust version (Huber 1964, 1973) with the Huber weight function. To do this we used the `rlm` function from the `MASS` package (Venables & Ripley 2002) in R (R Core Team 2013). M1-trim and M2-trim refer to Methods 1 and 2 where 10% of observations with the largest Cook's distance have been trimmed prior to the least squares step. This is one of the trimming procedures from Prendergast (2008). The iterative estimation scheme for this model and methods did not provide improved results so for simplicity the results are not presented here. Not surprisingly, OLS and the $M$-estimation equivalent completely fail even for large $n$ due to symmetric dependency. Methods 1 and 2 perform much better but can still struggle as is made evident by the moderate average squared correlations and large standard deviations. On the other hand PHD performs well, in particular for the larger sample size settings. For Methods 1 and 2 coupled with $M$-estimation, we see improved performance over PHD for both methods. These results suggest that, by using the good PHD results in the transformation step to remove the symmetric dependency problem and then $M$-estimation to protect against large response values, excellent results can be achieved. For the trimming approaches, improvements have been found in comparison with standard Methods 1 and 2. However, the results are a little worse than PHD and much worse than the transformation plus $M$-estimation methods.

## 5. Inverse regression methods and multiple direction OLS

### 5.1. Inverse regression approaches

The transformations discussed in Section 3.2 are certainly not limited to OLS and PHD. Here we briefly discuss the use of sliced inverse regression (SIR, Li 1991) and sliced average variance estimates (SAVE, Cook & Weisberg 1991). For brevity, we only briefly discuss these methods here and the reader is directed to the aforementioned articles for more detail. Let $S_1, \ldots, S_H$ denote $H$ non-overlapping yet collectively exhaustive intervals covering the range of $Y$. Let $\boldsymbol{\mu}_h = \mathrm{E}(\boldsymbol{x}|Y \in S_h)$ $(h = 1, \ldots, H)$ denote slice means and consider the matrix

$$\boldsymbol{V} = \boldsymbol{\Sigma}^{-1/2} \sum_{h=1}^{H} p_h (\boldsymbol{\mu}_h - \boldsymbol{\mu})(\boldsymbol{\mu}_h - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1/2}$$

where $p_h = \Pr(Y \in S_h)$. Let $\boldsymbol{\gamma}$ be an eigenvector of $\boldsymbol{V}$ corresponding to a non-zero eigenvalue. Li (1991) showed that $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\gamma}$ is an element of the span of the $\boldsymbol{\beta}_k$s from the model in (2), provided that a $K$-direction version of Condition 1 holds. Consequently, if $\boldsymbol{V}$ is of rank $K$, then SIR can recover a complete basis for the dimension reduction directions. However, SIR suffers from the same problems with symmetric dependency as OLS (Li 1991; Cook & Weisberg 1991) and is therefore a candidate for the same type of transformation.

Cook & Weisberg (1991) introduced SAVE, which does not suffer from symmetric dependency issues. It does, however, require that $\mathrm{var}(\boldsymbol{x}|\boldsymbol{\beta}_1^\top \boldsymbol{x}, \ldots, \boldsymbol{\beta}_K^\top \boldsymbol{x})$ is constant in addition to Condition 1. Both conditions are satisfied when $\boldsymbol{x}$ is normally distributed, although both will often approximately hold in practice (Hall & Li 1993; Leeb 2013). Let $\boldsymbol{\Sigma}_h = \mathrm{var}(\boldsymbol{x}|Y \in S_h)$ $(h = 1, \ldots, H)$ denote slice covariance matrices. Then SAVE is carried out similarly to SIR but where $\boldsymbol{M} = \sum_{h=1}^{H} p_h (\boldsymbol{I} - \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_h\boldsymbol{\Sigma}^{-1/2})^2$ is used instead of $\boldsymbol{V}$. For some models SAVE requires large sample sizes to achieve good results. However, we have found that a variation of SAVE that was proposed by Zhu, Zhu & Feng (2010) called Cumulative Variance Estimation (CUVE) often provides excellent results. For $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})$, CUVE estimates $\mathrm{E}((P(Y \leq \widetilde{Y})\boldsymbol{I} - \mathrm{var}(\boldsymbol{z}\, I(Y \leq \widetilde{Y})))^2)$ where $\widetilde{Y}$ is an independent copy of $Y$ and $I(\cdot)$ is the indicator function taking the value 1 if its argument is true or zero otherwise. Similarly, Zhu, Zhu & Feng (2010) provided Cumulative Mean Estimation (CUME) which is a variation of SIR based on $\mathrm{E}(\mathrm{E}(\boldsymbol{z}I(Y \leq \widetilde{Y}))\mathrm{E}(\boldsymbol{z}I(Y \leq \widetilde{Y}))^\top)$. Neither method requires choosing $H$, and Shaker & Prendergast (2011) showed that they can be successfully combined to obtain excellent results. Consequently, we will also consider CUME, following a transformation based on the first CUVE direction.

