HARVARD UNIVERSITY

An Overview of Sliced Inverse Regression for Dimension Reduction

Russell Kunes

November 2017

1 INTRODUCTION

Sliced inverse regression (SIR) is a dimension reduction technique proposed by Ker-Chau Li in a 1990 paper [1]. In this paradigm, we assume an extremely general regression model where $Y = f(\beta_1\mathbf{X},\ldots,\beta_K\mathbf{X},\epsilon)$ with $\mathbf{X} \in \mathbb{R}^p$ and each β_i is a $1 \times p$ row vector. The function f and $\beta_1,\ldots,\beta_K$ are unknown. The main innovation of this method is considering the inverse regression curve $\mathbb{E}(\mathbf{X}|Y)$ rather than the usual $\mathbb{E}(Y|\mathbf{X})$. As it turns out, this inverse regression curve is usually located within the vector space spanned by a transformation of the set of β_i's. So, by estimating this curve, we can try to recover the subspace spanned by the β_i's, which is the most important subspace of $\mathbb{R}^p$ for analyzing the relationship between Y and $\mathbf{X}$.

1.1 MOTIVATION

When it comes to dimensionality reduction, one of the first things that comes to mind is principal components analysis (PCA). In the context of linear regression, one might hope to address the problem of multicollinearity by reducing the dimensionality of the data such that the new data has orthogonal columns. To review the basic concepts of PCA:

- We are interested in a set of n p-dimensional variables $\{\mathbf{X}_i\}_{i=1}^n$, organized into an n-by-p design matrix $\mathbf{X}$ (note a slight abuse of notation since we use $\mathbf{X}$ as a vector outside of this section). The goal is to find some transformed matrix $\mathbf{X}^*$ such that $C(\mathbf{X}) \approx C(\mathbf{X}^*)$
- Perform a spectral decomposition on the sample covariance matrix $\mathbf{X}^\top\mathbf{X}$ into $\mathbf{Q}\Lambda\mathbf{Q}^\top$ with $\Lambda = diag(\{\lambda_1,\ldots,\lambda_p\})$, sorted eigenvalues $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p$ and $\mathbf{Q}$ orthogonal.
- Take $W_k = \mathbf{X}\mathbf{Q}_{1:k}$ as the new data matrix corresponding to the first k principal components of $\mathbf{X}^\top\mathbf{X}$.

Another way of thinking of this is that the first principal component direction $\mathbf{q}^*$ is chosen such that

$$\mathbf{q}^* = \text{argmax}_{||\mathbf{q}||=1} \mathbf{q}^\top \mathbf{X}^\top \mathbf{X}\mathbf{q}$$

That is, choosing $\mathbf{q}^*$ as the direction along which there is the most variance in the sample. Upon finding the first principal component, we find the second by optimizing over the same criterion among vectors that are orthogonal to $\mathbf{q}^*$, and repeat until we have a set of principal component directions of size K that are orthogonal with norm 1. After projecting each $\mathbf{X}_i$ onto these directions, this is equivalent to the above procedure. One hopes that the most interesting information in the data shows up in these first principal components. The result of applying PCA to linear regression problems is called Principal Components Regression (PCR). Though it is sometimes helpful, often it is not, perhaps due to the fact that the dimensionality reduction is performed completely independent of the output variable Y. By reducing the dimensionality of $\mathbf{X}$ completely independently of Y, we might be throwing away important information about Y. The motivation behind SIR is that to effectively reduce dimensionality in a regression problem, one must consider the relationship between X and Y.

1.2 Model

Throughout this project, we will assume the following regression model [1, 2]:

$$Y = f(\beta_1 \mathbf{X}, \ldots, \beta_K \mathbf{X}, \epsilon) \tag{1.1}$$

Where Y is a one-dimensional response variable, $\mathbf{X}$ is a random vector in $\mathbb{R}^p$, each β_i is a $1 \times p$ row vector ,and $K \leq p$ so that $\mathbf{X}$ as it relates to Y can be summarized by the $K-$dimensional vector $\{\beta_1 \mathbf{X}, \ldots, \beta_K \mathbf{X}\}$. Also assume that the noise term satisfies $\mathbb{E}(\epsilon) = 0$. Note that this is much more general than the regression models discussed in class. For one thing, f is completely general, and is treated as an unknown function. Figure (1.1) shows a graphical representation of dependence structure of this model. Importantly, we note that Y and $\mathbf{X}$ are conditionally independent given $\{\beta_1 \mathbf{X}, \ldots, \beta_K \mathbf{X}\}$, representing the assumption that there is a lower dimensional vector that captures all information about Y that is contained in $\mathbf{X}$. Also, introduce the notation that:

$$\mathbf{Z} = V^{-\frac{1}{2}}(\mathbf{X} - \mathbb{E}(\mathbf{X})))$$

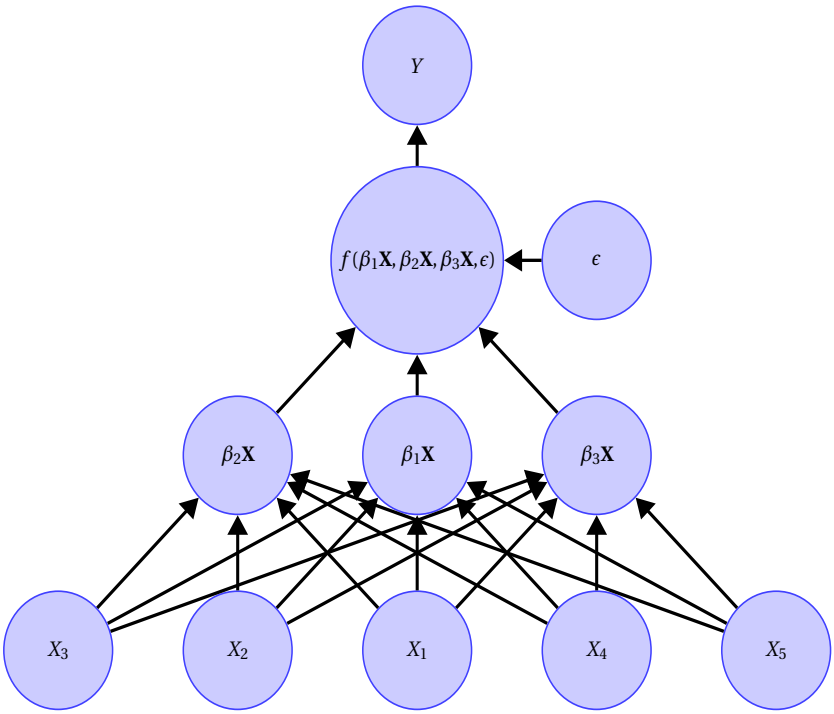
where $V = \text{Cov}(\mathbf{X})$ is positive definite, so that $\text{Cov}(\mathbf{Z}) = I_p$, and $\mathbb{E}(\mathbf{Z}) = \mathbf{0}$. Finally we note that this model is a generalization of the linear regression models from class, where $K = 1$ with $f(\beta \mathbf{X}_i, \epsilon_i) = \mathbf{X}_i \beta + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The assumption of this model is an extremely weak form of the hope that Y depends on $\mathbf{X}$ only through a lower dimensional projection of the data. So, a reasonable goal would be to estimate $\{\beta_1, \ldots, \beta_K\}$. But how is this possible, given that f is an unknown function? It turns out that though it is impossible to retrieve $\{\beta_1, \ldots, \beta_K\}$ exactly, we can often estimate the vector subspace of $\mathbb{R}^p$ that they span.

1.3 The Effective Dimension Reduction Directions

It seems pretty clear that since f is unknown, that we won't be able to estimate $\{\beta_1, \ldots, \beta_K\}$. To make this more concrete, $\{\beta_1, \ldots, \beta_K\}$ are not *identifiable*. For instance, consider a different basis of the subspace spanned by $\{\beta_1, \ldots, \beta_K\}$, and let this basis be called $\{\alpha_1, \ldots, \alpha_K\}$. Then, since $\{\alpha_i\}$ span, $\beta_i = \sum_j w_j^{(i)} \alpha_j$. So, it's clear that for a different function f', y is a function of $\alpha_i \mathbf{X}$'s:

$$y = f(\sum_j w_j^{(1)} \alpha_j \mathbf{X}, \ldots, \sum_j w_j^{(K)} \alpha_j \mathbf{X}, \epsilon) = f'(\alpha_1 \mathbf{X}, \ldots, \alpha_K \mathbf{X}, \epsilon) \tag{1.2}$$

Figure 1.1: Diagram of SIR Model Dependence Structure



Also, f can be written as any additive or multiplicative shift $f(\beta_1\mathbf{X},\ldots,\beta_K\mathbf{X},\epsilon) = f'(a_1\beta_1\mathbf{X}+c_1,\ldots,a_K\beta_K\mathbf{X}+c_K,\epsilon)$ where a_i and c_i are scalar constants, simply by multiplying and adding the appropriate constants to cancel out c_i and a_i. As we shall see later, this will allow us to change variables to standardized versions of $\mathbf{X}$. However, (1.2) held because $\{\alpha_i\}_{i=1}^K$ spanned the same vector subspace as $\{\beta_i\}_{i=1}^K$. Consider $\{\alpha_1,\ldots,\alpha_K\}$ that do not span the subspace generated by $\{\beta_1,\ldots,\beta_K\}$. Then at least one β_i cannot be written as a linear combination of α_i's, and we cannot reparameterize the function in terms of $\{\alpha_1\mathbf{X},\ldots,\alpha_K\mathbf{X}\}$. So the subspace generated by $\{\beta_1,\ldots,\beta_K\}$ is identifiable. A note here is that we assume that the $\{\beta_1,\ldots,\beta_K\}$ are linearly independent since otherwise we could just write f as a function of some proper subset of the $\beta_i\mathbf{X}$'s. This discussion motivates the following definitions:

Definition. *Under model (1.1), the vector subspace $\mathscr{B} \subseteq \mathbb{R}^p$ generated by $\{\beta_1,\ldots,\beta_K\}$ is called an e.d.r. space. Any non-zero vector in the e.d.r space is called an e.d.r. direction.*

So, since in general $\mathscr{B}$ is identifiable, our goal in analysis is to estimate this linear subspace. Once we have a good estimate, much like with principal components analysis, we can project our data onto this lower dimensional space and change to a $K-$dimensional basis. Under model (1.1), the position of $\mathbf{X}$ in this lower dimensional vector space contains all information from $\mathbf{X}$ about the response variable Y.

Remark. *Let $\mathbf{Z}$ be the standardized $\mathbf{X}$ variables. We can reparameterize f such that $y = f(\eta_1\mathbf{Z},\ldots,\eta_K\mathbf{Z},\epsilon)$*

Since $\mathbf{Z} = V^{-\frac{1}{2}}(\mathbf{X}-\mathbb{E}(\mathbf{X}))$, we can define η_i as

$$\eta_i = \beta_i V^{\frac{1}{2}}$$

And performing a change of variables, we get

$$y = f(\beta_1\mathbf{X},\ldots,\beta_K\mathbf{X},\epsilon) = f(\beta_1 V^{\frac{1}{2}} V^{-\frac{1}{2}}\mathbf{X},\ldots,\beta_K V^{\frac{1}{2}} V^{-\frac{1}{2}}\mathbf{X},\epsilon)$$

Adding the appropriate constant to each term:

$$y = f'(\beta_1 V^{\frac{1}{2}} V^{-\frac{1}{2}}\mathbf{X} - \beta_1 V^{\frac{1}{2}} V^{-\frac{1}{2}}\mathbb{E}(X), \ldots, \beta_K V^{\frac{1}{2}} V^{-\frac{1}{2}}\mathbf{X} - \beta_K V^{\frac{1}{2}} V^{-\frac{1}{2}}\mathbb{E}(X), \epsilon) = f'(\eta_1\mathbf{Z}, \ldots, \eta_K\mathbf{Z}, \epsilon)$$

Using these standardized versions of $\mathbf{X}$, we can estimate the span of $\{\eta_i\}$, and then transform back to $\{\beta_i\}$ since $\beta_i = \eta_i V^{-\frac{1}{2}}$. Another word on notation is that we will take η to be the matrix in $\mathbb{R}^{p\times K}$ such that column i is $\eta_i^\top$.

Having established all of this, a reasonable dimensionality reduction procedure involves estimating the span of $\{\eta_1, \ldots, \eta_K\}$. How can we do this? That is where the curve $\mathbb{E}(\mathbf{Z}|Y)$ comes in. As we will see, under certain conditions, the curve $\mathbb{E}(\mathbf{Z}|Y)$ actually resides in the span of $\{\eta_1, \ldots, \eta_K\}$ that we're interested in!

1.4 Main Theorem [1, 2]

The following theorem is the main result of SIR, which enables the dimension reduction algorithm. We will go through some of the mathematical details, but the upshot here is that, under the condition of (1.3) below, $\mathbb{E}(\mathbf{Z}|Y)$ is contained in $\mathscr{B}$,the e.d.r space, and even when the condition is broken, it tends to be close.

Theorem 1. *: Given the model, if*

$$\forall \gamma \in \mathbb{R}^p \ \exists (c_0, \ldots, c_k) \text{ s.t. } E(\gamma^\top \mathbf{X} | \beta_1\mathbf{X}, \ldots, \beta_k\mathbf{X}) = c_0 + \sum_{i=1}^{k} c_i \beta_i \mathbf{X} \tag{1.3}$$

then, the curve parametrized by Y, $g(Y) = E(\mathbf{X}|Y) - E(\mathbf{X})$, $g : \mathbb{R} \to \mathbb{R}^p$ is contained in the linear subspace span$(\beta_1 V, \ldots, \beta_k V)$ where $V = \text{Cov}(\mathbf{X})$.