For SIR and SAVE the user must specify the subranges of $Y$ used for 'slicing'. One simple approach is to set $H$ equally-probable slices. In practice this is equivalent to ordering the data by the magnitude of the response and allocating an (approximately) equal number of

observations to each slice. For estimation we use the R `dr` package (Weisberg 2002) which, by default, uses $\max(8, p+3)$ for $H$. However, another option would be, for example, to seek an optimal choice for $H$ (and simultaneously a value for $K$) using a bootstrap approach as given by Liquet & Saracco (2012).

******** Table 3 about here ********

An advantage of SIR, SAVE, CUME and CUVE is that the $y_i$s are used only to allocate $\boldsymbol{x}_i$s. Consequently, we would not expect extremely large $y_i$s to have the same detrimental effect on estimation as they do for OLS. To highlight this we reconsider Model 3 and adapt Method 2 as follows. SAVE is used to get the initial estimate to the direction of $\boldsymbol{\beta}$. SIR is then used on either the transformed $y_i$s (Method 1) or $\boldsymbol{x}_i$s (Method 2), where the SAVE direction has been used to facilitate the transformations. Similarly, we use CUME following a transformation using the CUVE estimated direction. We also consider iterative estimation procedures for both. In Table 3 we provide the results from 10,000 simulations where $p = 20$ and $n = 100$ or 200. Due to symmetry, both SIR and CUME fail to estimate the direction of $\boldsymbol{\beta}$. SAVE also has trouble estimating the direction of $\boldsymbol{\beta}$, especially for $n = 100$, although improvements are found for $n = 200$ and we observed good results for $n = 500$ (not shown). In contrast CUVE performs well, even for $n = 100$. The results also indicate that transformation Method 2 results in improved estimation, although the combination of SAVE and SIR is only moderately successful for the larger sample size. The combination of CUVE and CUME, however, provides excellent results even for $n = 100$. For both approaches, the iterative estimation scheme also provides improvements, as evidenced by the increased mean squared correlations.

### 5.2. Detecting multiple directions with OLS

We show here that a simple two-step estimation procedure for OLS can work exceptionally well when OLS is faced with the task of finding two directions, one of which is expected to be non-detectable due to symmetric dependency. For comparison we also consider PHD|OLS which is an iterative version of PHD and OLS considered by Shaker & Prendergast (2011). Here the first direction estimated is the OLS slope. Then PHD is used, conditional on this OLS slope estimate already detected, so that only new information is found in the second direction. The model that we focus on, given below, was also considered by Shaker & Prendergast (2011).

**Model 4.** $Y = \sin(0.5\boldsymbol{\beta}_1^\top \boldsymbol{x}) + \cos(0.5\boldsymbol{\beta}_2^\top \boldsymbol{x}) + 0.3\varepsilon$ *where* $\boldsymbol{\beta}_1 = (1, 2, -3, 0, \ldots, 0)^\top$ *and* $\boldsymbol{\beta}_2 = (1, 1, 0, -2, 0, \ldots, 0)^\top$.

For the model above we will consider the performance of SIR, PHD, PHD|OLS and three new approaches based on Methods 1 and 2. For these new approaches we will use the OLS slope vector as the first estimated direction and then use the transformation methods to estimate a second direction.

******** Table 4 about here ********

In Table 4 we provide the simulated average first and second canonical correlations between $\boldsymbol{X}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and $\boldsymbol{X}\widehat{\boldsymbol{B}}$ where $\widehat{\boldsymbol{B}}$ is a $p \times 2$ matrix consisting of the first and second estimated directions. A large average first canonical correlation, $\bar{r}_1$, indicates that the approach successfully detects the first direction. Similarly, a large $\bar{r}_2$ is indicative of good performance in detecting the second direction. SIR and PHD are both capable of finding one of the directions: for SIR the direction it is expected to find is $\boldsymbol{\beta}_1$ and for PHD it is $\boldsymbol{\beta}_2$. However, these methods do not perform well at finding the other direction. PHD|OLS is expected to find both and the average canonical correlations indicate this, although the method has some trouble in estimating the second direction for $n = 100$. The transformation methods provide improvements in estimating both directions, certainly in the case of SIR and PHD and give marginally better results than even PHD|OLS.

We could similarly use robust $M$-estimation regression methods here too. Rather than repeat the simulation for similar results, we choose this approach for the example considered in Section 6.