Proof:
We modified the proof of this proposition, borrowing many ideas from the original paper [1], but taking a slightly different approach. First note that without loss of generality we can assume $\mathbb{E}(\mathbf{X}) = 0$, since as we have seen $f(\beta_1\mathbf{X}, \ldots, \beta_K\mathbf{X})$ is invariant to shifts, so assume we have already subtracted off the mean of $\mathbf{X}$. Also, introduce the notation that $\mathbf{B}$ is the $K \times p$ matrix such that the k'th row of $\mathbf{B}$ is β_k. We'll begin by applying the condition of Theorem 1, when γ is the i'th standard unit vector,i.e. $(0, \ldots, 1, \ldots, 0)^\top$ with the 1 in the i'th position, to get (we denote $\mathbf{X}_i$ as the i^{th} entry of $\mathbf{X}$):

$$\mathbb{E}(\mathbf{X}_i|\mathbf{B}\mathbf{X}) = c_{i0} + \sum_{k=1}^{K} C_{ik}\beta_k\mathbf{X}$$

where C is a $p \times K$ matrix of constants that stores the constants c for $i = 1, \ldots, p$ and $k = 1, \ldots, K$. Also, clearly $\mathbf{B}\mathbf{X}$ is the set $\{\beta_1\mathbf{X}, \ldots, \beta_K\mathbf{X}\}$ stacked into a $K \times 1$ column vector. Taking the expectation of both sides:

$$\mathbb{E}(\mathbf{X}_i) = c_{i0}$$

and by the assumption that $\mathbb{E}(\mathbf{X}) = 0$ we have $c_{i0} = 0$ for all i. We can reduce the right side of the expression as:

$$\sum_{k=1}^{K} C_{ik}\beta_k\mathbf{X} = \sum_{k=1}^{K} C_{ik} \sum_{j=1}^{p} \mathbf{B}_{kj}\mathbf{X}_j = \sum_{j=1}^{p}\sum_{k=1}^{K} C_{ik}\mathbf{B}_{kj}\mathbf{X}_j = \sum_{j=1}^{p} \mathbf{X}_j \sum_{k=1}^{K} C_{ik}\mathbf{B}_{kj} = \sum_{j=1}^{p} \mathbf{X}_j (C\mathbf{B})_{ij} = (C\mathbf{B}\mathbf{X})_i$$

Each equality here is just repeated application of the summation definition of matrix multiplication. Thus, we see that:

$$\mathbb{E}(\mathbf{X}_i|\mathbf{BX}) = (C\mathbf{BX})_i$$

and so:

$$\mathbb{E}(\mathbf{X}|\mathbf{BX}) = C\mathbf{BX}$$

With the above expression we multiply both sides on the right by $(\mathbf{BX})^\top$ and then take expectation of both sides:

$$\mathbb{E}(\mathbb{E}(\mathbf{X}|\mathbf{BX})(\mathbf{BX})^\top) = \mathbb{E}((C\mathbf{BX})(\mathbf{BX})^\top)$$

Since we standardized $\mathbf{X}$, the right side has the familiar expression for Cov($\mathbf{BX}$), where $\mathbf{BX}$ is the $K \times 1$ vector as before. To make sense of the left side we'll apply Adam's law:

$$\mathbb{E}(\mathbb{E}(\mathbf{X}|\mathbf{BX})(\mathbf{BX})^\top) = \mathbb{E}(\mathbb{E}(\mathbb{E}(\mathbf{X}|\mathbf{BX})(\mathbf{BX})^\top|\mathbf{X})) = \mathbb{E}(\mathbb{E}(\mathbf{X}(\mathbf{BX})^\top|\mathbf{X})) = \mathbb{E}(\mathbf{X}(\mathbf{BX})^\top)$$

The second inequality comes from the fact that after conditioning on $\mathbf{X}$, $\mathbf{X}$ is a constant and its expectation is itself regardless of the conditional. The final inequality is just putting everything back together with Adam's law. So we are left with:

$$\text{Cov}(\mathbf{X}, \mathbf{BX}) = C\text{Cov}(\mathbf{BX}) \implies C = \text{Cov}(\mathbf{X}, \mathbf{BX})\text{Cov}(\mathbf{BX})^{-1}$$

All resulting in: $\mathbb{E}(\mathbf{X}|\mathbf{BX}) = \text{Cov}(\mathbf{X}, \mathbf{BX})\text{Cov}(\mathbf{BX})^{-1}\mathbf{BX}$. To complete the proof we solve for $\mathbb{E}(\mathbf{X}|Y)$ and apply Adam's law and conditional independence:

$$\mathbb{E}(\mathbf{X}|Y) = \mathbb{E}(\mathbb{E}(\mathbf{X}|\mathbf{BX}, Y)|Y) = \mathbb{E}(\mathbb{E}(\mathbf{X}|\mathbf{BX})|Y) = \mathbb{E}(\text{Cov}(\mathbf{X}, \mathbf{BX})\text{Cov}(\mathbf{BX})^{-1}\mathbf{BX}|Y) = \text{Cov}(\mathbf{X}, \mathbf{BX})h(Y)$$

The reason that we pulled out Cov($\mathbf{X}, \mathbf{BX}$) and left everything else as $h(Y)$ (a function from $\mathbb{R}$ to $\mathbb{R}^K$) is that Cov($\mathbf{X}, \mathbf{BX}$) = $V\mathbf{B}^\top$, where V = Cov($\mathbf{X}$). So

$$\mathbb{E}(\mathbf{X}|Y) = V\mathbf{B}^\top h(Y)$$

is a linear combination of the columns of $V\mathbf{B}^\top$, or the rows of $\mathbf{B}V$, which are $\{\beta_1 V, \ldots, \beta_k V\}$ as desired.

A corollary of this theorem is that if condition (1.3) is satisfied, and $\mathbf{X}$ is standardized to $\mathbf{Z}$, i.e. $\mathbf{Z} = V^{-1/2}(\mathbf{X} - E\mathbf{X})$, and $\eta_k = \beta_k V^{1/2}$, then $E(\mathbf{Z}|y) \subseteq \text{span}(\eta_1, \ldots, \eta_k)$

1.5 Elliptical Symmetry

Upon first glance, condition (1.3) seems like it might be quite restrictive. However, this requirement turns out to be satisfied by $\mathbf{X}$ coming from the class of distributions with elliptical symmetry. These distributions extend the class of multivariate normal distributions to allow for heavier tailed distributions and lighter tailed distributions, while preserving the property of having elliptical equidensity contour lines [7]. It is somewhat reassuring that this includes the familiar multivariate normal distributions. We illustrate this point here, following work done by Cook and Weisburg(1991) as well as Eaton (1986) [6, 4]:

Definition: A random variable $\mathbf{X} \in \mathbb{R}^n$ has a spherically symmetric distribution if it's CDF is invariant to orthogonal linear transformations, that is:

$$F_X(\mathbf{x}) = F_{\Gamma X}(\mathbf{x}), \forall \Gamma \in O(n) \tag{1.4}$$

Definition: A random variable $\mathbf{Y} \in \mathbb{R}^n$ has an elliptically symmetric distribution if $Y = \mathbf{TX} + \mathbf{c}$ where $\mathbf{X}$ is spherically symmetric for some linear transformation, $\mathbf{T}$. This is not the most general definition of elliptical symmetry [7], but for our purposes we will take this as a definition because of it's simplicity and intuitive interpretation as a symmetric distribution being "stretched" by linear transformation and then shifted.

Using these two definitions, the goal of this section will be to show that Theorem 1 holds when the distribution of $\mathbf{X}$ is elliptically symmetric (Cook and Weisburg, 1991 [6]).

Proof: First, we note that if a random vector $\mathbf{Z} \in \mathbb{R}^p$ has a mean vector and that for any pair of orthogonal vectors $\mathbf{v}, \mathbf{u} \neq 0$,

$$\mathbb{E}(\mathbf{u}^\top \mathbf{Z} | \mathbf{v}^\top \mathbf{Z}) = 0$$

then $\mathbf{Z}$ is spherical (Eaton, 1986 [4]). The converse is also true (Cambanis, Huang, Simons, 1981, [7]) To see why this is true consider the characteristic function of $\mathbf{Z}$, $\phi_Z(t) = \mathbb{E}(\exp(it^\top \mathbf{Z}))$ and the characteristic function of an orthogonal transformation Γ of $\mathbf{Z}$, $\phi_{\Gamma Z}(t) = \phi_Z(\Gamma t)$. By assumption, $\mathbf{Z}$ has a finite first moment, so [3]:

$$\nabla \phi_Z(t) = i\mathbb{E}(\mathbf{Z} \exp(it^\top \mathbf{Z}))$$

Now consider:

$$\mathbf{u}^\top \nabla \phi_Z(\mathbf{v}) = \mathbb{E}(i\mathbf{u}^\top \mathbf{Z} \exp(i\mathbf{v}^\top \mathbf{Z})) = \mathbb{E}(\mathbb{E}(i\mathbf{u}^\top \mathbf{Z} \exp(i\mathbf{v}^\top \mathbf{Z}))|\mathbf{v}^\top \mathbf{Z})) = \mathbb{E}(i \exp(i\mathbf{v}^\top Z))\mathbb{E}(\mathbf{u}^\top \mathbf{Z} | \mathbf{v}^\top \mathbf{Z})) = 0$$

Where the second equality is by Adam's law, and the last equality is because we assumed $\mathbb{E}(\mathbf{u}^\top \mathbf{Z} | \mathbf{v}^\top \mathbf{Z}) = 0$. Following the proof in Eaton's paper, we introduce a smooth mapping $c : (0,1) \to \{x \in \mathbb{R}^p : ||x||^2 = r^2\}$ where $r = ||t||$. Let $c(\alpha_1) = t$ and $c(\alpha_2) = \Gamma t$ with $\alpha_1, \alpha_2 \in (0,1)$, noting that

$$||\Gamma t|| = r$$

since Γ is orthogonal. Since c has been constructed to have constant norm with respect to α, that is $||c(\alpha)||^2 = r^2, \forall \alpha \in (0,1)$, we have $\frac{\partial}{\partial \alpha}(c(\alpha)^\top c(\alpha)) = 0$. By the chain rule:

$$\frac{\partial}{\partial \alpha}(c(\alpha)^\top c(\alpha)) = 0 \implies 2(\frac{\partial}{\partial \alpha} c(\alpha))^\top c(\alpha) = 0$$

indicating that $\frac{\partial}{\partial \alpha} c(\alpha)$ is orthogonal to $c(\alpha)$. Finally we note that for $\alpha \in (0,1)$:

$$\frac{\partial}{\partial \alpha} \phi_Z(t) = \frac{\partial}{\partial \alpha} \phi_Z(c(\alpha)) = (\frac{\partial}{\partial \alpha} c(\alpha))^\top \nabla(\phi_Z(c(\alpha))) = 0$$

where the second equality comes from the chain rule, and the last equality is due to the fact that $\frac{\partial}{\partial \alpha} c(\alpha) \perp c(\alpha)$ and $\mathbf{u}^\top \nabla \phi_Z(\mathbf{v}) = 0$ for $\mathbf{u} \perp \mathbf{v}$. Thus, $\phi_Z(t) = \phi_Z(\Gamma t) = \phi_{\Gamma Z}(t)$. By the uniqueness of the characteristic function, $\mathbf{Z}$ and $\Gamma \mathbf{Z}$ have the same distribution.

Now, applying this result to our problem, consider the standardized inverse regression curve $\mathbb{E}(\mathbf{Z}|Y)$. Recall that $\{\eta_1, \ldots, \eta_K\}$ refer to the standardized e.d.r. directions, $\eta_k = \beta_k V^{\frac{1}{2}}$. Let η denote the matrix with column i equal to $\eta_i^\top$. By Adam's law, $\mathbb{E}(\mathbf{Z}|Y) = \mathbb{E}(\mathbb{E}(\mathbf{Z}|\eta^\top \mathbf{Z}, Y)|Y) = \mathbb{E}(\mathbb{E}(\mathbf{Z}|\eta^\top \mathbf{Z})|Y)$. The last inequality follows from the fact that $\mathbf{Z}$ and Y are conditionally independent given $\eta^\top \mathbf{Z}$. Next, let P denote the projection matrix onto $Col(\eta)$ and let $Q = I - P$ be the projection onto $Col(\eta)^\perp$. Then

$$\mathbb{E}(\mathbb{E}(\mathbf{Z}|\eta^\top \mathbf{Z})|Y) = \mathbb{E}(\mathbb{E}(P\mathbf{Z} + Q\mathbf{Z}|\eta^\top \mathbf{Z})|Y) = \mathbb{E}(P\mathbf{Z}|Y) + \mathbb{E}(\mathbb{E}(Q\mathbf{Z}|\eta^\top \mathbf{Z})|Y)$$

This last equality follows since $P\mathbf{Z} = (\eta^\top\eta)^{-1}\eta^\top\mathbf{Z}$ so that $\mathbb{E}(P\mathbf{Z}|\eta^\top\mathbf{Z}) = P\mathbf{Z}$. Finally note that $\mathbb{E}(P\mathbf{Z}) = P\mathbb{E}(\mathbf{Z}) \in Col(\eta)$. So $E(\mathbf{Z}|Y) \in Col(\eta)$ provided that $\mathbb{E}(\mathbb{E}(Q\mathbf{Z}|\eta^\top\mathbf{Z})|Y) = 0$. This is true in the case where $\mathbf{Z}$ has a spherical distribution since $Q\eta = 0$ (Q projects onto the orthogonal complement of $Col(\eta)$). Therefore, this is true when $\mathbf{X}$ has elliptical symmetry, since $\mathbf{X}$ is an affine transformation of $\mathbf{Z}$.