### 5.3. Estimating $K$

So far, we have not discussed how to choose, or estimate, a suitable $K$. One approach is exploratory, in which ESSPs can be viewed to determine whether associated estimated vectors of predictor coefficients provide any information regarding the response. For this to work properly, the true $K$ should be one or two, so that two- or three-dimensional ESSPs can be insightful. However it may be preferable to apply tests concerning the value of $K$ so as to avoid confusion between what may be simply artifacts the data, as opposed to valid inference about $K$. This apporach is also more suitable when $K > 2$, since a series of ESSPs may not be adequately informative. With respect to the methods we have employed throughout and which allow for the estimation of multiple directions (e.g., SIR, PHD, CUVE etc), there are several choices for the testing of $K$ based on the method's respective eigenvalues. However, as we will show here, tests for $K$ need to be treated with caution.

For SIR, SAVE and PHD we employ tests for a candidate choice of $K$, denoted $K^*$. These tests are based on asymptotic chi-squared distributions for sums of estimates (squared estimates in the case of PHD) of $p$ hypothesised $K^*$ zero-valued eigenvalues (Li 1991, 1992).

For CUME and CUVE we use the approach due to Zhu, Zhu & Feng (2010) which opts for the $K^*$ that maximises a criterion function. If $\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_p \geq 0$ is the estimated eigenvalues for either CUME or CUVE then the criterion function is $n \sum_{k=1}^{K^*} \widehat{\lambda}_k^2 / \sum_{k=1}^{p} \widehat{\lambda}_k^2 - c_n K^*(K^* + 1)/2$ where the second term provides a penalty for larger $K^*$ and where we follow the advice of Zhu, Zhu & Feng (2010) and choose $c_n = 2n^{3/4}/p$. It would also be possible to employ this approach for SIR, SAVE and PHD but for simplicity we choose not to do so. However, it should be noted that the minimum choice for $K$ using the criterion is one, whereas zero can be chosen via the testing approach. In all cases we let $\widehat{K}$ denote the chosen $K$ based on either the tests or the criterion.

******** Table 5 about here ********

When faced with symmetric dependency, care needs to be taken when estimating $K$. This is highlighted in Table 5 where we simulated data 1000 times from each of Models 2, 3 and 4 and estimated $K$ using the approaches described above and for two different sample sizes choices of $n = 200$ and $n = 500$. For each of these models the rank of the SIR matrix is equal to $K - 1$ and, consequently, $K$ is typically underestimated. The test for PHD performs well for Model 2, but poorly for Model 3 where the dimension is often underestimated. This is likely due to a tendency for some extremely large response values to be generated from this model. These values are harmful to PHD but not to the slicing approaches. PHD also often underestimates $K$ for Model 4, which is expected since the rank of the Hessian matrix can be shown to equal to one for this model. Overall, CUVE appears to perform very well for Model 2 both when $n = 200$ and 500 and for Model 3 when $n = 500$. For Model 2, the rank of the CUVE matrix is one so that $K$ is expected to be underestimated. However, for the smaller sample size of $n = 200$ the correct $K = 2$ was chosen close to half of the time, presumably due to the zero-valued eigenvalues being estimated comparatively poorly (compared to when $n = 500$) and the penalty not being large enough.

As can be seen above, estimation of $K$ needs to be treated with some caution. One approach would be to compute several estimates of $K$ and to explore any discrepancies. Additionally, it may be possible that estimates of $K$ are identical using two different methods, yet the associated directions are different. For example, for Model 4 both SIR and PHD together estimated $K$ to be just one (underestimates of the desired $K = 2$) in 52.3% of the trials. However, among these 52.3% of trials, the average squared correlations between the dimension reduced predictors (i.e. between the $\widehat{\boldsymbol{b}}^\top \boldsymbol{x}_i$s for each of the methods) was just 0.057, indicating that different information regarding the $\boldsymbol{\beta}_k$s was typically found from the two methods. It is not difficult to compute the correlations from different methods with a single data set, and this therefore provides a logical way in which to proceed. However in future

work we may investigate some form of meta-estimates of $K$ that take into consideration the directions estimated using different methods.

## 6. The Ozone data example

Li (1992) considered the Ozone data from Breiman & Friedman (1985) which consists of 330 observations and eight predictors (e.g. wind speed, humidity etc.; refer to Table 4 of Li 1992 for the full list of predictors). The response is atmospheric ozone concentration. Li (1992) noted that SIR found a quadratic relationship (although not one that includerds symmetric dependency) between the response and eight predictors and that an almost identical relationship could be found using least squares. Using PHD, another direction was found that eluded SIR, which provided an ESSP that exhibited symmetric dependency. Conflicting estimates of $K$ were also obtained based on the SIR and PHD eigenvalues, where SIR chose $K = 1$ and PHD chose $K = 2$.

We now consider the Ozone data example further. Let $Y$ denote the response variable and $\boldsymbol{x}$ denote the eight-dimensional vector of predictor variables. As previously we let the sample data be denoted by $\{y_i, \boldsymbol{x}_i\}_{i=1}^{330}$. We base our model on $\sqrt{Y}$ which, as we will see shortly, allows methods such as OLS, $M$-estimator regression methods and SIR to detect a linear relationship between the response and predictors. We also use $M$-estimation with the Huber weight function as a robust least-squares method. For convenience we will refer to this method as RR.

******** Figure 3 about here ********

Let $\widehat{\boldsymbol{b}}_1$ be the estimated slope for the RR regression of the $\sqrt{y_i}$s on the $\boldsymbol{x}_i$s. A plot of the $\sqrt{y_i}$s versus the $\widehat{\boldsymbol{b}}_1^\top \boldsymbol{x}_i$s in Plot A of Figure 3 shows a linear relationship between the response and the dimension reduced predictors (labelled "1st RR dr predictor" on the plot). Using least squares regression, a simple linear model using the $\widehat{\boldsymbol{b}}_1^\top \boldsymbol{x}_i$s explains 72.1% of the variation in the square root of ozone. We now use transformation Method 1 with the first PHD direction but with RR replacing OLS, and let $\widehat{\boldsymbol{b}}_2$ denote this new estimate. Plot B shows that RR has now found another direction exhibiting a quadratic (or higher) relationship which is further emphasised by a fitted spline curve (black line). A least squares analysis based on a quadratic polynomial fit of just the $\widehat{\boldsymbol{b}}_2^\top \boldsymbol{x}_i$s results in a highly significant model that explains 33.1% of the variation in the response.

We now use OLS to fit a model to the $\widehat{\boldsymbol{b}}_1^\top \boldsymbol{x}_i$s, the $\widehat{\boldsymbol{b}}_2^\top \boldsymbol{x}_i$s and the square of each of these (we did not include the product of the two for a full quadratic model since this made little contribution). The estimated model is

$$\widehat{Y}^{1/2} = -99.33 + 1.15 \times (\widehat{\boldsymbol{b}}_1^\top \boldsymbol{x}) - 7.72 \times (\widehat{\boldsymbol{b}}_2^\top \boldsymbol{x}) + 0.17 \times (\widehat{\boldsymbol{b}}_1^\top \boldsymbol{x})^2 - 0.14 \times (\widehat{\boldsymbol{b}}_2^\top \boldsymbol{x})^2.$$

The above fitted model explains approximately 76.4% of the variation in the square root of the response, indicating a good fit, and all of the terms in the model were highly significant. In Plots C and D we provide the residuals versus fits plot for the fit and also the quantile-quantile plot to check to see whether one could assume something close to a normal error term for the underlying model. These plots are excellent. They evince no evidence that the assumption of normally distributed error terms with homogeneous variance does not hold, at least approximately. In summary, we have successfully used RR twice to find two directions that can be used to construct a simple model with simple error term properties.

## 7. Discussion

This paper shows that simple response and predictor transformations can be used to remove the problem of symmetric dependency that affects some dimension reduction methods. While we initially show that OLS and PHD can be successfully employed in tandem for improved estimates, our approaches need not be limited to these methods. To highlight this, we also show that the popular robust $M$-estimation methods can be used, as well as sliced inverse regression in conjunction with sliced average variance estimates and associated cumulative slicing approaches. These approaches are particularly useful when the data include very large response values that can be detrimental to OLS estimation. Another interesting discovery presented in this paper is the ability of OLS, and robust equivalents, to find more than one direction.

Further research is envisaged. For example, we show that estimating $K$ can be problematic in the sense that some directions may be "hidden" to some methods, resulting in an underestimation of $K$. However, different dimension reduction methods may find different directions and this possibility could be potentially be exploited to obtain a combined or meta estimate of $K$. Another possibility is to explore transformations for linear discriminant analysis (LDA). When $Y$ is discrete, SIR is equivalent to LDA (see, e.g., Li 2000; Cook & Yin 2001). If $g$ is the number of sub-populations (distinct values in the domain of $Y$) then LDA can find at most $g - 1$ directions. Consequently predictor transformations may be useful for recovering additional directions. Finally, an exploration of sensitivity to distributional assumptions could also be considered. It is possible that, although normality of the predictor may fail to hold, thus affecting PHD estimation, the estimated PHD direction may still be

adequate to the extent that OLS (which has less restrictive conditions) can benefit following transformation.