Here, it seems like elliptically symmetric distribution may be the only distributions such that Theorem 1 will hold. However, notice that we only needed $\mathbb{E}(Q\mathbf{Z}|\eta^\top\mathbf{Z}) = 0$ for the e.d.r. directions η and not all pairs of orthogonal vectors $\mathbf{u}, \mathbf{v}$. However since η is unknown, one might take the conservative approach that we must have $\mathbb{E}(Q\mathbf{Z}|\eta^\top\mathbf{Z}) = 0$ for arbitrary Q and η, thus implying that $\mathbf{Z}$ is spherically symmetric by the work done above.

Simulations in later studies have shown that SIR is not overly sensitive to condition (1.3), (rejoinder in Li, 1991). Hall and Li (1993) showed that the set of β directions for which (1.3) approximately holds covers "almost all directions of $\mathbb{R}^p$ as p grows."

2 An SIR Algorithm

Theorem 1 suggests the following algorithm for estimating the subspace spanned by $\{\beta_1, \ldots, \beta_K\}$. Suppose data are sampled as:

$$(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$$

The algorithm involves forming an estimate of the inverse regression curve $\mathbb{E}(\mathbf{Z}|Y)$ where $\mathbf{Z}$ is the standardized version of $\mathbf{X}$. If the condition of Theorem 1 is satisfied, then $\mathbb{E}(\mathbf{Z}|Y)$ will fall in the standardized e.d.r. space span$(\{\eta_1, \ldots, \eta_K\})$. The estimate is done by slicing: first the Y values are sorted, and placed into buckets. Taking the sample mean of $\mathbf{Z}$ inside each bucket, we approximate $\mathbb{E}(\mathbf{Z}|Y = y)$, viewing each slice mean as a realization of this curve. A principal components analysis will recover the most important K-dimensional linear subspace for the estimated curve. In the ideal scenario, the variation of $\mathbb{E}(\mathbf{Z}|Y)$ in directions orthogonal to the standardized e.d.r space will be 0, and the estimated eigenvalues for these directions will be small. We take the K eigenvectors corresponding to the K largest eigenvalues as estimates of the standardized e.d.r. space. Finally, we transform to the original scale of $\mathbf{X}$ by setting $\hat{\beta}_i = \hat{\eta}_i \hat{V}^{-\frac{1}{2}}$. Introducing some notation:

- Let H be the number of slices
- Let $\mathbf{z}_i$ be the standardized version of $\mathbf{x}_i$
- Let $\hat{\mathbf{m}}_h$ be the mean value of $\mathbf{z}_i$ within slice h
- Let $\hat{p}_h = \frac{1}{n}\sum_{i=1}^{n}(I(y_i \in I_h))$, the proportion of datapoints in slice h.
- Let Σ be the weighted covariance matrix estimate of the slice means: $\sum_{h=1}^{H} \hat{p}_h \mathbf{m}_h \mathbf{m}_h^\top$

Data: $\{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)\}$, K, H
Result: $\{\beta_1, \ldots, \beta_K\}$
sort $\{Y_1, \ldots, Y_n\}$;
estimate $V \leftarrow$ sample covariance matrix of $\mathbf{X}$;
divide the sorted dataset into H approximately equally sized slices, $\{I_1, \ldots, I_H\}$;
while $i \leq n$ **do**
  $\mathbf{z}_i \leftarrow V^{-\frac{1}{2}}(\mathbf{x}_i - \mathbf{x})$
end
while $h \leq H$ **do**
  $\hat{\mathbf{m}}_h \leftarrow n_h^{-1} \sum_{i \in I_h} \mathbf{z}_i$
end
estimate $\Sigma \leftarrow \sum_{h=1}^{H} \hat{p}_h \mathbf{m}_h \mathbf{m}_h^\top$;
solve *the eigendecomposition* $\Sigma \hat{\eta}_i = \lambda_i \hat{\eta}_i$; $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$;
return $\{\hat{\beta}_1, \ldots, \hat{\beta}_K\} = \{\eta_1 V^{-\frac{1}{2}} \ldots, \eta_K V^{-\frac{1}{2}}\}$ *corresponding the K largest eigenvalues*

Algorithm 1: Basic Sliced Inverse Regression

Note that if $\mathbb{E}(\mathbf{X}|Y)$ is contained in a proper subspace of the e.d.r. space, then estimating all of the e.d.r. directions is not possible since some directions will be orthogonal to $\mathbb{E}(\mathbf{X}|Y)$. Such cases can be dealt with by SIR II, which considers $\mathrm{Cov}(\mathbf{X}|Y)$ instead.

3 Some Simulation Examples

For each of the following simulations, we used our own implementation of SIR with $H = 10$.

3.1 Linear Regression Model

First we look at a linear regression model, with $\mathbf{X} \in \mathbb{R}^5$, and each $X_i \sim Unif(0, 10)$, with the added noise term following a Gaussian distribution: $\epsilon \sim \mathcal{N}(0, 1)$. Note that we have violated the elliptical symmetry condition on $\mathbf{X}$.

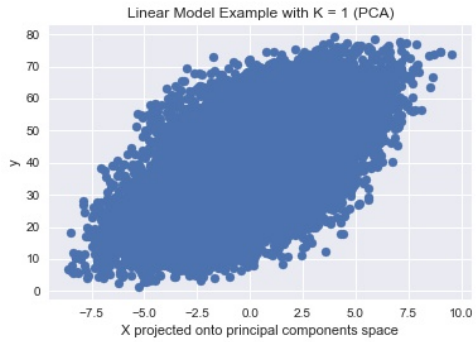
$$Y = 2X_1 + 1X_3 + 5X_5 + \epsilon$$

Running our implementation of SIR with $K = 1$, we recover the following e.d.r direction:

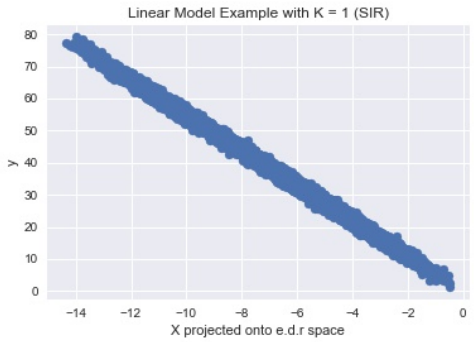
$$\hat{\beta}_1 = (-0.363, -0.002, -0.179, 0.0, -0.914)$$

which up to a sign is quite close to true β_1, $(0.365, 0, 0.913, 0, 0.183)$. Furthermore, we project $\mathbf{X}$ onto the estimated e.d.r. space, and plot y against $\beta_1 \mathbf{X}$ in Figure 3.1.