## A. Proof of Theorem 1

Throughout let $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}$, $\mathrm{var}(\boldsymbol{x}) = \boldsymbol{\Sigma}$, $\boldsymbol{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K]$ and recall that $\boldsymbol{v} \in \mathcal{S}$. It can be shown (see, e.g., Prendergast 2005) that Condition 1 is equivalent to

$$\mathrm{E}(\boldsymbol{x}|\boldsymbol{B}^\top \boldsymbol{x}) = \boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}^\top(\mathrm{E}(\boldsymbol{x}|\boldsymbol{B}^\top \boldsymbol{x}) - \boldsymbol{\mu}). \tag{11}$$

Since $\boldsymbol{v} \in \mathcal{S}$ so that $\mathrm{E}(\mathrm{sign}(\boldsymbol{v}^\top \boldsymbol{x} - \boldsymbol{v}^\top \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})) = \mathrm{E}\{\mathrm{sign}(\boldsymbol{v}^\top \boldsymbol{x} - \boldsymbol{v}^\top \boldsymbol{\mu})\mathrm{E}(\boldsymbol{x} - \boldsymbol{\mu}|\boldsymbol{B}^\top \boldsymbol{x})\}$, it follows from (11) that $\boldsymbol{\Sigma}^{-1}\mathrm{E}(t_x(\boldsymbol{x} - \boldsymbol{\mu}; \boldsymbol{v}))$ is equal to

$$\boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathrm{sign}(\boldsymbol{v}^\top \boldsymbol{x} - \boldsymbol{v}^\top \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})) = \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}^\top \mathrm{E}(t_x(\boldsymbol{x} - \boldsymbol{\mu}; \boldsymbol{v})) \in \mathcal{S}. \tag{12}$$

Similarly, by noting that

$$\mathrm{E}(\mathrm{sign}(\boldsymbol{v}^\top \boldsymbol{x} - \boldsymbol{v}^\top \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})Y) = \mathrm{E}\left(\mathrm{sign}(\boldsymbol{v}^\top \boldsymbol{x} - \boldsymbol{v}^\top \boldsymbol{\mu})Y\mathrm{E}(\boldsymbol{x} - \boldsymbol{\mu}|\boldsymbol{B}^\top \boldsymbol{x})\right)$$

since, from (2), $Y$ is a function of $\boldsymbol{B}^\top \boldsymbol{x}$ and $\varepsilon$ where $\varepsilon$ is independent of $\boldsymbol{x}$, we can also show that

$$\boldsymbol{\Sigma}^{-1}\mathrm{cov}(\mathrm{sign}(\boldsymbol{v}^\top \boldsymbol{x} - \boldsymbol{v}^\top \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu}), Y) \in \mathcal{S}. \tag{13}$$

This shows that (6) holds.

For simplicity in what follows, let $\boldsymbol{e}_t = \mathrm{E}(\mathrm{sign}(\boldsymbol{v}^\top \boldsymbol{x} - \boldsymbol{v}^\top \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu}))$. Now, using (12),

$$\mathrm{var}(t_x(\boldsymbol{x} - \boldsymbol{\mu}; \boldsymbol{v})) = \boldsymbol{\Sigma} - \boldsymbol{e}_t\boldsymbol{e}_t^\top$$

since $\mathrm{E}(t_x(\boldsymbol{x} - \boldsymbol{\mu}; \boldsymbol{v})t_x(\boldsymbol{x} - \boldsymbol{\mu}; \boldsymbol{v})^\top) = \mathrm{E}((\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top) = \boldsymbol{\Sigma}$. Therefore (e.g., use the Small Rank Adjustment Lemma, Horn & Johnson 1985, page 19)

$$(\mathrm{var}(t_x(\boldsymbol{x} - \boldsymbol{\mu}; \boldsymbol{v})))^{-1} = \boldsymbol{\Sigma}^{-1} + \frac{1}{1 - \boldsymbol{e}_t^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{e}_t}\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_t\boldsymbol{e}_t^\top\boldsymbol{\Sigma}^{-1}$$

In conjunction with (13), this shows that (7) holds, completing the proof since $\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_t \in \mathcal{S}$.

## References


BREIMAN, L. & FRIEDMAN, J.H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80**, 580–619. With discussion and with a reply by the authors.

BRILLINGER, D.R. (1977). The identification of a particular nonlinear time series system. *Biometrika* **64**, 509–515.

BRILLINGER, D.R. (1983). A generalized linear model with "Gaussian" regressor variables. In *A Festschrift for Erich L. Lehmann*. Wadsworth Statist./Probab. Ser., Belmont, CA: Wadsworth, pp. 97–114.

COOK, R.D. (1998a). Principal Hessian directions revisited. *J. Amer. Statist. Assoc.* **93**, 84–100. With comments by Ker-Chau Li and a rejoinder by the author.

COOK, R.D. (1998b). *Regression graphics*. New York: John Wiley & Sons Inc. Ideas for studying regressions through graphics.

COOK, R.D. & WEISBERG, S. (1991). Discussion of "Sliced inverse regression for dimension reduction". *J. Amer. Statist. Assoc.* **86**, 328–332.

COOK, R.D. & YIN, X. (2001). Dimension reduction and visualization in discriminant analysis (with discussion). *Aust. NZ. J. Stat.* **43**, 147–199.

FARAWAY, J.J. (2006). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. Boca Raton, FL: Chapman & Hall/CRC press.

FOX, J. & WEISBERG, S. (2011). *An R Companion to Applied Regression*. Thousand Oaks, CA: SAGE Publications.

GARNHAM, A.L. & PRENDERGAST, L.A. (2013). A note on least squares sensitivity in single-index model estimation and the benefits of response transformations. *Electron. J. Stat.* **7**, 1983–2004.

HALL, P. & LI, K.C. (1993). On almost linearity of low-dimensional projections from high-dimensional data. *Ann. Statist.* **21**, 867–889.

HORN, R.A. & JOHNSON, C.A. (1985). *Matrix Analysis*. New York: Cambridge University Press.

HUBER, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35**, 73–101.

HUBER, P.J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1**, 799–821.

LEEB, H. (2013). On the conditional distributions of low-dimensional projections from high-dimensional data. *Ann. Stat.* **41**, 464–483.

LI, K.C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86**, 316–342. With discussion and a rejoinder by the author.

LI, K.C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Amer. Statist. Assoc.* **87**, 1025–1039.

LI, K.C. (2000). High dimensional data analysis via sir/phd approach. http://www.stat.ucla.edu/ kcli/sir-PHD.pdf.

LI, K.C. & DUAN, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17**, 1009–1052.

LIQUET, B. & SARACCO, J. (2012). A graphical tool for selecting the number of slices and the dimension of the model in SIR and SAVE approaches. *Comput. Statist.* **27**, 103–125.

LUE, H.H. (2001). A study of sensitivity analysis on the method of principal hessian directions. *Comput. Stat.* **16**, 109–130.

PRENDERGAST, L.A. (2005). Influence functions for sliced inverse regression. *Scand. J. Statist.* **32**, 385–404.

PRENDERGAST, L.A. (2008). Trimming influential observations for improved single-index model estimated sufficient summary plots. *Comput. Statist. Data Anal.* **52**, 5319–5327.

PRENDERGAST, L.A. & SHEATHER, S. (2013). On sensitivity of inverse response plot estimation and the benefits of a robust estimation approach. *Scand. J. Stat.* **40**, 219–237.

PRENDERGAST, L.A. & SMITH, J.A. (2010). Influence functions for dimension reduction methods: an example influence study of principal Hessian direction analysis. *Scand. J. Stat.* **37**, 588–611.

R CORE TEAM (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

SHAKER, A.J. & PRENDERGAST, L.A. (2011). Iterative application of dimension reduction methods. *Electron. J. Stat.* **5**, 1471–1494.

VENABLES, W.N. & RIPLEY, B.D. (2002). *Modern Applied Statistics with S*. New York: Springer, 4th edn. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.

WEISBERG, S. (2002). Dimension reduction regression in R. *J. Stat. Softw.* **7**, 1–22.

ZHU, L.P., ZHU, L.X. & FENG, Z.H. (2010). Dimension reduction in regressions through cumulative slicing estimation. *J. Amer. Statist. Assoc.* **105**, 1455–1466.

Figure 1. Plots of the $y_i$s versus the $\boldsymbol{\beta}^\top \boldsymbol{x}_i$s (True Views) and the $y_i$s versus the $\widehat{\boldsymbol{b}}^\top \boldsymbol{x}_i$s (Estimated Views - ESSPs) for 100 observations generated for Model 1 (Plots A and B) and Model 2 (Plots C and D). OLS was used to estimate the direction of $\boldsymbol{\beta}$.

Figure 2. Under the assumption of zero error in Model 2 and choosing $\boldsymbol{v} = \boldsymbol{\beta}$, Plot A provides the plot of the transformed $Y$ versus $\boldsymbol{\beta}^\top \boldsymbol{x}$ and Plot B provides the plot of $Y$ versus the transformed $\boldsymbol{\beta}^\top t_x(\boldsymbol{x} - \boldsymbol{\mu}; \boldsymbol{v})$.

Figure 3. Plots of (A) the ESSP found by RR, (B) an ESSP created using a second direction found using RR following transformation Method 1, (C) residuals versus fitted values from a least squares fit to the dimension reduced predictors from Plots (A) and (B) and the square of these dimension reduced predictors and (D) the corresponding normal quantile-quantile plot.