Figure 3.1



(a) PCA with Linear Model



(b) SIR with Linear Model

3.2 Nonlinear Regression Model

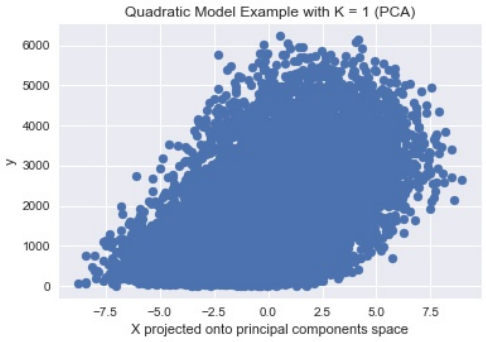
Now we turn our attention to a model with f a quadratic function:

$$Y = (2X_1 + 1X_3 + 5X_5 + \epsilon)^2$$

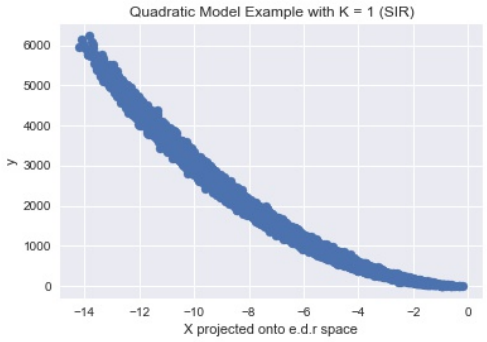
where the distribution of $\mathbf{X}$ is as before. Running our implementation of SIR with $K = 1$, we recover the following e.d.r direction:

$$\hat{\beta} = (-0.348, 0.001, -0.176, 0.001, -0.921)$$

which as before, is quite close to true β. This is unsurprising since this case involves a monotone transformation of the linear example. The result is plotted in Figures 3.2 and 3.3, alongside the result of PCA.



(a) Figure 3.2: PCA with Quadratic Model



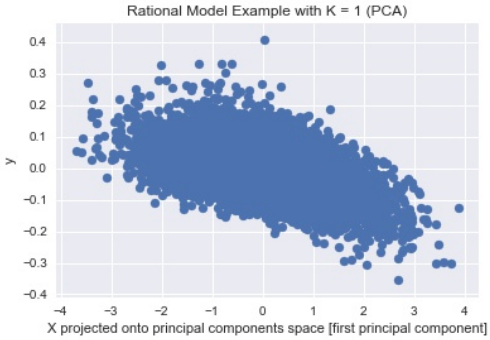
(b) Figure 3.3: SIR with Quadratic Model

3.3 Rational Function with K = 2

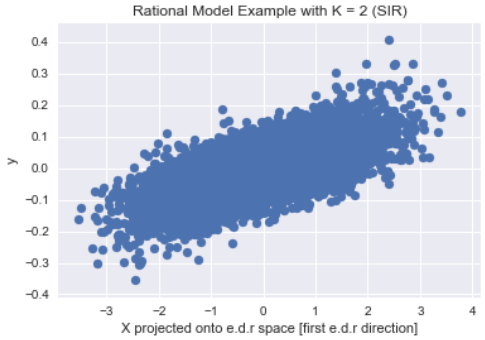
In the following example we consider a rational function where $K = 2$, with $\mathbf{X} \in \mathbb{R}^5$ as before, namely:

$$Y = \frac{X_1}{6.0 + (X_2 - 4.0)^2} + \epsilon$$

Again, SIR is able to recover approximately the correct e.d.r. directions $(1,0,0,0,0)$ and $(0,1,0,0,0)$. The result is plotted in Figures 3.4 and 3.5.



(a) Figure 3.4: PCA with Rational Model



(b) Figure 3.5: SIR with Rational Model

4 Intuition

In this section, we provide some intuition for why we are able to recover the e.d.r. space without any knowledge of the underlying function f. This is best illustrated by contour plots. Consider $\mathbf{X} \in \mathbb{R}^2$, so that $Y = g(X_1, X_2)$ and ignore the fact that there is usually a noise term. When $K = 1$, we have a function of the form:

$$Y = h(\alpha_1 X_1 + \alpha_2 X_2)$$

The contours are given by $\alpha_1 X_1 + \alpha_2 X_2 = c$, so we see that they are lines perpendicular to the vector $(\alpha_1, \alpha_2)^\top$ (Figure 4.1). If $\mathbf{X}$ follows a spherical distribution, then the data points will be symmetrically scattered within the contour lines. When SIR is used, slicing the data across Y values will create bins similar to the contour lines in Figure 4.1. The mean within each bin will fall approximately on the line $X_2 = \frac{\alpha_2}{\alpha 1} X_1$. Finally, a principal components analysis on the slice means will successfully estimate the orientation of the line. Note that if the covariance is not symmetric, they will not fall on this line, but will fall on a different line, which is accounted for by standardizing $\mathbf{X}$ to $\mathbf{Z}$, and after the eigendecomposition step transforming back by multiplying by a factor of $V^{-\frac{1}{2}}$.

5 Sampling Properties

Assume here that $\mathbf{X}_i$ has been standardized to $\mathbf{Z}_i$. Let $p_h = P(Y \in I_h)$ and $m_h = E(\mathbf{Z}|Y \in I_h)$. We can show that the sample mean of the $\mathbf{Z}_i's$ in each slice, $\hat{m}_h = \frac{1}{n\hat{p}_h} \sum_{Y \in I_h} \mathbf{Z}_i$, converges to m_h at rate $\frac{1}{\sqrt{n}}$ using the law of large numbers. If we denote $\Sigma = \sum_{h=1}^{H} p_h m_h m_h^T$, then the weighted covariance matrix $\hat{\Sigma} = \sum_{h=1}^{H} \hat{p}_h \hat{m}_h \hat{m}_h^T$ in the weighted principal component analysis part of the SIR procedure converges to Σ at the root n rate, and hence $\hat{\Sigma}^{-1/2}$ also converges to $\Sigma^{-1/2}$. From the corollary to the main theorem, we can see that the first K eigenvectors of Σ belong to the standardized effective dimension reduction space span$(\eta_1, \ldots, \eta_k)$. The output $\hat{\beta}_k = \hat{\eta}_k \hat{\Sigma}^{-1/2}$ therefore converges to a corresponding eigenvector in the effective dimension reduction space with rate root n, by the continuous mapping theorem. When we have the following set of intervals (in this case the range of each slice varies to allow for even distribution of observations):

$$I_h = (F_y^{-1}((h-1)/H), F_y^{-1}(h/H))$$

where F_y is the CDF of Y (to ensure an even distribution of observations), the root n consistency result still holds.