TABLE 1

*Average* $cor^2(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}\widehat{\boldsymbol{b}})$ *across 10,000 simulated runs for Model 2 with different choices of* $n$ *and* $p$*, where* $\widehat{\boldsymbol{b}}$ *is the estimate from one of five methods: OLS, PHD, Method 1 (M1) and Method 2 (M2). The designations M1-it and M2-it refer to the iterative estimation scheme for Methods 1 and 2. Standard deviations are in italics.*

| Method | $p = 10$ | | | | $p = 20$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $n = 50$ | $n = 100$ | $n = 200$ | $n = 500$ | $n = 50$ | $n = 100$ | $n = 200$ | $n = 500$ |
| OLS | 0.006 | 0.150 | 0.147 | 0.150 | 0.084 | 0.081 | 0.08 | 0.078 |
| | *0.171* | *0.167* | *0.165* | *0.165* | *0.105* | *0.101* | *0.099* | *0.095* |
| PHD | 0.716 | 0.897 | 0.958 | 0.985 | 0.258 | 0.686 | 0.893 | 0.965 |
| | *0.206* | *0.068* | *0.023* | *0.008* | *0.212* | *0.170* | *0.046* | *0.013* |
| M1 | 0.813 | 0.968 | 0.989 | 0.996 | 0.31 | 0.833 | 0.97 | 0.991 |
| | *0.234* | *0.038* | *0.007* | *0.002* | *0.271* | *0.177* | *0.017* | *0.003* |
| M1-it | 0.871 | 0.977 | 0.990 | 0.996 | 0.373 | 0.905 | 0.978 | 0.992 |
| | *0.211* | *0.024* | *0.005* | *0.002* | *0.306* | *0.147* | *0.010* | *0.003* |
| M2 | 0.833 | 0.967 | 0.989 | 0.996 | 0.332 | 0.831 | 0.969 | 0.991 |
| | *0.217* | *0.041* | *0.007* | *0.002* | *0.279* | *0.178* | *0.019* | *0.004* |
| M2-it | 0.875 | 0.977 | 0.990 | 0.996 | 0.372 | 0.887 | 0.978 | 0.992 |
| | *0.199* | *0.029* | *0.005* | *0.002* | *0.295* | *0.168* | *0.011* | *0.003* |

TABLE 2

*Average* $cor^2(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}\widehat{\boldsymbol{b}})$ *across 10,000 simulated data sets for Model 3 with different choices of* $n$*, where* $\widehat{\boldsymbol{b}}$ *is the estimate from various methods. The designations RR, RM1 and RM2 refer to Methods OLS, M1 and M2 but where robust regression* $M$*-estimation has been used in the regression step with the Huber weight function. The designations M1-trim and M2-trim refer to Methods M1 and M2 but where 10% of observations with the largest Cook's distance were trimmed. Standard deviations are in italics.*

| Method | $n = 100$ | $n = 200$ | $n = 500$ | Method | $n = 100$ | $n = 200$ | $n = 500$ |
|---|---|---|---|---|---|---|---|
| OLS | 0.007 | 0.004 | 0.002 | M3 | 0.164 | 0.235 | 0.268 |
| | *0.013* | *0.009* | *0.005* | | *0.124* | *0.137* | *0.146* |
| RR | 0.034 | 0.035 | 0.035 | RM1 | 0.718 | 0.955 | 0.987 |
| | *0.051* | *0.050* | *0.047* | | *0.257* | *0.063* | *0.018* |
| PHD | 0.662 | 0.920 | 0.978 | RM2 | 0.752 | 0.961 | 0.982 |
| | *0.229* | *0.035* | *0.008* | | *0.244* | *0.015* | *0.006* |
| M1 | 0.271 | 0.518 | 0.705 | M1-trim | 0.655 | 0.924 | 0.970 |
| | *0.243* | *0.311* | *0.317* | | *0.264* | *0.106* | *0.044* |
| M2 | 0.437 | 0.582 | 0.662 | M2-trim | 0.613 | 0.875 | 0.927 |
| | *0.226* | *0.212* | *0.197* | | *0.241* | *0.088* | *0.049* |

TABLE 3

*Average* $cor^2(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}\widehat{\boldsymbol{b}})$ *across 10,000 simulated data sets for Model 3 with two different choices of* $n$, $p = 20$, *where* $\widehat{\boldsymbol{b}}$ *is the estimate from various methods. The designation M2 refers to transformation Method 2 and M2-it refers to this transformation with iterative estimation. Standard deviations are in italics.*

| Method | $n = 100$ | $n = 200$ | Method | $n = 100$ | $n = 200$ |
|---|---|---|---|---|---|
| SIR | 0.057 | 0.059 | SAVE SIR (M2) | 0.121 | 0.494 |
| | *0.086* | *0.088* | | *0.175* | *0.365* |
| CUME | 0.058 | 0.059 | SAVE SIR (M2-it) | 0.162 | 0.686 |
| | *0.078* | *0.078* | | *0.243* | *0.429* |
| SAVE | 0.110 | 0.327 | CUVE CUME (M2) | 0.905 | 0.983 |
| | *0.134* | *0.241* | | *0.124* | *0.007* |
| CUVE | 0.808 | 0.937 | CUVE CUME (M2-it) | 0.940 | 0.988 |
| | *0.130* | *0.026* | | *0.114* | *0.005* |

TABLE 4

*Average first and second canonical correlations* $(\bar{r}_1, \bar{r}_2)$ *between* $\boldsymbol{X}[\boldsymbol{\beta}_1, \boldsymbol{\beta}_2]$ *and* $\boldsymbol{X}\widehat{\boldsymbol{B}}$ *across 10,000 simulated runs for data generated from Model 4 with different choices of* $n$ *and* $p$ *where* $\widehat{\boldsymbol{b}}$ *is the estimate from one of five methods: OLS, PHD, Method 1 (M1) and Method 2 (M2). Standard deviations are in italics.*

| Method | $n = 100$ | | $n = 200$ | | $n = 500$ | | $n = 1000$ | |
|---|---|---|---|---|---|---|---|---|
| | $\bar{r}_1$ | $\bar{r}_2$ | $\bar{r}_1$ | $\bar{r}_2$ | $\bar{r}_1$ | $\bar{r}_2$ | $\bar{r}_1$ | $\bar{r}_2$ |
| SIR | 0.743 | 0.264 | 0.844 | 0.296 | 0.939 | 0.393 | 0.971 | 0.53 |
| | *0.154* | *0.19* | *0.103* | *0.212* | *0.037* | *0.254* | *0.016* | *0.271* |
| PHD | 0.912 | 0.386 | 0.961 | 0.403 | 0.985 | 0.407 | 0.993 | 0.409 |
| | *0.059* | *0.226* | *0.024* | *0.232* | *0.008* | *0.235* | *0.004* | *0.233* |
| PHD\|OLS | 0.914 | 0.632 | 0.961 | 0.805 | 0.985 | 0.918 | 0.993 | 0.958 |
| | *0.059* | *0.219* | *0.023* | *0.125* | *0.008* | *0.045* | *0.004* | *0.022* |
| OLS,M1 | 0.927 | 0.671 | 0.972 | 0.827 | 0.99 | 0.923 | 0.995 | 0.96 |
| | *0.056* | *0.192* | *0.017* | *0.098* | *0.006* | *0.041* | *0.003* | *0.021* |
| OLS,M2 | 0.933 | 0.671 | 0.974 | 0.827 | 0.991 | 0.923 | 0.996 | 0.96 |
| | *0.054* | *0.194* | *0.016* | *0.099* | *0.005* | *0.042* | *0.002* | *0.021* |

TABLE 5

*Proportion of times that* $\widehat{K} = 0, 1, 2$ *or* $\geq 3$ *when using the estimated eigenvalues from the methods SIR, PHD, CUME and CUVE for Models 2-4 and with* $n = 200$ *and 500 for 1000 simulation runs.*

| | | $n = 200$ | | | | $n = 500$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Method | $\widehat{K} = 0$ | $\widehat{K} = 1$ | $\widehat{K} = 2$ | $\widehat{K} \geq 3$ | $\widehat{K} = 0$ | $\widehat{K} = 1$ | $\widehat{K} = 2$ | $\widehat{K} \geq 3$ |
| Model 2 | SIR | 0.950 | 0.049 | 0.001 | 0.000 | 0.951 | 0.047 | 0.002 | 0.000 |
| $(K = 1)$ | PHD | 0.000 | 0.973 | 0.026 | 0.001 | 0.000 | 0.952 | 0.042 | 0.006 |
| | CUME | - | 0.736 | 0.263 | 0.001 | - | 0.764 | 0.236 | 0.000 |
| | CUVE | - | 0.996 | 0.004 | 0.000 | - | 1.000 | 0.000 | 0.000 |
| Model 3 | SIR | 0.954 | 0.046 | 0.000 | 0.000 | 0.960 | 0.038 | 0.002 | 0.000 |
| $(K = 1)$ | PHD | 0.785 | 0.215 | 0.000 | 0.000 | 0.755 | 0.245 | 0.000 | 0.000 |
| | CUME | - | 0.452 | 0.547 | 0.001 | - | 0.462 | 0.535 | 0.003 |
| | CUVE | - | 0.007 | 0.747 | 0.246 | - | 0.983 | 0.017 | 0.000 |
| Model 4 | SIR | 0.628 | 0.344 | 0.025 | 0.003 | 0.092 | 0.828 | 0.076 | 0.004 |
| $(K = 2)$ | PHD | 0.011 | 0.894 | 0.091 | 0.004 | 0.000 | 0.867 | 0.122 | 0.011 |
| | CUME | - | 0.997 | 0.003 | 0.000 | - | 1.000 | 0.000 | 0.000 |
| | CUVE | - | 0.526 | 0.468 | 0.006 | - | 0.993 | 0.007 | 0.000 |