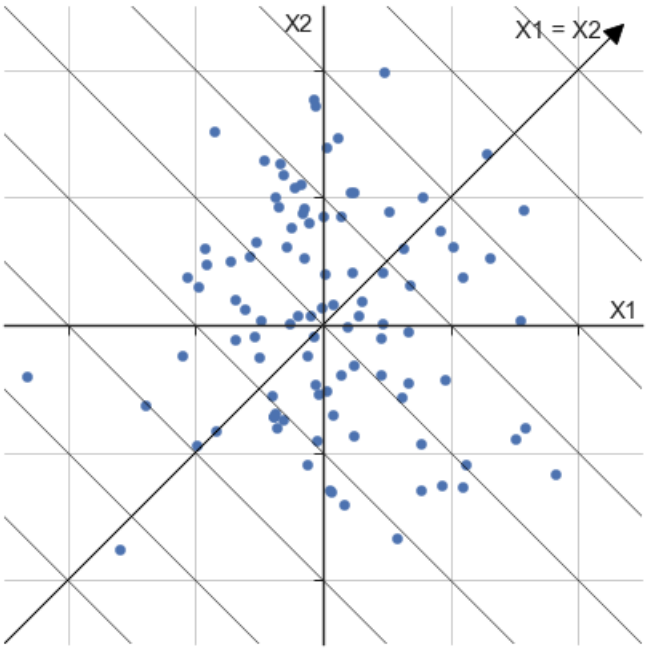


Figure 4.1: Contour Plot of the model $Y = g(X_1 + X_2)$ where X has a spherically symmetric distribution and g is unknown. Slicing across y values will estimate the contour lines.

An important question arises: when the number of slices becomes arbitrarily big and the observations are too sparse for $\hat{m}_h$ to consistently estimate m_h, can we still obtain reasonable estimates from the SIR procedure? As a matter of fact, the SIR procedure is still consistent when the number of slices H is large. One way to see this is through examining the identity

$$E(\text{Cov}(\mathbf{Z}|Y)) = \text{Cov}(\mathbf{Z}) - \text{Cov}(E(\mathbf{Z}|Y)) = I - \text{Cov}(E(\mathbf{Z}|Y))$$

After an eigenvalue decomposition of $E(\text{Cov}(\mathbf{Z}|Y))$, we can find standardized e.d.r directions from the eigenvectors associated with the smallest K eigenvalues. We can estimate $E(\text{Cov}(\mathbf{Z}|y))$ by first introducing a large number of slices for splitting up the range of y. Afterward, within each slice we find the sample covariance of the $\tilde{x}_i$'s that belong to that slice, and then we find the average of these estimated conditional covariances. When the number of slices is very large, the last step of averaging the estimated conditional covariances will make the final estimate relatively stable, although the sampling variance in each estimate of $\text{Cov}(\mathbf{Z}|y)$ is not guaranteed to become smaller. Therefore, even in the extreme case where the number of slices is $\frac{n}{2}$ so that each slice consists of only two observations, we can expect that the resulting estimate is root n consistent. The connection this has for SIR is that the estimate of $E(\text{Cov}(\mathbf{Z}|Y))$ using this procedure is proportional to $I - \hat{\Sigma}$ due to the identity above. Conducting principal component analysis on this estimate of $E(\text{Cov}(\mathbf{Z}|Y))$ is therefore equivalent to conducting principal component analysis on $\hat{\Sigma}$. The choice of H thus has minimal impact on the consistency of the SIR procedure. (K.C. Li, 1991)

To be more mathematically precise, consider the case $H = \frac{n}{L}$ so there are L observations in each slide. We fix L and examine what happens when n goes to infinity. We start with the identity

$$\hat{V} = \hat{\Sigma}_\eta + \hat{\Sigma}_a$$

where $\hat{\Sigma}_a$ is the average of estimated conditional covariances:

$$\hat{\Sigma}_a = \frac{1}{H}\sum_{h=1}^{H}\left(\frac{\sum_{i \text{ in slice } h}(\mathbf{X}_i - \bar{\mathbf{X}}_h)(\mathbf{X}_i - \bar{\mathbf{X}}_h)^T}{L}\right)$$

and

$$\hat{V} = \frac{\sum_i (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T}{n}$$

$$\hat{\Sigma}_\eta = \frac{\sum_{h=1}^{H}(\bar{\mathbf{X}}_h - \bar{\mathbf{X}})(\bar{\mathbf{X}}_h - \bar{\mathbf{X}})^T}{H}$$

This identity is equivalent to (just multiply $LH = n$ on both sides):

$$\sum_i (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T = L\sum_{h=1}^{H}(\bar{\mathbf{X}}_h - \bar{\mathbf{X}})(\bar{\mathbf{X}}_h - \bar{\mathbf{X}})^T + \sum_{h=1}^{H}\sum_{i \text{ in slice } h}(\mathbf{X}_i - \bar{\mathbf{X}}_h)(\mathbf{X}_i - \bar{X}_h)^T$$

The left hand side is equal to:

$$\sum_i \mathbf{X}_i\mathbf{X}_i^T - n\bar{\mathbf{X}}\bar{\mathbf{X}}^T$$

and the right hand side is equal to

$$L\sum_{h=1}^{H}\bar{\mathbf{X}}_h\bar{\mathbf{X}}_h^T - LH\bar{\mathbf{X}}\bar{\mathbf{X}}^T + \sum_i \mathbf{X}_i\mathbf{X}_i^T - L\sum_{h=1}^{H}\bar{\mathbf{X}}_h\bar{\mathbf{X}}_h^T = \sum_i \mathbf{X}_i\mathbf{X}_i^T - n\bar{\mathbf{X}}\bar{\mathbf{X}}^T$$

so both sides are indeed equal. $\hat{\Sigma}_a$ can be shown to converge to

$$\frac{L-1}{L}E(\text{Cov}(\mathbf{X}|Y)) = \frac{L-1}{L}\Sigma_a$$

We also have $V = \Sigma_\eta + \Sigma_a$, or equivalently:

$$\text{Cov}(\mathbf{X}) = \text{Cov}(E(\mathbf{X}|Y)) + E(\text{Cov}(\mathbf{X}|Y))$$

From the strong law of large numbers, it is clear that $\hat{\Sigma}_\mathbf{X}$ converges to $\Sigma_\mathbf{X}$. Hence, we have that $\hat{\Sigma}_\eta$ converges to (by continuous mapping theorem)

$$V - \frac{L-1}{L}\Sigma_a = \frac{1}{L}\Sigma_\mathbf{X} + \frac{L-1}{L}\Sigma_\eta$$

By eigendecomposition we obtain

$$\Sigma_\eta\beta_i = \lambda_i V\beta_i$$

and so

$$\left(L^{-1}V + \frac{L-1}{L}\Sigma_\eta\right)\beta_i = \left(L^{-1} + \frac{L-1}{L}\lambda_i\right)V\beta_i$$

Hence, the eigenvectors obtained from SIR are the same as H goes to infinity in a way such that L is fixed.

The SIR procedure can potentially handle high-dimensional data. When sample size and the dimension of $\mathbf{X}$ increase but K and the number of slices H are kept constant, assuming the condition that the largest singular value for $\hat{\Sigma} - \Sigma$ converges to 0 in probability, $\hat{\eta}_k$ "converges" to η_k (i.e. the angle between the two goes to zero). The above condition occurs when $\frac{p}{n}$ goes to zero in a way such that the difference between the largest eigenvalue of $\hat{V}$ and that of V converges to zero, and the eigenvalues of Σ are bounded away from 0 and infinity as n increases.

6 A USEFUL PROPERTY FOR NORMAL DATA

Practically, when performing SIR we would like the first K eigenvalues of $\hat{\Sigma}$ to be non-zero. We would like to be able to test whether our assumption of K is reasonable. The following theorem may be useful, so we will state it without proof. Let $\bar{\lambda}_{p-k}$ be the average of the smallest $p-k$ eigenvalues of $\hat{\Sigma}$ (Kato, Chapter 2, 1976).

Theorem 2. *Under model (1.1),* $\mathbf{X} \sim MVN(\mu, V) \implies n(p-K)\bar{\lambda}_{p-K} \xrightarrow{D} \chi^2_{(p-K)(H-K-1)}$

Using this result, we can test whether our assumption of K is reasonable by using the the average of the $p-K+1$ smallest eigenvalues, $\bar{\lambda}_{p-K+1}$, and comparing this to the quantiles of the corresponding χ^2 distribution. A p-value of $< \alpha$ will give us evidence that there are likely at least K linearly independent e.d.r. directions. This test is not entirely satisfying since working with the averages of the smallest eigenvalues makes it overly conservative.

7 REAL DATA EXAMPLE

In this section, we will compare using SIR to PCA for dimensionality reduction in a linear regression problem. We will model death rate in a Metropolitan Statistical Area as a function of 15 demographic and geographical variables (see link under Code). First we standardize the columns to have sample mean 0 and sample variance 1. Since SIR is a supervised procedure, we only consider accuracy metrics that are done out of sample. As an evaluation metric, we report the out of sample R^2 using a K-Fold cross validation with 5 folds. We set K, the number of dimensions of the reduced set of variables to be 2 and compare the out-of-sample performance of Principal Components Regression to that of reducing dimensionality with SIR before performing regression, making sure to estimate the e.d.r. directions and principal components using only training data. The following table lists the out-of-sample R^2 scores:

Fold	PCR	SIR + OLS	OLS
1	-0.287	0.301	0.398
2	-0.265	0.474	0.652
3	0.071	0.351	0.053
4	0.400	0.634	0.718
5	0.059	0.613	0.543

The mean R^2 across folds for principal components regression with two principal components was -0.004, the mean R^2 across folds for SIR using two e.d.r. directions followed by ordinary least squares regression was 0.475, and the mean R^2 score across folds using ordinary least squares regression and all 15 columns was 0.473. So, we see that for this particular problem, SIR allows us to reduce the dimension of the data from 15 to 2 with very little sacrifice in out of sample performance.

After this analysis, we might also wonder what a reasonable value of K is. In figure 7.1, we include a plot of the eigenvalues in sorted descending order (H was set to 20 to make this plot). This may give us an ad hoc sense of what values of K are reasonable, seeing that there is a drop off between the 6th and 7th eigenvalue. Li's original paper notes that 0.25 seems to be a reasonable threshold [1].

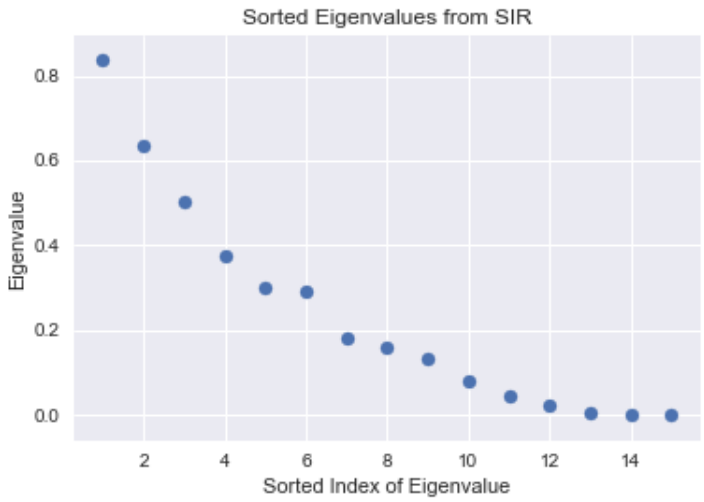


Figure 7.1: Eigenvalues from SIR eigendecomposition step in descending order

8 Extensions and Conclusion

Our discussion of SIR was motivated by the problem in class where we dealt with multicollinearity among features of our dataset by applying a dimension reduction procedure. The method that we've discussed here takes the approach that in order to tackle this problem, we should assume that our features affect the response variable through much smaller set of linear combinations of features. Since SIR was introduced in 1991, other similar techniques, called Sufficient Dimension Reduction techniques, have been proposed to estimate the subspace spanned by $\{\beta_1, \ldots, \beta_K\}$ under the same model as SIR. Though sufficient dimension reduction techniques are widely used, SIR remains the most popular due to its simplicity and computational efficiency (Liu, 2016). However, the consistency results and sampling properties of SIR relied on n, the number of observations, growing faster than p, the number of features. In this big data era, often we are faced with datasets where the number of features can far exceed the number of observations, leading to data analysis problems that are quite intractable. A current area of research involves studying the behavior of SIR under circumstances where p is large relative to n. Lin (2015) showed that the e.d.r. space estimated by SIR is consistent if and only if $\lim \frac{p}{n} \to 0$. In other cases, certain restrictions need to be imposed, such as the number of active variables being much less than p and n. For example, Li and Nachstein proposed a method, using the shrinkage idea of LASSO regression, that reduces the number of non-zero elements of each β_i [8]. Exploring this area of research would be an interesting extension to this project.

9 Code

The code used for this project can be found at:

https://github.com/russellkune/SIR

The data used for the real data example can be found here:

https://people.sc.fsu.edu/~jburkardt/datasets/regression/x28.txt

REFERENCES

[1] Li, *Sliced Inverse Regression for Dimension Reductions*, Journal of the American Statistical Association, 1991.

[2] Li, *Lecture Notes for Statistics 216: High Dimensional Data Analysis* UCLA, 2000.

[3] Joseph K. Blitzstein, Carl N. Morris, *Probability for Statistical Science* (draft version)

[4] Morris L. Eaton, *A Characterization of Spherical Distributions*, Journal of Multivariate Analysis, 1986.

[5] Hardle and Simar, *Applied Multivariate Statistical Analysis*, Springer-Verlag Berlin Heidelberg, 2012.

[6] Cook and Weisburg, *Sliced Inverse Regression for Dimension Reduction: Comment*, Journal of the American Statistical Association, 1991.

[7] Cambanis, Huang, Simons *On the Theory of Elliptically Contoured Distributions* Journal of Multivariate Analysis, 1981.

[8] Li and Nachtsheim, *Sparse Sliced Inverse Regression* Technometrics, 2006

[9] Lin, Zhao, and Liu, *Sparse Sliced Inverse Regression for High Dimensional Data*, arxiv preprint, 2016